The Gender Gap in Confidence: Expected but Not Accounted For[†]

By Christine L. Exley and Kirby Nielsen*

We investigate how the gender gap in confidence affects the views that evaluators (e.g., employers) hold about men and women. We find the confidence gap is contagious, causing evaluators to form overly pessimistic beliefs about women. This result arises even though the confidence gap is expected and even though the confidence gap shouldn't be contagious if evaluators are Bayesian. Only an intervention that facilitates Bayesian updating proves (somewhat) effective. Additional results highlight how similar findings follow even when there is no room for discriminatory motives or differences in priors because evaluators are asked about arbitrary, rather than gender-specific, groups. (JEL D82, D83, D91, J16, J22, M51)

Women are underrepresented and underpaid in many areas of the labor market, especially in male-stereotyped fields (Bertrand, Goldin, 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, Fox, and Punćcohar 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, Collis, and Kulkarni 2019b; 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, Wiswall, and Zafar 2017), being less likely to enter competitive fields (Niederle and Vesterlund 2007; Buser, Niederle, and Oosterbeek 2014), being less likely to speak up (Coffman 2014), and being less likely to apply for challenging work (Coffman, Collis, and Kulkarni 2019a). 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 under-confidence 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, Imas, and Rosenberg 2019; Coffman, Exley, and Niederle 2021).[1] We conduct an experiment that investigates whether individuals expect and account for the confidence gap to test between these hypotheses.

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 all 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 percent of women believe they have a "poor performance" (i.e., a performance that is indicative of poor math and science skills) while only 56 percent 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 indicative of poor math and science skills.

Then, to investigate how this confidence gap affects others' beliefs about men and women, we incentivize "evaluators" to provide accurate beliefs about workers' performance both before and after they learn how workers answer the main self-evaluation question. Specifically, after evaluators learn whether they will be asked to provide 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 percent 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 percent 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.

According to their prior beliefs, i.e., 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. However, this expected performance gap is small ($\sim$3.9 percentage points) and is not statistically different from the true gap ($\sim$1.7 percentage points).

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'

---

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

self-evaluations, evaluators hold an 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 the gap in 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.

In considering what prevents evaluators from accounting for the confidence gap when forming their posterior beliefs, one possibility relates to an "attention" problem: evaluators may simply fail to attend to the confidence gap when forming their posterior beliefs. To investigate this possibility, we test a light-touch intervention. In the *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 relates to a "calculation" problem: evaluators may be either unable or unwilling to do the necessary calculations and Bayesian updating required to accurately account for the confidence gap. To investigate this possibility, we test a much more extensive intervention. In the *Calculation* treatment, to alleviate any difficulty with Bayesian updating, we provide evaluators with their implied Bayesian posterior beliefs before eliciting their posteriors. This intervention proves effective: the expected performance gap shrinks and is only marginally significantly different from the true performance gap. As additional evidence of the calculation problem, we also see that the extent to which evaluators' posteriors disfavor women is positively and significantly correlated with evaluators exhibiting base rate neglect.

To further investigate the calculation problem—and the extent to which our results are specific to gender—we examine whether similar results follow when we ask evaluators about arbitrary, rather than gender-specific, groups. In particular, in the *Unknown Gender* treatments, we ask evaluators 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). This maintains the confidence gap between these two groups of workers while allowing differences between evaluators' posterior beliefs and their implied Bayesian posterior beliefs to reflect failures in Bayesian updating but not any gender-specific biases. We find that the results from these treatments are indistinguishable from the results where evaluators know the workers' gender, which shows that the confidence gap is contagious even when it is a calculation problem about arbitrary groups and cannot reflect discriminatory motives or differences in priors.

To summarize, our main results show that the confidence gap is contagious—causing the expected performance gap in which evaluators have overly pessimistic beliefs about women relative to men—and specifically point towards the role of failures in Bayesian updating rather than other sources such as taste-based discrimination against women per se. Three results support this conclusion, the latter two of which are explained in detail above. First, counter to what one may expect but consistent with prior findings in Card et al. (2020) in which the gender of the referee does not significantly affect the relative assessment of economics papers written by

men versus women, we find that female evaluators are just as likely as male evaluators to hold posterior beliefs that significantly disfavor women. Second, the only intervention that is somewhat successful at shrinking the expected performance gap is the *Calculation* treatment that assists evaluators with Bayesian updating. Third, the expected performance gap persists in the *Gender Unknown* treatments, i.e., the treatment in which we remove gender labels so the expected performance gap may reflect failures in Bayesian updating but not discriminatory motives against women.

In addition to the robustness of the expected performance gap evident from the above, we conclude with a few more notes on robustness. First, since how we classify workers as having a "poor performance" may contribute to the complexity involved in evaluators forming their posterior beliefs, we show that our results are robust to using a simpler performance outcome in which evaluators provide beliefs about whether a worker has performance in the top half—see Section VA for more details. Second, and related to recent work on understanding experts' beliefs (DellaVigna and Pope 2018a, b), we replicate our results among a pool of participants for whom one may posit this type of problem is less complex: professional evaluators who self-report hiring and managerial experience. Third, our results persist across a variety of types of evaluators and across a variety of conditions, including when evaluators know additional information on the workers and when workers face strategic incentives.

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 et al. 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, Sapienza and Zingales 2014), and that the confidence gap conveyed with employees' self-evaluations does not influence employers' relative ratings of their male and female employees (Bohnet, Hauser, and Kristal 2022).

Relative to this literature, part of our main contribution lies in eliciting a variety of incentivized beliefs that allow us to cleanly document and narrow in on *why* individuals do not account for the confidence gap.[3] Indeed, our evidence makes clear that it is not simply an attention problem and instead points towards a calculation problem. This connects our work to the extensive literature on errors in Bayesian updating (see Benjamin 2019 for a review). For instance, the fact that our evaluators react *too much* to self-evaluation information relative to the Bayesian posterior is consistent with a growing literature that shows overinference from weak signals (Edwards 1968; Augenblick, Lazarus, and Thaler 2023; Ba, Bohren, and Imas 2023). Our results also relate to early (Kahneman and Tversky 1972b, 1973; Grether

---

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

1980; Koehler 1996) and more recent (Esponda, Vespa, and Yuksel 2024) work on base-rate neglect, work that documents a relationship between non-Bayesian updating and cognitive uncertainty (Enke and Graeber 2023), and other related belief updating biases such as correlation neglect (Enke and Zimmermann 2019), learning from missing information (Enke 2020), and failure to unlearn from retracted signals (Gonçalves, Libgober, and Willis 2021). In addition, that errors in Bayesian updating can contribute to worse beliefs about women relative to men also connects our findings to the work on discrimination that is reflective of *inaccurate* beliefs (Bordalo et al. 2019; Bohren et al. 2019). More generally, our results point towards the need for more extensive interventions that directly help individuals with Bayesian updating in order to account for the confidence gap. We discuss possibilities along these lines and other directions for future work in Section VI.

I. Experimental Design

A. *Design Overview*

Our experimental design involves two main types of participants: "workers" and "evaluators." The workers are incentivized to accurately answer self-evaluations about their performance on a test, and the evaluators are incentivized to accurately provide beliefs about the workers before and after the evaluators are given information on the workers' self-evaluations. In this way, our design is akin to a situation where evaluators (such as managers) hold beliefs about their workers and these beliefs may then be affected by self-evaluations that workers provide (such as in interviews and performance reviews).

An important question in our experimental design is what type of self-evaluation to focus on. Should we examine self-evaluations in which workers are asked about their absolute performance (e.g., the number questions they got right on the test), their relative performance (e.g., whether their test performance was in the top half among all workers), or their subjective performance (e.g., whether their test performance was "poor")? Absolute, relative, and subjective performance outcomes all have been used in prior work and are relevant in many contexts outside of the laboratory. Thus, our design approach is three-fold.

First, our *Worker Study* asks workers to complete 17 self-evaluations about their performance, including absolute, relative, and subjective performance questions. This allows us to examine whether there are gender differences in self-evaluations—i.e., whether the confidence gap arises—in a wide range of different types of self-evaluations.

Second, our *Evaluator Study* asks evaluators to provide beliefs—incentivized for accuracy—about these workers both before and after being provided with information about just *one* type of self-evaluation. Specifically, to avoid confusing the evaluators and to allow for clean Bayesian benchmarks, our *Evaluator Study* elicits beliefs about our *main* self-evaluation. In choosing which self-evaluation should be our "main" self-evaluation, we chose a self-evaluation question in which workers provide beliefs about a *binary and subjective* performance outcome (detailed more in Section IB). The binary nature of this performance outcome facilitates Bayesian benchmarks, and the subjective nature of this performance outcome allows us to

build upon prior related work (Exley and Kessler 2022) and connect to important contexts outside of the laboratory where individuals complete self-evaluations and discuss their performance in more subjective ways.[4]

Third, we show that our focus on *one* type of self-evaluation in the *Evaluator Study* is not driving our results by documenting the robustness of our results in other study versions. In particular, all of our results—for both workers and evaluators— are robust to considering different self-evaluation questions, including a simpler measure of whether a worker's performance is in the top half among other workers (see Section VA).

Below, to explain our main design most concisely, we will refrain from discussing these other self-evaluation questions for now and will instead focus on our *main* self-evaluation question. In particular, Section IB describes our *Worker Study*, with specific attention paid to the main self-evaluation question even though workers are asked to answer 17 self-evaluation questions. Section IC describes *Evaluator Study*, which only relates to our main self-evaluation question. Following this, Section ID briefly details our recruitment and implementation details, including for other studies. In total, we recruited 7,694 participants—mostly on Prolific and as detailed later.

B. *Design for the Worker Study*

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 20 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, then workers earn $0.10 for each question they answer correctly.

After part 1 but before part 2, workers report an answer 0–10 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] Online Appendix Table A.4 details all 17 of the self-evaluation questions, and the *main self-evaluation question* is as follows:

---

[4]The subjective nature also allows us to control information on *objective* performance (see Section VJ).

[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 online Appendix Table A.4.

[7]See the table note of online Appendix Table A.4 for details on randomization and incentives.

- **Main Self-Evaluation Question:** Did your classifier describe your performance on the math and science test as indicative of poor math and science skills?

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 that 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 online Appendix Table A.4.[9] We will also use the shorthand of "good performance" to refer to the opposite.

C. *Design for the Evaluator Study*

In the *Evaluator Study*, evaluators are randomly assigned into one of six treatments. We will detail the *Baseline* treatment below, and we will describe the the additional five treatments later as they become relevant (see also online Appendix Figures A.1–A.3 for an overview of each of these treatments).

The *Baseline* treatment of the *Evaluator Study* elicits four beliefs: prior beliefs, posterior beliefs, underconfidence beliefs, and overconfidence beliefs. See online Appendix Table A.5 for the exact wording of the belief questions. In addition to a \$2 completion fee for a 10-minute study, evaluators may earn \$1 in bonus payment because they are incentivized to accurately provide these beliefs, as detailed below. Each of these four beliefs relates to whether "their worker" has a poor performance, defined in the same manner as noted above in Section IB.[10] 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"). Therefore, each evaluator is only asked about female workers *or* male workers.

To examine evaluators' beliefs before they learn any information on workers' self-evaluations, we first elicit an evaluator's *prior belief* about the percent chance

---

[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 online Appendix Table A.4. 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 online Appendix Table A.5).

that their worker has a poor performance by asking them the "Prior Belief" question noted below.

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

Next, to examine how evaluators' beliefs are influenced by information on workers' self-evaluations, we provide them with *accurate* information on the workers' self-evaluations and then elicit their *posterior beliefs*. Specifically, from the available pool of workers from which the worker could be randomly selected, evaluators are accurately informed that 80 percent of female workers thought they had a poor performance if their worker is a randomly selected female worker, or instead are accurately informed that 56 percent of male workers thought they had a poor performance if their worker is a randomly selected male worker. We then elicit evaluators' *posterior belief* about the percent chance that their worker has a poor performance by asking them the "Posterior Belief" question noted below.

- **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, to assess how likely evaluators think it is that their worker is overconfident or underconfident, we elicit evaluators' *overconfidence belief* and *underconfidence belief* by asking them, via a strategy-method style elicitation, the following "Overconfidence Belief" question and "Underconfidence Belief" question.

- **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?

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 provide beliefs 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 workers who had performances in the "middle," or in the twenty-fifth–seventy-fifth percentile, in the *Worker Study*. This restricted worker pool allows us to ensure that there are no gender differences in the actual performance of workers, but it does introduce some complexity in terms of how we describe the available pool of workers to evaluators.[11] We thus emphasize

---

[11] Specifically, 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

that, as an important robustness check, we show that similar results persist when we remove this restriction and instead ask evaluators to provide beliefs about the full pool of workers in the *Evaluator* (*Full Distribution*) *Study* (see Section VB). We also note that other 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—see the *Evaluator* (*Professional Evaluators*) *Study* in Section VC, the *Evaluator* (*Additional Demographics*) *Study* in Section VI, and the *Evaluator* (*Known Performance*) *Study* in Section VJ.

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 over-confidence 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 percent) and may earn a $1 bonus according to an incentive-compatible Becker-DeGroot-Marschak (BDM) procedure.[12] In addition, at the end of the study after they have provided all of the above beliefs, participants are surprised with the opportunity to earn $1 if they correctly answer one question, selected at random, out of five additional questions. These five additional questions test for common cognitive biases that might correlate with belief updating behavior (see Section IVA for details).

D. *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.

For our *Worker Study*, 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. For an overview of this study (referred to as the *Worker Study— Baseline Treatment*) as well as additional study versions that involve workers, see online Appendix Table A.1.

For the *Evaluator* study, 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 of our *Evaluator Study*: the *Baseline* treatment ($n = 402$), the *Attention* treatment ($n = 403$), the *Calculation* treatment

---

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 percent 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). Future work may also examine the robustness of these results to instead eliciting beliefs as frequencies.

[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 percent 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.

($n = 405$), the *Baseline, Unknown Gender* treatment ($n = 405$), the *Attention, Unknown Gender* treatment ($n = 392$), and the *Calculation, Unknown Gender* treatment ($n = 393$). For an overview of these six treatments in our *Evaluator Study*, see online Appendix Table A.2. For an overview of additional study versions that involve evaluators, see online Appendix Table A.3.

In total, in addition to our *Worker Study* and our *Evaluator Study*, we recruited an additional 1,091 workers and 3,800 evaluators to complete additional study versions, which we will discus as they become relevant in this paper.[14] All of our data are available in the repository Exley and Nielsen (2024).

II. Worker Results

To establish the confidence gap, we first examine data from the *Worker Study*.

Table 1 presents results on how male and female workers answer the *main self-evaluation question* by showing the likelihood that a worker believes that they have a poor performance (the dependent variable equals 1 if a worker believes that they have a poor performance and 0 otherwise) regressed on *Female*, which is an indicator for female workers.

The estimates in column 1 show a clear confidence gap among the full pool of workers: 57 percent of male workers believe they have a poor performance (see the coefficient estimate on the constant) while 73 percent of female workers believe they have a poor performance (note the sum of the coefficient estimates on the constant and *Female*). This confidence gap arises despite the fact that the actual likelihood of a poor performance is 53 percent among female workers and 47 percent among male workers ($p = 0.09$).[15] In addition, the inclusion of performance fixed effects in column 2 reveals that this confidence gap is statistically significant when comparing equally performing men and women.

The estimates in column 3 also show a clear confidence gap among the available pool of workers that evaluators are asked about (i.e., workers who had performances in the middle): 56 percent of male workers in the available pool of male workers believe they have a poor performance while 80 percent of female workers in the available pool of female workers believe they have a poor performance.[16] This confidence gap arises despite the fact that the actual likelihood of a poor performance is 50 percent among these female workers and 48 percent among these male workers ($p = 0.56$). The inclusion of performance fixed effects in column 4 reveals that this confidence gap remains statistically significant when comparing equally performing men and women.

---

[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 35 in the online Appendix.

[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.

[16] One might wonder whether this result arises from differences in beliefs about absolute performance or about differences in beliefs in the poor performance "standards" of the classifiers. Controlling for performance, or performance and beliefs about performance, we find no significant gender difference in how workers answer the classifier questions ($p < 0.1$). We do not elicit workers' beliefs about the classifier thresholds of others.

Table 1—Self-Evaluations in the *Baseline* Treatment of the *Worker Study*

| | Workers' answer to main self-evaluation question | | | |
| | 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)	
Observations	393	393	249	249
Perf fixed effects	No	Yes	No	Yes

Notes: Standard errors 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 fixed effects 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 twenty-fifth–seventy-fifth percentile.

While the above focuses on documenting the confidence gap in response to the main self-evaluation question, recall that workers answered 16 other self-evaluation questions as well. Online Appendix Table B.1 presents the regression results of all self-evaluations. These results reveal that the confidence gap is robust in response to all 17 self-evaluation questions, and significantly so in 16 of the 17.[17] Specifically, the column headers in online Appendix Table B.1 refer to the relevant self-evaluation question label that is detailed in online Appendix Table A.4. For instance, Column "0" of online Appendix Table B.1 presents the workers' responses to Self-Evaluation Question "0" as labeled in online Appendix Table A.4.

Focusing on absolute performance outcomes, panel A of online Appendix Table B.1 shows that women believe they got fewer questions correct than men in absolute terms (column 0), believe that they are less likely than men to have answered at least 3 questions correctly (columns 1B and 1C), to have answered at least 5 questions correctly (columns 2B and 2C), and to have answered at least 7 questions correctly (columns 3B and 3C).

Focusing on relative performance outcomes, panel B of online Appendix Table B.1 shows that women are less likely than men to believe they scored in the top half relative to all other participants who took the study (columns 4B and 4C), relative to women who took the study (columns 5B and 5C), and relative to men who took the study (columns 6B and 6C).

Finally, focusing on the subjective performance outcomes, panel C of online Appendix Table B.1 shows that the results are robust to different types of subjective performance outcomes (columns 7B–8C).

---

[17] The confidence gap is not statistically significant in column 3C, statistically significant at the $p < 0.1$ level in column 3B, and statistically significant at the $p < 0.01$ in all 15 other columns. The results in columns 3B and 3C may in part reflect that even male workers thought it was very unlikely to have answered at least 7 questions correctly. Indeed, in response to the binary self-evaluation question in column 3B, only 13 percent of male workers thought they got 7+ questions right. In response to the percent chance self-evaluation question in column 3C, the average believed percent chance of getting 7+ questions right was 26 percent among male workers.

Thus, taken together, our results reveal that the gender gap in confidence persists when we ask workers about simple and objective performance outcomes (i.e., the absolute and relative performance outcomes) *and* when we ask workers about subjective outcomes that could reflect—as is often the case in self-evaluations and communications about one's performance and ability in practice—workers' beliefs about their absolute and relative performance as well as their subjective assessments of what constitutes a "poor performance."

III. Evaluator Results

While the gender gap in confidence persists across the various self-evaluation questions we asked workers (as just shown in Section II), recall that our *Evaluator Study* focuses only on our main self-evaluation question (and we return to the other self-evaluation questions later in Section VA).

A. *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*.

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 a poor performance. According to their priors, evaluators believe that there is a 42.97 percent chance of a female worker having a poor performance and there is a 39.08 percent chance of a male working having a poor performance. That is, evaluators believe that female workers are 3.89 percentage points more likely to have a poor performance than male workers. 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).

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 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).

Column 4 (Implied Bayesian Posteriors) of Table 2 presents evaluators *implied Bayesian posterior beliefs*, which we define to equal what Bayesian evaluators would believe is the likelihood that a worker has a poor performance after they are provided with the information on workers' self-evaluations. As detailed in online

Table 2—Evaluators' Beliefs in the *Baseline* Treatment of the *Evaluator Study*

	Prior (1)	Over-confidence (2)	Under-confidence (3)	Implied Bayesian Posterior (4)	Posterior (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
SE of Δ	(1.87)	(2.27)	(2.06)	(1.87)	(1.78)
Panel B. Evaluators' beliefs − truth					
$B(F) - Truth(F)$	−6.56	24.51	−19.12	−5.70	12.32
$B(M) - Truth(M)$	−8.71	9.05	−6.53	−7.72	3.57
$\Delta - Truth(\Delta)$	2.15	15.46	−12.59	2.03	8.75
SE of $\Delta - Truth(\Delta)$	(1.87)	(2.27)	(2.06)	(1.87)	(1.78)
Observations	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
$Truth(\Delta)$	1.74	−23.70	22.65	1.74	1.74

Notes: Standard errors are robust and shown in parentheses. See online Appendix Table A.5 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 $B(M)$), the difference in these averages (see Δ), and the standard error on the difference in these averages (see SE of 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 $B(M) - Truth(M)$), the difference in these demeaned averages (see $\Delta - Truth(\Delta)$), and the standard error on the difference in these demeaned averages (see SE of $\Delta - Truth(\Delta)$). 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 $Truth(\Delta)$). When considering evaluators' prior, implied Bayesian posterior, and posterior beliefs, we define these truth values as the actual likelihood of a randomly selected male/female worker having a poor performance. When considering evaluators' overconfidence and underconfidence beliefs, we define these truth values 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 (see equations E4 and E5, respectively, in online Appendix E.E3). Data are from the 402 participants in the *Baseline* treatment of the *Evaluator Study*.

Appendix E, we can calculate each evaluator's implied Bayesian posterior belief given the three evaluator beliefs discussed so far: an evaluator's prior belief, overconfidence belief, and underconfidence belief. This is because evaluators' overconfidence and underconfidence beliefs determine their beliefs about the accuracy of the workers' self-evaluation information and hence how much they should update their prior beliefs after learning it. The implied Bayesian posterior beliefs reveal that—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 after learning the workers' self-evaluation information—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). This results from the fact that, in our data and as detailed in online Appendix Section E.E4, evaluators believe that workers are sufficiently miscalibrated in their self-evaluations such that a Bayesian evaluator would update very little from this information.

Another way to summarize the implied Bayesian posterior beliefs is as follows: according to evaluators' implied Bayesian posterior beliefs, the confidence gap should not be contagious. That is, 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, even though this self-evaluation information conveys a large confidence gap. However, an examination of evaluators' posterior beliefs shows that this is not the case.

Specifically, column 5 (Posteriors) of Table 2 presents evaluators' posterior beliefs and shows that—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. They expect that female workers are 10.49 percentage points more likely to have a poor performance than male workers. 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 six 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.[18]

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 IVC). 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 online Appendix Figure B.1 for histograms).

B. *Results from Attention and Calculation Treatments*

One hypothesis as to why evaluators fail to accurately account for the confidence gap in the *Baseline* treatment—detailed above in Section IIIA—relates to an "attention" problem. For instance, since evaluators' overconfidence and underconfidence beliefs do reveal an expected gender gap 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, in the *Baseline* treatment, we elicit evaluators' overconfidence and underconfidence beliefs only *after* they provide their posterior beliefs. Thus, to investigate the attention problem via a light-touch intervention, we ran the *Attention* treatment. The *Attention* treatment elicits evaluators' overconfidence and underconfidence beliefs *before* they provide their posterior beliefs.[19] Compare

---

[18] 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 standard errors.

[19] Specifically, 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. By contrast, in

Panel A. Prior beliefs

Panel B. Posterior beliefs
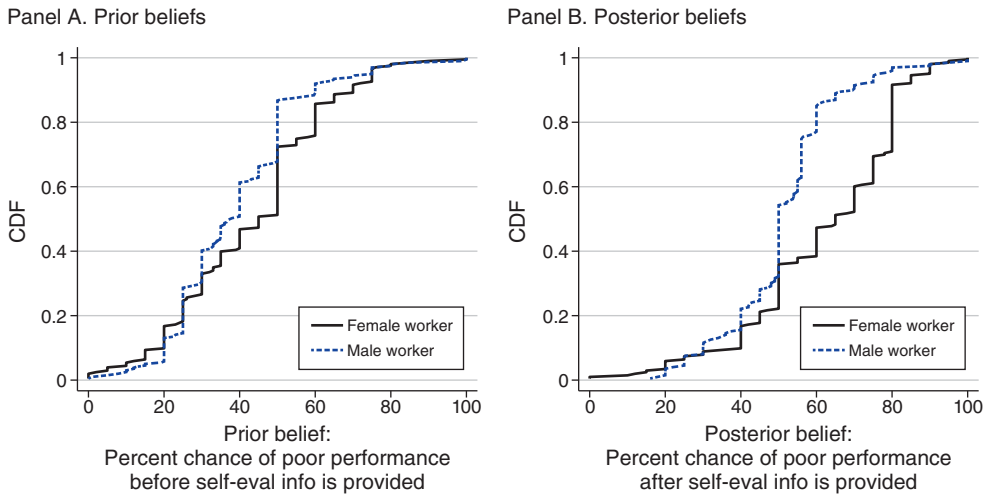
Figure 1. Evaluators' Beliefs in Baseline Treatment

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

online Appendix Figures A.1 and A.2 for a visual representation of this change between the *Baseline* treatment and *Attention* treatment.

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.[20] 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, the *Attention* treatment does not significantly reduce the extent to which evaluators' posterior beliefs indicate an expected performance gap.

Another hypothesis for evaluators' failure to accurately account for the confidence gap relates to a "calculation" problem. For example, evaluators may be unable or unwilling to do the necessary calculations and Bayesian updating required to accurately account for the confidence gap. 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

---

the *Attention* treatment, the order changes to be the following: (i) their prior beliefs, (ii) their overconfidence and underconfidence beliefs, and then (iii) their posterior beliefs.

[20] Following the structure of Table 2, online 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—EVALUATORS' BELIEFS IN THE *BASELINE*, *ATTENTION*, AND *CALCULATION* TREATMENT OF THE *EVALUATOR STUDY*

	Prior (1)	Over-confidence (2)	Under-confidence (3)	Implied Bayesian Posterior (4)	Posterior (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)
Observations	1,210	1,210	1,210	1,209	1,210
Condition fixed effects	Yes	Yes	Yes	Yes	Yes
Truth(Δ)	1.74	−23.70	22.65	1.74	1.74

Notes: Standard errors are robust and shown in parentheses. See online Appendix Table A.5 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 1,210 participants in the *Baseline*, *Attention*, or *Calculation* treatment of the *Evaluator Study*. Sample size differs slightly in column 4 as some evaluators' beliefs imply a Bayesian posterior that is undefined.

they provide their posterior beliefs. See online Appendix Figure A.3 for a visual representation of this treatment.

Table 3 directly compares the *Baseline* treatment to the *Calculation* treatment and reveals one set of significant differences: 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.[21] 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.

[21] Following the structure of Table 2, online 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.

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. It could 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 could 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. We also note that exploring other types of calculation interventions and assistance in forming Bayesian posteriors, including ones that would be more subtle, is an interesting avenue for future work.

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 percent) 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 persist when only considering this 61 percent of participants.[22] Third, as will become evident in Section IIIC, we will be able to show that—to the extent experimenter demand effects or social desirability 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.

C. *Results from Unknown Gender Treatments*

To further investigate the calculation problem and understand the extent to which our results are specific to gender, we ran three additional treatments in which the gender of workers is not known. Specifically, for $X \in \{Baseline, Attention, Calculation\}$, the X, *Unknown Gender* treatment is the same as the X treatment except that instead of providing beliefs about male or female workers, evaluators provide beliefs 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 question in our follow-up survey, 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 exactly the same set as our male workers and group-2 workers are exactly the same set as our female workers. This maintains the confidence gap between these two groups of workers while allowing differences between evaluators' posterior beliefs and their implied Bayesian posterior beliefs to reflect failures in Bayesian updating but not any gender-specific biases.

Following the structure of Table 2, online 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

---

[22] For the 61 percent 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.

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 that a group of workers is relatively more underconfident and relatively less overconfident when they learn that 80 percent of workers in that group believe they have a poor performance compared to when they learn that 56 percent of workers in that group believe they have 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 than group-1 (male) workers.[23] Thus, the confidence gap is contagious even when it can only reflect a calculation problem about arbitrary groups and cannot reflect discriminatory motives or differences in priors. Indeed, the posterior beliefs in these *Unknown Gender* treatments are statistically indistinguishable from the posterior beliefs in the comparable treatments in which gender is known. Specifically, each column in Table 4 presents the posterior beliefs from a pair of treatments that compares the *X* treatment and the *X, Unknown Gender* treatment for $X \in \{Baseline, Attention, Calculation\}$. Across all three pairs of treatments, Table 4 reveals no significant differences in posteriors in the *Unknown Gender* treatments compared to those where gender is known (see the coefficients on $\Delta \times Unknown\ Gender$).[24]

IV. Heterogeneity

To provide further insight into our results, we now turn to a series of additional results that are facilitated via heterogeneity analyses. For conciseness, we will focus on evaluators' posterior beliefs from the *Baseline* treatment, while showing how they are similar in the *Attention* treatment and indicate less of an expected performance gap between men and women in the *Calculation* treatment.

A. *Are Our Results Driven by Evaluators Who Exhibit Other Cognitive Biases?*

Motivated by the evidence in support of the calculation problem, one might expect a correlation between our results and well-known cognitive biases. To investigate this, we incentivize evaluators to correctly answer five additional questions

---

[23] In all treatments, evaluators' posterior beliefs are significantly different than the truth. In addition, again pointing to the role of the calculation problem, evaluators' posterior beliefs are significantly different than their implied Bayesian posteriors in the *Baseline, Unknown Gender* and *Attention, Unknown Gender* treatment, but not in the *Calculation, Unknown Gender* treatment. Also, see online Appendix Table B.7 to compare the beliefs across the three *Unknown Gender* treatments.

[24] 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.

Table 4—Evaluators' Posterior Beliefs about Workers According to Whether or Not They Are in an *Unknown Gender* Treatment of the *Evaluator Study*

	Evaluators' posterior beliefs in X and X, *Unknown Gender* condition given $X =$		
	Baseline (1)	Attention (2)	Calculation (3)
Panel A. Evaluators' beliefs			
Δ	10.49	10.85	4.92
	(1.78)	(1.73)	(1.81)
$\Delta \times$ *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)
$\Delta \times$ *Unknown gender*	0.57	−0.29	−0.05
	(2.40)	(2.45)	(2.53)
Observations	807	795	798
Condition fixed effects	Yes	Yes	Yes
Truth(Δ)	1.74	1.74	1.74

Notes: Standard errors 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 fixed effects) 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 ($\Delta \times$ *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 2,400 participants in the *Evaluator Study*, split across the three columns according to the relevant treatments.

at the end of the study: a standard Bayesian updating question, a question designed to detect base rate neglect (Kahneman and Tversky 1972a), and the three-question cognitive reflection test (CRT) (Frederick 2005), all presented in random order.[25]

Table 5 presents results on how these measures correlate with the extent to which evaluators expect a performance gap, according to their posterior beliefs in the *Baseline* treatment in panel A, the *Attention* treatment in panel B, and the *Calculation* treatment in panel C. Counter to cognitive errors or general updating failures explaining our results, the expected performance gap is directionally larger for evaluators with higher cognitive ability scores (see the coefficients on $\Delta \times$ X in column 1) and is directionally smaller for evaluators who give a response farther from the Bayesian posterior in the Bayesian updating question (see the coefficients on $\Delta \times$ X in 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 (see the coefficients on $\Delta \times$ X in column 2) or who give a response farther from the Bayesian posterior in the base rate neglect question (see the coefficients on $\Delta \times$ X in column 3). In addition, consistent with the *Calculation*

[25] For full question text, see supplemental online Appendix Figures G.1.8–G.1.12.

Table 5—By Cognitive Ability Measures: Evaluators' Posterior Beliefs about Workers in *Evaluator Study* in the *Baseline*, *Attention*, and *Calculation* Treatments

$X =$	Evaluators' posterior beliefs			
	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)
$\Delta \times X$	0.96	3.15	0.35	−0.09
	(1.48)	(3.83)	(0.17)	(0.09)
Observations	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)
$\Delta \times X$	0.90	7.91	0.36	−0.16
	(1.47)	(3.86)	(0.14)	(0.09)
Observations	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)
$\Delta \times X$	1.60	3.83	−0.04	−0.10
	(1.51)	(4.18)	(0.14)	(0.08)
Observations	405	405	405	405
Suppressed X	Yes	Yes	Yes	Yes
Truth(Δ)	1.74	1.74	1.74	1.74

Notes: Standard errors 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 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)).

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

B. *Are Our Results Driven by Evaluators with Certain Demographic Characteristics?*

There are many reasons to expect our result to potentially correlate with different demographic groups. For instance, one may expect that, relative to female evaluators, male evaluators form more pessimistic posterior beliefs about women because they may have less experience with the confidence gap themselves or because of an in-group bias or discriminatory motives. This proves not to be the case. Table 6, which reproduces column 5 of Table 3 for male evaluators in column 1 and female evaluators in column 2, shows that male evaluators, if anything, hold *less*

Table 6—By Demographics: Evaluators' Posterior Beliefs about Workers in *Evaluator Study* When Gender Is Known

	Evaluators' posterior beliefs	
	Men	Women
	(1)	(2)
Δ	9.31	11.56
	(3.02)	(2.27)
$\Delta \times$ *Attention*	−2.05	1.76
	(4.18)	(3.09)
$\Delta \times$ *Calculation*	−4.31	−5.83
	(4.08)	(3.36)
Observations	507	669
Condition fixed effects	Yes	Yes
Truth(Δ)	1.74	1.74

Notes: Standard errors 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 and are women in column 2. Each column presents an OLS of evaluators' posterior beliefs on (i) suppressed indicators (i.e., Condition fixed effects) 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 ($\Delta \times$ 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)).

pessimistic posterior beliefs about women than female evaluators do. This adds to the evidence on situations in which believed gender differences are not only driven by men (see also Abrevaya and Hamermesh 2012; Babcock et al. 2017; Card et al. 2020; and Exley et al. 2022 for related results), and suggests the fact that evaluators' posteriors disfavor women in our study is not due to gender-specific bias or discrimination by male evaluators.[26]

In addition, online Appendix Table C.4, which reproduces column 5 of Table 3 for various other demographics groups, shows that—regardless of whether we consider evaluators who are split according to their educational attainment, income, age or political affiliation—it is always the case that evaluators hold pessimistic posterior beliefs about women.

C. *Do Our Results Persist for Evaluators Who Expect the Confidence Gap?*

One could worry that our confidence elicitation is complicated or noisy or otherwise does not capture evaluators' true expectations about gender differences in confidence.[27] To provide additional evidence of the confidence gap being expected—and

[26] There is, of course, a vast literature that often shows such an in-group bias does exist (see Tajfel et al. 1979; Chen and Li 2009; Chen and Chen 2011; Ioannou, Qi, and Rustichini 2016; Carlsson and Eriksson 2019 among many others). That said, even when an in-group bias is observed, future work may examine whether this in-group bias is specific to gender per se. Indeed, evidence from Coffman, Exley, and Niederle (2021) reveals that in-group preferences that are specific to gender and in-group preferences that are instead based on arbitrary groups can give rise to a similar pattern of discrimination in hiring decisions.

[27] While we did not directly elicit confidence in one's beliefs, we find that over/underconfidence beliefs typically do not fall at 50 percent (see the histograms in online Appendix Figure B.2), which might have been an indicator of evaluators being entirely unsure about the confidence of men and women.

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 percent of evaluators expect no gender difference in confidence, nearly all of remaining evaluators expect the confidence gap: 51 percent believe that women are less confident and only 3 percent believe that men are less confident. The second question asks specifically about confidence in math and science tasks, and similar results follow: while 42 percent of evaluators expect no gender difference in confidence, 51 percent believe that women are less confident while only 7 percent believe that men are less confident.

Online 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—and our results persist even when we only consider evaluators who directly say that there is a confidence gap. In addition, as shown in online Appendix Tables D.10 and D.11, discussed in online Appendix D.D8, we can show—in a different study version in which we incentivize 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.

D. *Do Our Results Persist for Evaluators Who Think They Accurately Accounted for the Confidence Gap?*

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 confidence in our study. To investigate this, we can turn to data from the following question that we ask in the follow-up survey of the known gender treatments: "When providing your predictions in this study, to what extent were you accounting for any gender differences in confidence?" 63 percent of evaluators answer "neither too little nor too much," 14 percent of evaluators answer "slightly or far too much," and 23 percent of evaluators answer "far or slightly too little."

Online 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 online Appendix Table C.3).

E. *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 percent—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 percent of evaluators have prior beliefs that fall right at 50 percent, as shown in online Appendix Figure B.1). For evaluators in the *Baseline* treatment, online Appendix Figure C.1 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.[28]

V. Robustness

In Section V, to investigate the robustness of our results, we turn to additional study versions.

A. *Are Our Results Robust to Evaluator Beliefs When Asked about Other Types of Performance Outcomes?*

As explained in Section IA, we chose to focus on *one* type of self-evaluation that we called our *main self-evaluation* question. To show that our results are robust to other types of self-evaluation questions—including those that relate to simpler performance outcomes—we ran two additional studies. Specifically, we recruited 400 new evaluators for the *Evaluator* (*Alternative Questions*) *Study* (see online Appendix Section D.D1 for results) and 400 new evaluators for the *Evaluator* (*Attention*, *Top Half*) *Study*.

In the *Evaluator* (*Alternative Questions*) *Study*, in addition to providing beliefs about the likelihood of a worker having poor performance in the manner defined in our main self-evaluation question (see online Appendix Table A.5), evaluators are also asked to provide beliefs about five other self-evaluations questions (see online Appendix Table A.6). This study is otherwise similar to the *Baseline* treatment of the *Evaluator Study*. As shown in online Appendix Table D.1, 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

---

[28] Online Appendix Figure C.2 shows that similar results follow in the *Attention* treatment. Online Appendix Figure C.3 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.

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. Specifically, evaluators' posteriors disfavor women according to two subjective classifications, and indicate that they expect that women are less likely to get 3+ questions right, less likely to get 5+ questions right, and less likely to perform in the top half. Only when asked about the percent chance of participants getting 7+ question right is a gender difference *not* expected—and this lack of a gender difference could reflect very few workers expecting to get 7+ questions right regardless of their gender (i.e., only 10 percent of male workers and 4 percent of female workers expected to get 7+ questions right).

While the *Evaluator* (*Alternative Questions*) *Study* shows the robustness of our results to other self-evaluations questions—including a simpler measure about whether a worker's performance is in the top half among other workers—the fact that we ask evaluators to provide beliefs about six self-evaluation questions could make the *Evaluator* (*Alternative Questions*) *Study* more complex than our main *Evaluator Study* in other ways. Thus, as an additional and important robustness check, we ran the *Evaluator* (*Attention*, *Top Half*) *Study* that *only* asks evaluators to provide beliefs about whether a worker's performance is in the top half among other workers and hence does *not* introduce any complexity by also asking evaluators to provide beliefs about other self-evaluation questions. In addition, motivated by a desire to mitigate the "attention problem" when considering evaluators' posterior beliefs, the *Evaluator* (*Attention*, *Top Half*) *Study* builds off of the *Attention*—rather than the *Baseline* treatment—of the *Evaluator Study*.

As shown in Table 7, even when evaluators are asked about a simple performance metric (i.e., being in the top half) and are in the *Attention* treatment, our results persist: evaluators' priors indicate no gender difference, 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 a large and significant expected performance gap.

B. *Are Our Main Results Robust to Asking Evaluators about the Full Distribution of Workers?*

As explained in Section IC, we chose to focus on asking evaluators about workers with performances in the middle, or in the twenty-fifth to seventy-fifth percentile. To examine whether our results are robust to instead asking evaluators about the full pool of workers, we recruited 400 new evaluators for the *Evaluator* (*Full Distribution*) *Study* (see online Appendix Section D.D2 for results). This study asks evaluators about either a female worker who is randomly selected from the entire pool of female workers who completed the *Worker Study* or a male worker who is randomly selected from the entire pool of male workers who completed the *Worker Study*, but is otherwise identical to the *Baseline* treatment of the *Evaluator Study*. As shown in online Appendix Table D.2, we find that our results persist with similar magnitudes and with statistical significance: when providing beliefs about the full pool of workers, evaluators' priors indicate no gender difference, 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

Table 7—Evaluators' Beliefs in the *Evaluator* (*Attention*, *Top Half*) *Study*

	Prior (1)	Over-confidence (2)	Under-confidence (3)	Implied Bayesian Posterior (4)	Posterior (5)
Panel A. Evaluators' beliefs					
$B(F)$	47.15	38.05	46.42	45.36	40.52
$B(M)$	46.35	48.05	39.96	47.08	46.89
Δ	0.80	−10.00	6.46	−1.72	−6.38
SE of Δ	(1.82)	(2.19)	(2.04)	(1.85)	(1.73)
Panel B. Evaluators' beliefs − truth					
$B(F) - Truth(F)$	0.81	24.41	−13.23	−0.98	−5.82
$B(M) - Truth(M)$	−2.06	15.74	0.62	−1.33	−1.52
$\Delta - Truth(\Delta)$	2.87	8.67	−13.85	0.35	−4.31
SE of $\Delta - Truth(\Delta)$	(1.82)	(2.19)	(2.04)	(1.85)	(1.73)
Observations	400	400	400	395	400
$Truth(F)$	46.34	13.64	59.65	46.34	46.34
$Truth(M)$	48.41	32.31	39.34	48.41	48.41
$Truth(\Delta)$	−2.07	−18.67	20.30	−2.07	−2.07

Notes: Standard errors are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 400 participants in the *Evaluator* (*Attention*, *Top Half*) *Study*. Note that the being in the "top half" meant that your score was greater than or *equal to* the scores of at least 50 percent of other participants, and these other participants are 50 randomly selected men and 50 randomly selected women.

performance gap, but their posteriors indicate a large and significant expected performance gap.

C. *Are Our Results Robust to Evaluators with Hiring and Managerial Experience Providing Beliefs about Typical Job Candidates?*

To investigate whether evaluators could better account 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, we ran two additional studies. Specifically, we recruited 409 new evaluators for the *Baseline* treatment of the *Evaluator* (*Professional Evaluators*) *Study*, and 391 new evaluators for the *Baseline*, *Unknown Gender* treatment of the *Evaluator* (*Professional Evaluators*) *Study* (see online Appendix Section D.D4 for results). These studies are similar to the *Baseline* and *Baseline*, *Unknown Gender* treatments of the *Evaluator Study* aside from recruitment details. Specifically, the evaluators in these studies were recruited so that—according to self-reported data—they met the following two criteria: (i) they have experience in making hiring decisions (i.e., have been responsible for hiring job candidates) and (ii) they have experience in a management position.[29] 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

---

[29] 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 percent of these participants responded "yes" when asked a different but similar

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. We thus also recruited 354 undergraduate students through Ohio State University for our *Worker* (*Undergraduates*) *Study* (see online Appendix Section D.D3 for results).

Following Table 1, online Appendix Table D.3 presents the results for these undergraduate students and confirms that the confidence gap persists for them.

Following Table 2, online Appendix Table D.4 presents the results for these "professional" evaluators in the *Baseline* treatment. Our main findings persist: these professional evaluators' priors indicate no gender difference, they 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 a large and significant expected performance gap.

Online Appendix Table D.5 presents parallel results for these professional evaluators in the *Baseline*, *Unknown Gender* treatment. The results in the *Baseline*, *Unknown Gender* treatment persist to a similar degree, showing that—even for professional evaluators and even when it is a calculation problem that cannot reflect discriminatory motives—the confidence gap is contagious.[30]

D. *Are Our Results Robust to Evaluators Gaining More Experience with Worker Self-Evaluations?*

To investigate whether evaluators could better account for the confidence gap if they had more experience with the exact type of self-evaluations in our study, we recruited 406 new evaluators for the *Baseline* treatment of the *Evaluator* (*Extended*) *Study* (see online Appendix Section D.D5 for results).

The *Baseline* treatment of the *Evaluator* (*Extended*) *Study* is similar to the *Baseline* treatment of the *Evaluator Study* except for the "experience" stage. Specifically, after providing their prior beliefs—but before providing their posterior beliefs—evaluators gain "experience" by providing 20 beliefs about specific workers after learning each worker's self-evaluation. For each of these 20 specific workers, evaluators are informed of the specific worker's reported percent chance of having a poor performance and then are asked to report a belief about the percent chance of that specific worker having a poor performance (see supplemental online Appendix Figure I.5.6 for an example).[31] As shown in online Appendix Table D.6, gaining experience with self-evaluations does not help evaluators to better account for the gender gap in confidence: experienced evaluators' priors indicate no gender difference, they expect that male workers are more likely to be overconfident and female workers are more likely to be underconfident, their implied Bayesian beliefs imply a small expected performance gap that does not differ from the truth, but their posteriors indicate a large and significant expected performance gap.

---

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?"

[30] The expected gender gap in performance according to evaluators' posterior beliefs—i.e., the estimates on Δ in columns 5 of online Appendix Tables D.4 and D.5—are statistically indistinguishable ($p > 0.1$).

[31] 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*.

E. *Are Our Results Robust to Beliefs about Specific Workers?*

To investigate whether 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 20 worker-specific beliefs that evaluators provide in the experience stage of the *Baseline* treatment of the *Evaluator* (*Extended*) *Study*, described above in Section VD (again, see online Appendix Section D.D5 for results).

As shown in the northeastern region of online Appendix Figure D.1, 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 percent chance of having a poor performance in their self-evaluation, the average evaluator believes there is a 74 percent chance of that worker having a poor performance if the worker is a man but only a 70 percent chance of that worker having a poor performance if that worker is a woman. Nonetheless, online Appendix Table D.7 shows that—even when asked about specific workers—evaluators' posteriors indicate a large and significant expected performance gap.

F. *Do Our Results Persist When Workers Face Strategic Incentives?*

To investigate whether our results are robust to workers having strategic incentives that may encourage them to inflate their self-evaluations to potential evaluators—in a manner that is akin those in Exley and Kessler (2022) and as is the case in many settings outside of the laboratory—we ran two additional studies. Specifically, we recruited 387 new workers for the *Strategic Incentives* treatment of the *Worker Study* (see online Appendix Section D.D6 for results) and 394 new evaluators for the *Strategic Incentives* treatment of the *Evaluator* (*Extended*) *Study* (see online Appendix Section D.D7 for results).

In the *Strategic Incentives* treatment of the *Worker Study*, workers face strategic incentives because they earn more money if they are hired by an "employer" who learns their self evaluation. This study is otherwise similar to the *Worker Study*. Following Table 1, online Appendix Table D.8 presents the results for these workers and confirms that the confidence gap persists for them.

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. This study is otherwise similar to the *Baseline* treatment of the *Evaluator* (*Extended*) *Study*. Following Table 2, online Appendix Table D.9 presents the results for these evaluators and shows that our results persist: even when workers face strategic incentives, evaluators' priors indicate no gender difference, they 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 performance gap, but their posteriors indicate a large and significant expected performance gap.

G. *Are Our Results Robust to Being Asked about Both Men and Women?*

To investigate whether evaluators could better account for the confidence gap if they were making explicit comparisons between male and female workers—inspired

by prior work that suggests judgments are less reasoned when comparison information is lacking (Bohnet, van Geen, and Bazerman 2016)—we ran two additional studies. Specifically, we recruited 205 new evaluators for the *Joint Evaluations* treatment of the *Evaluator* (*Extended*) *Study* (see online Appendix Section D.D8 for results) and 195 new evaluators for the *Joint Evaluations*, *Strategic Incentives* treatment of the *Evaluator* (*Extended*) *Study* (see online Appendix Section D.D8 for results). In these treatments, we asked evaluators to provide beliefs about a male worker and a female worker *on the same decision screen*. These studies are otherwise similar to the *Baseline* treatment and *Strategic Incentives* treatment of the *Evaluator* (*Extended*) *Study*, respectively.

Following the same specifications as those in Table 2, online Appendix Table D.10 presents results for evaluators in the *Joint Evaluations* treatment and online Appendix Table D.11 presents results for evaluators in the *Joint Evaluations*, *Strategic Incentives* treatment. In both cases, our results persist: when providing joint evaluations—for workers who faced strategic incentives or otherwise—evaluators' priors indicate little to no gender differences, they expect that male workers are more likely to be overconfident and female workers are more likely to be underconfident, their implied Bayesian beliefs imply little to no expected performance gap, but their posteriors indicate a large and significant expected performance gap.

H. *Are Our Results Robust to Considering Evaluators' Beliefs at the Individual Level?*

Since the evaluators from our main *Evaluator Study* provide beliefs about only male or female workers as discussed in Section IC, those 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 and *Joint Evaluations*, *Strategic Incentives* treatment allow for such classifications.

Figure 2 presents the results from the *Joint Evaluations* treatment. 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 likely to have a poor performance is 39 percent, 43 percent, and 18 percent 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 percent, 11 percent, and 10 percent respectively. Thus, even when considering the individual-level results, the confidence gap is contagious. See online Appendix Figure D.2 for similar results from the *Joint Evaluations*, *Strategic Incentives* treatment.

I. *Are Our Results Robust to Conveying Gender More Subtly?*

To investigate whether our results are robust to conveying gender more subtly—and relatedly, to ensure that our results are not driven by experimenter demand
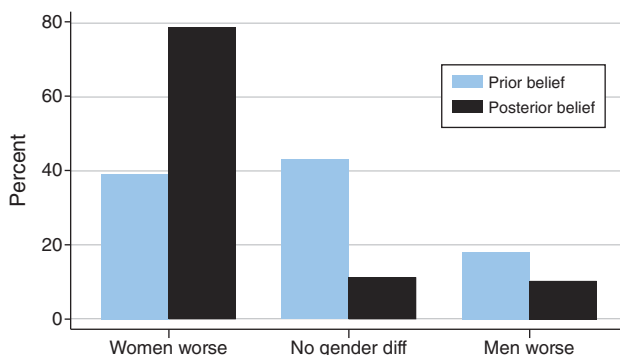
FIGURE 2. *JOINT EVALUATIONS TREATMENT:* CLASSIFYING EVALUATORS ACCORDING TO THEIR BELIEFS

Notes: 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, middle, and right pair of bars, respectively. Data are from the *Joint Evaluations* treatment of the *Evaluator (Extended) Study*.

effect or social desirability bias—we recruited 198 new evaluators for the *Evaluator (Additional Demographics) Study* (see online Appendix Section D.D9 for results). Specifically, in this study, we tell 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.[32] This study is otherwise similar to the *Baseline* treatment of the *Evaluator Study*.

Following Table 2, online Appendix Table D.12 presents results for the evaluators in the *Evaluator (Additional Demographics) Study* and shows that our results persist: evaluators' priors indicate no gender difference, they (directionally) 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 a large and significant expected performance gap.

J. *Are Our Results Robust to Situations Where Absolute Performance Information Is Known about the Workers?*

To investigate the robustness of our results to a situation where absolute performance information is known, we recruited 198 new evaluators for the *Evaluator (Known Performance) Study* (see online Appendix Section D.D10 for results). Specifically, in this study, we tell evaluators 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. This study is otherwise similar to the *Baseline* treatment of the *Evaluator Study*.

---

[32] These demographics were modal in the *Worker Study*, with modal age being the modal generation. When comparing these groups of workers, male workers have a 43 percent chance of poor performance compared to 35 percent for female workers ($p = 0.51$); nevertheless, as with our prior result, these female workers report significantly more pessimistic self-evaluations: 77 percent of female workers in this group believe they have a poor performance while only 38 percent of male workers do.

Following Table 2, online Appendix Table D.13 presents the results from the *Evaluator* (*Known Performance*) *Study* and shows that our results persist: evaluators' priors indicate no gender difference, they (directionally) 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 a large and significant expected performance gap.

We make one additional note on these results. One might wonder why evaluators in the *Evaluator* (*Known Performance*) *Study* update *at all* in response to the provided self-evaluation information about whether a worker believes they have a poor performance, given that evaluators know the worker got 5 questions right on the math and science test. If evaluators had a stable mapping from absolute performance outcomes (i.e., the number of questions the worker got right) to the subjective performance outcome of interest (i.e., whether the worker has a poor performance), then learning workers' self-evaluation information would not affect evaluators' beliefs. Alternatively, if evaluators' beliefs about subjective performance outcomes depend on more than the number of questions the worker got right—e.g., they also are influenced by others' views about what constitutes a poor performance—then such a stable mapping may not exist. That there need not be a stable mapping between absolute and subjective performance outcomes is one of the reasons we chose to focus on a subjective performance outcome in our main self-evaluation question. Ultimately, because of this, we can show that the confidence gap is contagious—causing overly pessimistic beliefs about women relative to men—even when women and men are known to have answered exactly the same number of questions right. More generally, since many factors could influence individuals' views of subjective performance outcomes (including absolute performance outcomes, other objective performance outcomes, other subjective criteria such as those relating to one's standards, or even confusion), these results suggest that the contagious confidence gap may arise even in situations in which ample information on a worker is known. We leave further investigation of this to future work.

VI. Conclusion

Through a series of experiments in which evaluators are incentivized to provide accurate beliefs, 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 about their performance on a math and science test—results in overly pessimistic beliefs about women relative to men. This "contagious" confidence gap arises even though it should not have if evaluators were Bayesian and even though the confidence gap is expected. Additional results support the interpretation of this contagious confidence gap reflecting more of a calculation problem, rather than an awareness or attention problem.

We see many important avenues for future work, four of which we mention here. One stream of future work may investigate ways to counter the contagious confidence gap, particularly since the confidence gap may be conveyed via self-evaluations in hiring, promotion, and pay decisions. Given the ineffectiveness of our *Attention* treatment and the fact that individuals expect the confidence gap, our results highlight how awareness or attention need not be sufficient. That said, future work may

reveal more effective attention interventions, which likely depend on the salience of the attention intervention and the context itself (e.g., an attention intervention in which employers reviewing job candidates view a pop-up window that says "remember that women are typically underconfident" may prove more effective than an attention intervention that elicits related beliefs as in our *Attention* treatment).[33] In addition, motivated by our results reflecting a calculation problem—notably including the effectiveness of the *Calculation* treatment and that similar results arise when evaluators are asked about arbitrary rather than gender-specific groups—our results suggest that a confidence gap, resulting from gender or other group differences, can cause non-Bayesian agents to be biased even absent any explicit discriminatory motives. This lends particular promise to strategies that help individuals overcome cognitive limitations, even when these limitations are not directly related to factors such as gender. For instance, while evaluators did not receive feedback in our experiment, future work may test the effectiveness of allowing evaluators to *learn* about their biases via iterative feedback. Also, given the positive correlation between the extent to which evaluators' posteriors disfavor women and the extent to which they exhibit base rate neglect, future work may test the effectiveness of strategies that build off of insights from the broader literature on cognitive limitations and behavioral biases.[34] When strategies are uncovered that do work, important questions will also relate to the *relative* effectiveness of these strategies and the conditions under which they work.

A second stream of future work may explore the impact of removing gender information from applications and various types of evaluations (see Kolev, Fuentes-Medel, and Murray 2019). 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, Babcock, and Lai 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.

A third stream of future work may explore how others form beliefs about men and women in settings in which the size, and magnitude, of the confidence gap—and the actual performance gap—vary due to any number of factors such as the relevance of stereotypes and the selection of individuals involved. Such future work may reveal: more situations in which the confidence gap is contagious and causes overly pessimistic beliefs about women relative to men, situations in which the confidence gap does not exist, and even situations in which the confidence gap causes overly *optimistic* beliefs about women relative to men. To better understand these situations, however, we hope future work considers how men and women's beliefs about

---

[33] The effectiveness of attention interventions could be different in settings involving more free-form communication (Coffman, Flikkema, and Shurchkov 2019) or that require updating from the lack of information (Enke 2020; Charness, Oprea, and Yuksel 2022; Agan, Cowgill, and Gee 2023). Also, the effect of the intervention could depend on the gender of individuals selecting into the context, see Exley, Niederle, and Vesterlund (2020) for an example of how selection influences when it is a good idea to negotiate.

[34] As an example of this in the broader literature on non-Bayesian updating, Gonçalves, Libgober, and Willis (2021) show that individuals can fail to "unlearn" from signals that are retracted. It would be interesting to see how these results carry over to an environment with self-evaluations; that is, if evaluators update from self-evaluations, but then are *told* that the self-evaluations are biased, do evaluators sufficiently unlearn the self-evaluation?

themselves—often communicated through self-evaluations—affect the beliefs that others hold about men and women.

A fourth stream of future work may examine how these results extend beyond gender and to other biases. On extending beyond gender, since the confidence gap is contagious even when evaluators are asked about arbitrary rather than gender-specific groups, future work may naturally investigate whether similar results follow whenever individuals are asked about two groups as long as one of those groups has lower confidence. On extending to other biases, future work may investigate whether biases—driven by different sources than the confidence gap—also result in similar findings. For instance, future work may explore whether individuals expect that certain groups face discrimination but nonetheless fail to account for discrimination when evaluating those groups. This future work may also explore whether expecting a bias 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 IVD, 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

Abrevaya, Jason, and Daniel S. Hamermesh. 2012. "Charity and Favoritism in the Field: Are Female Economists Nicer (To Each Other)?" *Review of Economics and Statistics* 94 (1): 202–207.

Agan, Amanda Y., Bo Cowgill, and Laura Gee. 2023. "Salary History and Employer Demand: Evidence from a Two-Sided Audit." Unpublished.

Augenblick, Ned, Eben Lazarus, and Michael Thaler. 2023. "Overinference from Weak Signals and Underinference from Strong Signals." Unpublished.

Ba, Cuimin, J. Aislinn Bohren, and Alex Imas. 2023. "Over- and Underreaction to Information." Unpublished.

Babcock, Linda, Maria P. Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability." *American Economic Review* 107 (3): 714–47.

Benjamin, Daniel J. 2019. "Errors in Probabilistic Reasoning and Judgment Biases." In *Handbook of Behavioral Economics*, edited by B. Douglas Bernheim, Stefano Dellavigna and David Laibson, 69–186. Amsterdam: Elsevier Press.

Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz. 2010. "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors." *American Economic Journal: Applied Economics* 2 (3): 228–55.

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

Biasi, Barbara, and Heather Sarsons. Forthcoming. "Flexible Pay, Bargaining, and The Gender Gap." *Quarterly Journal of Economics*.

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends. and Explanations." *Journal of Economic Literature* 55 (3): 789–865.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2016. "When Performance Trumps Gender Bias: Joint vs. Separate Evaluation." *Management Science* 62 (5): 1225–34.

Bohnet, Iris, Oliver P. Hauser, and Ariella Kristal. 2022. "Can Gender and Race Dynamics in Performance Appraisals Be Disrupted? The Case of Anchoring." Unpublished

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination." Unpublished.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109 (10): 3395–3436.

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

Bowles, Hannah Riley, Linda Babcock, and Lei 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 (1): 84–103.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices." *Quarterly Journal of Economics* 129 (3): 1409–47.

Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Irriberri. 2020. "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics* 135 (1): 269–327.

Carlsson, Magnus, and Stefan Eriksson. 2019. "In-Group Gender Bias in Hiring: Real-World Evidence." *Economics Letters* 185: 108686.

Charness, Gary, Ryan Oprea, and Sevgi Yuksel. 2022. "How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment." *Journal of the European Economic Association* 19 (3): 1656–91.

Chen, Yan, and Roy Chen. 2011. "The Potential of Social Identity for Equilibrium Selection." *American Economic Review* 101 (6): 2562–89.

Chen, Yan, and Sherry X. Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99 (1): 431–57.

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *Quarterly Journal of Economics* 129 (4): 1625–60.

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov. 2019. "Gender Stereotypes in Deliberation and Team Decisions." Unpublished.

Coffman, Katherine B., Manuela R. Collis, and Leena Kulkarni. 2019a. "When to Apply?" Unpublished.

Coffman, Katherine, Manuela Collis, and Leena Kulkarni. 2019b. "Stereotypes and Belief Updating." Unpublished.

Coffman, Katherine B., Christine L. Exley, and Muriel Niederle. 2021. "The Role of Beliefs in Driving Gender Discrimination." *Management Science* 67 (6): 3551–69.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–74.

DellaVigna, Stefano, and Devin Pope. 2018a. "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy* 126 (6): 2410–56.

DellaVigna, Stefano, and Devin Pope. 2018b. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

Edwards, Ward. 1968. "Conservatism in Human Information Processing." In *Formal Representation of Human Judgment,* edited by Benjamin Kleinmuntz.

Enke, Benjamin. 2020. "What You See Is All There Is." *Quarterly Journal of Economics* 135 (3): 1363–98.

Enke, Benjamin, and Florian Zimmermann. 2019. "Correlation Neglect in Belief Formation." *Review of Economic Studies* 86 (1): 313–32.

Enke, Benjamin, and Thomas Graeber. 2023. "Cognitive Uncertainty." *Quarterly Journal of Economics* 138 (4): 2021–67.

Erkal, Nisvan, Lata Gangadharan, and Boon Han Koh. 2021. "Gender Biases in Performance Evaluation: The Role of Beliefs versus Outcomes." Unpublished.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel. 2024. "Mental Models and Learning: The Case of Base-Rate Neglect." *American Economic Review* 114 (3): 752–83.

Exley, Christine, Oliver P. Hauser, Molly Moore, and John-Henry Pezzuto. 2022. "Beliefs about Gender Differences in Social Preferences." Unpublished.

Exley, Christine L., and Judd B. Kessler. 2022. "The Gender Gap in Self-Promotion." *Quarterly Journal of Economics* 137 (3): 1345–81.

Exley, Christine L., Muriel Niederle, and Lise Vesterlund. 2020. "Knowing When to Ask: The Cost of Leaning-in." *Journal of Political Economy* 128 (3): 816–54.

Exley, Christine L., and Kirby Nielsen. 2024. "Replication data for: The Gender Gap in Confidence: Expected but Not Accounted For." American Economic Association [Publisher], Inter-University Consortium for Political and Social Research [Distributor] https://doi.org/10.3886/E194446V1.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19 (4): 25–42.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104 (4): 1091–1119.

Gonçalves, Duart, Jonathan Libgober, and Jack Willis. 2021. "Learning versus Unlearning: An Experiment on Retractions." NBER Working Paper 29512.

Grether, David M. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *Quarterly Journal of Economics* 95 (3): 537–57.

Grossman, Philip J., Catherine Eckel, Mana Komai, and Wei Zhan. 2019. "It Pays to be a Man: Rewards for Leaders in a Coordination Gam." *Journal of Economic Behavior and Organization* 161: 197–215.

Hernandez-Arenaz, Iñigo, and Nagore Iriberri. 2019. "A Review of Gender Differences in Negotiation." *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/acref ore/9780190625979.013.464.

Huber, Christoph, and Jürgen Huber. 2020. "Bad Bankers No More? Truth-Telling and (Dis) Honesty in the Finance Industry." *Journal of Economic Behavior and Organization* 180: 472–93.

Ioannou, Christos A., Shi Qi, and Aldo Rustichini. 2015. "Group Payoffs as Public Signals." *Journal of Economic Psychology* 48: 89–105.

Kahneman, Daniel, and Amos Tversky. 1972a. "On Prediction and Judgment." *ORI Research Monograph* 12.

Kahneman, Daniel, and Amos Tversky. 1972b. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3 (3): 430–54.

Kahneman, Daniel, and Amos Tversky. 1973. "On the Psychology of Prediction." *Psychological Review* 80 (4): 237–51.

Koehler, Jonathan J. 1996. "The Base Rate Fallacy Reconsidered: Descriptive, Normative, and Methodological Challenges." *Behavioral and Brain Sciences* 19 (1): 1–53.

Kolev, Julian, Yuly Fuentes-Medel, and Fiona Murray. 2019. "Is Blinded Review Enough? How Gendered Outcomes Arise Even Under Anonymous Evaluation." *Academy of Management Proceedings* 1.

Lundeberg, Mary A., Paul W. Fox, and Judith Punćcohaŕ. 1994. "Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments." *Journal of Educational Psychology* 86 (1): 114–21.

Michelmore, Katherine, and Sharon 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 (4): 194–215.

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2022. "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science* 68 (11): 7793–7817.

Murciano-Goroff, Raviv. 2021. "Missing Women in Tech: The Labor Market for Highly Skilled Software Engineers." *Management Science* 68 (5): 3262–81.

Niederle, Muriel. 2016. "Gender." In *Handbook of Experimental Economics*. Vol. 2, edited by John Kagel and Alvin E. Roth, 481–553. Princeton, NJ: Princeton University Press.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3): 1067–1101.

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and Competition." *Annual Review of Economics* 3 (1): 601–30.

Reuben, Ernesto, Pedro Rey-Biel, Paola Sapienza, and Luigi Zingales. 2012. "The Emergence of Male Leadership in Competitive Environments." *Journal of Economic Behavior and Organization* 83 (1): 111–17.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences* 111 (12): 4403–4408.

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. 2017. "Preferences and Biases in Educational Choices and Labor Market Expectations: Shrinking the Black Box of Gender." *Economic Journal* 604 (127): 2153–86.

Riach, P. A., and J. Rich. 2002. "Field Experiments of Discrimination in the Market Place." *Economic Journal* 112 (483): 480–518.

Roussille, Nina. 2021. "The Central Role of the Ask Gap in Gender Pay Inequality." Unpublished.

Rudman, Laurie A., and Julie E. Phelan. 2008. "Backlash Effects for Disconfirming Gender Stereotypes in Organizations." *Research in Organizational Behavior* 28: 61–79.

Rudman, Laurie A., and Kimberly Fairchild. 2004. "Reactions to Counterstereotypic Behavior: The Role of Backlash in Cultural Stereotype Maintenance." *Journal of Personality and Social Psychology* 87 (2): 157–76.

Saccardo, Silvia, and Marta Serra-Garcia. 2022. "Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment." Unpublished.

Sarsons, Heather, and Xu Guo. 2021. "Confidence Men? Evidence on Confidence and Gender among Top Economists." *American Economic Association Papers and Proceedings* 111: 65–68.

Tajfel, Henri, John C. Turner, William G. Austin, and Stephen Worchel. 1979. "An Integrative Theory of Intergroup Conflict." In *Organizational Identity: A Reader*, edited by Mary Jo Hatch and Majken Schultz, 56–65. Oxford University Press.

Online Appendix for
The Gender Gap in Confidence:
Expected But Not Accounted For

By: Christine L. Exley and Kirby Nielsen

Appendix Table of Contents

A   Additional Design Details

Figure A.1: Timeline of *Baseline* and *Baseline, Unknown Gender* treatments of the *Evaluator Study*

```
┌──────────────┐   ┌──────────────┐   ┌──────────────┐   ┌──────────────────┐
│ Prior Belief │   │ Receive Info │   │  Posterior   │   │ Overconfidence/  │
│ that worker  │→  │ of workers'  │→  │   Belief     │→  │ Underconfidence  │
│  had poor    │   │  own belief  │   │ that worker  │   │     Beliefs      │
│ performance  │   │  of poor     │   │  had poor    │   │ belief of info   │
│              │   │ performance  │   │ performance, │   │    accuracy      │
└──────────────┘   └──────────────┘   │  given info  │   └──────────────────┘
                                       └──────────────┘
```

Implied Bayesian Posterior Belief
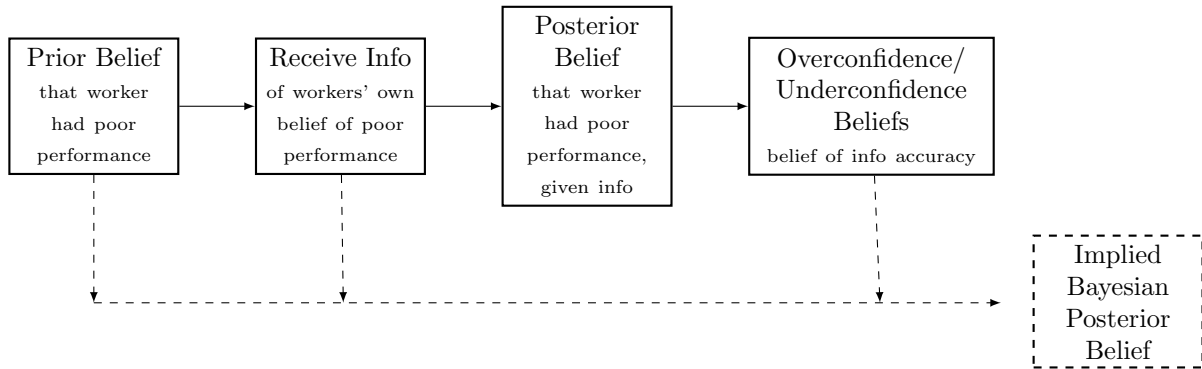
In the *Baseline* and *Baseline, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, we 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.2:** Timeline of *Attention* and *Attention, Unknown Gender* treatments of the *Evaluator Study*
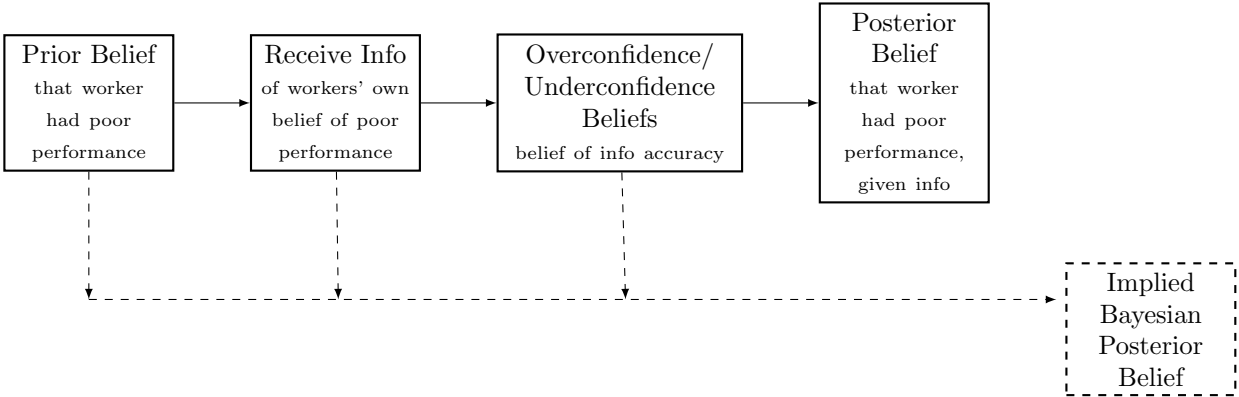
In the *Attention* and *Attention, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, we 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.3: Timeline of *Calculation* and *Calculation, Unknown Gender* treatments of the *Evaluator Study*
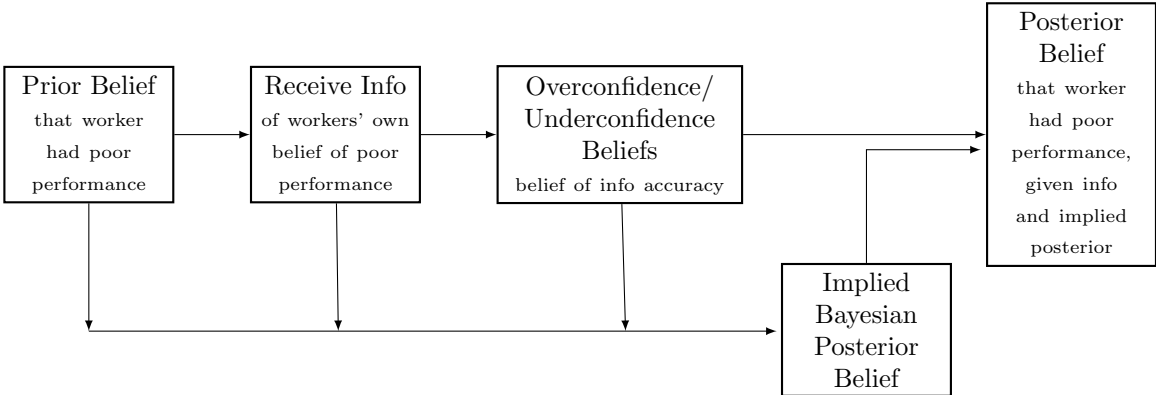
In the *Calculation* and *Calculation, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, 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: 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.4	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 6.6
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 guess they make in Part 1, they have a chance of earning \$10 for each guess 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.4 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 6.3

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.

Table A.2: Overview of *The Evaluator Study* Treatments

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.3: 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 6.3
Evaluator (Professional Evaluators) Study – Baseline, Unknown Gender Treatment	Same as the Evaluator (Professional Evaluators) Study – Baseline Treatment except the gender of workers is unknown	N=391, September 2022	Section 6.3
Evaluator (Extended) Study – Baseline Treatment	Same as Evaluator Study – Baseline Treatment except that, before providing posterior belief, evaluators provide 20 beliefs about specific workers after learning each of those workers' self-evaluations	N=406, May 2022	Sections 6.4 6.5
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 6.6
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 6.7
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 6.7
Evaluator (Alternative Questions) Study	Same as Evaluator Study – Baseline Treatment except that, rather than only answering the belief questions in Appendix Table A.5, evaluators also answer the belief questions in Appendix Table A.6	N=400, May 2022	Section 6.1
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 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 6.9
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 6.10
Evaluator (Attention, Top Half) Study	Same as Evaluator Study – Attention Treatment except that, rather than answering the belief questions in Appendix Table A.5, evaluators answer Prior (top half), Over/underconfidence (Top Half), and Posterior (Top half) from Appendix Table A.6	N=400, March 2023	Section 6.1
Evaluator (Full Distribution) Study	Same as Evaluator Study – Baseline Treatment except that, rather than providing beliefs about male or female workers with performances in the middle, evaluators provide beliefs about all male or female workers	N=400, March 2023	Section 6.2

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.4: 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. Our main self-evaluation question corresponds to self-evaluation 8B.

Table A.5: 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.4), 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.6: 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 half)	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.

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

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.4 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*. Our main self-evaluation question corresponds to self-evaluation 8B.

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

DV:	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
SE of Δ	(1.83)	(2.20)	(2.08)	(1.80)	(1.73)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-7.12	27.34	-22.03	-5.84	9.39
B(M) - Truth(M)	-8.80	8.23	-9.21	-7.64	0.28
Δ - Truth(Δ)	1.67	19.11	-12.82	1.80	9.11
SE of Δ - Truth(Δ)	(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

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*

DV:	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
SE of Δ	(1.82)	(2.27)	(1.98)	(1.75)	(1.81)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-7.81	24.35	-19.74	-7.05	-1.47
B(M) - Truth(M)	-9.14	10.06	-8.81	-8.42	-4.64
Δ - Truth(Δ)	1.33	14.29	-10.93	1.37	3.18
SE of Δ - Truth(Δ)	(1.82)	(2.27)	(1.98)	(1.75)	(1.81)
N	405	405	405	404	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

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*

DV:	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
SE of Δ	(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
SE of Δ - Truth(Δ)	(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

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*

DV:	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
SE of Δ	(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
SE of Δ - Truth(Δ)	(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

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*

DV:	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
SE of Δ	(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
SE of Δ - Truth(Δ)	(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

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*

| DV:	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

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.

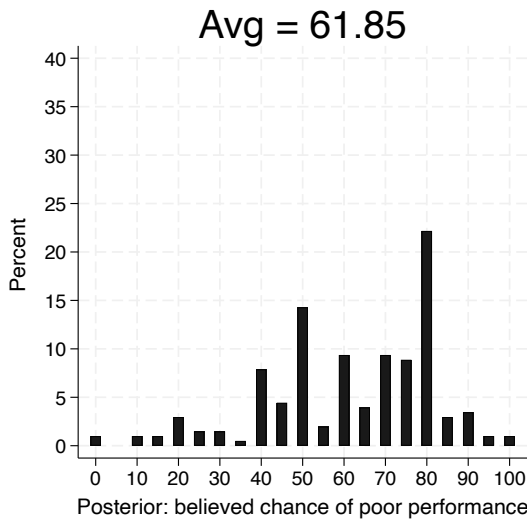**Figure B.1:** *Baseline Treatment*: Prior and Posterior Beliefs
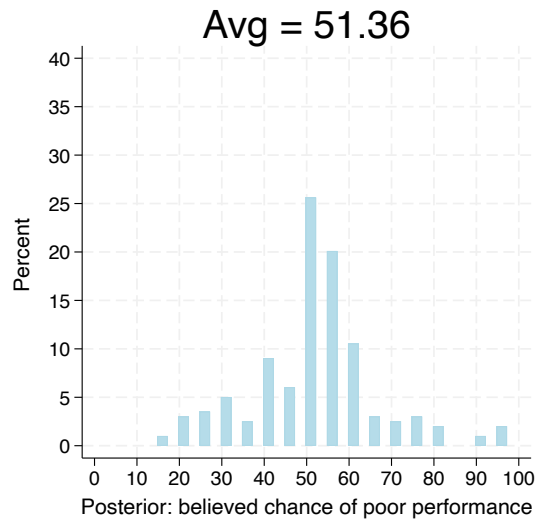
A: Prior Beliefs about Women
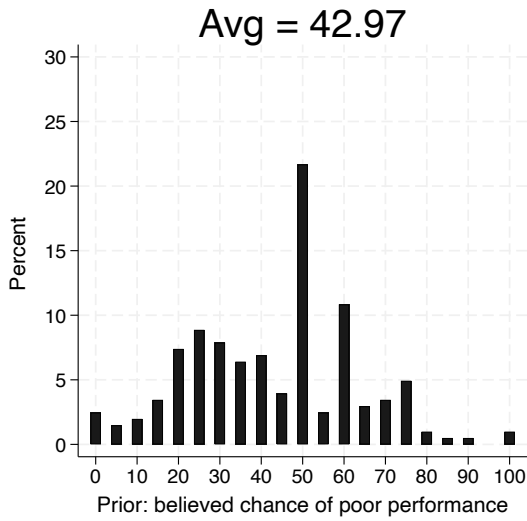
B: Prior Beliefs about Men
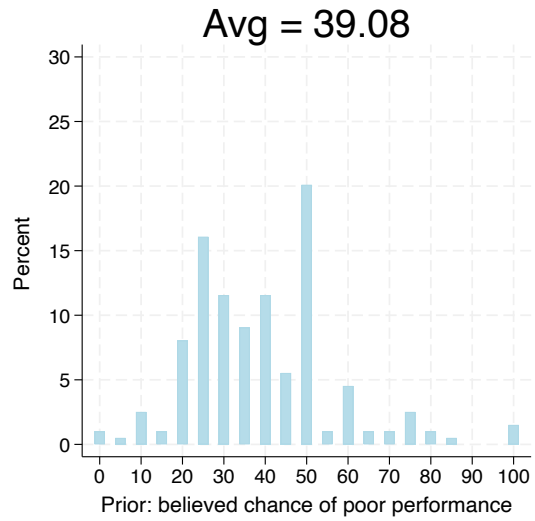
C: Posterior Beliefs about Women

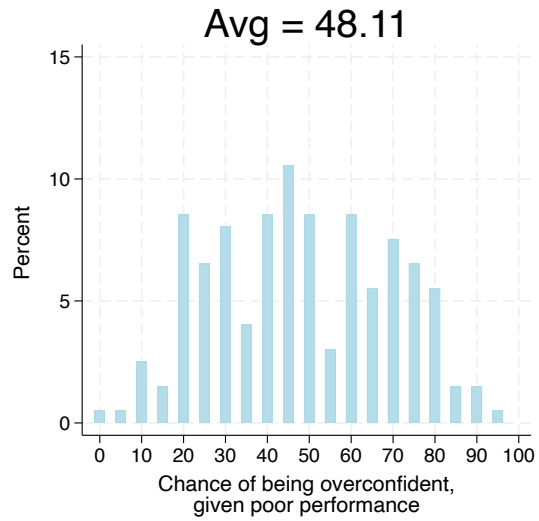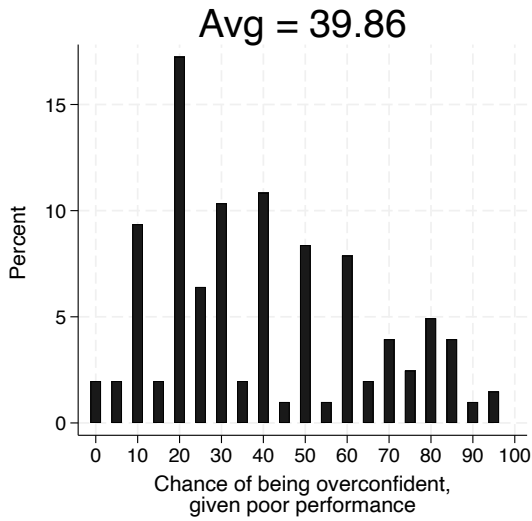**D**: Posterior Beliefs about Men

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

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

A: Overconfidence Beliefs about Women

Avg = 39.86

B: Overconfidence Beliefs about Men

Avg = 48.11

C: Underconfidence Beliefs about Women

Avg = 55.68

D: Underconfidence Beliefs about Men

Avg = 45.61
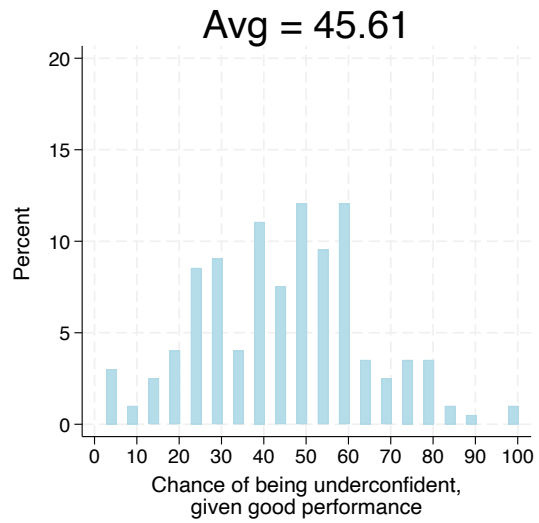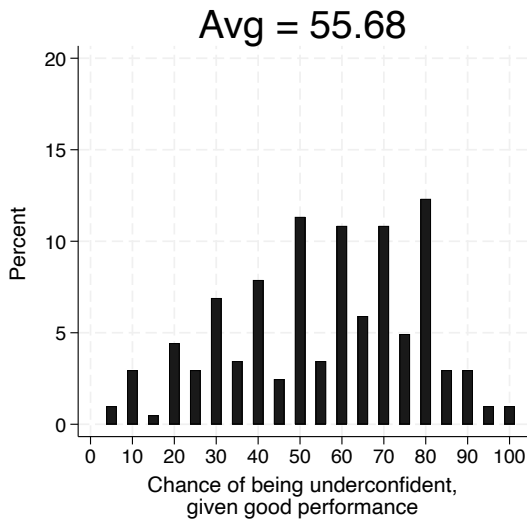
Data are from the *Baseline* treatment of the *Evaluator Study*.

C   Additional Heterogeneity Results

Figure C.1: *Baseline Treatment*: Posterior Beliefs as a Function of Their Other Beliefs
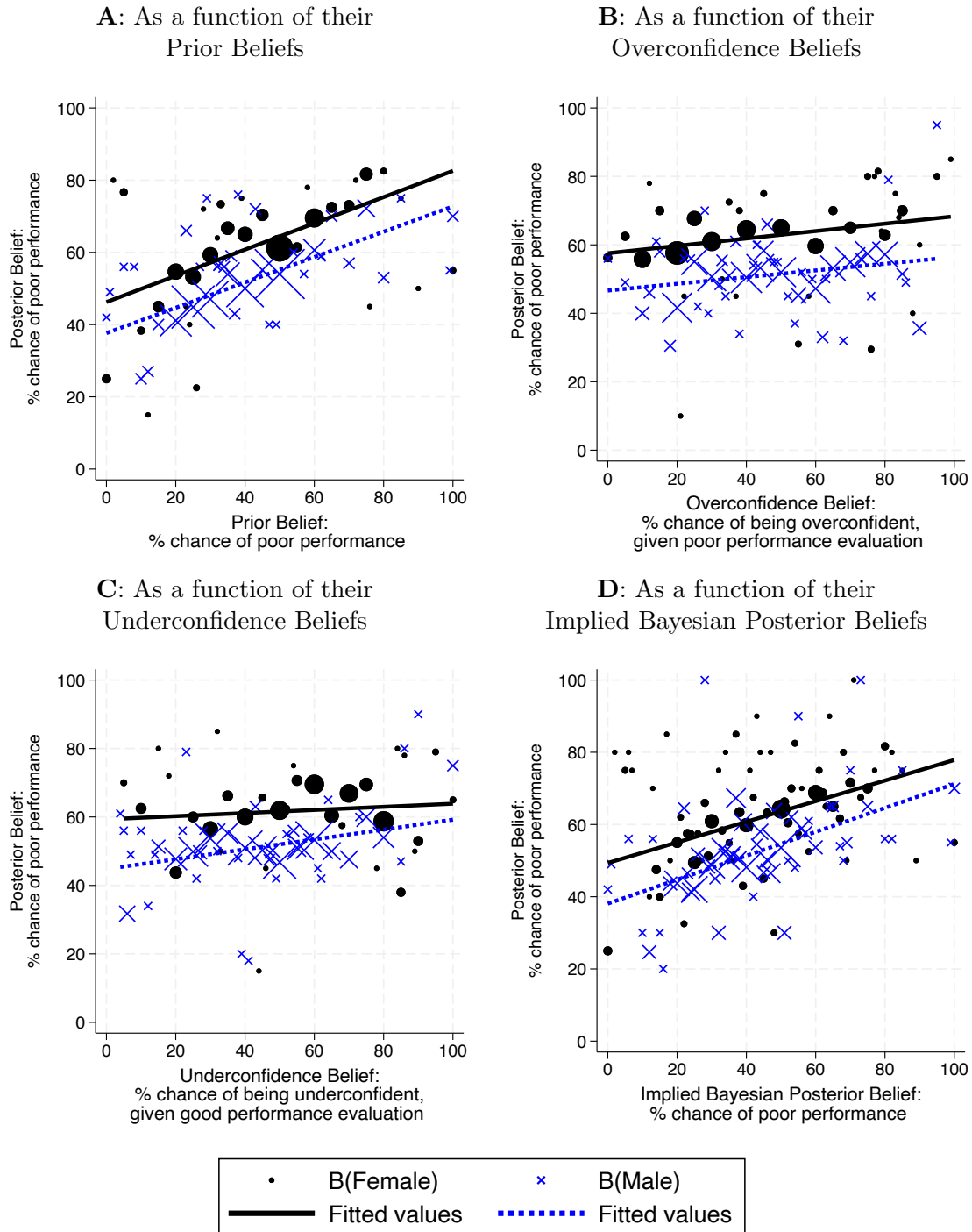
Graphs show a scatter plot (dots weighted by sample size) of evaluators' posterior beliefs as a function of their beliefs noted on the horizontal axis. Data are from the *Baseline* treatment of the *Evaluator Study*.

Figure C.2: *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

•	B(Female)	×	B(Male)
——	Fitted values	······	Fitted values
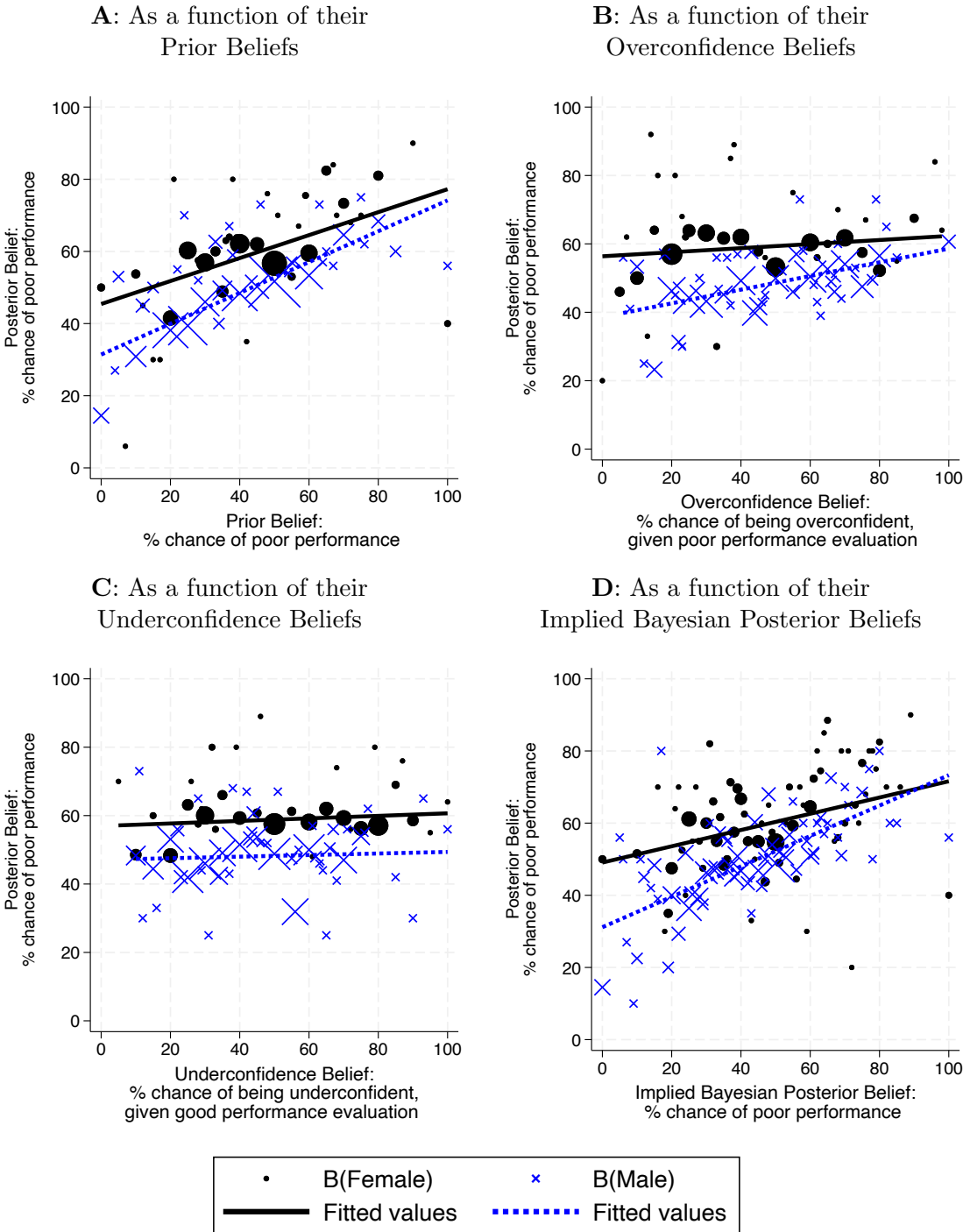
See Figure C.1 for a description of the graphs above. Data are from the *Attention* treatment of the *Evaluator Study*.

Figure C.3: *Calculation 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
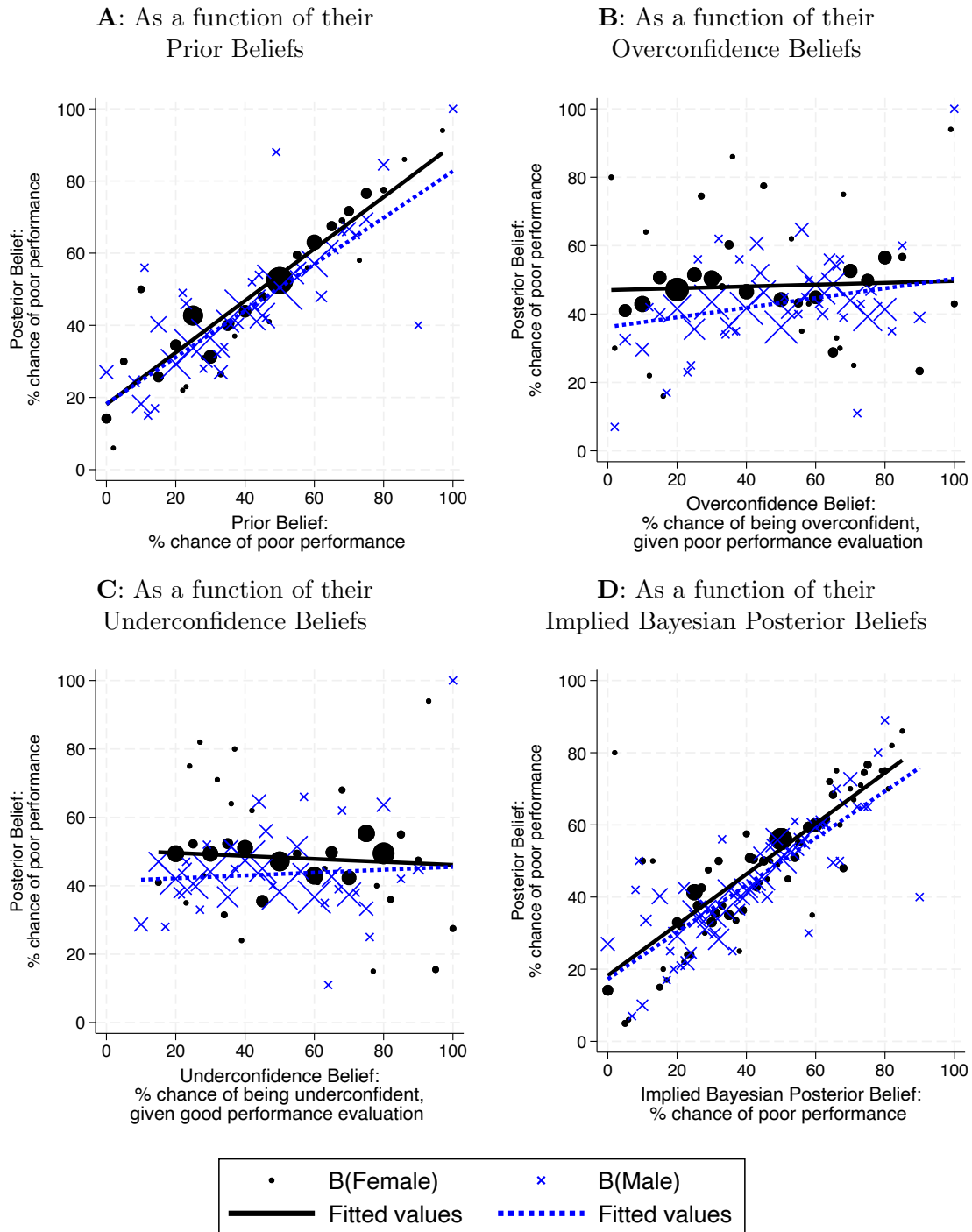
See Figure C.1 for a description of the graphs above. Data are from the *Calculation* treatment of the *Evaluator Study*.

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

	DV: Evaluators' Posterior Beliefs					
	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

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 6.

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

	DV: Evaluators' Posterior Beliefs		
	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

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 6.

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

	DV: Evaluators' Posterior Beliefs		
	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

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 6.

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

	Low Education (1)	High Education (2)	Low Income (3)	High Income (4)	Younger (5)	Older (6)	Favor Democrats (7)	Favor Republicans (8)
				DV: Evaluators' Posterior Beliefs				
Δ	11.20	9.90	11.33	9.94	9.19	12.04	9.78	11.81
	(2.61)	(2.45)	(3.00)	(2.20)	(2.32)	(2.74)	(2.15)	(3.17)
Δ *Attention	-0.75	1.37	-0.47	0.95	0.53	0.29	1.09	-0.79
	(3.65)	(3.41)	(3.96)	(3.20)	(3.28)	(3.80)	(2.99)	(4.53)
Δ *Calculation	-5.80	-5.84	-5.24	-5.93	-2.71	-9.37	-4.96	-6.97
	(3.78)	(3.44)	(4.03)	(3.31)	(3.26)	(4.01)	(3.02)	(4.63)
N	572	638	531	679	691	519	826	384
Condition FE	yes	yes	yes	yes	yes	yes	yes	yes
Truth(Δ)	1.74	1.74	1.74	1.74	1.74	1.74	1.74	1.74

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: have an educational attainment of an Associate's Degree or less in Column 1, have an educational attainment of Bachelor's Degree or more in Column 2, have a reported annual income of below $50,000 in Column 3, report annual income equal to or exceeding $50,000 in Column 4, are 18-35 year old in Column 5, are 36 years or older in Column 6, indicate that they feel more favorably about Democrats than Republicans in Column 7, and indicate that they feel (weakly) more favorably about Republicans than Democrats in Column 8. The regression specifications are the same as in Appendix Table 6.

D   Additional Robustness Results

In this Appendix, we present results from several additional study versions. See Section D.1 for the *Evaluator (Alternative Questions) Study*, Section D.2 for the *Evaluator (Full Distribution) Study*, Section D.3 for the *Worker (Undergraduates) Study*, Section D.4 for the corresponding *Evaluator (Professional Evaluators) Study*, Section D.5 for the *Baseline* treatment of the *Evaluator (Extended) Study*, Section D.6 for the *Strategic Incentives* treatment of the *Worker Study*, Section D.7 for the corresponding *Strategic Incentives* treatment of the *Evaluator (Extended) Study*, Section D.8 for the *Joint Evaluations* and *Joint Evaluations, Strategic Incentives* treatments of the *Evaluator (Extended) Study*, Section D.9 for the *Evaluator (Additional Demographics) Study*, and Section D.10 for the *Evaluator (Known Performance) Study*.

D.1   The *Evaluator (Alternative Questions) Study*

Appendix Table D.1 presents the results from the *Evaluator (Alternative Questions) Study*, as discussed in Section 6.1. Note that, for priors (shown in Column 1) and posteriors (shown in Column 5), the expected performance gap is in the direction of evaluators believing that male workers performed better than female workers for all performance outcomes, but this presents as a *positive* coefficient on Δ for the performance outcomes in Panels A and B and presents as a *negative* coefficient on Δ for the performance outcomes in Panels C–F.

26

Table D.1: Evaluators' Beliefs in the *Evaluator (Alternative Questions) Study*

DV:	Prior	Over-confidence	Under-confidence	Implied Bayesian Belief	Posterior
	(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	-4.12	-11.73	5.25	-1.47	6.51
SE of Δ	(1.68)	(2.21)	(2.15)	(1.70)	(1.74)
N	400	400	400	396	400
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
SE of Δ	(1.76)	(2.26)	(2.14)	(1.76)	(1.82)
N	400	400	400	396	400
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
SE of Δ	(1.70)	(2.93)	(2.86)	(1.93)	(1.42)
N	400	400	400	394	400
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
SE of Δ	(1.87)	(2.24)	(2.14)	(1.89)	(1.68)
N	400	400	400	398	400
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	3.55	0.15	-0.40
SE of Δ	(2.21)	(2.74)	(2.50)	(2.56)	(1.97)
N	400	400	400	397	400
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
SE of Δ	(1.81)	(2.30)	(2.18)	(1.80)	(1.49)
N	400	400	400	396	400

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.6. Data are from the 400 participants in the *Evaluator (Alternative Questions) Study*. See Appendix Tables A.5 and A.6 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.2 The *Evaluator (Full Distribution) Study*

Appendix Table D.2 presents the results from the *Evaluator (Full Distribution) Study*, as discussed in Section 6.2. Since there is a true performance gap of 5.69 percentage points (i.e., women actually are 5.69 percentage points more likely to have a poor performance), a few word on the results in Panel B which present the evaluators' beliefs minus the "truth" are warranted. Column 1 of Panel B, reveals that, according to their priors, evaluators expect women to be *less* likely to have a poor performance relative to the truth. Similarly, Column 4 of Panel B, reveals that evaluators—if they are Bayesians—should expect women to be *less* likely to have a poor performance relative to the truth. Yet, even so, Column 5 of Panel B reveals that evaluators according to their posteriors, expect that women are *more* likely to have a poor performance relative to the truth.

Table D.2: Evaluators' Beliefs' in the *Evaluator (Full Distribution) Study*

DV:	Prior	Over-confidence	Under-confidence	Implied Bayesian Posterior	Posterior
	(1)	(2)	(3)	(4)	(5)
Panel A: Evaluators' Beliefs					
B(F)	41.30	36.33	54.31	43.43	62.93
B(M)	42.19	49.17	43.68	42.98	52.90
Δ	-0.90	-12.83	10.64	0.45	10.03
SE of Δ	(1.75)	(2.20)	(2.04)	(1.73)	(1.59)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-11.79	22.81	-9.34	-9.65	9.84
B(M) - Truth(M)	-5.20	17.12	-3.99	-4.41	5.51
Δ - Truth (Δ)	-6.59	5.69	-5.35	-5.24	4.33
SE of Δ - Truth(Δ)	(1.75)	(2.20)	(2.04)	(1.73)	(1.59)
N	400	400	400	398	400
Truth(F)	53.08	13.52	63.66	53.08	53.08
Truth(M)	47.39	32.04	47.67	47.39	47.39
Truth(Δ)	5.69	-18.51	16.00	5.69	5.69

SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 400 participants in the *Evaluator (Full Distribution) Study*.

D.3   The *Worker (Undergraduates) Study*

Appendix Table D.3 presents the results from the *Worker (Undergraduates) Study*, as discussed in Section 6.3. We excluded 4 of the 354 recruited participants—because they neither identify as men nor women and we are under-powered to consider this group—resulting in a sample of 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.1 for a discussion of the minor differences between the *Worker (Undergraduates) Study* and *Worker Study*.

Table D.3: Self-Evaluations in the *Baseline* treatment of the *Worker (Undergraduates) Study*

| | DV: Binary guess of "poor performance" | | | |
| | 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

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.

D.4 The *Baseline* and *Baseline, Unknown Gender* Treatments of The *Evaluator (Professional Evaluators) Study*

Appendix Table D.4 presents the results from the *Baseline* treatment of the *Evaluator (Professional Evaluators) Study*, and Appendix Table D.5 presents results from the *Baseline, Unknown Gender* treatment of the *Evaluator (Professional Evaluators) Study*, as discussed in Section 6.3.

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.

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

DV:	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
SE of Δ	(1.89)	(2.16)	(2.00)	(1.83)	(1.48)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	8.59	-1.79	-5.53	9.33	21.09
B(M) - Truth(M)	8.90	5.88	14.35	9.38	8.36
Δ - Truth(Δ)	-0.30	-7.67	-19.88	-0.05	12.73
SE of Δ - Truth(Δ)	(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	27.35	27.35
Truth(Δ)	1.91	-3.16	34.95	1.91	1.91

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

DV:	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
SE of Δ	(1.97)	(2.22)	(2.05)	(1.91)	(1.49)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	9.98	2.30	-8.39	10.95	21.19
B(M) - Truth(M)	10.68	0.17	16.34	9.14	9.26
Δ - Truth(Δ)	-0.70	2.14	-24.73	1.81	11.92
SE of Δ - Truth(Δ)	(1.97)	(2.22)	(2.05)	(1.91)	(1.49)
N	391	391	391	388	391
Truth(F)	29.27	40.57	58.17	29.27	29.27
Truth(M)	27.35	43.73	23.22	27.35	27.35
Truth(Δ)	1.91	-3.16	34.95	1.91	1.91

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.5 The *Evaluator (Extended) Study*

Appendix Table D.6 presents the results from the *Evaluator (Extended) Study*, as discussed in Section 6.4.

Appendix Figure D.1 and Appendix Table D.7 show how evaluators' beliefs respond to individual worker's self-evaluations, as discussed in Section 6.5.

Table D.6: Evaluators' Beliefs in the *Baseline* Treatment of the *Evaluator (Extended) Study*

DV:	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
SE of Δ	(1.65)	(2.27)	(2.14)	(1.68)	(1.49)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-9.32	22.86	-22.28	-6.75	16.19
B(M) - Truth(M)	-9.44	6.85	-8.68	-8.09	3.18
Δ - Truth(Δ)	0.12	16.02	-13.61	1.34	13.01
SE of Δ - Truth(Δ)	(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

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.

Figure D.1: Evaluators' Beliefs About Specific Workers as a Function of Worker's Self-Evaluation
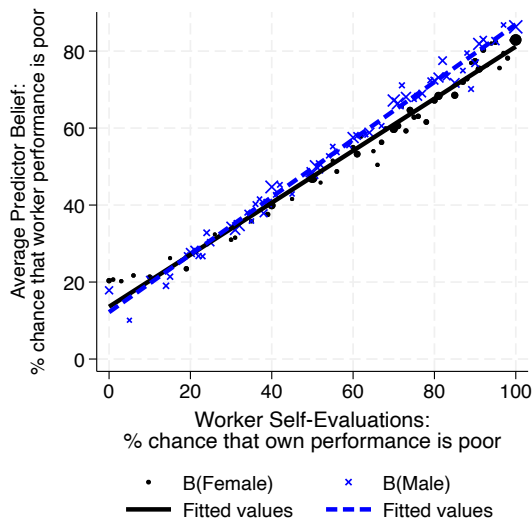
Graph shows a scatterplot of the average believed chance that a worker had a poor performance against that worker's believed percent chance that they had a poor performance. Data are from the *Evaluator (Extended) Study*.

Table D.7: Evaluators' Beliefs about Specific Workers in the *Baseline* treatment of the *Evaluator Study*

	DV: Evaluators' Posterior Beliefs	
	(1)	(2)
Δ	4.65	4.68
	(1.11)	(1.11)
Constant	55.08	
	(0.72)	
N	8120	8120
Performance FE	no	yes

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*.

D.6 The *Strategic Incentives* Treatment of the *Worker Study*

Appendix Table D.8 presents the results from the *Strategic Incentives* treatment of the *Worker Study*, as discussed in Section 6.6. 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 35 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.[35]

Appendix Table D.8 presents results on these workers, as discussed in Section 6.6[36] 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).

---

[35]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.4) and the possible absolute performance guesses that the worker could have given (Question 0 in Appendix Table A.4). Employers do not know workers' gender. We find that, for all binary self-evaluations, employers 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.

[36]Similar results follow from the other self-evaluation questions as well. Specifically, results in 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.8: Self-Evaluations in the *Strategic Incentives* treatment of the *Worker Study*

	DV: Binary guess of "poor performance"			
	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

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.

D.7 The *Evaluator (Extended, Strategic Incentives) Study*

Appendix Table D.9 presents the results from the *Evaluator (Extended, Strategic Incentives) Study*, as discussed in Section 6.3.

Table D.9: Evaluators' Beliefs' about Workers in the *Strategic Incentives* treatment of the *Evaluator (Extended) Study*

| DV:	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
SE of Δ	(1.71)	(2.22)	(2.03)	(1.65)	(1.31)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-10.42	12.86	-17.75	-9.56	11.95
B(M) - Truth(M)	-10.08	10.07	-7.51	-8.38	4.24
Δ - Truth(Δ)	-0.35	2.79	-10.24	-1.18	7.72
SE of Δ - Truth(Δ)	(1.71)	(2.22)	(2.03)	(1.65)	(1.31)
N	394	394	394	393	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

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

Appendix Tables D.10 and D.11 present the results from the *Joint Evaluations* treatment and the *Joint Evaluations, Strategic Incentives* treatment of *Evaluator (Extended) Study*, as discussed in Section 6.7.

Figure D.2 presents additional individual-level results from the *Joint Evaluations, Strategic Incentives* treatment of *Evaluator (Extended) Study*, as discussed in Section 6.8.

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

DV:	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
SE of Δ	(1.51)	(2.17)	(2.14)	(1.54)	(1.27)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-7.74	16.16	-20.40	-4.63	18.65
B(M) - Truth(M)	-8.99	10.90	-17.74	-6.00	5.66
Δ - Truth(Δ)	1.25	5.26	-2.66	1.37	12.99
SE of Δ - Truth(Δ)	(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

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

DV:	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
SE of Δ	(1.59)	(2.15)	(2.06)	(1.52)	(1.19)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	-9.92	8.25	-17.64	-8.12	11.78
B(M) - Truth(M)	-11.07	14.35	-15.62	-8.32	2.28
Δ - Truth(Δ)	1.14	-6.10	-2.01	0.21	9.50
SE of Δ - Truth(Δ)	(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

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.

Figure D.2: *Joint Evaluations, Strategic Incentives Treatment*: Classifying Evaluators According to Their Beliefs
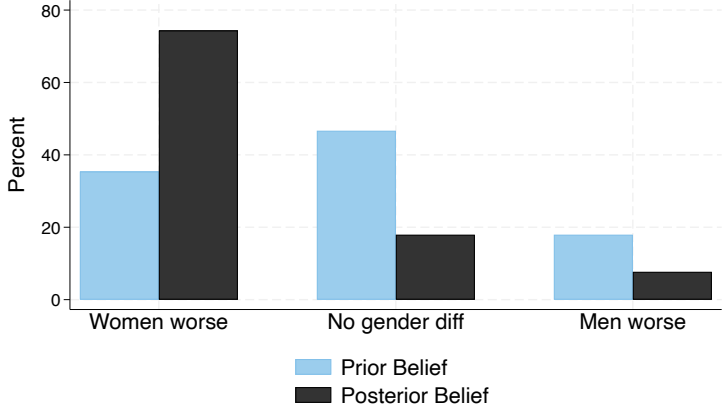
This graph shows the percent of evaluators who, given their prior or posterior beliefs, believe that women–relative to men—are more, equally, or less likely to have a poor performance in the first two, middle two, and right two bars, respectively. Data are from the *Joint Evaluations, Strategic Incentives* treatment of the *Evaluator (Extended) Study.*

D.9 The *Evaluator (Additional Demographics) Study*

Appendix Table D.12 presents the results from the *Evaluator (Additional Demographics) Study*, as discussed in Section 6.9.

Since there is a true performance gap of -7.77 percentage points (i.e., women actually are 7.77 percentage points *less* likely to have a poor performance), it is important to pay close attention to the relative magnitude of the results in Panel B. Column 1 of Panel B, reveals that, according to their priors, evaluators expect women are *10.34 percentage points more* likely to have a poor performance relative to the truth. Similarly, Column 4 of Panel B reveals that evaluators—if they are Bayesians—should (similarly) expect women to be *11.80 percentage points more* likely to have a poor performance relative to the truth. But, Column 5 of Panel B reveals a much larger expected performance gap according to evaluators' posteriors: evaluators expect that women are *30.44 percentage points more* likely to have a poor performance relative to the truth. That is, even though evaluators always directionally expect that women are more likely to have a poor performance relative to the truth (driven by the truth being that women are less likely to have a poor performance), it is still the case that evaluators' posteriors indicate that they expect a much larger performance gap relative to the truth than they should if they were Bayesian.

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

DV:	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
SE of Δ	(2.45)	(3.20)	(2.89)	(2.52)	(2.13)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	8.65	32.79	-18.96	9.78	27.81
B(M) - Truth(M)	-1.69	-14.48	2.07	-2.02	-2.60
Δ - Truth(Δ)	10.34	47.27	-21.03	11.80	30.42
SE of Δ - Truth(Δ)	(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)	-7.77	-52.27	32.37	-7.77	-7.77

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*.

D.10 The *Evaluator (Known Performance) Study*

Appendix Table D.13 presents the results from the *Evaluator (Known Performance) Study*, as discussed in Section 6.10.

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

DV:	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
SE of Δ	(3.38)	(3.04)	(2.62)	(3.30)	(2.52)
Panel B: Evaluators' Beliefs - Truth					
B(F) - Truth(F)	1.41	12.36	-14.18	1.61	18.60
B(M) - Truth(M)	1.88	-11.38	5.02	1.41	4.75
Δ - Truth(Δ)	-0.46	23.74	-19.20	0.20	13.85
SE of Δ - Truth(Δ)	(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

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 8B and the main Evaluator questions), or having a classifier who described the worker's performance as poor (corresponding to Worker Question 7B 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":

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

To see this:

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

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

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

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

$$(B) = P(Poor|SE^{Good})$$

$$= \frac{P(Poor \cap SE^{Good})}{P(SE^{Good})}$$

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

$$= \frac{\text{prior belief} * \text{overconfidence belief}}{\text{prior belief} * \text{overconfidence belief} + (1 - \text{prior belief}) * (1 - \text{underconfidence belief})}$$

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:

$$P(Good) \equiv \% \text{ chance that the worker had a good performance}$$

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

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

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

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

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

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 an 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

A: Beliefs About Male Workers   **B**: Beliefs About Female Workers
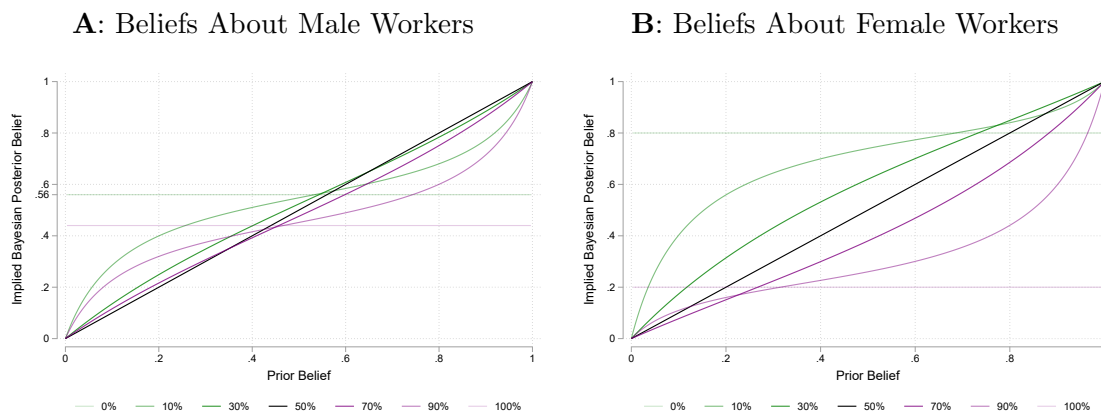
Graphs show the implied Bayesian posterior, across priors, for the overconfidence and underconfidence beliefs noted in the legend (assuming, for simplicity, that the level of the overconfidence and underconfidence belief is the same). Bayesian updating is done separately for male workers and female workers 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 5.3. Similarly, in our follow-up survey, we ask evaluators if they think that they accounted for the gender gap in confidence when providing their beliefs, and our results persist among the group of individuals who believe they did; see Section 5.4.

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

A: Beliefs About Male Workers
Given Evaluators' Confidence Beliefs

C: Beliefs About Female Workers
Given Evaluators' Confidence Beliefs
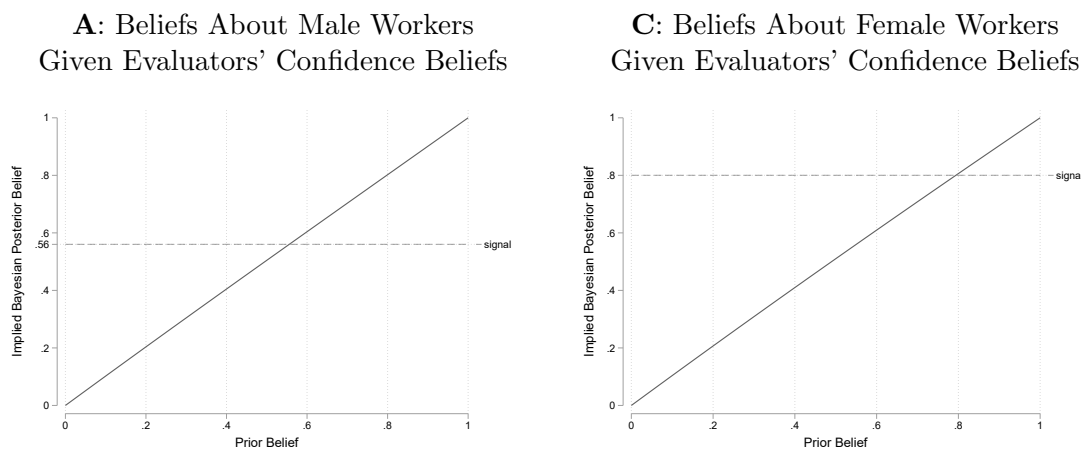
Graphs show the implied Bayesian posterior, across priors, given evaluators' beliefs about the likelihood that workers were over- and underconfident in the *Baseline* treatment of the *Evaluator Study*. Evaluators believed there to be a 39.86% chance that female workers were overconfident and a 48.11% 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.

Online Appendix for
"The Gender Gap in Confidence: Expected But Not Accounted For"

By Christine L. Exley and Kirby Nielsen

Experimental Instructions

Table Of Contents

This paper involved four main study waves. Section G presents the full instructions for the main *Evaluator Study* and its treatments. Section F presents the full instructions for the *Worker Study* and its treatments. Section I presents the full instructions for additional *Evaluator* studies. Section H presents the full instructions for additional *Worker* studies.

F   Full Instructions for the *Worker Study*

F.1   Instructions for the *Baseline* Treatment of *Worker Study*

After consenting to participate in the study, each participant is informed of the $3 study completion fee and of the opportunity to earn additional payment. Figure F.1.1 shows the overview participants are given and the corresponding comprehension question they must answer correctly in order to proceed. Then, participants proceed to Part 1, which involves a 10-item Math and Science Test. Figure F.1.2 shows the Part 1 instructions and the corresponding comprehension question they must answer correctly in order to proceed.

After completing the Math and Science Test, participants are then asked two questions about what would characterize poor performance and poor math and science skills (Classifier Question 1 and Classifier Question 2), as shown in Figure F.1.3.

Participants then proceed to the Part 2 instructions, which are related to predicting their own performance on the test via a series of self-evaluation questions. Figure F.1.4 shows the Part 2 instructions and the corresponding comprehension questions that participants need to answer correctly in order to proceed. Participants answer 17 self-evaluation questions (see Appendix Table A1 for corresponding labels of these self-evaluation questions), which are presented in randomized order (Figures F.1.5-F.1.13).

After completing Part 2, participants complete a short follow-up survey that collects additional control and demographic information.

Figure F.1.1: Study Overview, the *Baseline* Treatment of *Worker Study*

Overview: This study will consist of 2 parts and a short follow-up survey. Following certain instructions, you will be asked understanding questions. You must answer these understanding questions correctly in order to proceed to complete the study.

Your Payment: For completing this study, you will receive $3 as a completion payment. In addition, one part out of the 2 parts will be randomly selected as the part-that-counts. Any amount you earn in the part-that-counts will be given to you as a bonus payment.

Understanding Question: Which of the following statements is true?

> For completing this study, I will receive nothing.

> For completing this study, I will receive $3 for sure, and I will have no chance of a bonus payment.

> For completing this study, I will receive $3 for sure. In addition, I will receive any amount I earn in the part-that-counts as a bonus payment.

Figure F.1.2: Part 1 Instructions, *Baseline* Treatment of the *Worker Study*

Instructions for Part 1 out of 2:

In Part 1, you will complete a math and science test. On the test, you will be asked to answer up to 10 questions. Each question will test your math and science skills. Specifically, you will be asked about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on this test is often used as a measure of cognitive ability by academic researchers.

You will be presented with each of the 10 questions on separate pages. You will be given up to 20 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 20 seconds are up.

If Part 1 is randomly selected as the part-that-counts, your additional payment will equal 10 cents times the number of questions you answer correctly on this test.

Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

> will not depend on how many questions you answer correctly on the test.

> will be lower if you answer more questions correctly on the test.

> will be higher if you answer more questions correctly on the test.

Figure F.1.3: Classifier Questions, *Baseline* Treatment of the *Worker Study*

Before proceeding to Part 2, please answer the following two questions:

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 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

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 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Figure F.1.4: Part 2 Instructions, *Baseline* Treatment of the *Worker Study*

Instructions for Part 2 out of 2:

In part 2, you will be asked to make 17 predictions related to **your performance on the math and science test** you took in part 1.

In some of these predictions, you will be asked to guess the right answer to a multiple-choice question. In each of those predictions, you will earn $1 if your guess is right.

In the other predictions, you will be asked to guess the percent chance of some outcome being true on a slider that ranges from 0% to 100%. In each of those predictions, to secure the largest chance of earning $1 from the prediction, you should report your most-accurate guess. To learn the precise rule that determines how much you earn from these predictions click here.

If part 2 is randomly selected as the part-that-counts, one of your 17 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Thus, to maximize your chance of earning an additional payment of $1, you should provide your most-accurate guess when making each prediction.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

It doesn't matter

As accurately as possible

Randomly

Understanding Question: If this part is randomly selected as the part-that-counts, how much additional payment will you receive?

I will receive what I earn from all predictions in this part.

I will receive what I earn from the prediction-that-counts in this part.

Nothing

Figure F.1.5: Self-Evaluation Question 0, *Baseline* Treatment of the *Worker Study*

Prediction X out of 17: Out of the 10 questions on the math and science test, what do you think is the number you answered correctly?

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Figure F.1.6: Self-Evaluation Questions 1B and 1C, *Baseline* Treatment of the *Worker Study*

Prediction X out of 17: Did you get 3 or more questions right out of the 10 questions on the math and science test?

| No | Yes |

Prediction X out of 17: What is the percent chance that you got 3 or more questions right out of the 10 questions on the math and science test?

Extremely unlikely 0   10   Somewhat unlikely 20   30   Neither likely nor unlikely 40   50   60   Somewhat likely 70   80   Extremely likely 90   100

% chance that you got 3 or more questions right

Figure F.1.7: Self-Evaluation Questions 2B and 2C, *Baseline* Treatment of the *Worker Study*

Prediction X out of 17: Did you get 5 or more questions right out of the 10 questions on the math and science test?

| No | Yes |

Prediction X out of 17: What is the percent chance that you got 5 or more questions right out of the 10 questions on the math and science test?

Extremely unlikely 0   10   Somewhat unlikely 20   30   Neither likely nor unlikely 40   50   60   Somewhat likely 70   80   Extremely likely 90   100

% chance that you got 5 or more questions right

Figure F.1.8: Self-Evaluation Questions 3B and 3C, *Baseline* Treatment of the *Worker Study*

Prediction X out of 17: Did you get **7 or more questions right** out of the 10 questions on the math and science test?

No	Yes

Prediction X out of 17: What is the **percent chance** that you got **7 or more questions right** out of the 10 questions on the math and science test?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely		Somewhat likely			Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that you got 7 or more questions right

Figure F.1.9: Self-Evaluation Questions 4B and 4C, *Baseline* Treatment of the *Worker Study*

Please review the below information to make the next two predictions:
You scored in the top half if your score (i.e., the number of questions you got right) is greater than or equal to the scores of at least 50% of other participants. These other participants involve 50 randomly-selected men and 50 randomly-selected women from the set of all other participants who take this study.

Prediction X out of 17: Did you **score in the top half** when compared to **other participants** who took the study?

No	Yes

Prediction X out of 17: What is the **percent chance** that you **scored in the top half** when compared to **other participants**?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely		Somewhat likely			Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that you scored in top half when compared to other participants

Figure F.1.10: Self-Evaluation Questions 5B and 5C, *Baseline* Treatment of the *Worker Study*

Please review the below information to make the next two predictions:
When compared to women, you scored in the top half if your score (i.e., the number of questions you got right) is greater than or equal to the scores of at least 50% of 100 women. These women involve 100 randomly-selected women from the set of all other participants who are women and take this study.

Prediction X out of 17: Did you score in the top half when compared to women?

No	Yes

Prediction X out of 17: What is the percent chance that you scored in the top half when compared to women?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely			Somewhat likely			Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that you scored in top half when compared to women

Figure F.1.11: Self-Evaluation Questions 6B and 6C, *Baseline* Treatment of the *Worker Study*

Please review the below information to make the next two predictions:
When compared to men, you scored in the top half if your score (i.e., the number of questions you got right) is greater than or equal to the scores of at least 50% of 100 men. These men involve 100 randomly-selected men from the set of all other participants who are men and take this study.

Prediction X out of 17: Did you score in the top half when compared to men?

No	Yes

Prediction X out of 17: What is the percent chance you scored in the top half when compared to men?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely			Somewhat likely			Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that you scored in top half when compared to men

Figure F.1.12: Self-Evaluation Questions 7B and 7C, *Baseline* Treatment of the *Worker Study*

Please review the below information to make the next two predictions:

Recall that, prior to beginning Part 2, you were asked a question about which scores you believed were poor. You will be matched with an "evaluator" who was also asked this question, and your evaluator is said to have described your performance as poor if they indicated that your score was poor. Your evaluator will be randomly selected from the set of all other workers who also complete the study and is equally likely to be a man or a woman.

Prediction X out of 17: Did your evaluator describe your performance on the math and science test as poor?

No	Yes

Prediction X out of 17: What is the percent chance that your evaluator described your performance on the math and science test as poor?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely			Somewhat likely			Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that your performance was described as poor

Figure F.1.13: Self-Evaluation Questions 8B and 8C, *Baseline* Treatment of the *Worker Study*

Please review the below information to make the next two predictions:
Recall that, prior to beginning Part 2, you were asked a question about which scores you believed were indicative of poor math and science skills. You will be matched with an "evaluator" who was also asked this question, and your evaluator is said to have described your performance as indicative of poor math and science skills if they indicated that your score was indicative of poor math and science skills. Your evaluator will be randomly selected from the set of all other workers who also complete the study and is equally likely to be a man or a woman.

Prediction X out of 17: Did your evaluator describe your performance on the math and science test as indicative of poor math and science skills?

No	Yes

Prediction X out of 17: 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?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely		Somewhat likely		Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that your performance was described as indicative of poor math and science skills

F.2   Instructions for the *Strategic Incentives* Treatment of the *Worker Study*

Relative to the *Baseline* treatment of the *Worker Study* (Section F.1), all that differs in the *Strategic Incentives* treatment of the *Worker Study* is the Part 2 instructions. In this condition, workers are informed that one of their answers may be shown to their employer who will determine how much they earn if Part 2 is randomly selected as the part-that-counts. New Figures F.2.1 and F.2.2 below show the Part 2 instructions and the corresponding comprehension questions that participants need to answer correctly in order to proceed. All other screens look identical to the *Baseline* treatment of the *Worker Study*, shown above.

Instructions for Part 2 out of 2:

In part 2, you will be asked to make 17 predictions related to **your performance on the math and science test** you took in part 1.

In some of these predictions, you will be asked to guess the right answer to a multiple-choice question. In the other predictions, you will be asked to guess the percent chance of some outcome being true on a slider that ranges from 0% to 100%.

One of your predictions will be randomly selected as the prediction-that-counts.

Your answer to the prediction-that-counts will be shown to "your employer," who will be another Prolific worker who completes a different version of this study. Your employer will decide whether to hire you.

Aside from your answer to the prediction-that-counts, your employer will not be provided with any information on you or on your performance. For instance, your employer will NOT be informed of any demographic information about you, and your employer will NOT be informed of how many questions you answered correctly on the math and science test.

If this part is randomly selected as the part-that-counts, the additional payment given to your employer and to you will be determined as follows:

- If your employer chooses NOT to hire you, your additional payment will equal 50 cents and your employer's additional payment will equal 50 cents.

- If your employer chooses to hire you, your additional payment will equal 100 cents and your employer's additional payment will equal 10 cents times the number of questions you answered correctly on the test.

Understanding Question: If Part 2 is randomly selected as the part-that-counts, your additional payment...

> will be higher if your predictions are more accurate

> will NOT depend on how accurate your predictions are

Understanding Question: If Part 2 is randomly selected as the part-that-counts, will your employer learn how many questions you answered correctly on the math and science test?

> No - they will only be provided with my answer to one of my predictions

> Yes

Figure F.2.2: Part 2 Comprehension Questions, *Strategic Incentives* Treatment of the *Worker Study*

Understanding Question: If Part 2 is randomly selected as the part-that-counts, your additional payment...

will be higher if you are hired by your employer

will NOT depend on whether you are hired by your employer

Understanding Question: If Part 2 is randomly selected as the part-that-counts and your employer hires you, your employer's additional payment...

will be higher if you have a good performance and lower if you have a bad performance on the math and science test

will NOT depend on your performance on the math and science test

G Full Instructions for the *Evaluator Study*

All participants in this study are randomized to be asked about male or female workers (or "group-1" or "group-2" workers in some conditions) and to be in one of six treatments described below.

G.1 Instructions for the *Baseline* Treatment of the *Evaluator Study*

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. Figures G.1.1, G.1.2, and G.1.3 show the overview and comprehension questions we give to participants who are randomized to evaluate **female workers**. They must answer comprehension questions correctly in order to proceed. Then, participants provide their prior beliefs (Figure G.1.4). Subsequently, they are provided with information on female workers' self-evaluations and asked to provide their posterior beliefs (Figure G.1.5). After this, they are asked to provide their overconfidence and underconfidence beliefs (Figure G.1.6). Finally, all participants take a short survey of five randomized bonus questions, as shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figures G.1.5 and G.1.6 changes from 80% to **56%**.

Figure G.1.1: Study Overview, *Baseline* Treatment of the *Evaluator Study*

Main Instructions (Page 1 out of 2)

Overview:
This study will consist of 3 predictions and a short follow-up survey. For completing this study, you are guaranteed to receive $2 within 24 hours. In addition, any additional payment you earn will be distributed to you as a bonus payment.

The Workers:
In a prior study, an approximately equal number of men and women (called "workers") completed a math and science test with 10 questions. Each question tested their math and science skills by asking them about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on questions like these is often used as a measure of cognitive ability by academic researchers. A worker's score on the test equals the number of questions they answer correctly, and a worker earns 10 cents times their score.

Your Predictions:
You will be asked to make 3 predictions related to the **performance of workers on the math and science test**.

To maximize your chance of earning an additional payment of $1, you should provide your most-accurate guess when making each prediction. This is because each prediction will ask you to guess the percent chance of some outcome being true. In each of those predictions, to secure the largest chance of earning $1 from the prediction, you should report your most-accurate guess. To learn the precise payment rule that determines how much you earn from these predictions click here.

One of your 3 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

| It doesn't matter |
| As accurately as possible |
| Randomly |

Understanding Question: How much additional payment will you receive?

| I will receive what I earn from all predictions in this part |
| I will receive what I earn from the prediction-that-counts |
| Nothing |

→

15

Figure G.1.2: Instructions about Female Workers, *Baseline* Treatment of the *Evaluator Study*

<u>**Main Instructions (Page 2 out of 2)**</u>

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

<u>**Your Workers**</u>
In each prediction, your worker will be randomly selected from the following group: all of the female workers who had performances in the "middle" (when compared to all female and male workers) on the math and science test. Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

Workers who had performances in the 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.

You will sometimes be provided with information on beliefs held by the relevant workers when they were asked to make predictions about their own performance.

<u>**Types of Predictions**</u>
In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described the worker's performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker had an evaluator who was randomly selected from the set of all other workers who completed the study and who was equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator choose how to describe a worker's performance, the evaluator effectively knew how many questions the worker got right but did not know anything else about the worker, such as the worker's gender.

<u>**Worker Predictions**</u>:
In some predictions, you will be informed of the average prediction made by the female workers who could be randomly selected to be your worker in that prediction

Thus, please note that: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most-accurate guess to maximize their chance of earning an additional bonus payment of $1.

Figure G.1.3: Comprehension Questions about Female Workers, *Baseline* Treatment of the *Evaluator Study*

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Understanding Question: In each prediction, my worker will be randomly selected from the following group:

all of the female workers

all of the female workers who had performances in the middle

all of the male workers who had performances in the middle

→

Figure G.1.4: Prior Belief about Female Workers, *Baseline* Treatment of the *Evaluator Study*

Prediction 1 out of 3:
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will be randomly selected from the following group: all of the female workers who had performances in the middle.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[ ]

→

Figure G.1.5: Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator Study*

Prediction 2 out of 3:
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female workers who had performances in the middle.

After completing the math and science test, 80% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[ ]

→

Figure G.1.6: Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator Study*

<u>Prediction 3 out of 3:</u>
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female workers who had performances in the middle.

After completing the math and science test, 80% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

<u>Overconfidence Prediction</u>: If your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your female worker is overconfident because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?

```
[          ]
```

<u>Underconfidence Prediction</u>: If your female worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your female worker is underconfident because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?

```
[          ]
```

→

19

Figure G.1.7: Bonus Questions Instructions, *Baseline* Treatment of the *Evaluator Study*

Bonus Questions

Recall that you are guaranteed to receive $2 within 24 hours. In addition, you will receive $1 as a bonus payment if you provided the correct answer to one randomly selected prediction out of the predictions you have already made.

Now, you have the chance to earn an additional $1 as bonus payment (for up to a total of $2 as bonus payment).

In particular, you will now be asked to answer 5 bonus questions. One of the bonus questions will be randomly selected as the bonus-question-that-counts. If you provide the correct answer in the bonus-question-that-counts, you will earn an additional $1 as bonus payment.

<div style="text-align:right">

→

</div>

Figure G.1.8: Bonus Question 1: Bayesian Updating, *Baseline* Treatment of the *Evaluator Study*

Bonus Question X out of 5:
There are two doctors at a hospital: Doctor Bailey and Doctor Grey.

- Doctor Bailey has 100 patients and 10% are female.
- Doctor Grey has 100 different patients and 70% are female.
- We put Doctor Bailey's and Doctor Grey's patient files together and randomly shuffle all 200 of them.
- We draw one file from the stack at random, and the patient from this file is male.

What is the percent chance that this patient is Doctor Bailey's patient?
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

[ ]

<div style="text-align:right">

→

</div>

Figure G.1.9: Bonus Question 2: CRT1, *Baseline* Treatment of the *Evaluator Study*

<u>**Bonus Question X out of 5**</u>:
A cookie and a peppermint cost $1.10 in total. The cookie costs a dollar more than the peppermint.

How much does the peppermint cost (in cents)?
Please omit the "cents" symbol and only write in the corresponding number (e.g., 0, 1, 2,...)

[   ]

→

Figure G.1.10: Bonus Question 3: CRT2, *Baseline* Treatment of the *Evaluator Study*

<u>**Bonus Question X out of 5**</u>:
If it takes 5 machines 5 minutes to make 5 microwaves, how many minutes would it take 100 machines to make 100 microwaves?
Please omit "minutes" from your answer and only write in the corresponding number (e.g., 0, 1, 2,...)

[   ]

→

Figure G.1.11: Bonus Question 4: CRT3, *Baseline* Treatment of the *Evaluator Study*

<u>**Bonus Question X out of 5**</u>:
A virus spreads through a population. Every day, the number of infected people doubles.

If it takes 48 days for the entire population to catch the virus, how many days would it take for half of the population to catch the virus?
Please omit "days" from your answer and only write in the corresponding number (e.g., 0, 1, 2,...)

[   ]

→

Figure G.1.12: Bonus Question 5: Base Rate Neglect, *Baseline* Treatment of the *Evaluator Study*

<u>Bonus Question X out of 5</u>:
A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the percent chance (rounded to the nearest whole number) that the cab involved in the accident was Blue rather than Green?
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

☐

→

G.2 Instructions for the *Attention* Treatment of the *Evaluator Study*

The *Attention* treatment of the *Evaluator Study* differs from the *Baseline* treatment of the *Evaluator Study* (Section G.1) only in the order of the predictions made by the participant.

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. The screenshots for this study are identical to those above in the *Baseline* Treatment but are shown in a different order. Figures G.1.1, G.1.2, and G.1.3 show the overview and comprehension questions we give to participants who are randomized to evaluate **female workers**. They must answer comprehension questions correctly in order to proceed. Then, participants provide their prior beliefs (Figure G.1.4). Next, they are provided with information on workers' self-evaluations and asked to provide their over/underconfidence beliefs (Figure G.1.6). After this, they are asked to provide their posterior beliefs (Figure G.1.5). Finally, all participants take a short survey of five randomized bonus questions, as previously shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figures G.1.5 and G.1.6 changes from 80% to **56%**.

G.3 Instructions for the *Calculation* Treatment of the *Evaluator Study*

The *Calculation* treatment of the *Evaluator Study* differs from the *Attention* treatment of the *Evaluator Study* (Section G.2) only in the decision screen that elicits their posterior beliefs, highlighted via the new Figure G.3.1 shown below.

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. Figures G.1.1, G.1.2, and G.1.3 show the overview and comprehension questions we give to participants who are randomized to evaluate **female workers**. They must answer comprehension questions correctly in order to proceed. Then, participants provide their prior beliefs (Figure G.1.4). Next, they are provided with information on workers' self-evaluations and asked to provide their over/underconfidence beliefs (Figure G.1.6). After this, they are asked to provide their posterior beliefs (new Figure G.3.1 below). Finally, all participants take a short survey of five randomized bonus questions, as previously shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figures G.1.6 and G.3.1 changes from 80% to **56%**.

Figure G.3.1: Posterior Belief about Female Workers, *Calculation* Treatment of the *Evaluator Study*

<u>**Prediction 3 out of 3:**</u>
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female workers who had performances in the middle.

After completing the math and science test, 80% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

There is a very well-known theory in probability and statistics (called Bayes' Rule) that gives a mathematical way to update your guess after receiving some new information. Given the information above on what female workers thought about their own performance, and given how likely you thought female workers are to be overconfident or underconfident, Bayes' Rule would say that your updated guess (from Prediction 1) would be X%.

We are telling you this just in case it is helpful for you. You do NOT have to use Bayes' Rule to update your guess.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

> [   ]

> [→]

G.4 Instructions for the *Baseline, Unknown Gender* Treatment of the *Evaluator Study*

The *Baseline, Unknown Gender* treatment differs from the *Baseline* treatment of the *Evaluator Study* (Section G.1) only in that participants are not told the gender of their worker and "group-1 workers" and "group-2 workers" replace "male workers" and "female workers," respectively.

After consenting to participate in the study, each participant is informed of the \$2 study completion fee and of the opportunity to earn additional payment. Figures G.1.1 (above), G.4.1, and G.4.2 show the overview participants who are randomized to evaluate **group-2 workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (new Figure G.4.3), are provided with information on group-2 workers' self-evaluations and asked to provide their posterior beliefs (new Figure G.4.4), and are asked to provide their overconfidence and underconfidence beliefs (new Figure G.4.5). Finally, all participants take a short survey of five randomized bonus questions, as previously shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **group-1 workers** (considered "male workers" in the *Baseline* Treatment of the *Evaluator Study* (Section G.1)), "group-2" is replaced by "group-1" everywhere, and the self-evaluation information provided in Figures G.4.4 and G.4.5 changes from 80% to **56%**.

Figure G.4.1: Instructions about Group-2 Workers, *Baseline, Unknown Gender* Treatment of the *Evaluator Study*

Main Instructions (Page 2 out of 2)

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your Workers

In each prediction, your worker will be randomly selected from the following group: all of the group-2 workers who had performances in the "middle" (when compared to all group-1 and group-2 workers) on the math and science test. Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

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.

We assigned each worker to group-1 or group-2 based on an answer they provided to a question in the follow-up survey. While you will not be informed of their answer to this follow-up survey question, you will sometimes be provided with information on beliefs held by the relevant workers when they were asked to make predictions about their own performance.

Types of Predictions

In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described the worker's performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker had an evaluator who was randomly selected from the set of all other workers who completed the study and who was equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator choose how to describe a worker's performance, the evaluator effectively knew how many questions the worker got right but did not know anything else about the worker, such as the worker's gender.

Figure G.4.2: Comprehension Questions about Group-2 Workers, *Baseline, Unknown* Treatment of the *Evaluator Study*

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Understanding Question: In each prediction, my worker will be randomly selected from the following group:

all of the group-2 workers

all of the group-2 workers who had performances in the middle

all of the group-1 workers who had performances in the middle

→

Figure G.4.3: Prior Belief about Group-2 Workers, *Baseline, Unknown* Treatment of the *Evaluator Study*

Prediction 1 out of 3:

Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your group-2 worker will be randomly selected from the following group: all of the group-2 workers who had performances in the middle.

What do you think is the percent chance that your group-2 worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

```
┌─────────────────┐
│                 │
└─────────────────┘
```

Figure G.4.4: Posterior Belief about Group-2 Workers, *Baseline, Unknown* Treatment of the *Evaluator Study*

Prediction 2 out of 3:

Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your group-2 worker will again be randomly selected from the following group: all of the group-2 workers who had performances in the middle.

After completing the math and science test, 80% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your group-2 worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

```
┌─────────────────┐
│                 │
└─────────────────┘
```

Figure G.4.5: Over/Underconfidence Beliefs about Group-2 Workers, *Baseline, Unknown* Treatment of the *Evaluator Study*

<u>**Prediction 3 out of 3:**</u>
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your group-2 worker will again be randomly selected from the following group: all of the group-2 workers who had performances in the middle.

After completing the math and science test, 80% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

<u>**Overconfidence Prediction**</u>**: If your group-2 worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your group-2 worker is overconfident because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?**

```
┌─────────────────┐
│                 │
└─────────────────┘
```

<u>**Underconfidence Prediction**</u>**: If your group-2 worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your group-2 worker is underconfident because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?**

```
┌─────────────────┐
│                 │
└─────────────────┘
```

G.5 Instructions for the *Attention, Unknown Gender* Treatment of the *Evaluator Study*

The *Attention, Unknown Gender* treatment differs from the *Attention* treatment of the *Evaluator Study* (Section G.2) in the same way that the *Baseline, Unknown Gender* treatment (Section G.4) differs from the *Baseline* treatment of the *Evaluator Study* (Section G.1). Participants are not told the gender of their worker and "group-1 workers" and "group-2 workers" replace "male workers" and "female workers," respectively.

G.6 Instructions for the *Calculation, Unknown Gender* Treatment of the *Evaluator Study*

The *Calculation, Unknown Gender* treatment differs from the *Calculation* treatment of the *Evaluator Study* (G.3) in the same way that the *Baseline, Unknown Gender* treatment (Section G.4) differs from the *Baseline* treatment of the *Evaluator Study* (Section G.1). Participants are not told the gender of their worker and "group-1 workers" and "group-2 workers" replace "male workers" and "female workers," respectively.

H Full Instructions for Additional Worker Studies

H.1 Instructions for the *Worker (Undergraduate Students) Study*

The *Worker (Undergraduate Students) Study* surveys undergraduate students of a university.

After consenting to participate in the study, each participant is informed of the $10 study completion fee and of the opportunity to earn additional payment. Figure H.1.1 shows the overview participants are given and the corresponding comprehension question they must answer correctly in order to proceed. Participants then proceed to Part 1. Figure H.1.2 shows the Part 1 instructions and the corresponding comprehension question they must answer correctly in order to proceed. After completing Part 1, participants are asked two questions about what would characterize poor test performance and poor math and science skills (Classifier Question 1 and Classifier Question 2), as previously shown in Figure F.1.3.

Participants then proceed to Part 2. Figure H.1.3 shows the Part 2 instructions and the corresponding comprehension questions that participants need to answer correctly in order to proceed. Participants then answer 13 self-evaluation questions (see Appendix Table A1 for corresponding labels of these self-evaluation questions). In addition to 7 self-evaluation questions of the *Baseline* treatment of the *Worker Study* (Figures F.1.5-F.1.8 above), participants were asked 6 more self-evaluation questions (additional Figures H.1.4-H.1.7 below; Figure H.1.5 shows the additional instructions and comprehension question for Figures H.1.6 and H.1.7). These self-evaluation questions are presented in a randomized order (with the constraint that Figure H.1.6 and H.1.7 are consecutive).

After completing Part 2, participants complete a short follow-up survey that collects additional control and demographic information.

Overview: This study will consist of 2 parts and a short follow-up survey. Following certain instructions, you will be asked understanding questions. You must answer these understanding questions correctly in order to proceed to complete the study.

Your Payment: For completing this study, you will receive an Amazon gift card that will be emailed to you.   The amount of your gift card is guaranteed to be at least $10. In addition, one part out of the 2 parts will be randomly selected as the part-that-counts. Any amount you earn in the part-that-counts will be added to the $10 to determine the total amount on your gift card.

Understanding Question: Which of the following statements is true?

> For completing this study, I will receive an Amazon gift card that is worth no more than $10.

> For completing this study, I will receive an Amazon gift card that is worth the sum of $10 and any amount I earn in the part-that-counts.

> For completing this study, I will receive an Amazon gift card that is worth the amount I earn in the part-that-counts.

<u>Instructions for Part 1 out of 2:</u>

In Part 1, you will complete a math and science test. On the test, you will be asked to answer up to 10 questions. Each question will test your math and science skills. Specifically, you will be asked about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on this test is often used as a measure of cognitive ability by academic researchers.

You will be presented with each of the 10 questions on separate pages. You will be given up to 15 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 15 seconds are up.

If Part 1 is randomly selected as the part-that-counts, your additional payment will equal $1 times the number of questions you answer correctly on this test.

Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

will not depend on how many questions you answer correctly on the test.

will be lower if you answer more questions correctly on the test.

will be higher if you answer more questions correctly on the test.

Figure H.1.3: Part 2 Instructions, *Worker (Undergraduate Students) Study*

Instructions for Part 2 out of 2:

In part 2, you will be asked to make 13 predictions related to **your performance on the math and science test** you took in part 1.

In some of these predictions, you will be asked to guess the right answer to a multiple-choice question. If each of those predictions, you will earn $5 if your guess is right.

In the other predictions, you will be asked to guess the percent chance of some outcome being true on a slider that ranges from 0% to 100%. In each of those predictions, to secure the largest chance of earning $5 from the prediction, you should report your most-accurate guess. To learn the precise rule that determines how much you earn from these predictions click here.

If part 2 is randomly selected as the part-that-counts, one of your 13 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Thus, to maximize your chance of earning an additional payment of $5, you should provide your most-accurate guess when making each prediction.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

> It doesn't matter

> As accurately as possible

> Randomly

Understanding Question: If this part is randomly selected as the part-that-counts, how much additional payment will you receive?

> I will receive what I earn from all predictions in this part.

> I will receive what I earn from the prediction-that-counts in this part.

> Nothing

35

Figure H.1.4: Self-Evaluation Questions New-1B and New-1C, *Worker (Undergraduate Students) Study*

Prediction X out of 13: Did you get **9 or more questions right** out of the 10 questions on the math and science test?

No	Yes

Prediction X out of 13: What is the **percent chance** that you got **9 or more questions right** out of the 10 questions on the math and science test?

Extremely unlikely		Somewhat unlikely		Neither likely nor unlikely			Somewhat likely			Extremely likely
0	10	20	30	40	50	60	70	80	90	100

% chance that you got 9 or more questions right

Figure H.1.5: Self-Evaluation Questions New-2B, 2C, 3B and C Instructions, *Worker (Undergraduate Students) Study*

Additional Instructions

Recall that, prior to beginning Part 2, you were asked the following two questoins:

- **Question 1:** An individual's performance on the math and science test was **poor** if the number of questions the individual answered correctly was **equal to or less than...**
- **Question 2:** 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 **equal to or less than...**

In the next 4 predictions, we will pair you with an evaluator and will ask you to predict how your evaluator described your performance on the math and science test given their answers to the above two questions. Your evaluator will be randomly selected from the set of other participants who completed this study and will be equally likely to be a man or a woman. Then, how your evaluator described your perfromance will be determined as follows:

- If the evaluator indicated that an individual who answered the same number of questions correctly on the test as you did had a performance that was poor in Question 1, then your evaluator has described your performance as poor.
- If the evaluator indicated that an individual who answered the same number of questions correctly on the test as you did had a performance that was indicative of poor math and sceience skills in Question 2, then your evaluator has described your performance as being indicative of poor math and science skills.

For example, if your evaluator indicated that a score of less than 5 was indicative of poor math and science skills, and if you scored a 4, then your evaluator has described your performance as being indicative of poor math and science skills.

Understanding Question: How is the evaluator selected?

The evaluator is randomly selected from the set of other participants who completed this study.

The evaluator is selected to have the same performance as you did on the math and science test.

The evaluator is selected in some other way.

37

Figure H.1.6: Self-Evaluation Questions New-2B and 2C, *Worker (Undergraduate Students) Study*

Prediction X out of 13: Did your evaluator describe your performance on the math and science test as **poor**?

No	Yes

Prediction X out of 13: What is the **percent chance** that your evaluator described your performance on the math and science test as **poor**?

Extremely unlikely   Somewhat unlikely   Neither likely nor unlikely   Somewhat likely   Extremely likely

0   10   20   30   40   50   60   70   80   90   100

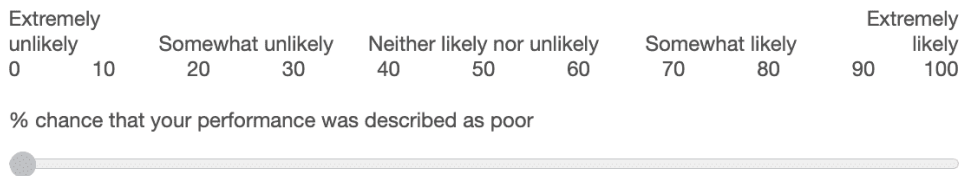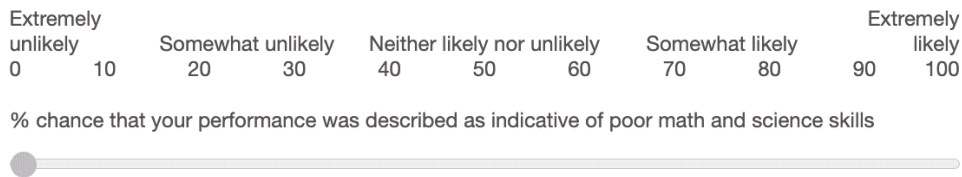% chance that your performance was described as poor

Figure H.1.7: Self-Evaluation Questions New-3B and 3C, *Worker (Undergraduate Students) Study*

Prediction X out of 13: Did your evaluator describe your performance on the math and science test as **indicative of poor math and science skills**?

No	Yes

Prediction X out of 13: 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**?

Extremely unlikely   Somewhat unlikely   Neither likely nor unlikely   Somewhat likely   Extremely likely

0   10   20   30   40   50   60   70   80   90   100

% chance that your performance was described as indicative of poor math and science skills

I Full Instructions for Additional Evaluator Studies

I.1 Full Instructions for the *Evaluator (Alternative Questions) Study*

All participants in this study are randomized to be asked about male or female workers.

After consenting to participate in the study, each participant is informed of the $3 study completion fee and of the opportunity to earn additional payment. Figures I.1.1-I.1.4 show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed.

Then, participants are provided with additional instructions about their prior beliefs (Figure I.1.5), are asked to provide their prior beliefs relating to six different outcomes that are presented in a random order (Figures I.1.6-I.1.11), are provided with additional instructions about their posterior beliefs (Figure I.1.12), are asked to provide their posterior beliefs relating to six different outcomes that are presented in a random order (Figures I.1.13-I.1.18), are provided with additional instructions about their overconfidence and underconfidence beliefs (Figure I.1.19), and are asked to provide their overconfidence and underconfidence beliefs relating to six different outcomes that are presented in a random order (Figures I.1.20-I.1.25). Finally, all participants complete a follow-up survey that collects additional control and demographic information.

For evaluators who are instead asked to evaluate **male workers**, "female" is replaced by "male" everywhere. In addition to this, see Figures I.1.26-I.1.31 for posterior belief questions about **male workers** and the corresponding self-evaluation information provided for each question.

Figure I.1.1: Study Overview, *Evaluator (Alternative Questions) Study*

Overview:
This study will consist of 18 predictions and a short follow-up survey. For completing this study, you are guaranteed to receive $3 within 24 hours. In addition, any additional payment you earn will be distributed to you as a bonus payment.

The Workers:
In a prior study, an approximately equal number of men and women (called "workers") completed a math and science test with 10 questions. Each question tested their math and science skills by asking them about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on questions like these is often used as a measure of cognitive ability by academic researchers. A worker's score on the test equals the number of questions they answer correctly, and a worker earns 10 cents times their score.

Your Predictions:
You will be asked to make 18 predictions related to the **performance of workers on the math and science test**.

To maximize your chance of earning an additional payment of $1, you should provide your most-accurate guess when making each prediction. This is because each prediction will ask you to guess the percent chance of of some outcome being true. In each of those predictions, to secure the largest chance of earning $1 from the prediction, you should report your most-accurate guess. To learn the precise payment rule that determines how much you earn from these predictions click here.

One of your 18 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

| It doesn't matter |
| As accurately as possible |
| Randomly |

Understanding Question: How much additional payment will you receive?

| I will receive what I earn from all predictions in this part |
| I will receive what I earn from the prediction-that-counts |
| Nothing |

40

Figure I.1.2: Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

Main Instructions (Page 2 out of 2)

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your workers
In each prediction, your worker will be randomly selected from the group of female workers who had performances in the "middle" (when compared to all male and female workers) on the math and science test. Specifically, your female worker will be randomly selected from the group of all female workers who had performances in the middle (when compared to all male and female workers). Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

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.

Figure I.1.3: Instructions about Female Workers, cont., *Evaluator (Alternative Questions) Study*

<u>**Types of Predictions**</u>
You will be asked to make four different types of predictions about the **percent chance that some outcome is true.**

In one set of predictions, you will be asked to predict the **percent chance that your worker in that prediction got at least some number of questions right** on the math and science test.

In a second set of predictions, you will be asked to predict the **percent chance that your worker in that prediction had an evaluator who described their performance as poor or as indicative of poor math and science skills**. Thus, please note the following:

- Each worker has an evaluator who is randomly selected from the set of all other workers who complete the study and who is equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were poor and a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as poor or as indicative of poor math and science skills if they indicated the worker's score was poor or was indicative of poor math and science skills, respectively.
- Thus, when an evaluator chooses how to describe a worker's performance, the evaluator effectively knows how many questions the worker got right but does not know anything else about the worker, such as the worker's gender.

In a third set of predictions, you will be asked to predict the **percent chance that your worker in that prediction scored in the top half.** Thus, please note the following:

- A worker scored in the top half if their score (i.e., the number of questions they got right) was greater than or equal to the scores of 50% of other participants.
- These other participants involve 50 randomly-selected men and 50 randomly-selected women from the set of all other participants who took this study.

In a fourth of predictions, you will be asked to predict the **percent chance that your worker in that prediction is overconfident or underconfident** when asked to make predictions about their own performance.

Figure I.1.4: Comprehension Questions about Female Workers, *Evaluator (Alternative Questions) Study*

Understanding Question: In each prediction, my worker will be randomly selected from...

the entire group of female workers

the group of female workers who had performances in the middle

the group of male workers who had performances in the middle

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Figure I.1.5: Prior Belief Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

Additional Instructions

The next set of predictions will ask you to make predictions about the percent chance that your female worker had some performance on the math and science test How we classify workers according to their performance will be defined on each decision screen.

Figure I.1.6: Prior Belief (3+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction got 3 or more questions right on the test?

Figure I.1.7: Prior Belief (5+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction got 5 or more questions right on the test?

Figure I.1.8: Prior Belief (7+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction got 7 or more questions right on the test?

Figure I.1.9: Prior Belief (poor-2) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as poor on the test?

[_____]

Figure I.1.10: Prior Belief (main self-evaluation) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills on the test?

[_____]

Figure I.1.11: Prior Belief (top-half) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction scored in the top half on the test?

[_____]

Figure I.1.12: Posterior Belief Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

Additional Instructions

In each of next predictions, you will be informed of the average prediction made by all of the female workers who could be randomly selected to be your worker in that prediction.

Worker Predictions:
After completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most-accurate guess to maximize their chance of earning an additional bonus payment of $1.

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Figure I.1.13: Posterior Belief (3+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 78% of female workers predicted that they got 3 or more questions right.

What do you think is the percent chance that your female worker in this prediction got 3 or more questions right?

Figure I.1.14: Posterior Belief (5+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 26% of female workers predicted that they got 5 or more questions right.

What do you think is the percent chance that your female worker in this prediction got 5 or more questions right?

Figure I.1.15: Posterior Belief (7+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 4% of female workers predicted that they got 7 or more questions right.

What do you think is the percent chance that your female worker in this prediction got 7 or more questions right?

Figure I.1.16: Posterior Belief (poor-2) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 81% of female workers predicted that they had an evaluator who described her performance as poor.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as poor?

Figure I.1.17: Posterior Belief (main self-evaluation) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 80% of female workers predicted that they had an evaluator who described her performance as indicative of poor math and science skills.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

Figure I.1.18: Posterior Belief (top-half) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 26% of female workers predicted that they scored in the top half.

What do you think is the percent chance that your female worker in this prediction scored in the top half?

Figure I.1.19: Over/Underconfidence Beliefs Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

Additional Instructions

In the next prediction, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from the next prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

Figure I.1.20: Over/Underconfidence Beliefs (3+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction got fewer than 3 questions right, what do you think is the percent chance that she is overconfident because she predicted that she got 3 or more questions right?

```

```

Underconfidence Prediction: If your female worker in this prediction got 3 or more questions right, what do you think is the percent chance that she is underrconfident because she predicted that she got fewer than 3 questions right?

```

```

Figure I.1.21: Over/Underconfidence Beliefs (5+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction got fewer than 5 questions right, what do you think is the percent chance that she is overconfident because she predicted that she got 5 or more questions right?

Underconfidence Prediction: If your female worker in this prediction got 5 or more questions right, what do you think is the percent chance that she is underrconfident because she predicted that she got fewer than 5 questions right?

Figure I.1.22: Over/Underconfidence Beliefs (7+) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction got fewer than 7 questions right, what do you think is the percent chance that she is overconfident because she predicted that she got 7 or more questions right?

Underconfidence Prediction: If your female worker in this prediction got 7 or more questions right, what do you think is the percent chance that she is underrconfident because she predicted that she got fewer than 7 questions right?

Figure I.1.23: Over/Underconfidence Beliefs (poor-2) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction had an evaluator who described her performance as poor, what do you think is the percent chance that she is overconfident because she predicted that she had an evaluator who did NOT describe her performance as poor?

[_____]

Underconfidence Prediction: If your female worker in this prediction had an evaluator who did NOT describe her performance as poor, what do you think is the percent chance that she is underrconfident because she predicted that she had an evaluator who described her performance as poor?

[_____]

Figure I.1.24: Over/Underconfidence Beliefs (main self-evaluation) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills, what do you think is the percent chance that she is overconfident because she predicted that she had an evaluator who did NOT describe her performance as indicative of poor math and science skills?

[_____]

Underconfidence Prediction: If your female worker in this prediction had an evaluator who did NOT describe her performance as indicative of poor math and science skills, what do you think is the percent chance that she is underrconfident because she predicted that she had an evaluator who described her performance as indicative of poor math and science skills?

[_____]

Figure I.1.25: Over/Underconfidence Beliefs (top-half) about Female Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction did NOT score in the top half, what do you think is the percent chance that she is overconfident because she predicted that she scored in the top half?

Underconfidence Prediction: If your female worker in this prediction scored in the top half, what do you think is the percent chance that she is underrconfident because she predicted that she did NOT score in the top half?

Figure I.1.26: Posterior Belief (3+) about Male Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 83% of male workers predicted that they got 3 or more questions right.

What do you think is the percent chance that your male worker in this prediction got 3 or more questions right?

Figure I.1.27: Posterior Belief (5+) about Male Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 47% of male workers predicted that they got 5 or more questions right.

What do you think is the percent chance that your male worker in this prediction got 5 or more questions right?

Figure I.1.28: Posterior Belief (7+) about Male Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 10% of male workers predicted that they got 7 or more questions right.

What do you think is the percent chance that your male worker in this prediction got 7 or more questions right?

Figure I.1.29: Posterior Belief (poor-2) about Male Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 60% of male workers predicted that they had an evaluator who described his performance as poor.

What do you think is the percent chance that your male worker in this prediction had an evaluator who described his performance as poor?

> [   ]

Figure I.1.30: Posterior Belief (main self-evaluation) about Male Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 56% of male workers predicted that they had an evaluator who described his performance as indicative of poor math and science skills.

What do you think is the percent chance that your male worker in this prediction had an evaluator who described his performance as indicative of poor math and science skills?

> [   ]

Figure I.1.31: Posterior Belief (top-half) about Male Workers, *Evaluator (Alternative Questions) Study*

Prediction X out of 18:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 46% of male workers predicted that they scored in the top half.

What do you think is the percent chance that your male worker in this prediction scored in the top half?

> [   ]

I.2 Full Instructions for the *Evaluator (Attention, Top Half) Study*

All participants in this study are randomized to be asked about male or female workers and are asked to evaluate male or female workers based on whether they scored in the top half of 50 randomly selected male workers and 50 randomly selected female workers.

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. Figures G.1.1 (above) and I.2.1 (below) show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed.

Participants provide their prior beliefs (Figure I.2.2), are provided with information on female workers' self-evaluations and asked to provide their overconfidence and underconfidence beliefs (Figure I.2.3), and then are asked to provide their posterior beliefs (Figure I.2.4). Finally, all participants take a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figures I.2.3 and I.2.4 changes from 26% to **46%**.

Figure I.2.1: Study Overview, *Evaluator (Attention, Top Half) Study*

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your Workers

In each prediction, your worker will be randomly selected from the following group: all of the female workers who had performances in the "middle" (when compared to all female and male workers) on the math and science test. Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

Workers who had performances in the 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.

You will sometimes be provided with information on beliefs held by the relevant workers when they were asked to make predictions about their own performance.

Types of Predictions

In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction scored in the top or bottom half. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Other "comparison" participants involve 50 randomly selected men and 50 randomly selected women from the set of all other participants who took this study.
- A worker scored in the top half if their score (i.e., the number of questions they got right) was greater than or equal to the scores of 50% of the other comparison participants.
- A worker scored in the bottom half if their score (i.e., the number of questions they got right) was less than the scores of 50% of the other comparison participants.

Worker Predictions:

In some predictions, you will be informed of the average prediction made by the female workers who could be randomly selected to be your worker in that prediction.

Thus, please note: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most accurate guess to maximize their chance of earning an additional bonus payment of $1.

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Understanding Question: In each prediction, my worker will be randomly selected from the following group:

all of the female workers

all of the female workers who had performances in the middle

all of the male workers who had performances in the middle

56

Figure I.2.2: Prior Belief about Female Workers, *Evaluator (Attention, Top Half) Study*

Prediction 1 out of 3:
Please provide an integer answer (0, 1, 2, ..., 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will be randomly selected from the following group: all of the female workers who had performances in the middle.

What do you think is the percent chance that your female worker in this prediction scored in the top half?

```
┌─────────────────┐
│                 │
└─────────────────┘
```

Figure I.2.3: Over/Underconfidence Beliefs about Female Workers, *Evaluator (Attention, Top Half) Study*

Prediction 2 out of 3:
Please provide an integer answer (0, 1, 2, ..., 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female workers who had performances in the middle.

After completing the math and science test, 26% of workers in that group predicted that they scored in the top half.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

Overconfidence Prediction: If your female worker in this prediction scored in the bottom half, what do you think is the percent chance that your female worker is overconfident because they predicted that they scored in the top half?

```
┌─────────────────┐
│                 │
└─────────────────┘
```

Underconfidence Prediction: If your female worker in this prediction scored in the top half, what do you think is the percent chance that your female worker is underconfident because they predicted that they scored in the bottom half?

```
┌─────────────────┐
│                 │
└─────────────────┘
```

Figure I.2.4: Posterior Belief about Female Workers, *Evaluator (Attention, Top Half) Study*

Prediction 3 out of 3:

Please provide an integer answer (0, 1, 2, ..., 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female workers who had performances in the middle.

After completing the math and science test, 26% of workers in that group predicted that they scored in the top half.

What do you think is the percent chance that your female worker in this prediction scored in the top half?

```

```

I.3 Full Instructions for *Evaluator (Full Distribution) Study*

In the *Evaluator (Full Distribution) Study*, all participants in this study are randomized to be asked about male or female workers and are asked to consider all male or female workers rather than only those with performances in the "middle."

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. Figures G.1.1 (above), I.3.1, and I.3.2 show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed.

Participants provide their prior beliefs (Figure I.3.3), are provided with information on female workers' self-evaluations and asked to provide their posterior beliefs (Figure I.3.4), and then are asked to provide their overconfidence and underconfidence beliefs (Figure I.3.5). Finally, all participants take a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figures I.3.4 and I.3.5 changes from 76% to **57%**.

Figure I.3.1: Study Overview, *Evaluator (Full Distribution) Study*

Main Instructions (Page 2 out of 2)

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your Workers:
In each prediction, your worker will be randomly selected from the following group: **all of the female workers** who took the math and science test.

Types of Predictions
In each prediction, you will be asked to predict the percent chance that some outcome is true. Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described the worker's performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker had an evaluator who was randomly selected from the set of all other workers who completed the study and who was equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator chose how to describe a worker's performance, the evaluator effectively knew how many questions the worker got right but did not know anything else about the worker, such as the worker's gender.

Worker Predictions:
In some predictions, you will be informed of the average prediction made by the female workers who could be randomly selected to be your worker in that prediction.

Thus, please note: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most accurate guess to maximize their chance of earning an additional bonus payment of $1.

Figure I.3.2: Comprehension Questions about Female Workers, *Evaluator (Full Distribution) Study*

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Understanding Question: In each prediction, my worker will be randomly selected from the following group:

all of the female workers

all of the male workers

Figure I.3.3: Prior Belief about Female Workers, *Evaluator (Full Distribution) Study*

Prediction 1 out of 3:
Please provide an integer answer (0, 1, 2, ..., 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will be randomly selected from the group of all female workers.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

Figure I.3.4: Posterior Belief about Female Workers, *Evaluator (Full Distribution) Study*

Prediction 2 out of 3:
Please provide an integer answer (0, 1, 2, ..., 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the group of all female workers.

After completing the math and science test, 76% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

Figure I.3.5: Over/Underconfidence Beliefs about Female Workers, *Evaluator (Full Distribution) Study*

<u>**Prediction 3 out of 3:**</u>
Please provide an integer answer (0, 1, 2, ..., 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the group of all female workers.

After completing the math and science test, 76% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

<u>**Overconfidence Prediction**</u>**: If your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your female worker is overconfident because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?**

```
┌─────────────────┐
│                 │
└─────────────────┘
```

<u>**Underconfidence Prediction**</u>**: If your female worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your female worker is underconfident because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?**

```
┌─────────────────┐
│                 │
└─────────────────┘
```

I.4   Full Instructions for the *Evaluator (Professional Evaluators) Study*

All participants in this study are randomized to be asked about male or female workers or about "group-1" or "group-2" workers.

I.4.1   Instructions for the *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

The *Baseline* treatment of the *Evaluator (Professional Evaluators) Study* is similar to the *Baseline* treatment of the *Evaluator Study* (Section G.1) with the major difference being that, in the *Evaluator (Professional Evaluators) Study*, the participants are asked about workers from the *Worker (Undergraduate Students) Study* (Section H.1) rather than other Prolific workers. In addition, participants in this study—according to self-reported data collected via Prolific's internal screening questions—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.

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. Figures I.4.1-I.4.3 below show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.4.4), are provided with information on female workers' self-evaluations and asked to provide their posterior beliefs (Figure I.4.5), and are asked to provide their overconfidence and underconfidence beliefs (Figure I.4.6). Finally, participants take a short survey of five randomized bonus questions, as previously shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figures I.4.5 and I.4.6 changes from 59% to **32%**.

<u>**Main Instructions (Page 1 out of 2)**</u>

<u>**Overview:**</u>
This study will consist of 3 predictions and a short follow-up survey. For completing this study, you are guaranteed to receive $2 within 24 hours. In addition, any additional payment you earn will be distributed to you as a bonus payment.

<u>**The Workers**</u>:
In a prior study, we recruited an approximately equal number of male and female undergraduate students from a large midwestern university. These students were assigned the role of "workers" and completed a math and science test with 10 questions. Each question tested their math and science skills by asking them about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on questions like these is often used as a measure of cognitive ability by academic researchers. A worker's score on the test equals the number of questions they answer correctly, and a worker earns 10 cents times their score.

<u>**Your Predictions**</u>:
You will be asked to make 3 predictions related to the **performance of workers on the math and science test**.

To maximize your chance of earning an additional payment of $1, you should provide your most-accurate guess when making each prediction. This is because each prediction will ask you to guess the percent chance of some outcome being true. In each of those predictions, to secure the largest chance of earning $1 from the prediction, you should report your most-accurate guess. To learn the precise payment rule that determines how much you earn from these predictions click here.

One of your 3 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

It doesn't matter

As accurately as possible

Randomly

Understanding Question: How much additional payment will you receive?

I will receive what I earn from all predictions in this part

I will receive what I earn from the prediction-that-counts

Nothing

Figure I.4.2: Instructions about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

<u>**Your Workers**</u>
In each prediction, your worker will be randomly selected from the following group: all of the female undergraduate students who completed the prior study and expect to graduate in Spring 2023. Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

You will sometimes be provided with information on beliefs held by the relevant workers when they were asked to make predictions about their own performance.

<u>**Types of Predictions**</u>
In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described the worker's performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker had an evaluator who was randomly selected from the set of all other workers who completed the prior study and who was equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator choose how to describe a worker's performance, the evaluator effectively knew how many questions the worker got right but did not know anything else about the worker, such as the worker's gender.

<u>**Worker Predictions**</u>:
In some predictions, you will be informed of the average prediction made by the female workers who could be randomly selected to be your worker in that prediction.

Thus, please note that: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most accurate guess to maximize their chance of earning an additional bonus payment of $1.

Figure I.4.3: Comprehension Questions about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Understanding Question: In each prediction, my worker will be randomly selected from the following group:

all of the female undergraduate students who completed the prior study

all of the female undergraduate students who completed the prior study and expect to graduate in Spring 2023

all of the male undergraduate students who completed the prior study and expect to graduate in Spring 2023

Figure I.4.4: Prior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Prediction 1 out of 3:

Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will be randomly selected from the following group: all of the female undergraduate students who completed the prior study and expected to graduate in Spring 2023.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[ ]

Figure I.4.5: Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Prediction 2 out of 3:

Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female undergraduate students who completed the prior study and expected to graduate in Spring 2023.

After completing the math and science test, 59% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that you female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[ ]

Figure I.4.6: Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

<u>**Prediction 3 out of 3:**</u>
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your female worker will again be randomly selected from the following group: all of the female undergraduate students who completed the prior study and expected to graduate in Spring 2023.

After completing the math and science test, 59% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

<u>**Overconfidence Prediction**</u>**: If your female worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your female worker is overconfident because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?**

```
┌─────────────────┐
│                 │
└─────────────────┘
```

<u>**Underconfidence Prediction**</u>**: If your female worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your female worker is underconfident because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?**

```
┌─────────────────┐
│                 │
└─────────────────┘
```

I.4.2   Instructions for the *Baseline, Unknown Gender* Treatment of the *Evaluator (Professional Evaluators) Study*

The *Baseline, Unknown Gender* treatment of the *Evaluator (Professional Evaluators) Study* is the same as the *Baseline* treatment of the *Evaluator (Professional Evaluators) Study* (Section I.4.1) except "male" and "female" are replaced with "group-1" and "group-2," respectively, and worker gender is unknown to participants.

After consenting to participate in the study, each participant is informed of the $2 study completion fee and of the opportunity to earn additional payment. Figures I.4.1 (shown above), I.4.7, and I.4.8 show the overview participants who are randomized to evaluate **group-2 workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.4.9), are provided with information on group-2 workers' self-evaluations and asked to provide their posterior beliefs (Figure I.4.10), and are asked to provide their overconfidence and underconfidence beliefs (Figure I.4.11). Finally, participants take a short survey of five randomized bonus questions, as previously shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **group-1 workers** (considered "male workers" in the Treatment of the *Evaluator (Professional Evaluators) Study* (Section I.4.1)), "group-2" is replaced by "group-1" everywhere, and the self-evaluation information provided in Figures I.4.10 and I.4.11 changes from 59% to **32%**.

Figure I.4.7: Instructions about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

<u>**Main Instructions (Page 2 out of 2)**</u>

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

<u>**Your Workers**</u>
In each prediction, your worker will be randomly selected from the following group: all of the group-2 undergraduate students who completed the prior study and expect to graduate in Spring 2023. Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

We assigned each worker to group-1 or group-2 based on an answer they provided to a question in the follow-up survey. While you will not be informed of their answer to this follow-up survey question, you will sometimes be provided with information on beliefs held by the relevant workers when they were asked to make predictions about their own performance.

<u>**Types of Predictions**</u>
In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described the worker's performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker had an evaluator who was randomly selected from the set of all other workers who completed the prior study and who was equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator choose how to describe a worker's performance, the evaluator effectively knew how many questions the worker got right but did not know anything else about the worker, such as the worker's gender.

<u>**Worker Predictions**</u>:
In some predictions, you will be informed of the average prediction made by the group-2 workers who could be randomly selected to be your worker in that prediction.

Thus, please note that: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most accurate guess to maximize their chance of earning an additional bonus payment of $1.

71

Figure I.4.8: Comprehension Questions about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Understanding Question: In each prediction, my worker will be randomly selected from the following group:

all of the group-2 undergraduate students who completed the prior study

all of the group-2 undergraduate students who completed the prior study and expect to graduate in Spring 2023

all of the group-1 undergraduate students who completed the prior study and expect to graduate in Spring 2023

Figure I.4.9: Prior Belief about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Prediction 1 out of 3:
Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your group-2 worker will be randomly selected from the following group: all of the group-2 undergraduate students who completed the prior study and expected to graduate in Spring 2023.

What do you think is the percent chance that your group-2 worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

Figure I.4.10: Posterior Belief about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Prediction 2 out of 3:

Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your group-2 worker will again be randomly selected from the following group: all of the group-2 undergraduate students who completed the prior study and expected to graduate in Spring 2023.

After completing the math and science test, 59% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that you group-2 worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[ ]

Figure I.4.11: Over/Underconfidence Beliefs about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Prediction 3 out of 3:

Please provide an integer answer (0, 1, 2,...100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your group-2 worker will again be randomly selected from the following group: all of the group-2 undergraduate students who completed the prior study and expected to graduate in Spring 2023.

After completing the math and science test, 59% of workers in that group predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

Overconfidence Prediction: If your group-2 worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your group-2 worker is overconfident because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?

Underconfidence Prediction: If your group-2 worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your group-2 worker is underconfident because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?

I.5 Full Instructions for the *Evaluator (Extended) Study*

All participants in this study are randomized to be asked about male or female workers and to be in one of four treatments described below.

I.5.1 Instructions for the *Baseline* Treatment of the *Evaluator (Extended) Study*

After consenting to participate in the study, each participant is informed of the \$3 study completion fee and of the opportunity to earn additional payment. Figures I.5.1-I.5.3 below show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.5.4), are provided with information on 20 female workers' self-evaluations and asked to provide their posterior beliefs (see Figures I.5.5 and I.5.6 for additional instructions and an example). Participants then are provided with additional instructions and asked to provide their posterior belief about the average self-evaluation of female workers (Figures I.5.7 and I.5.8) and are asked to provide their overconfidence and underconfidence beliefs (Figures I.5.9 and I.5.10). Finally, all participants complete a follow-up survey that collects additional control and demographic information.

For evaluators who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figure I.5.8 changes from 80% to **56%**.

Main Instructions (Page 1 out of 2)

Overview:

This study will consist of a series of predictions and a short follow-up survey. For completing this study, you are guaranteed to receive $3 within 24 hours. In addition, any additional payment you earn will be distributed to you as a bonus payment.

The Workers:

In a prior study, an approximately equal number of men and women (called "workers") completed a math and science test with 10 questions. Each question tested their math and science skills by asking them about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on questions like these is often used as a measure of cognitive ability by academic researchers. A worker's score on the test equals the number of questions they answer correctly, and a worker earns 10 cents times their score.

Your Predictions:

You will be asked to make 23 predictions related to the **performance of workers on the math and science test**.

To maximize your chance of earning an additional payment of $1, you should provide your most-accurate guess when making each prediction. This is because each prediction will ask you to guess the percent chance of of some outcome being true. In each of those predictions, to secure the largest chance of earning $1 from the prediction, you should report your most-accurate guess. To learn the precise payment rule that determines how much you earn from these predictions click here.

One of your 23 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

It doesn't matter

As accurately as possible

Randomly

Understanding Question: How much additional payment will you receive?

I will receive what I earn from all predictions in this part

I will receive what I earn from the prediction-that-counts

Nothing

Figure I.5.2: Instructions about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Main Instructions (Page 2 out of 2)

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your workers

In each prediction, your worker will be randomly selected from the group of female workers who had performances in the "middle" (when compared to all male and female workers) on the math and science test. Specifically, your female worker will be randomly selected from the group of all female workers who had performances in the middle (when compared to all male and female workers). Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

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.

Types of Predictions

In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described their performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker has an evaluator who is randomly selected from the set of all other workers who complete the study and who is equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator chooses how to describe a worker's performance, the evaluator effectively knows how many questions the worker got right but does not know anything else about the worker, such as the worker's gender.

Figure I.5.3: Comprehension Questions about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Understanding Question: In each prediction, my worker will be randomly selected from...

the entire group of workers

the group of female workers with performances in the middle

the group of male workers with performances in the middle

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Figure I.5.4: Prior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

Figure I.5.5: Additional Instructions about Worker-Specific Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Additional Instructions

In each of the next predictions, you will be informed of the prediction made by your worker when that worker was asked to make a prediction about their own performance.

Worker Predictions:

After completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most-accurate guess to maximize their chance of earning an additional bonus payment of $1.

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Figure I.5.6: Example of Worker-Specific Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:

In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, **your female worker** in this prediction predicted that there is a **50%** chance that her evaluator described her performance as indicative of poor math and science skills.

What do you think is the percent chance that **your female worker** in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

Figure I.5.7: Additional Instructions about Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Additional Instructions

In the next prediction, rather than being informed of the prediction made by your worker in that prediction, you will be informed of the average prediction made by all of the female workers who could be randomly selected to be your worker in that prediction.

Figure I.5.8: Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 80% of female workers predicted that their evaluator described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

Figure I.5.9: Additional Instructions about Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Additional Instructions

In the next prediction, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from the next prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

Figure I.5.10: Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills, what do you think is the **percent chance that she is overconfident** because she predicted that her evaluator did NOT describe her performance as indicative of poor math and science skills?

```
┌─────────────────────┐
│                     │
└─────────────────────┘
```

Underconfidence Prediction: If your female worker in this prediction had an evaluator who did NOT describe her performance as indicative of poor math and science skills, what do you think is the **percent chance that she is underconfident** because she predicted that her evaluator described her performance as indicative of poor math and science skills?

```
┌─────────────────────┐
│                     │
└─────────────────────┘
```

I.5.2 Instructions for the *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

For the *Evaluator (Extended) Study*, the *Joint Evaluations* treatment differs from the *Baseline* treatment (Section I.5.1) by asking about both a male worker and a female worker on each decision screen.

After consenting to participate in the study, each participant is informed of the $3 study completion fee and of the opportunity to earn additional payment. Figures I.5.1 (shown above) and I.5.11 (below) show the overview participants who are randomized to evaluate **female and male workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.5.12), are provided with the self-evaluations of 20 female and 20 male workers and asked to provide their posterior beliefs (see Figure I.5.13 for an example), are asked for their posterior belief about male and female workers' average self-evaluations (Figure I.5.14), and are asked to provide their overconfidence and underconfidence beliefs (Figure I.5.15). Finally, all participants complete a follow-up survey that collects additional control and demographic information.

<u>Main Instructions (Page 2 out of 2)</u>

In each prediction set, we will ask you to make make two predictions: one about a female worker and one about a male worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

<u>Your workers</u>
In each prediction set, your two workers will be randomly selected from the group of male and female workers who had performances in the "middle" (when compared to all male and female workers) on the math and science test. Specifically, your male worker will be randomly selected from the group of all male workers who had performances in the middle (when compared to all male and female workers), and your female worker will be randomly selected from the group of all female workers who had performances in the middle (when compared to all male and female workers). Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

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.

<u>Types of Predictions</u>
In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described their performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker has an evaluator who is randomly selected from the set of all other workers who complete the study and who is equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator chooses how to describe a worker's performance, the evaluator effectively knows how many questions the worker got right but does not know anything else about the worker, such as the worker's gender.

Understanding Question: In each prediction, my female/male worker will be randomly selected from...

the entire group of female/male workers

the group of female/male workers with performances in the middle

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Figure I.5.12: Prior Belief about Female and Male Workers, *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

<u>Prediction Set X out of 13</u>:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your male worker in this prediction set had an evaluator who described his performance as indicative of poor math and science skills?

> [   ]

What do you think is the percent chance that your female worker in this prediction set had an evaluator who described her performance as indicative of poor math and science skills?

> [   ]

Figure I.5.13: Example of Worker-Specific Posterior Belief about Female and Male Workers, *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

<u>**Prediction Set X out of 13**</u>:

In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test:

- **your female worker in this prediction set predicted that there is a 85% chance that her evaluator described her performance as indicative of poor math and science skills, and**
- **your male worker in this prediction set predicted that there is a 92% chance that his evaluator described his performance as indicative of poor math and science skills.**

For your workers in this prediction set, what do you think is the percent chance that their evaluator described their performance as indicative of poor math and science skills?

Percent chance your female worker in this prediction set had an evaluator who described her performance as indicative of poor math and science skills:

[   ]

Percent chance your male worker in this prediction set had an evaluator who described his performance as indicative of poor math and science skills:

[   ]

Figure I.5.14: Posterior Belief about Female and Male Workers, *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

<u>**Prediction Set X out of 13**</u>:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test:

- **56% of male workers predicted** that their evaluator described their performance as indicative of poor math and science skills, and
- **80% of female workers predicted** that their evaluator described their performance as indicative of poor math and science skills.

For your workers in prediction set, what do you think is the percent chance that their evaluator described their performance as indicative of poor math and science skills?

Percent chance your male worker in this prediction set had an evaluator who described his performance as indicative of poor math and science skills:

☐

Percent chance your female worker in this prediction set had an evaluator who described her performance as indicative of poor math and science skills:

☐

Figure I.5.15: Over/Underconfidence Beliefs about Female and Male Workers, *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

<u>**Prediction Set X out of 13**</u>:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

<u>**Overconfidence Prediction**</u>: **If your male/female worker in prediction set had an evaluator 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 evaluator did NOT describe their performance as indicative of poor math and science skills?**

Percent chance your male worker **in this prediction set is** overconfident:

[ ]

Percent chance your female worker **in this prediction set is** overconfident:

[ ]

<u>**Underconfidence Prediction**</u>: **If your male/female worker in this prediction set had an evaluator 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 evaluator described their performance as indicative of poor math and science skills?**

Percent chance your male worker **in this prediction set is** underconfident:

[ ]

Percent chance your female worker **in this prediction set is** underconfident:

[ ]

I.5.3 Instructions for the *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

For the *Evaluator (Extended) Study*, the *Strategic Incentives* treatment differs from the *Baseline* treatment (Section I.5.1) only in that participants are instead asked about workers who face strategic incentives.

After consenting to participate in the study, each participant is informed of the $3 study completion fee and of the opportunity to earn additional payment. Figures I.5.1 (shown above) and I.5.16 (below) show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.5.17), are provided with information on 20 female workers' self-evaluations and asked to provide their posterior beliefs (see new Figures I.5.18 and I.5.19 below for additional instructions and an example). Participants then are provided with additional instructions and asked to provide their posterior belief about the average self-evaluation of female workers (Figures I.5.20 and I.5.21) and are asked to provide their overconfidence and underconfidence beliefs (Figures I.5.22 and I.5.23). Finally, all participants complete a follow-up survey that collects additional control and demographic information.

For evaluators who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figure I.5.21 changes from 74% to **57%**.

Figure I.5.16: Study Overview, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

In each of the next predictions, you will be informed of the prediction made by your worker when that worker was asked to make a prediction about their own performance.

Worker Predictions:
After completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that an employer would decide whether to hire them after the employer learned one of their predictions. They knew that the employer would only learn this prediction before deciding to hire them or not—the employer would not learn any demographic information about the worker and would not learn the worker's true score on the math and science test.

Specifically, a worker knew that:
- If their employer chooses NOT to hire them, then the worker would earn 50 cents and their employer would earn 50 cents.
- If their employer chooses to hire them, then the worker would earn 100 cents and their employer would earn 10 cents times the number of questions that the worker answered correctly on the test.

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Figure I.5.17: Prior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

Figure I.5.18: Additional Instructions about Worker-Specific Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

<u>**Additional Instructions**</u>

In each of the next predictions, you will be informed of the prediction made by your worker when that worker was asked to make a prediction about their own performance.

<u>**Worker Predictions**</u>:
After completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that an employer would decide whether to hire them after the employer learned one of their predictions. They knew that the employer would only learn this prediction before deciding to hire them or not—the employer would not learn any demographic information about the worker and would not learn the worker's true score on the math and science test.

Specifically, a worker knew that:
- If their employer chooses NOT to hire them, then the worker would earn 50 cents and their employer would earn 50 cents.
- If their employer chooses to hire them, then the worker would earn 100 cents and their employer would earn 10 cents times the number of questions that the worker answered correctly on the test.

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

91

Figure I.5.19: Worker-Specific Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, your female worker in this prediction predicted that there is a 50% chance that her evaluator described her performance as indicative of poor math and science skills.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

[ ]

Figure I.5.20: Additional Instructions about Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

Additional Instructions

In the next prediction, rather than being informed of the prediction made by your worker in that prediction, you will be informed of the average prediction made by all of the female workers who could be randomly selected to be your worker in that prediction.

Figure I.5.21: Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test, 74% of female workers predicted that their evaluator described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

[ ]

Figure I.5.22: Additional Information about Over/Underconfidence Beliefs about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

Additional Instructions

In the next prediction, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from the next prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

Figure I.5.23: Over/Underconfidence Beliefs about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:

In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills, what do you think is the percent chance that she is overconfident because she predicted that her evaluator did NOT describe her performance as indicative of poor math and science skills?

[ ]

Underconfidence Prediction: If your female worker in this prediction had an evaluator who did NOT describe her performance as indicative of poor math and science skills, what do you think is the percent chance that she is underconfident because she predicted that her evaluator described her performance as indicative of poor math and science skills?

[ ]

I.5.4 Instructions for the *Joint Evaluations, Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

The *Joint Evaluations, Strategic Incentives* treatment differs from the *Joint Evaluations* treatment (Section I.5.2) in the same way as the *Strategic Incentives* treatment (Section I.5.3) differs from the *Baseline* treatment (Section I.5.1). Participants are asked about workers who face strategic incentives and are asked about both a male worker and a female worker on each decision screen. See Figures I.5.24, I.5.25, I.5.26, and I.5.27 for the prior belief, worker-specific posterior belief, posterior belief about average self-evaluations, and overconfidence and underconfidence beliefs questions about **female workers** and **male workers**, respectively.

Figure I.5.24: Prior Belief about Male and Female Workers, *Strategic Incentives, Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:

In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

What do you think is the percent chance that your female worker in this prediction had an evaluator who described her performance as indicative of poor math and science skills?

[   ]

What do you think is the percent chance that your male worker in this prediction had an evaluator who described his performance as indicative of poor math and science skills?

[   ]

Figure I.5.25: Worker-Specific Posterior Belief about Male and Female Workers, *Strategic Incentives, Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:

In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test:

- **your female worker in this prediction set predicted that there is a 90% chance that her evaluator described her performance as indicative of poor math and science skills, and**
- **your male worker in this prediction set predicted that there is a 80% chance that his evaluator described his performance as indicative of poor math and science skills.**

For your workers in this prediction set, what do you think is the percent chance that their evaluator described their performance as indicative of poor math and science skills?

Percent chance your female worker in this prediction set had an evaluator who described her performance as indicative of poor math and science skills:

[   ]

Percent chance your male worker in this prediction set had an evaluator who described his performance as indicative of poor math and science skills:

[   ]

Figure I.5.26: Posterior Belief about Male and Female Workers, *Strategic Incentives, Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

Prediction Set X out of 13:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

After completing the math and science test:

- **57% of male workers predicted** that their evaluator described their performance as indicative of poor math and science skills, and
- **74% of female workers predicted** that their evaluator described their performance as indicative of poor math and science skills.

For your workers in prediction set, what do you think is the percent chance that their evaluator described their performance as indicative of poor math and science skills?

Percent chance your male worker in this prediction set had an evaluator who described his performance as indicative of poor math and science skills:

☐

Percent chance your female worker in this prediction set had an evaluator who described her performance as indicative of poor math and science skills:

☐

Figure I.5.27: Over/Underconfidence Beliefs about Male and Female Workers, *Strategic Incentives, Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

Prediction X out of 23:
In each prediction, please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

Overconfidence Prediction: If your female/male worker in prediction set had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your female/male worker is overconfident because they predicted that their evaluator did NOT describe their performance as indicative of poor math and science skills?

Percent chance your female worker in this prediction set is overconfident:

[   ]

Percent chance your male worker in this prediction set is overconfident:

[   ]

Underconfidence Prediction: If your female/male worker in this prediction set had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your female/male worker is underconfident because they predicted that their evaluator described their performance as indicative of poor math and science skills?

Percent chance your female worker in this prediction set is underconfident:

[   ]

Percent chance your male worker in this prediction set is underconfident:

[   ]

I.6 Full Instructions for the *Evaluator (Additional Demographics) Study*

All participants in this study are randomized to be asked about male or female workers.

After consenting to participate in the study, each participant is informed of the $1.50 study completion fee and of the opportunity to earn additional payment. Figures I.6.1-I.6.3 show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.6.4), posterior beliefs (Figure I.6.5), and their overconfidence and underconfidence beliefs (Figure I.6.6). Finally, all participants take a short survey of five randomized bonus questions, as shown above in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figure I.6.5 changes from 68% to **38%**.

Main Instructions (Page 1 out of 2)

Overview:
This study will consist of 3 predictions and a short follow-up survey. For completing this study, you are guaranteed to receive $1.50 within 24 hours. In addition, any additional payment you earn will be distributed to you as a bonus payment.

The Workers:
In a prior study, an approximately equal number of men and women (called "workers") completed a math and science test with 10 questions. Each question tested their math and science skills by asking them about general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. Performance on questions like these is often used as a measure of cognitive ability by academic researchers. A worker's score on the test equals the number of questions they answer correctly, and a worker earns 10 cents times their score.

Your Predictions:
You will be asked to make 3 predictions related to the **performance of workers on the math and science test**.

To maximize your chance of earning an additional payment of $1, you should provide your most-accurate guess when making each prediction. This is because each prediction will ask you to guess the percent chance of of some outcome being true. In each of those predictions, to secure the largest chance of earning $1 from the prediction, you should report your most-accurate guess. To learn the precise payment rule that determines how much you earn from these predictions click here.

One of your 3 predictions will be randomly selected as the prediction-that-counts, and your additional payment will equal the amount you earn in the prediction-that-counts.

Understanding Question: To maximize your chance of additional bonus payment, how should you make predictions in this part?

It doesn't matter

As accurately as possible

Randomly

Understanding Question: How much additional payment will you receive?

I will receive what I earn from all predictions in this part

I will receive what I earn from the prediction-that-counts

Nothing

Figure I.6.2: Instructions about Female Workers, *Evaluator (Additional Demographics) Study*

Main Instructions (Page 2 out of 2)

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your workers
In each prediction, your worker will be randomly selected from the group of workers who:

- **work full time,**
- **are between 26 and 40 years old**
- **live in the Southern region of the United States,**
- **have at least completed some college education, and**
- **are female.**

Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

Figure I.6.3: Instructions about Female Workers cont., *Evaluator (Additional Demographics) Study*

Types of Predictions

In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described their performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker has an evaluator who is randomly selected from the set of all other workers who complete the study and who is equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator chooses how to describe a worker's performance, the evaluator effectively knows how many questions the worker got right but does not know anything else about the worker, such as the worker's gender.

Worker Predictions:

In some predictions, you will be informed of the average prediction made by the group of workers who could be randomly selected to be your worker in that prediction.

Thus please note that: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most-accurate guess to maximize their chance of earning an additional bonus payment of $1.

Figure I.6.4: Prior Belief about Female Workers, *Evaluator (Additional Demographics) Study*

Prediction 1 out of 3:

Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your worker will be randomly selected from the group of workers who:

- work full time,
- are between 26 and 40 years old
- live in the Southern region of the United States,
- have at least completed some college education, and
- are female.

What do you think is the percent chance that your worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[   ]

Figure I.6.5: Posterior Belief about Female Workers, *Evaluator (Additional Demographics) Study*

Prediction 2 out of 3:

Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your worker will again be randomly selected from the group of workers who:

- work full time,
- are between 26 and 40 years old
- live in the Southern region of the United States,
- have at least completed some college education, and
- are female.

After completing the math and science test, 77% of workers from the above group of workers predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

[   ]

Figure I.6.6: Over/Underconfidence Beliefs about Female Workers, *Evaluator (Additional Demographics) Study*

<u>**Prediction 3 out of 3:**</u>
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your worker will again be randomly selected from the group of workers who:

- work full time,
- are between 26 and 40 years old
- live in the Southern region of the United States,
- have at least completed some college education, and
- are female.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

<u>**Overconfidence Prediction:**</u> **If your worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the percent chance that your worker is overconfident because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?**

☐

<u>**Underconfidence Prediction:**</u> **If your worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the percent chance that your worker is underconfident because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?**

☐

I.7 Full Instructions for the *Evaluator (Known Performance) Study*

All participants in this study are randomized to be asked about male or female workers.

After consenting to participate in the study, each participant is informed of the $1.50 study completion fee and of the opportunity to earn additional payment. Previous Figure I.6.1 and Figures I.7.1-I.7.3 show the overview participants who are randomized to evaluate **female workers** are given and the corresponding comprehension questions they must answer correctly in order to proceed. Then, participants provide their prior beliefs (Figure I.7.4), posterior beliefs (Figure I.7.5), and their overconfidence and underconfidence beliefs (see Figure I.7.6). Finally, all participants take a short survey of five randomized bonus questions, as previously shown in Figures G.1.7-G.1.12, and a follow-up survey that collects additional control and demographic information.

For participants who are randomized to be asked about **male workers**, "female" is replaced by "male" everywhere, and the self-evaluation information provided in Figure I.7.5 changes from 68% to **41%**.

Figure I.7.1: Study Overview, *Evaluator (Known Performance) Study*

In each prediction, we will ask you to make a prediction about the performance of your worker. Below please learn more about your workers and the types of predictions we will ask you to make about your workers.

Your workers
In each prediction, your worker will be randomly selected from the group of all female workers who got 5 questions right on the math and science test.

Your worker in one prediction will never be the same as your worker in another prediction. **Thus, you will never be asked about the same worker twice.**

Figure I.7.2: Instructions about Female Workers, *Evaluator (Known Performance) Study*

Types of Predictions

In each prediction, you will be asked to predict the **percent chance that some outcome is true.** Sometimes, you will be asked to predict the percent chance that your worker in that prediction had an evaluator who described their performance as indicative of poor math and science skills. Other times, you will be asked to predict the percent chance that your worker is overconfident or underconfident when asked to make predictions about their own performance. Thus, please note the following:

- Each worker has an evaluator who is randomly selected from the set of all other workers who complete the study and who is equally likely to be a man or a woman. The evaluators answered a question about which scores they believed were indicative of poor math and science skills.
- An evaluator is said to have described a worker's performance as indicative of poor math and science skills if they indicated the worker's score was indicative of poor math and science skills.
- Thus, when an evaluator chooses how to describe a worker's performance, the evaluator effectively knows how many questions the worker got right but does not know anything else about the worker, such as the worker's gender.

Worker Predictions:

In some predictions, you will be informed of the average prediction made by the group of workers who could be randomly selected to be your worker in that prediction.

Thus please note that: after completing the math and science test, workers made predictions about their performance on the math and science test. When making these predictions, they were not given any information on their own performance and knew that they that should report their most-accurate guess to maximize their chance of earning an additional bonus payment of $1.

Figure I.7.3: Comprehension Questions about Female Workers, *Evaluator (Known Performance) Study*

Understanding Question: When an evaluator describes a worker's performance, do they know the gender of the worker?

Yes

No

Understanding Question: A worker should expect to earn more...

if they provided more accurate predictions

if they got hired (regardless of how accurate their predictions were)

Figure I.7.4: Prior Belief about Female Workers, *Evaluator (Known Performance) Study*

Prediction 1 out of 3:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your worker will be a randomly selected from the group of female workers who got 5 questions right on the math and science test.

What do you think is the percent chance that your worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

Figure I.7.5: Posterior Belief about Female Workers, *Evaluator (Known Performance) Study*

Prediction 2 out of 3:
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your worker will again be a randomly selected from the group of female workers who got 5 questions right on the math and science test.

After completing the math and science test, 68% of female workers who got 5 questions right predicted that they had an evaluator who described their performance as indicative of poor math and science skills.

What do you think is the percent chance that your worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills?

Figure I.7.6: Over/Underconfidence Beliefs about Female Workers, *Evaluator (Known Performance) Study*

<u>**Prediction 3 out of 3:**</u>
Please provide an integer answer (from 0 to 100) and please omit the percent sign in your answer. For example, please type 0 if your answer is 0%, 100 if your answer is 100%, etc.

In this prediction, your worker will again be a randomly selected from the group of female workers who got 5 questions right on the math and science test.

Below, we will ask you to make predictions about the percent chance that your worker is overconfident or underconfident, depending on that worker's performance. Specifically, we will ask you to make one guess for each performance that your worker could have had. **For determining how much money you earn from this prediction, we will then only consider your guess that corresponds to the prediction that your worker actually made.**

<u>**Overconfidence Prediction:**</u> **If your worker in this prediction had an evaluator who described their performance as indicative of poor math and science skills, what do you think is the** percent chance that your worker is overconfident **because they predicted that they had an evaluator who did NOT describe their performance as indicative of poor math and science skills?**

```
┌──────────────┐
│              │
└──────────────┘
```

<u>**Underconfidence Prediction:**</u> **If your worker in this prediction had an evaluator who did NOT describe their performance as indicative of poor math and science skills, what do you think is the** percent chance that your worker is underconfident **because they predicted that they had an evaluator who described their performance as indicative of poor math and science skills?**

```
┌──────────────┐
│              │
└──────────────┘
```