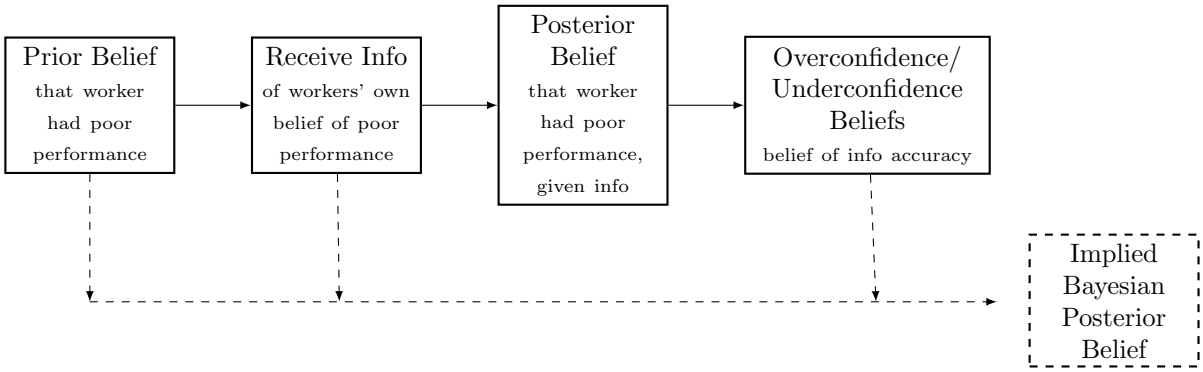
Online Appendix for The Gender Gap in Confidence: Expected But Not Accounted For

By: Christine L. Exley and Kirby Nielsen

Appendix Table of Contents

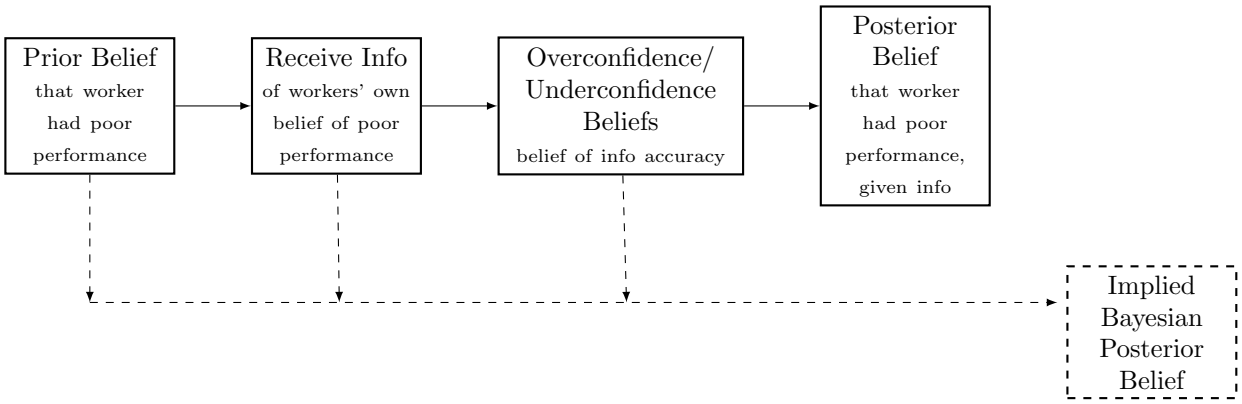
A Additional Design Details

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



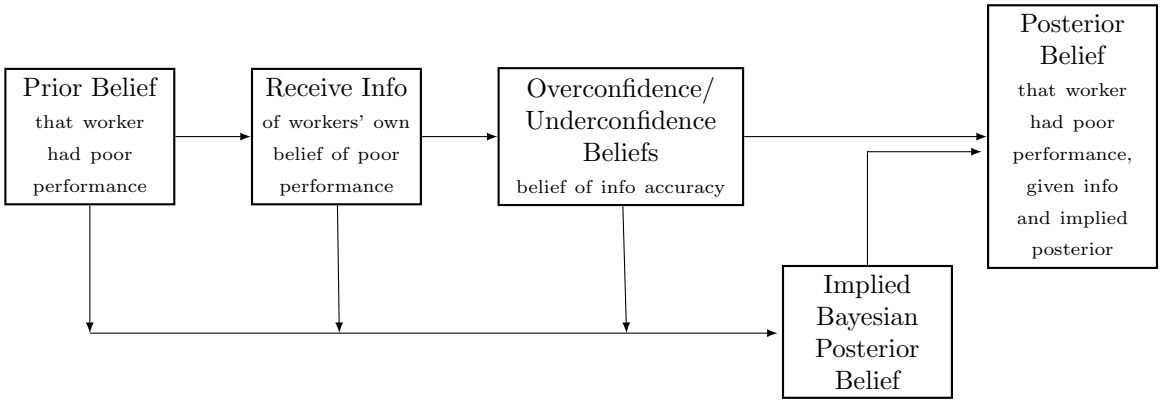
In the *Baseline* and *Baseline, Unknown Gender* treatments, we elicit an evaluator's prior belief that a randomly selected male or female worker had a poor performance. Then, we provide evaluators with the percentage of male or female workers who believed they had a poor performance. After this, we 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.

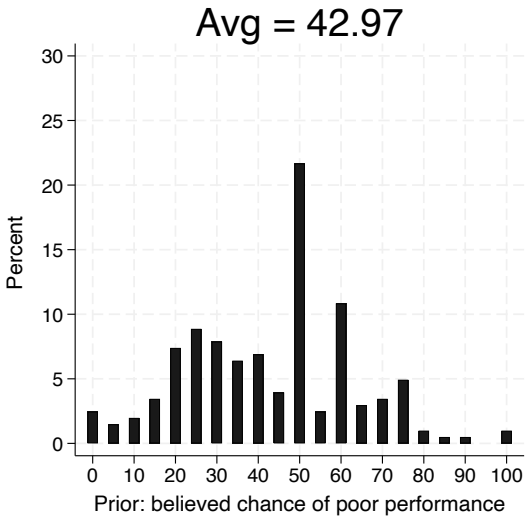
Table B.7: Evaluators' Beliefs in the *Baseline, Unknown Gender*, *Attention, Unknown Gender* and *Calculation, Unknown Gender* treatment of the *Evaluator Study*

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

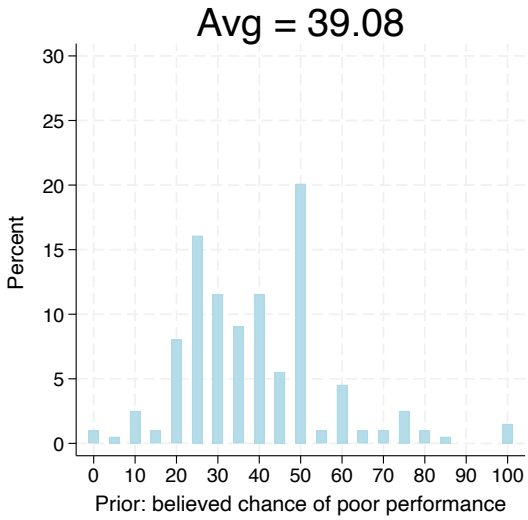
SEs are robust and shown in parentheses. Results follow the structure of Table 3. Data are from the 1190 participants in the *Baseline, Unknown Gender* treatment, the *Attention, Unknown Gender* or the *Calculation, Unknown Gender* treatment of *Evaluator Study*. Sample size differs slightly in column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

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

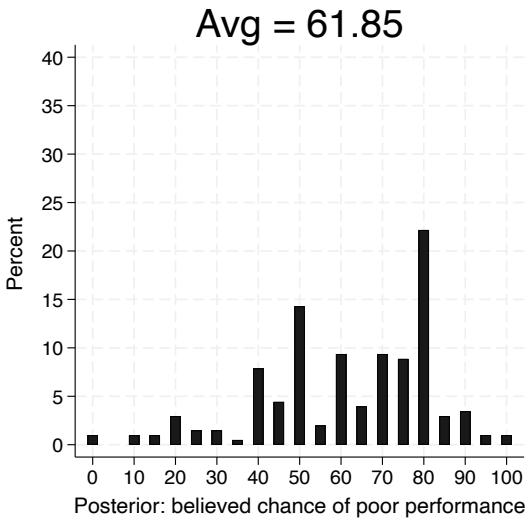
A: Prior Beliefs about Women



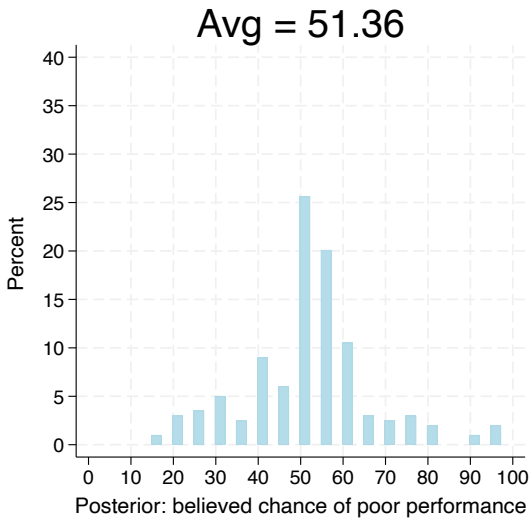
B: Prior Beliefs about Men



C: Posterior Beliefs about Women



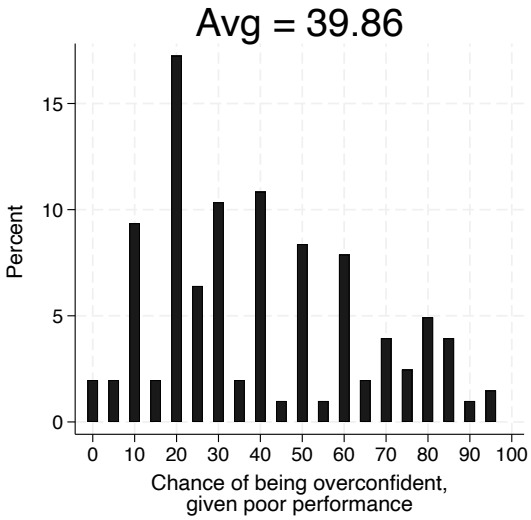
D: Posterior Beliefs about Men



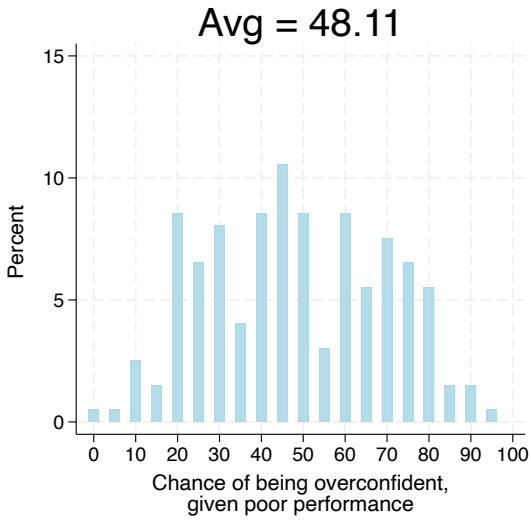
Data are from the *Baseline* treatment of the *Evaluator Study*.

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

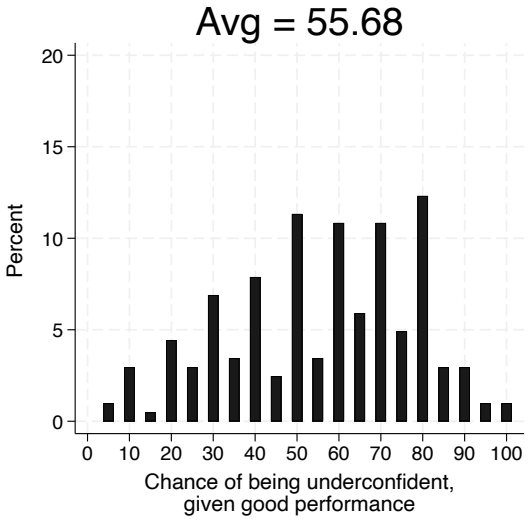
A: Overconfidence Beliefs about Women



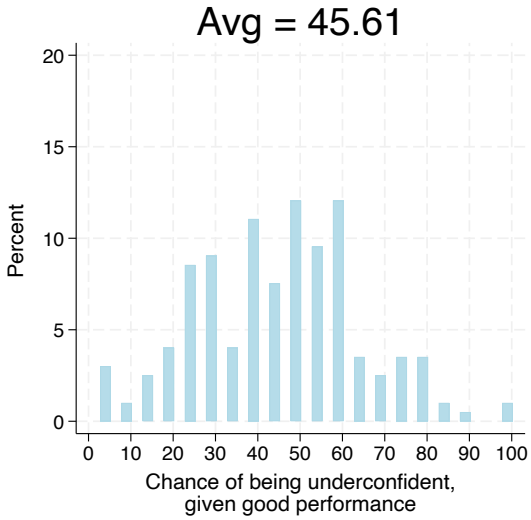
B: Overconfidence Beliefs about Men



C: Underconfidence Beliefs about Women



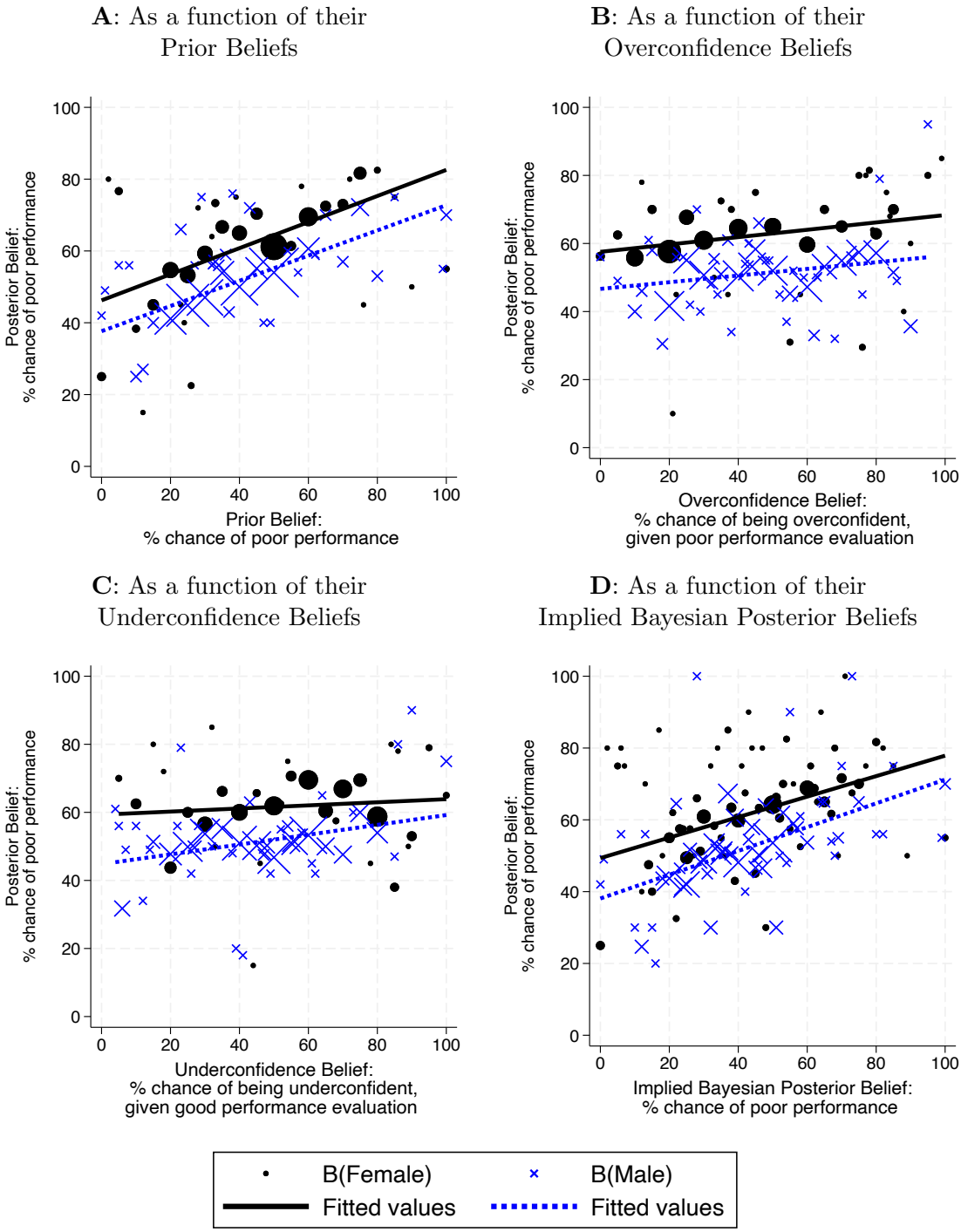
D: Underconfidence Beliefs about Men



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

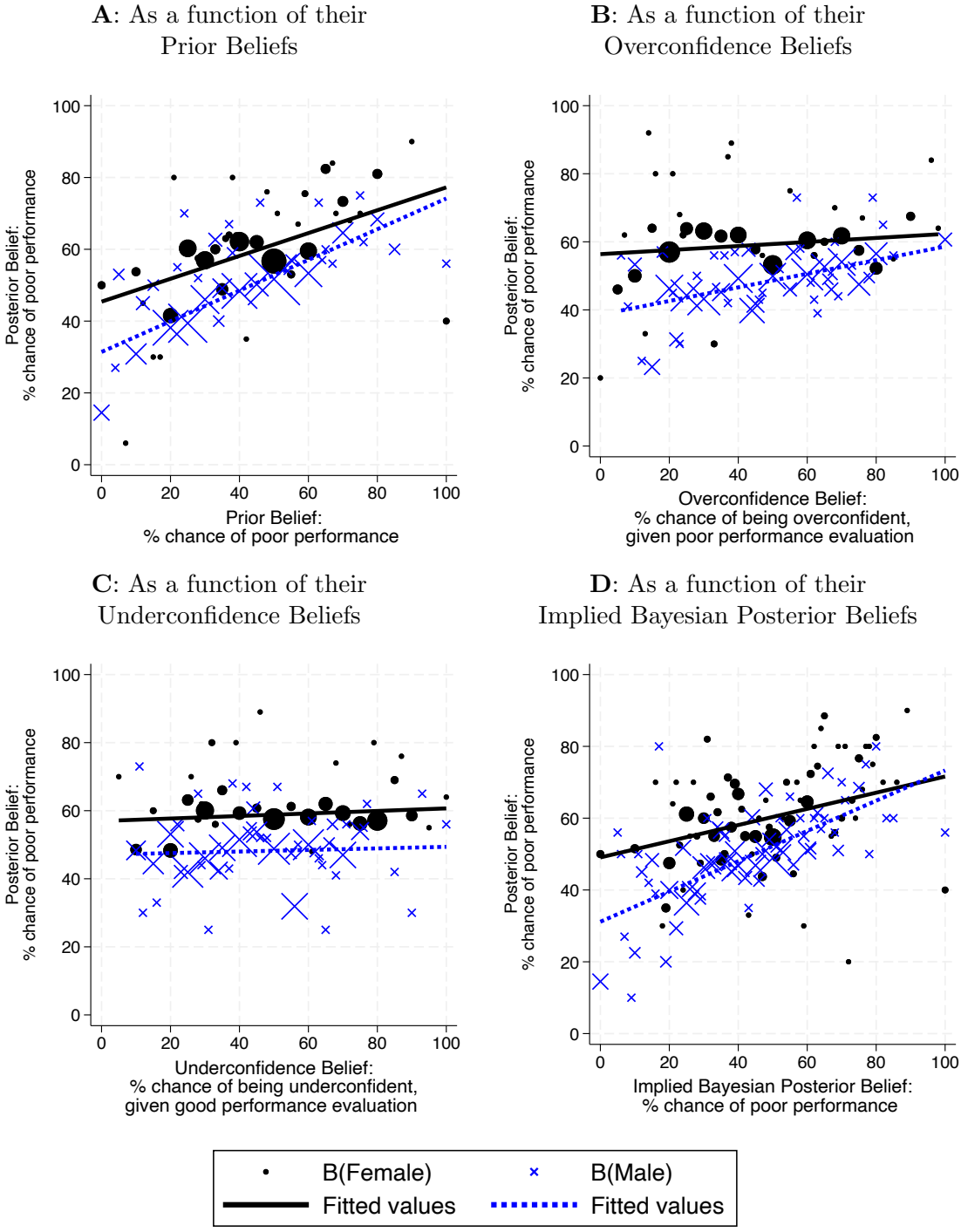
C Additional Heterogeneity Results

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



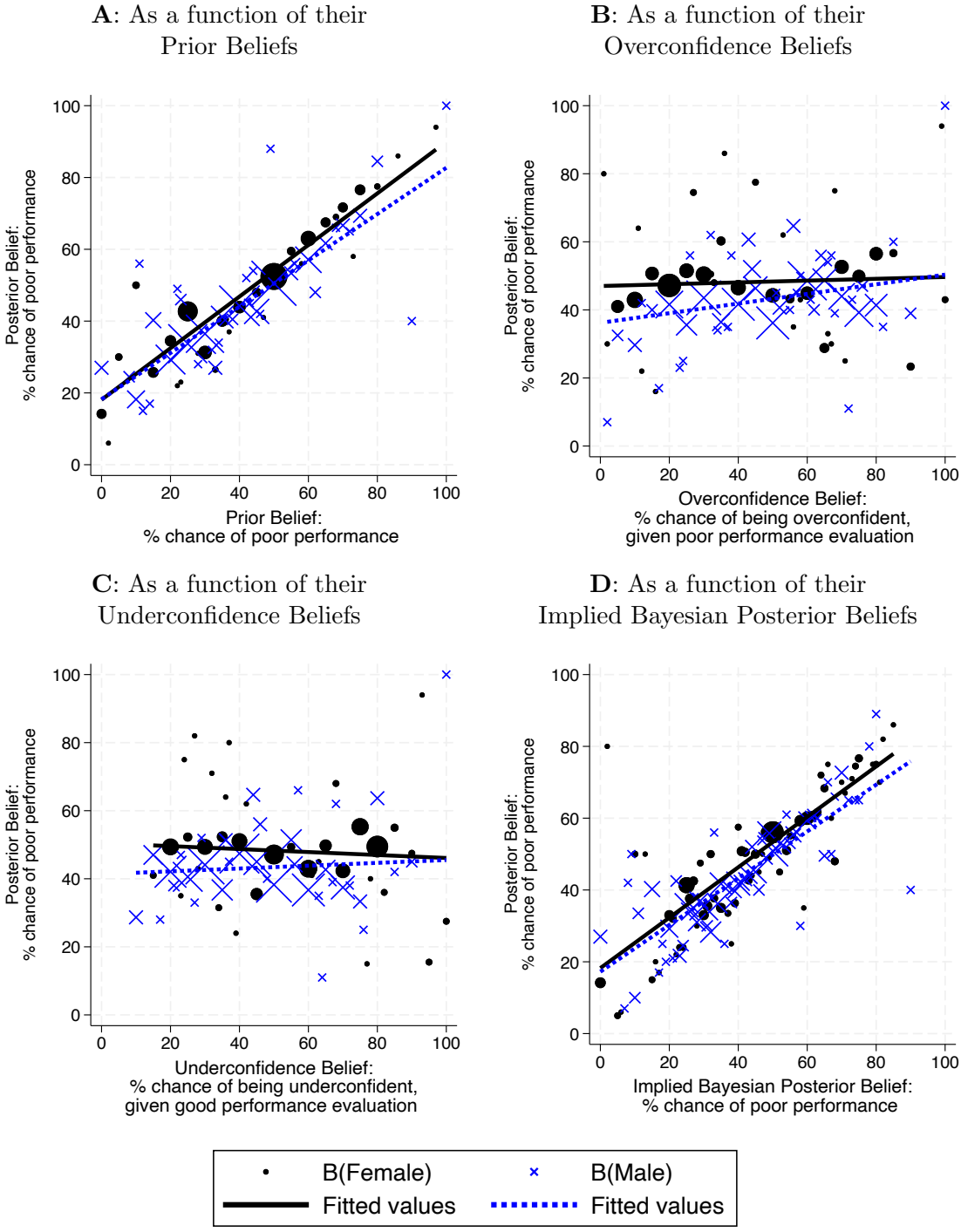
Graphs show a scatter plot (dots weighted by sample size) of evaluators' posterior beliefs as a function of their beliefs noted on the horizontal axis. Data are from the *Baseline* treatment of the *Evaluator Study*.

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



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



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

	DV: Evaluators' Posterior Beliefs							
	Low Education (1)	High Education (2)	Low Income (3)	High Income (4)	Younger (5)	Older (6)	Favor Democrats (7)	Favor Republicans (8)
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	-0.75	1.37	-0.47	0.95	0.53	0.29	1.09	-0.79
*Attention	(3.65)	(3.41)	(3.96)	(3.20)	(3.28)	(3.80)	(2.99)	(4.53)
Δ	-5.80	-5.84	-5.24	-5.93	-2.71	-9.37	-4.96	-6.97
*Calculation	(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.

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.

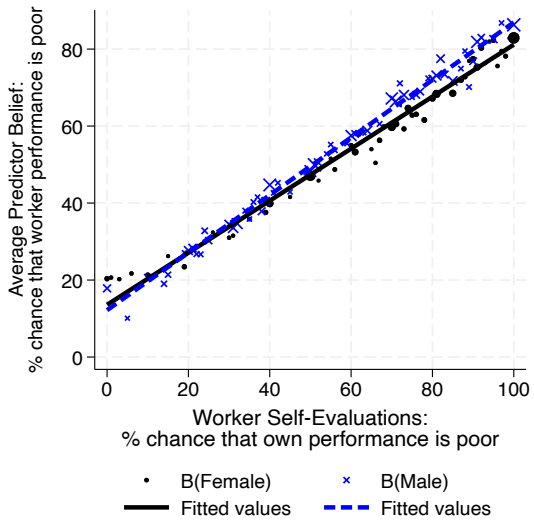
Appendix Figure D.1 and Appendix Table D.7 show how evaluators' beliefs respond to individual worker's self-evaluations, as discussed in Section 6.5.

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

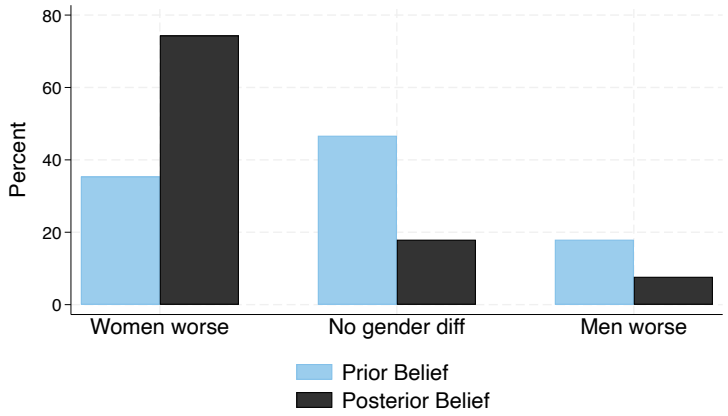
SEs are robust and shown in parentheses. Results follow the structure of Table 2. Data are from the 410 participants in the *Joint Evaluations* treatment of the *Evaluator (Extended) Study*. Sample size differs slightly in column (4) as some evaluators' beliefs imply a Bayesian posterior that is undefined.

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":

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

To see this:

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

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

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

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

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

E.2 Bayes of Good Performance

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

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

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

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

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

To see this:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E.4 Bayesian Posterior Beliefs As A Function of Confidence

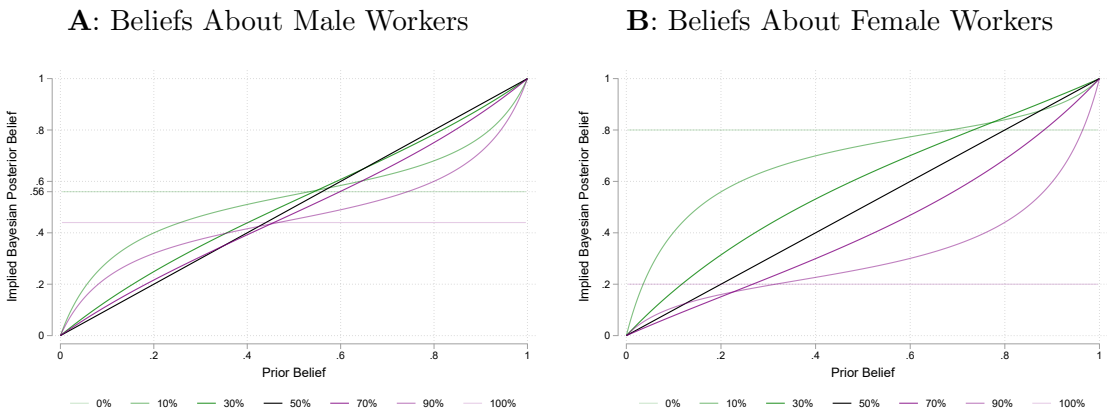
Appendix Figure E.1 shows how the levels of overconfidence and underconfidence beliefs affect the implied Bayesian posterior belief. These graphs plot the equation from Appendix Section E.1 as a function of the prior belief, overconfidence belief, and underconfidence belief. Panel A shows the implied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

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

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

the implied Bayesian posterior belief should be equal to the prior for all prior beliefs. As over- and underconfidence beliefs increase away from 0% toward 50%, the implied Bayesian posterior beliefs move toward the perfectly uninformative posterior. As 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



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.

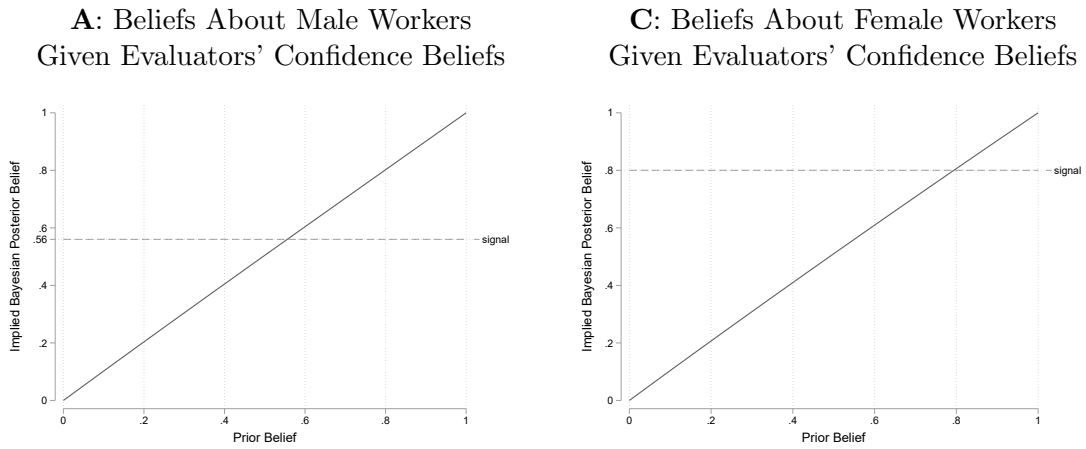
This is particularly striking in the context of our experiment. It implies that evaluators believe the signal to be as good as noise and therefore should discard it, but instead they incorporate it too much into their posterior beliefs. As a result, the gender gap in believed performance emerges from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



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*

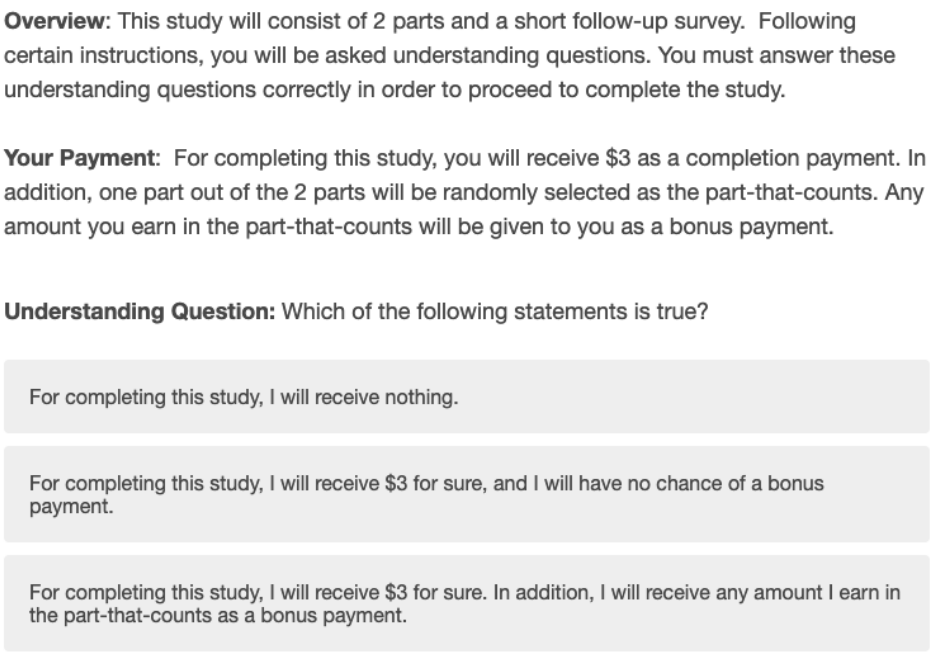


Figure F.1.2: Part 1 Instructions, *Baseline* Treatment of the *Worker Study*

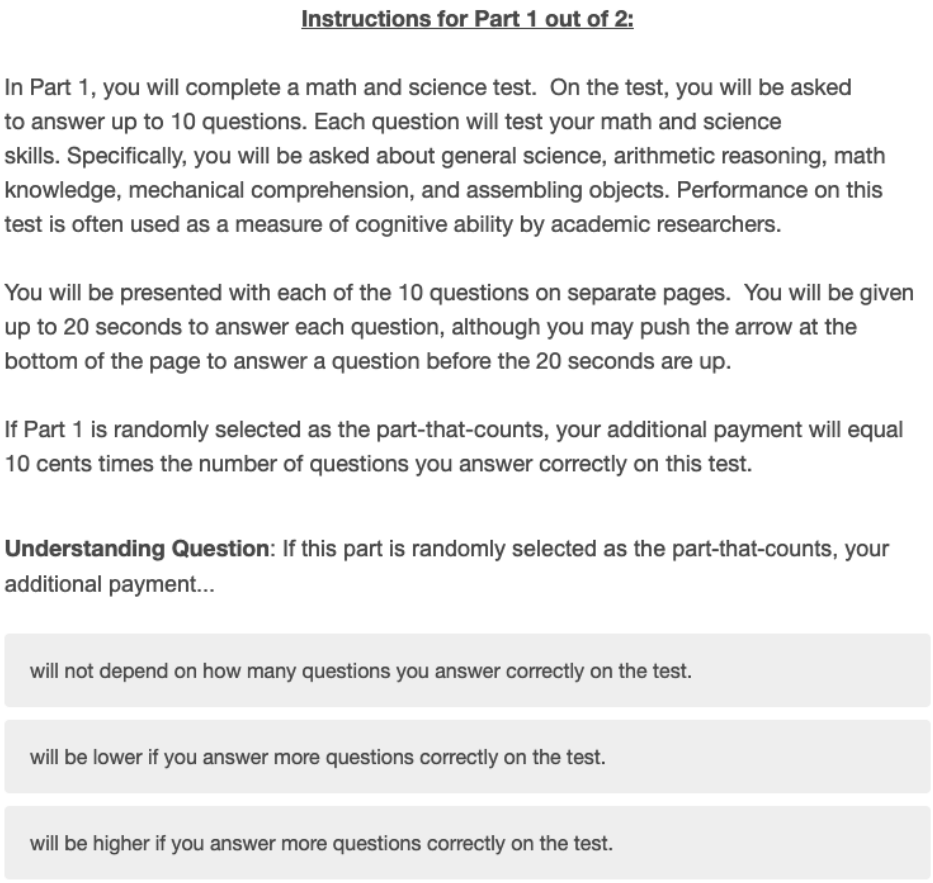


Figure F.1.3: Classifier Questions, *Baseline* Treatment of the *Worker Study*

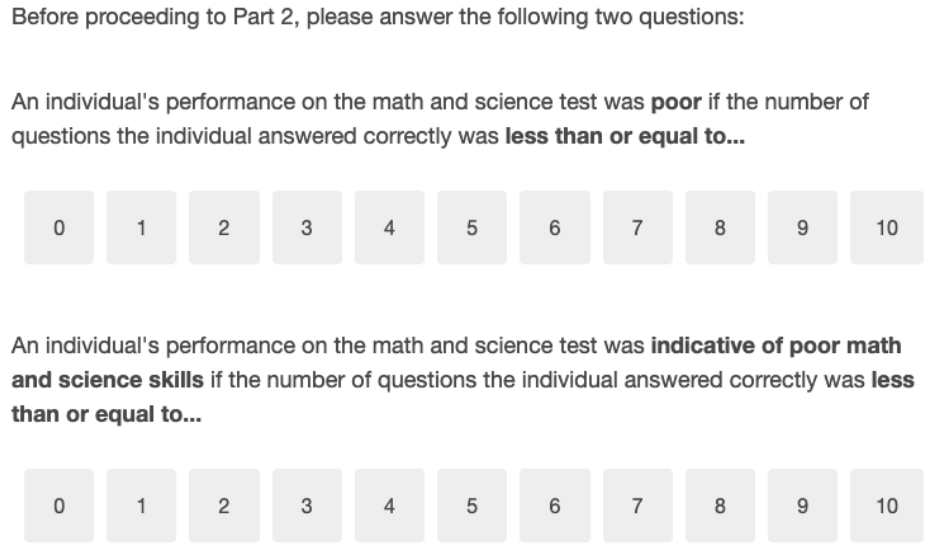


Figure F.1.4: Part 2 Instructions, *Baseline* Treatment of the *Worker Study*

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*

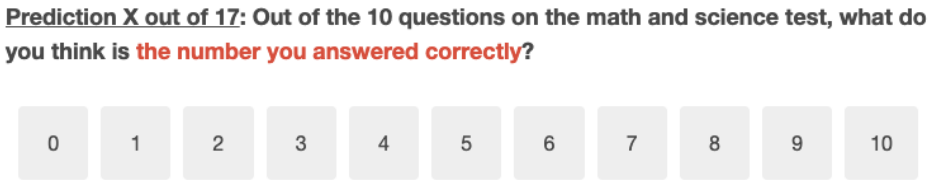


Figure F.1.6: Self-Evaluation Questions 1B and 1C, *Baseline* Treatment of the *Worker Study*

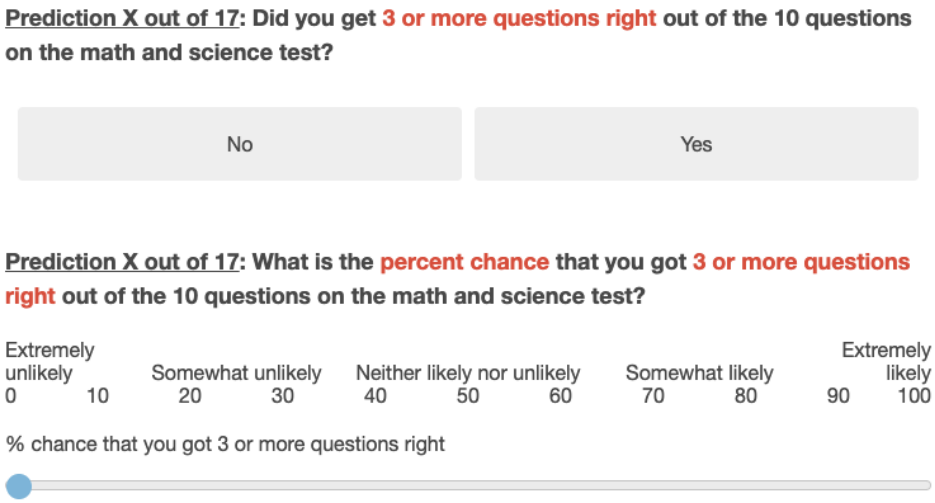


Figure F.1.7: Self-Evaluation Questions 2B and 2C, *Baseline* Treatment of the *Worker Study*

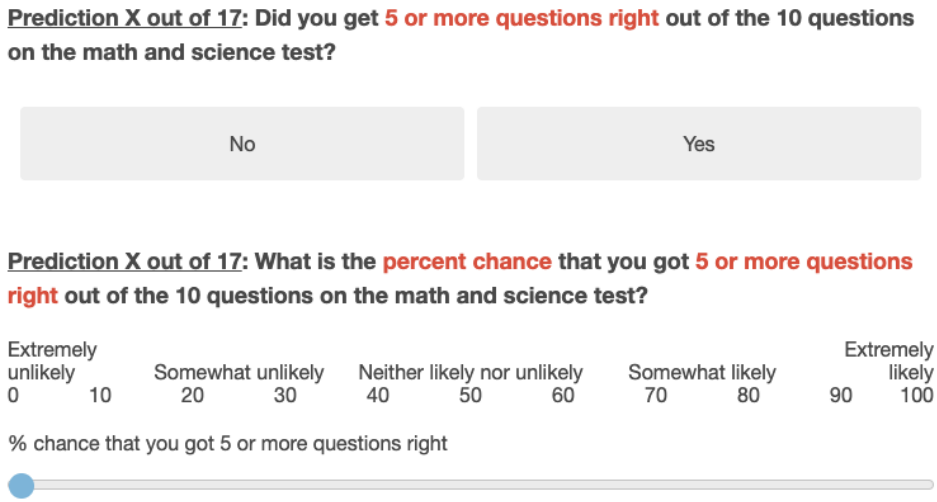


Figure F.1.8: Self-Evaluation Questions 3B and 3C, *Baseline* Treatment of the *Worker Study*

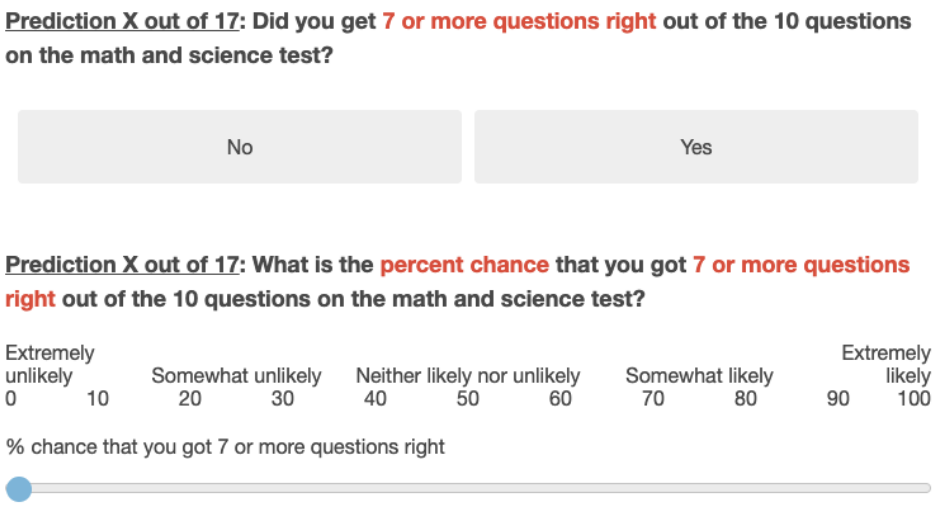


Figure F.1.9: Self-Evaluation Questions 4B and 4C, *Baseline* Treatment of the *Worker Study*

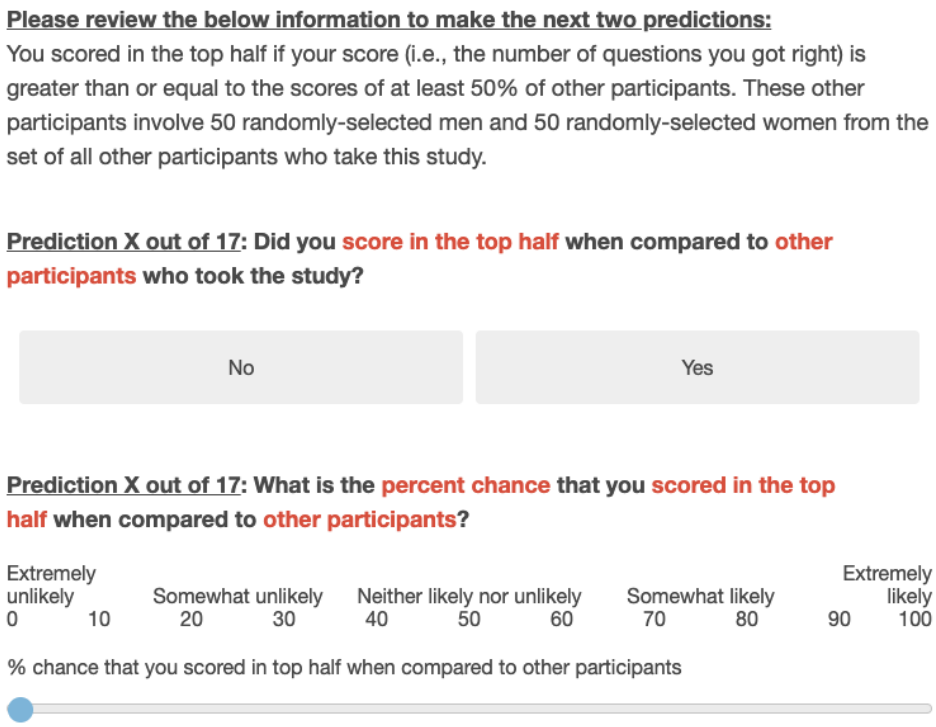


Figure F.1.10: Self-Evaluation Questions 5B and 5C, *Baseline* Treatment of the *Worker Study*

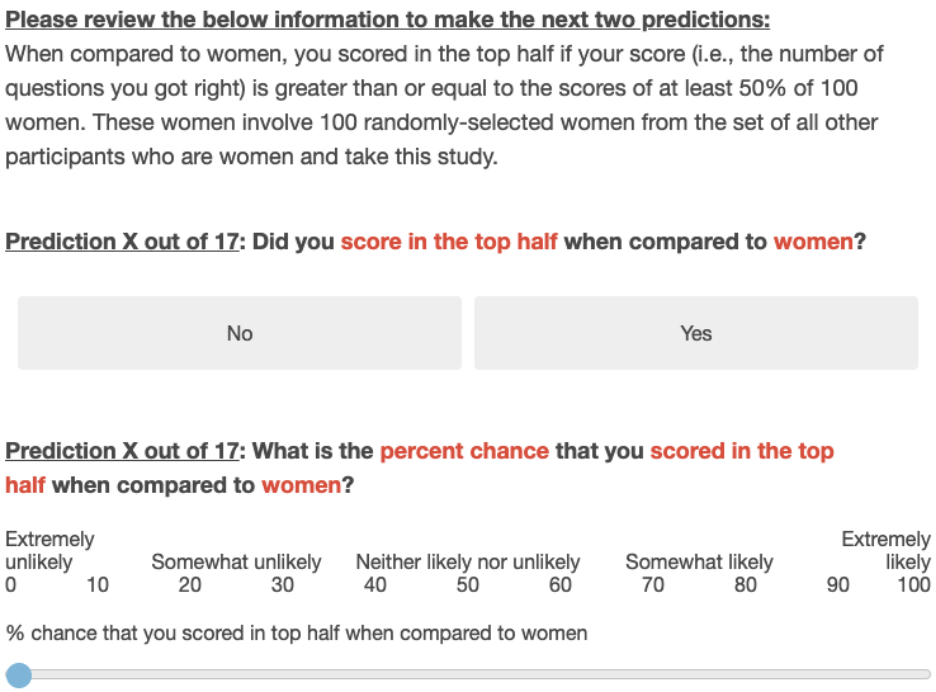


Figure F.1.11: Self-Evaluation Questions 6B and 6C, *Baseline* Treatment of the *Worker Study*

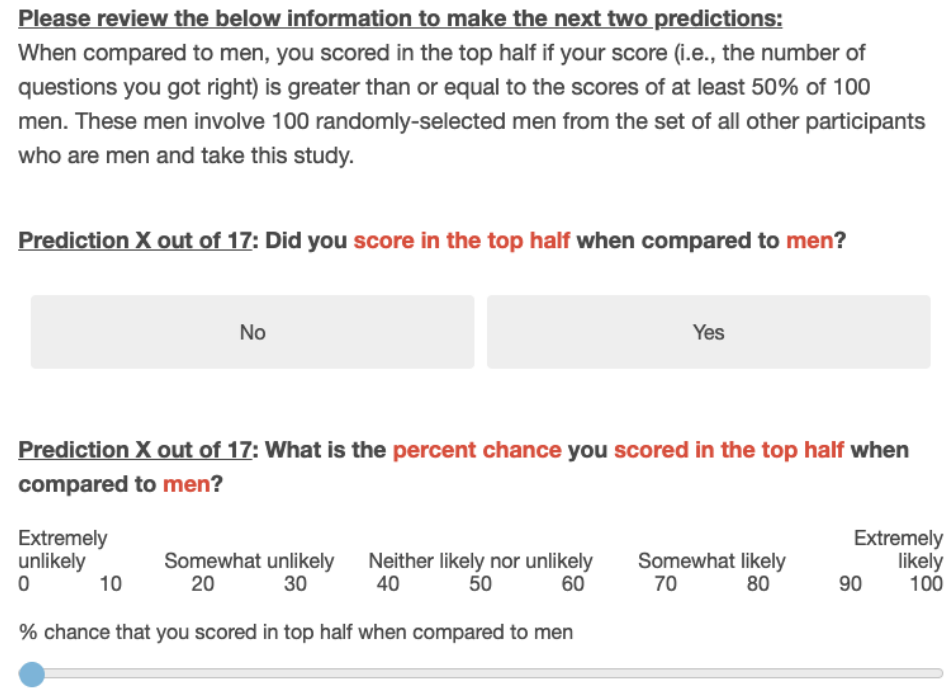


Figure F.1.12: Self-Evaluation Questions 7B and 7C, *Baseline* Treatment of the *Worker Study*

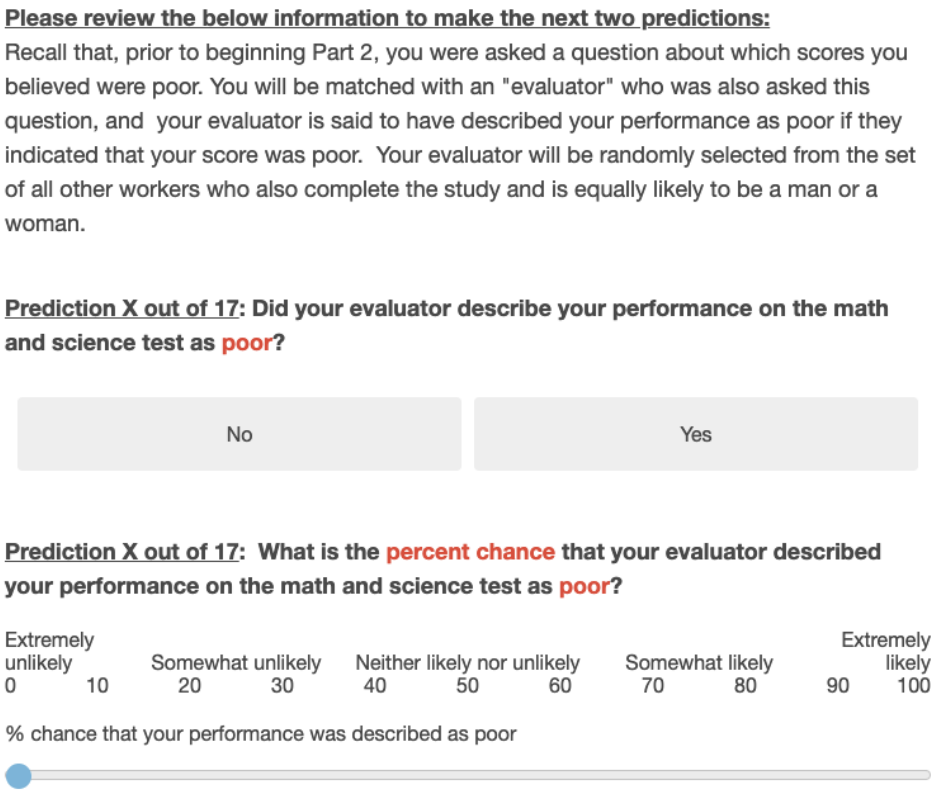
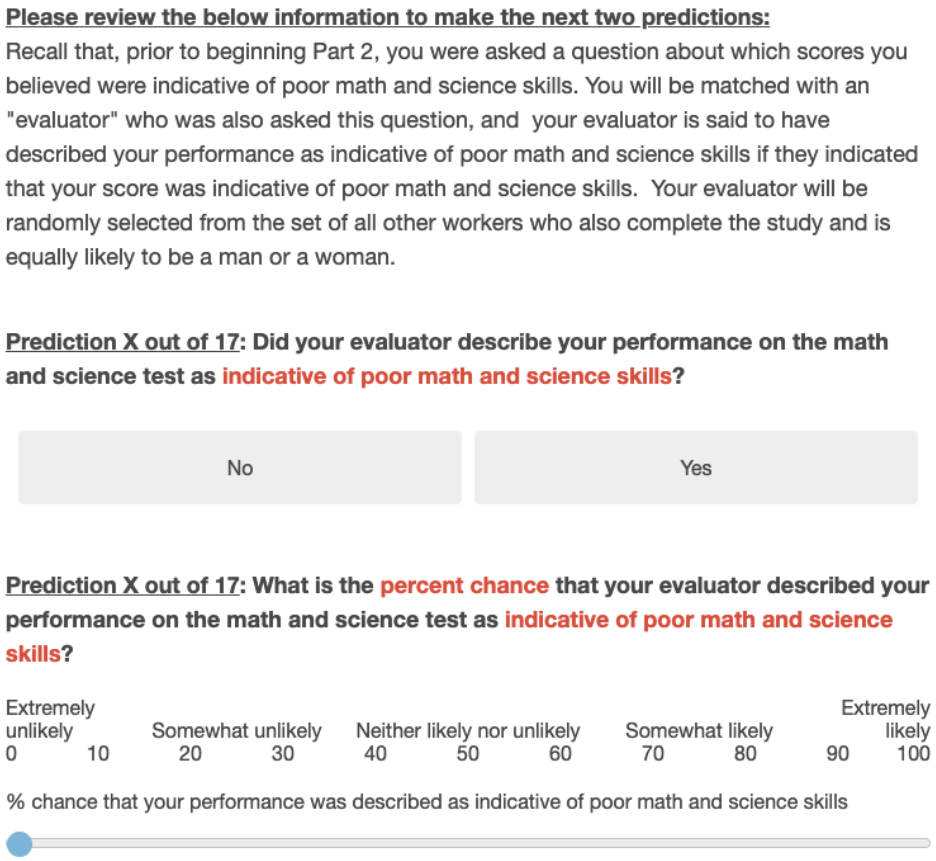


Figure F.1.13: Self-Evaluation Questions 8B and 8C, *Baseline* Treatment of the *Worker Study*



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.

Figure F.2.1: Part 2 Instructions, *Strategic Incentives* 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 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

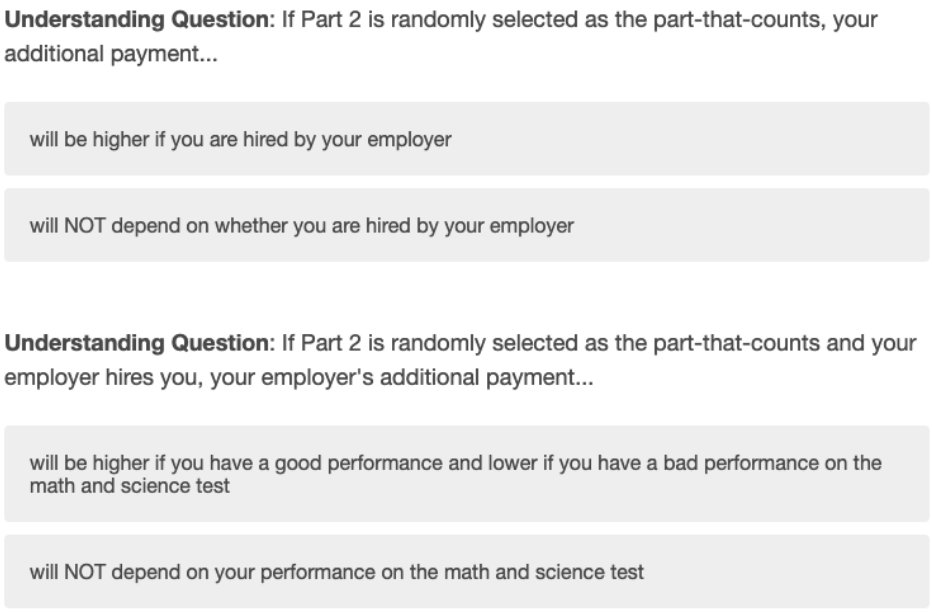
will NOT depend on how accurate your predictions are

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*



G Full Instructions for the *Evaluator Study*

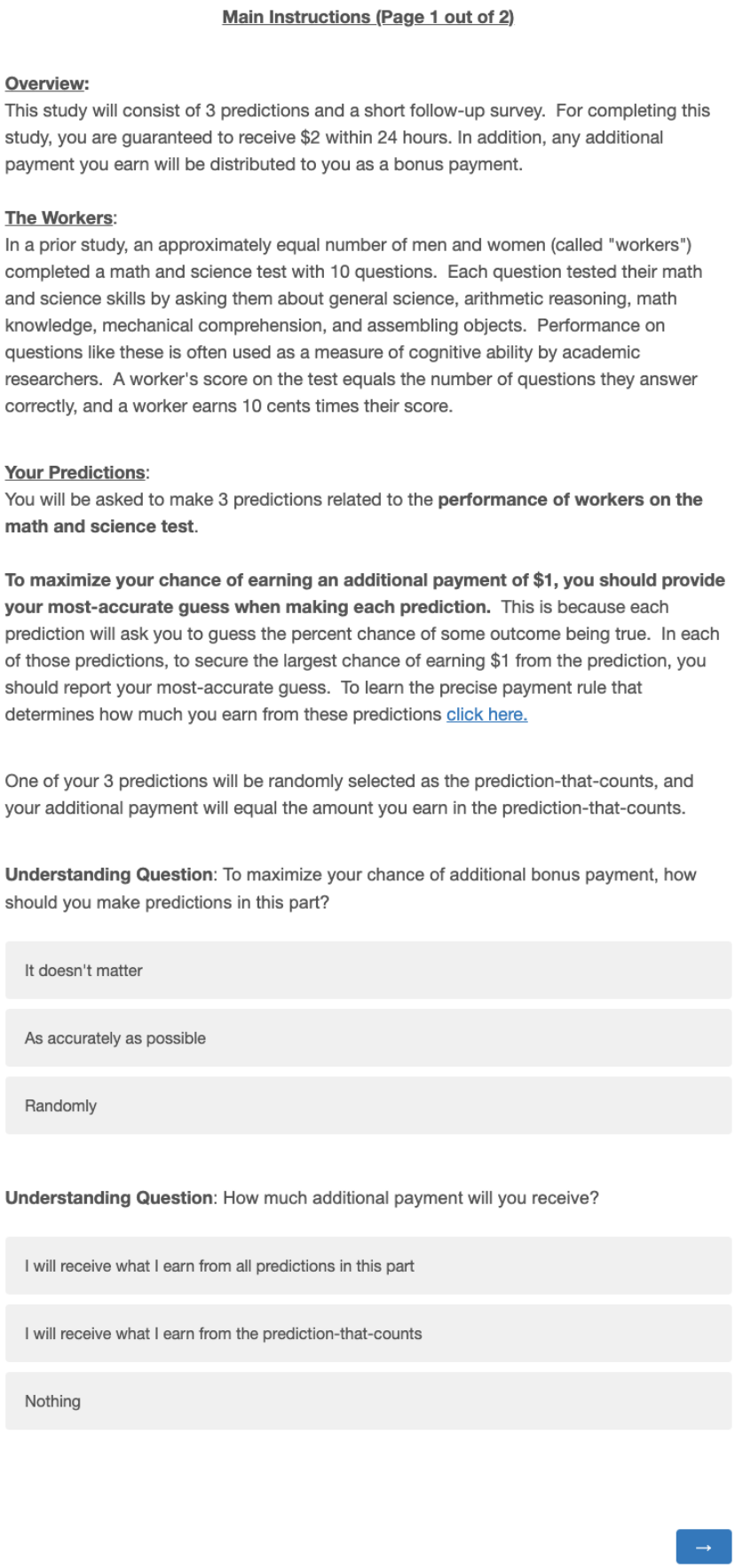
All participants in this study are randomized to be asked about male or female workers (or "group-1" or "group-2" workers in some conditions) and to be in one of six treatments described below.

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

→

Female Workers

Figure G.1.2: Instructions about Female Workers, *Baseline* 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 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 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.

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 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*

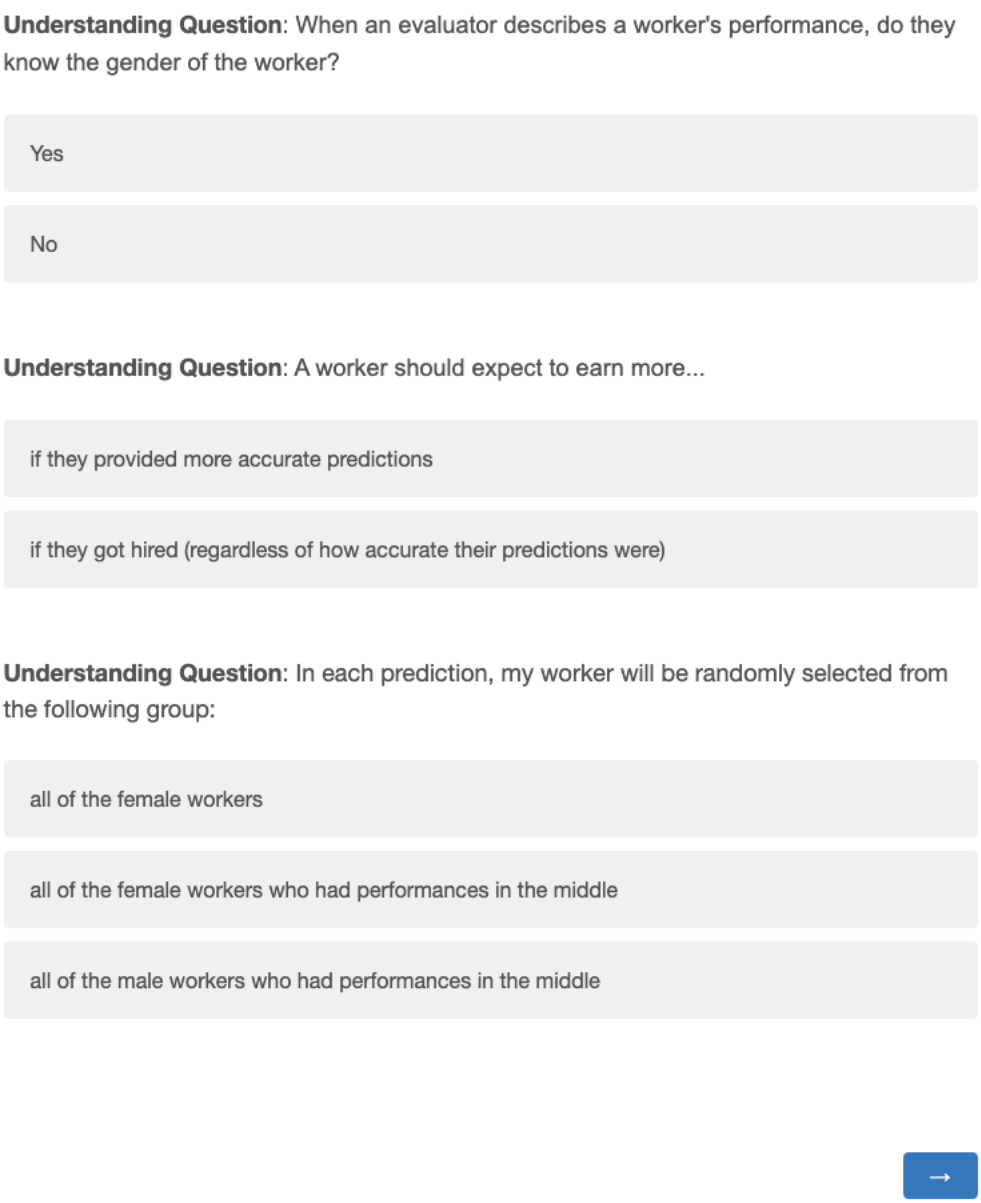


Figure G.1.4: Prior Belief about Female Workers, *Baseline* Treatment of the *Evaluator Study*

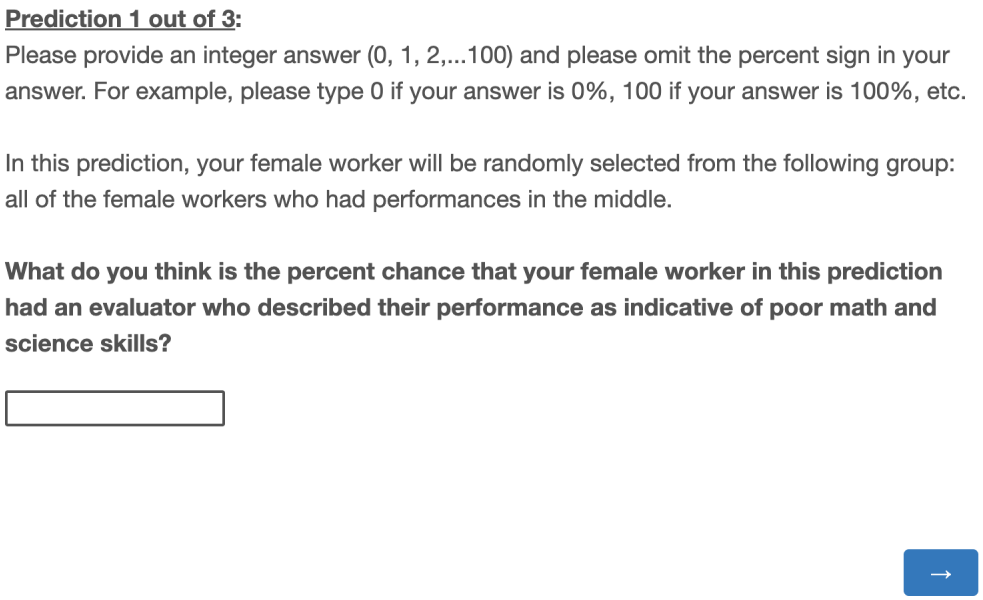


Figure G.1.5: Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator Study*

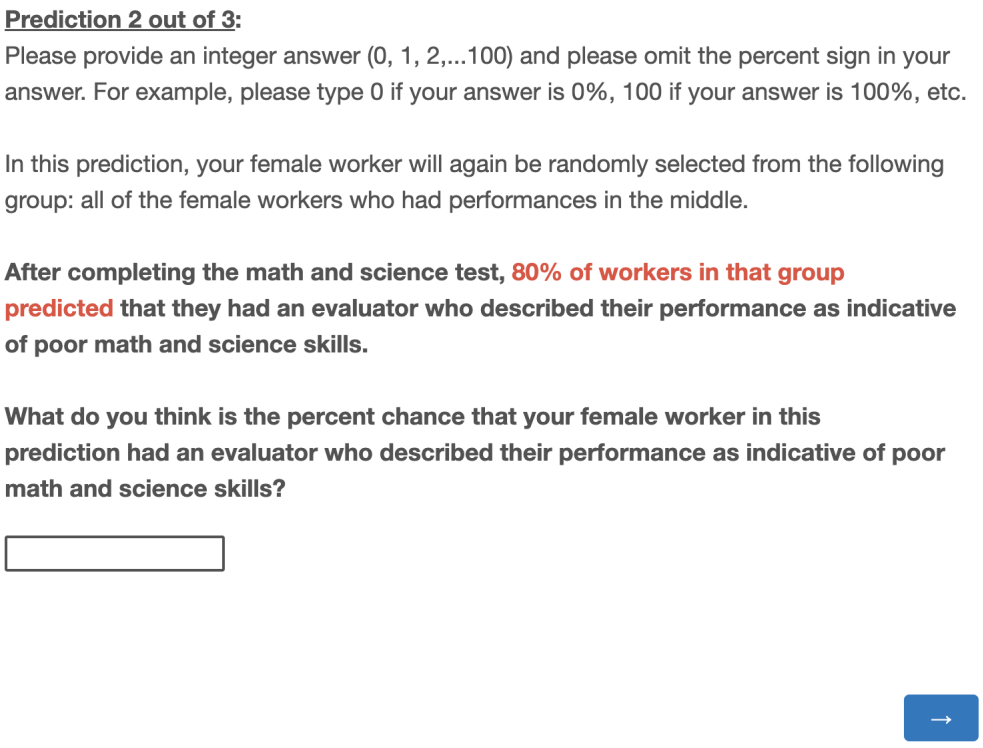


Figure G.1.6: Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator Study*

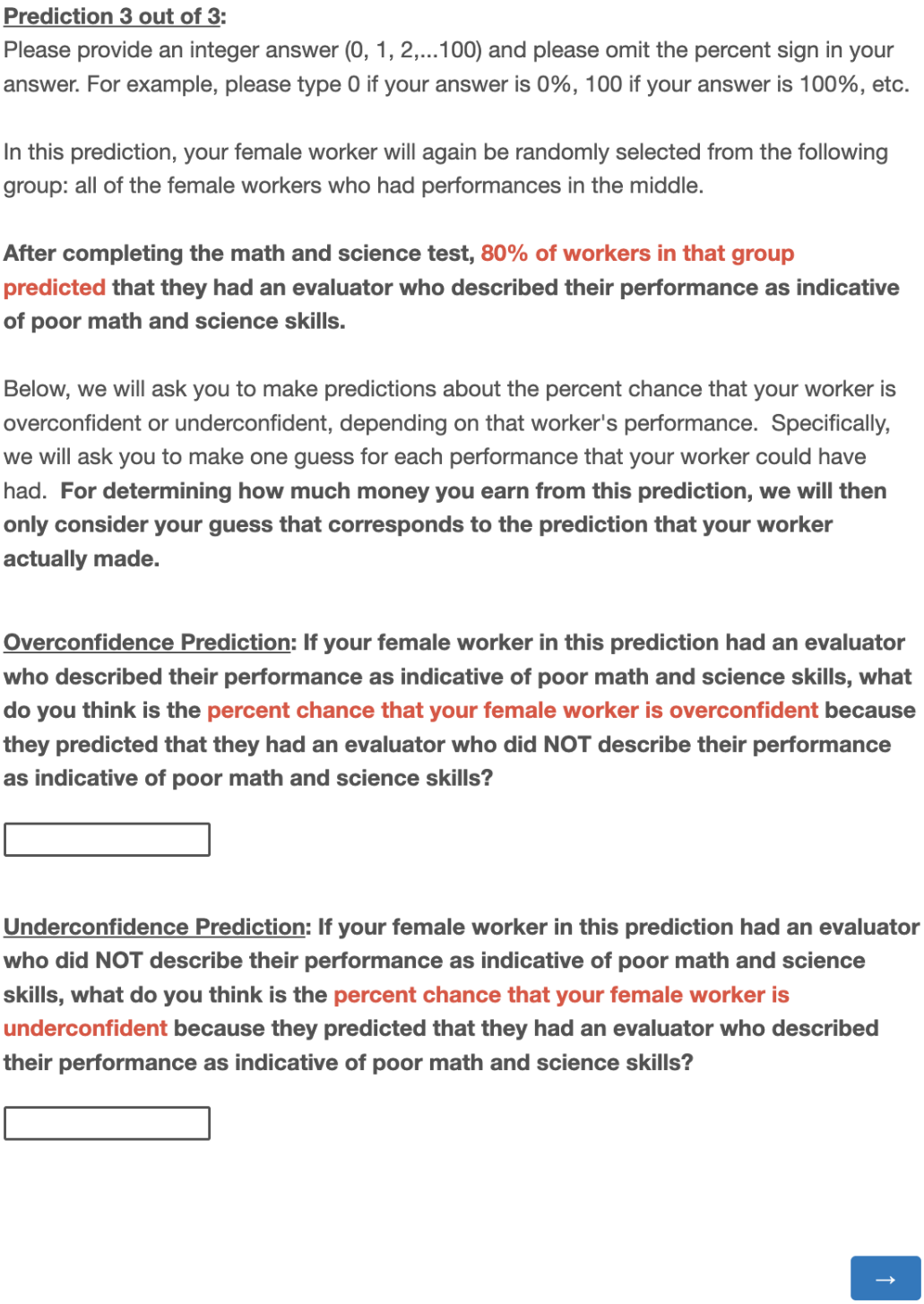


Figure G.1.7: Bonus Questions Instructions, *Baseline* Treatment of the *Evaluator Study*

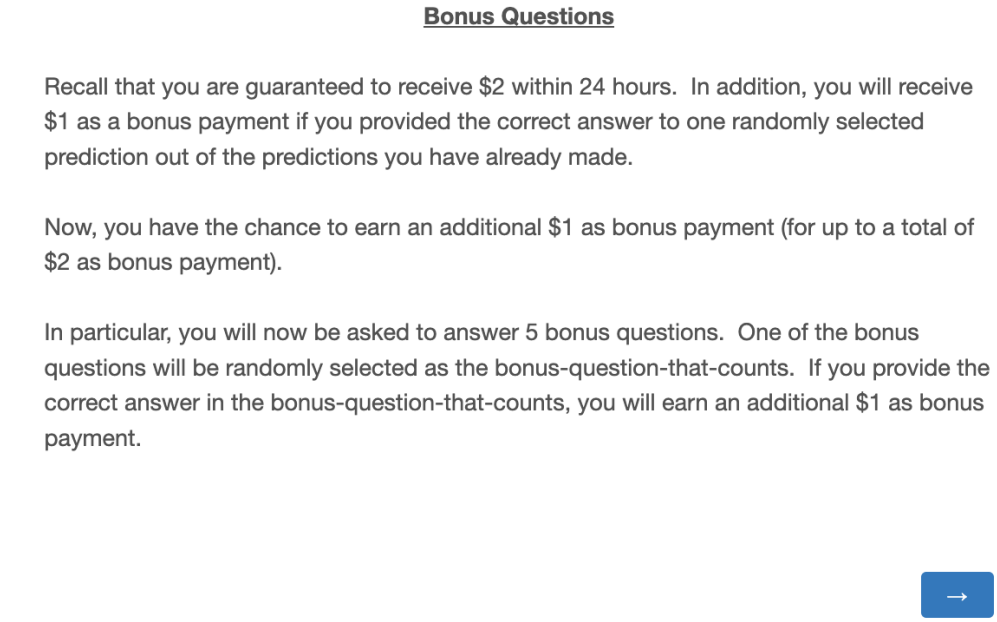


Figure G.1.8: Bonus Question 1: Bayesian Updating, *Baseline* Treatment of the *Evaluator Study*

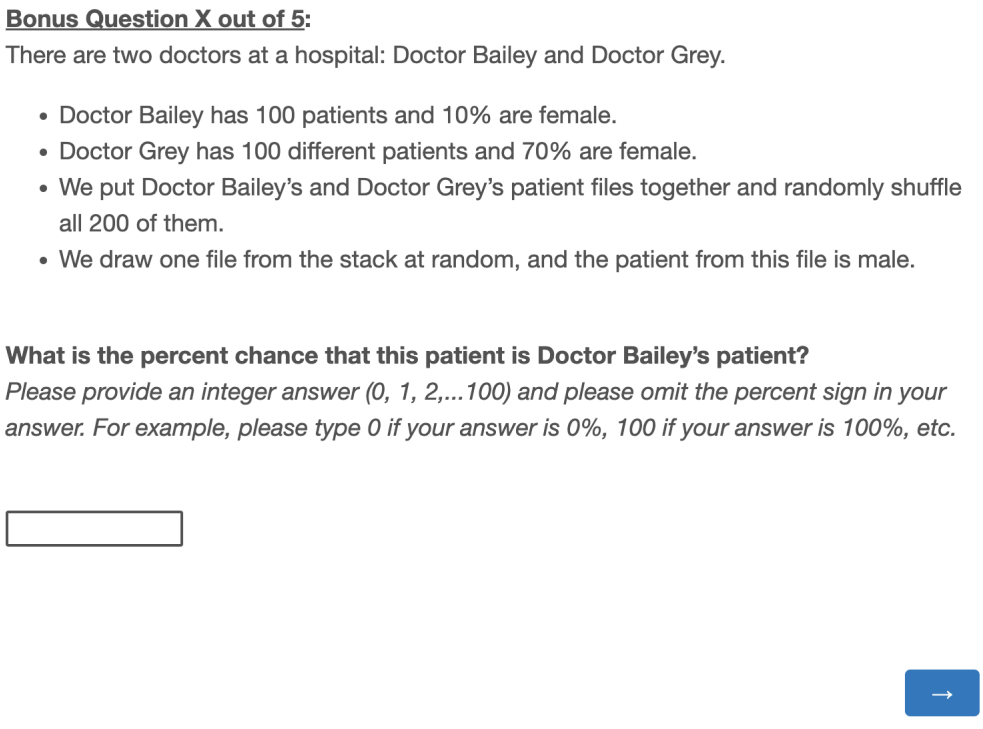


Figure G.1.9: Bonus Question 2: CRT1, *Baseline* Treatment of the *Evaluator Study*

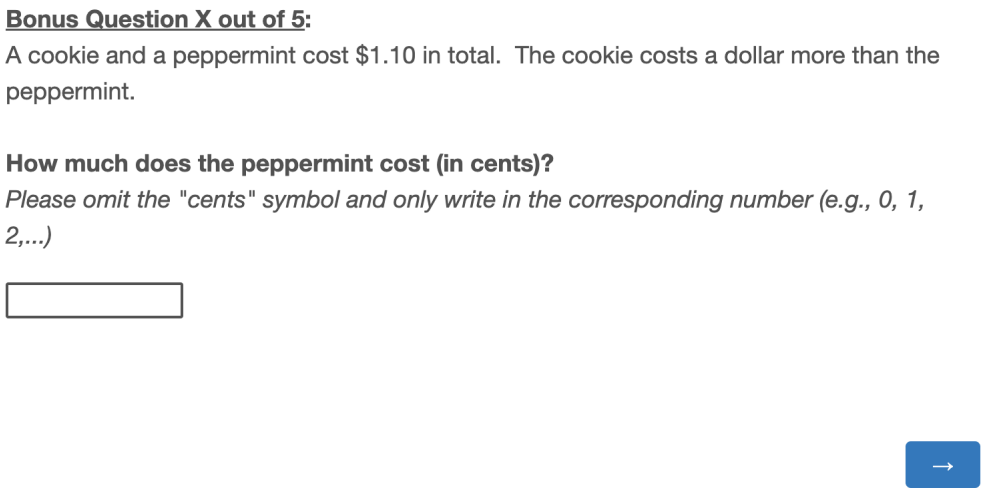


Figure G.1.10: Bonus Question 3: CRT2, *Baseline* Treatment of the *Evaluator Study*

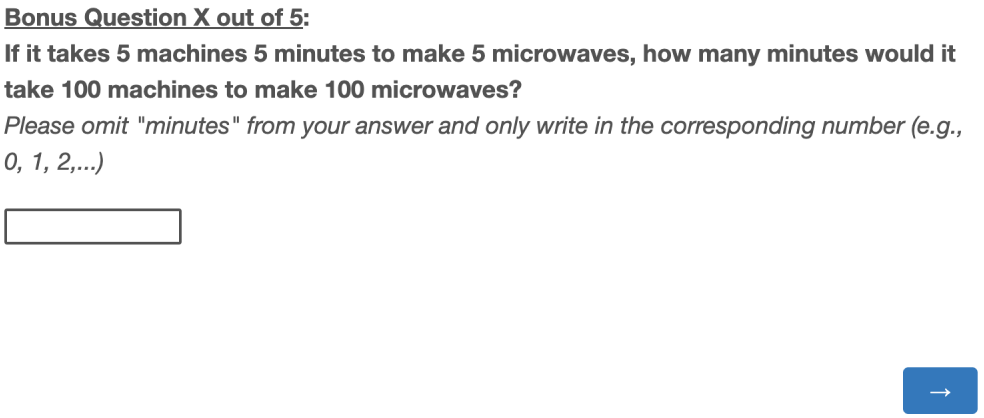


Figure G.1.11: Bonus Question 4: CRT3, *Baseline* Treatment of the *Evaluator Study*

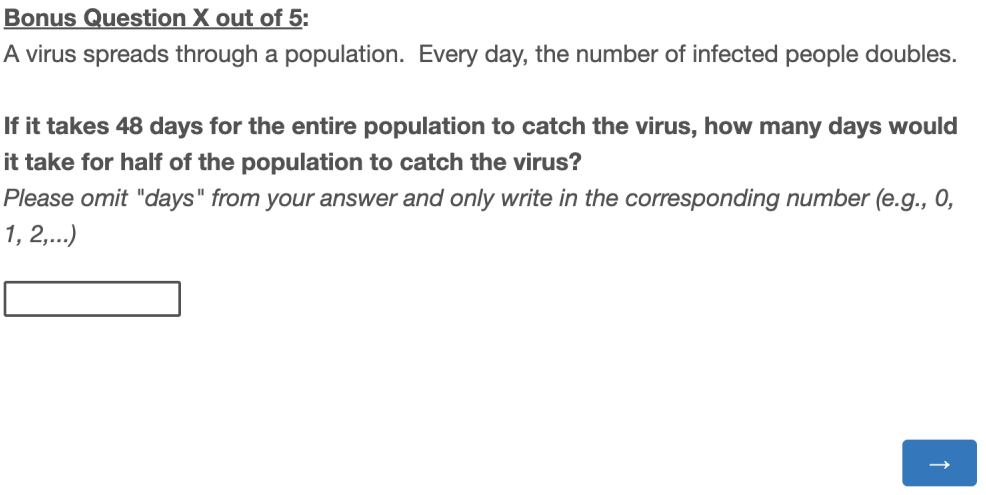


Figure G.1.12: Bonus Question 5: Base Rate Neglect, *Baseline* Treatment of the *Evaluator Study*

Bonus Question X out of 5:

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*

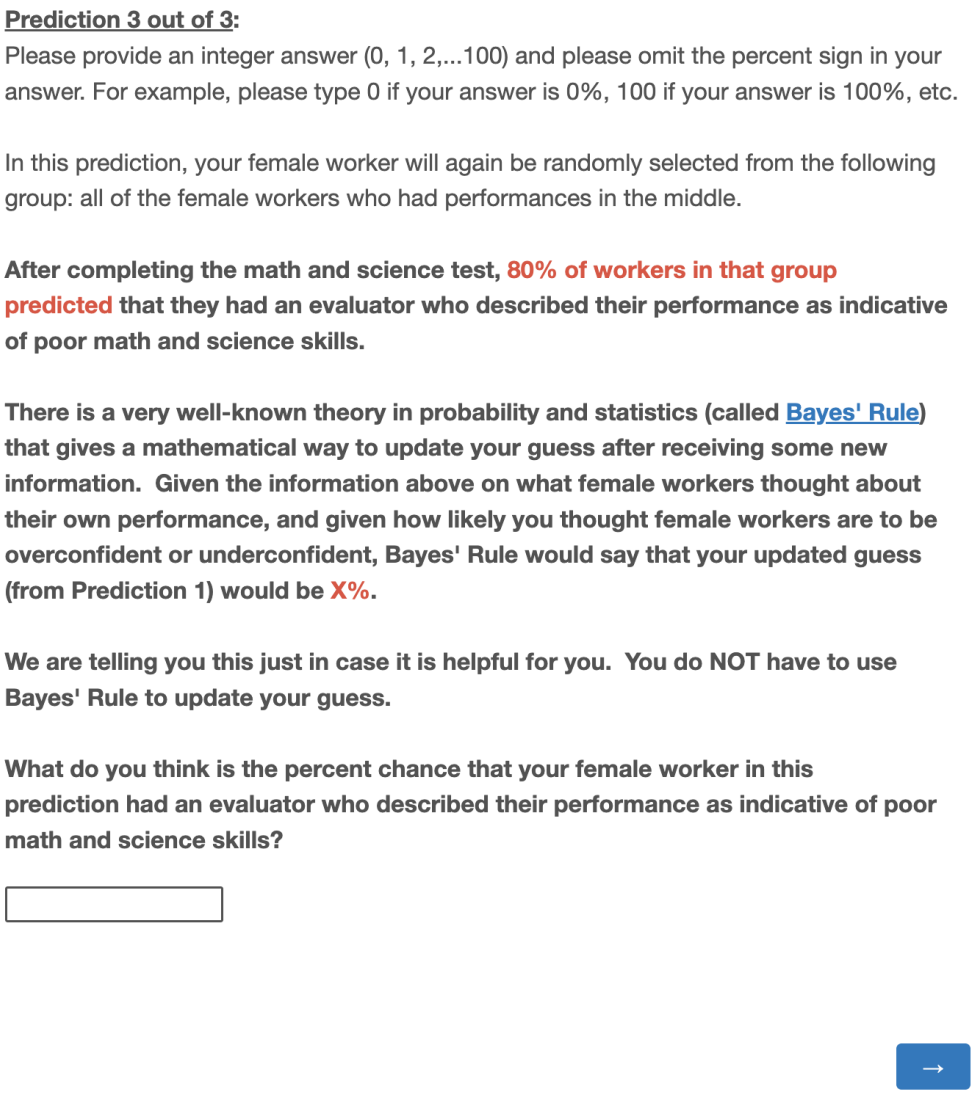
The *Calculation* treatment of the *Evaluator Study* differs from the *Attention* treatment of the *Evaluator Study* (Section G.2) only in the decision screen that elicits their posterior beliefs, highlighted via the new Figure G.3.1 shown below.

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%**.

Female Workers

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



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%**.

Group-2 (Female) Workers

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*

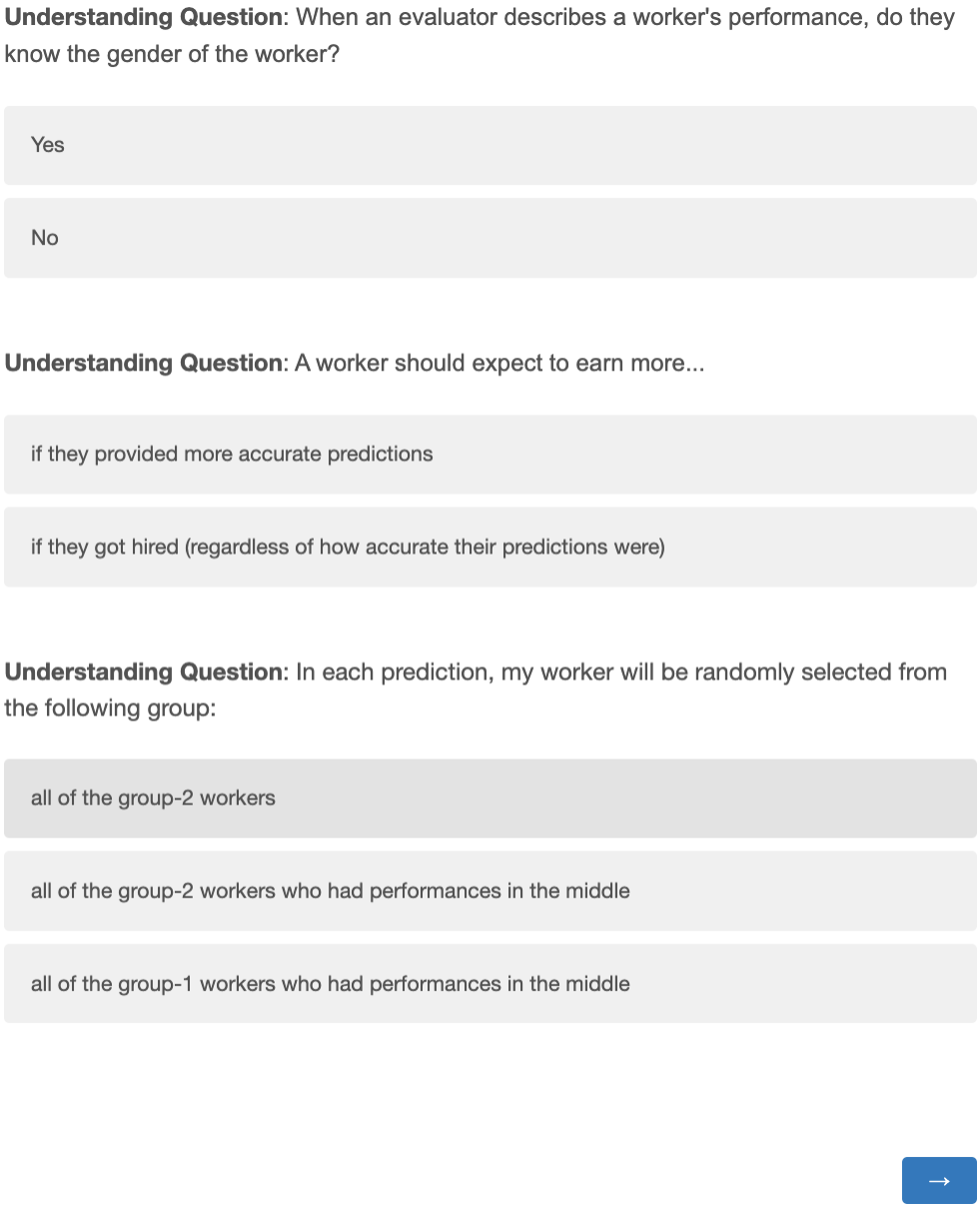


Figure G.4.3: Prior Belief about Group-2 Workers, *Baseline, Unknown* Treatment of the *Evaluator Study*

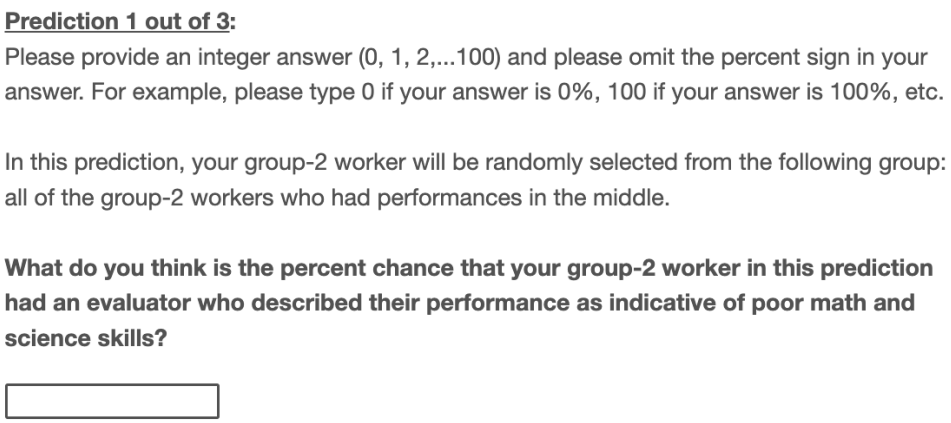


Figure G.4.4: Posterior Belief about Group-2 Workers, *Baseline, Unknown* Treatment of the *Evaluator Study*

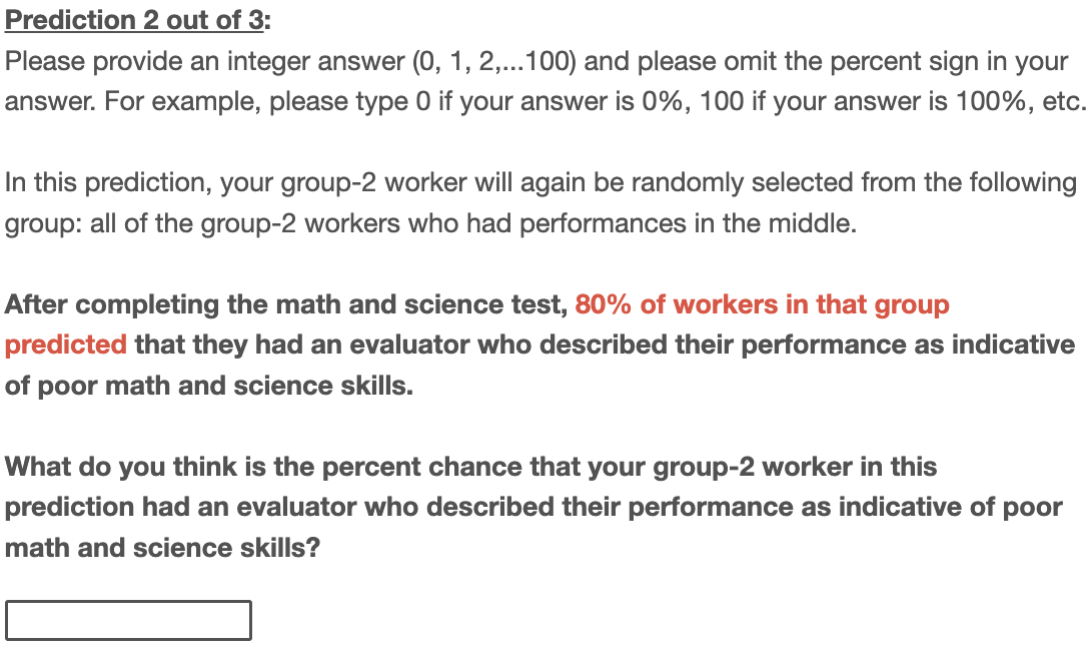


Figure G.4.5: Over/Underconfidence Beliefs about Group-2 Workers, *Baseline*, *Unknown* Treatment of the *Evaluator Study*

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 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.**

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?

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.

Figure H.1.1: Study Overview, *Worker (Undergraduate Students) 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 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.

Figure H.1.2: Part 1 Instructions, *Worker (Undergraduate Students) 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 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

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

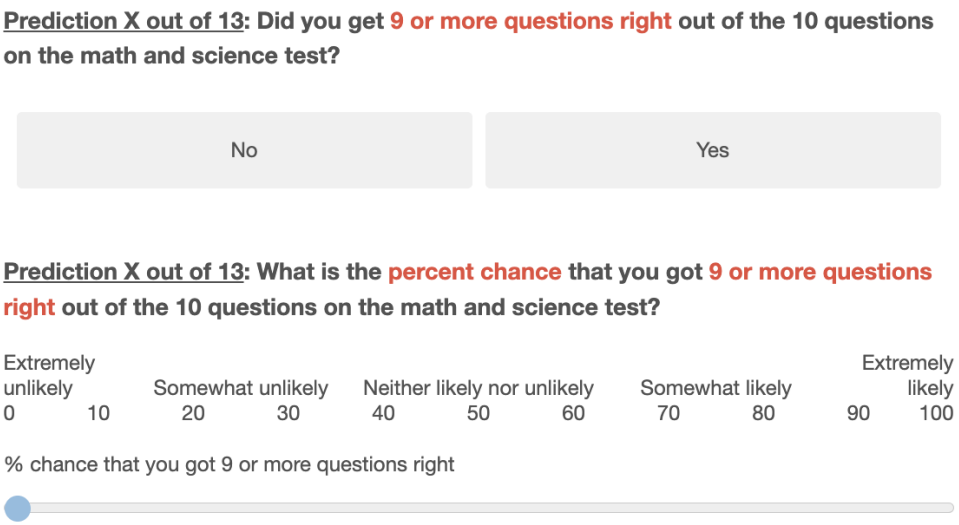


Figure H.1.5: Self-Evaluation Questions New-2B, 2C, 3B and C Instructions, *Worker (Undergraduate Students) Study*

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.

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

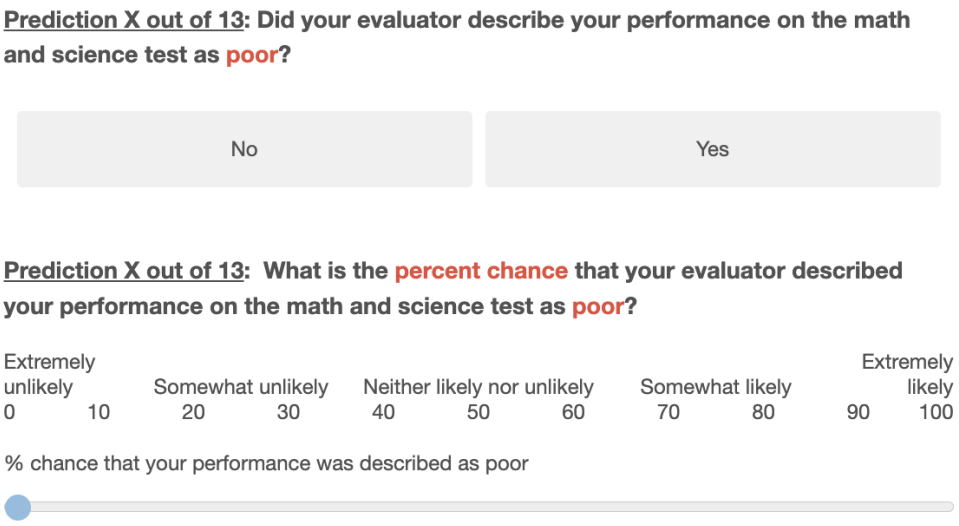
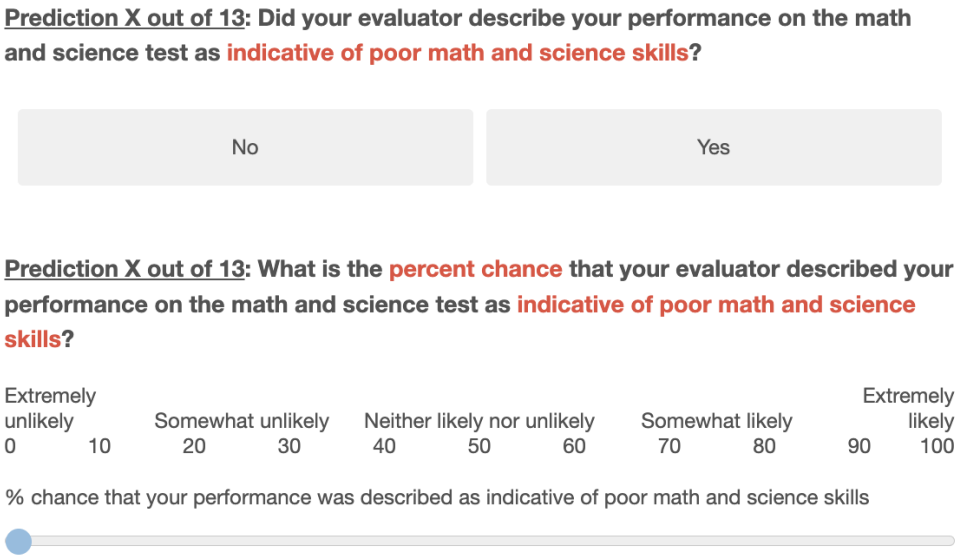


Figure H.1.7: Self-Evaluation Questions New-3B and 3C, *Worker (Undergraduate Students) Study*



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*

Main Instructions (Page 1 out of 2)

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

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*

Types of Predictions

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*

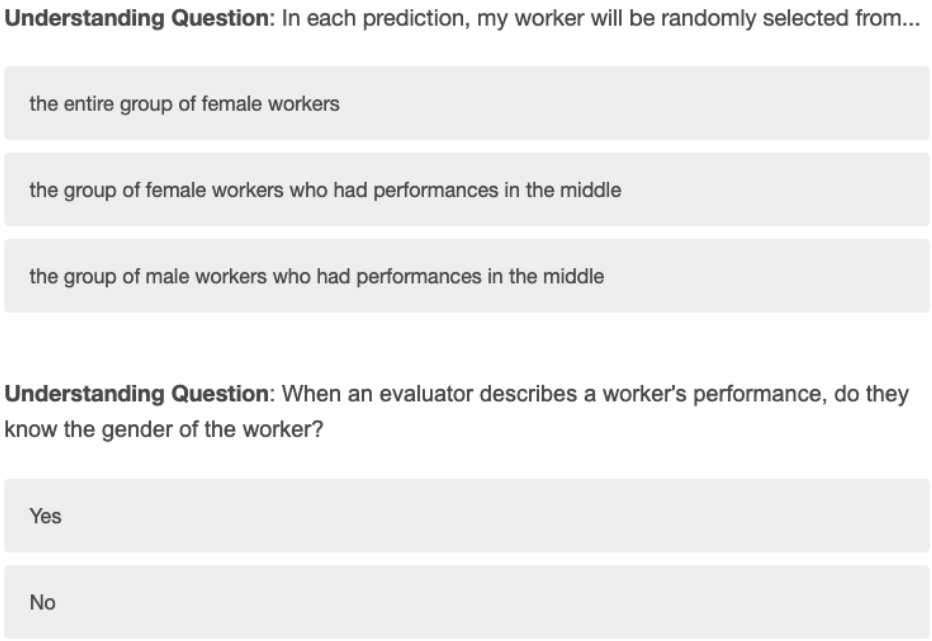


Figure I.1.5: Prior Belief Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

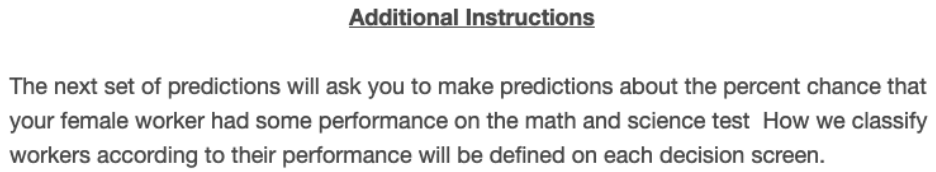


Figure I.1.6: Prior Belief (3+) about Female Workers, *Evaluator (Alternative Questions) Study*

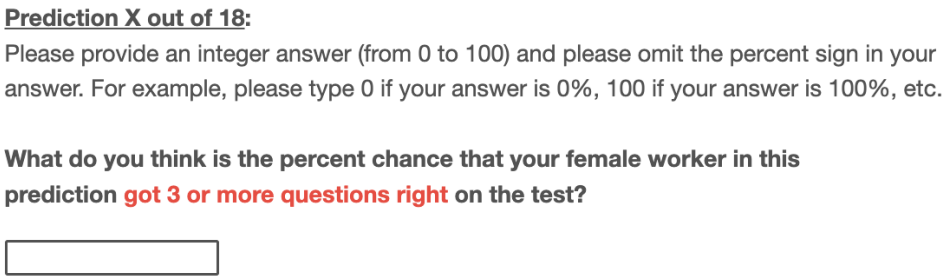


Figure I.1.7: Prior Belief (5+) about Female Workers, *Evaluator (Alternative Questions) Study*

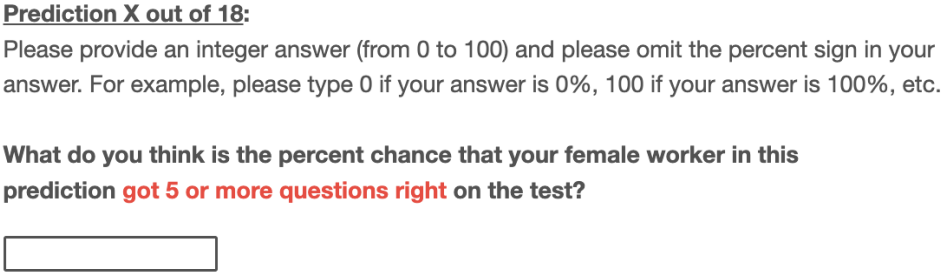


Figure I.1.8: Prior Belief (7+) about Female Workers, *Evaluator (Alternative Questions) Study*

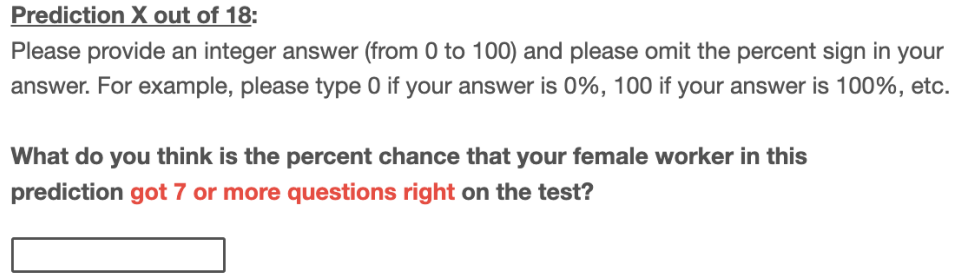


Figure I.1.9: Prior Belief (poor-2) about Female Workers, *Evaluator (Alternative Questions) Study*

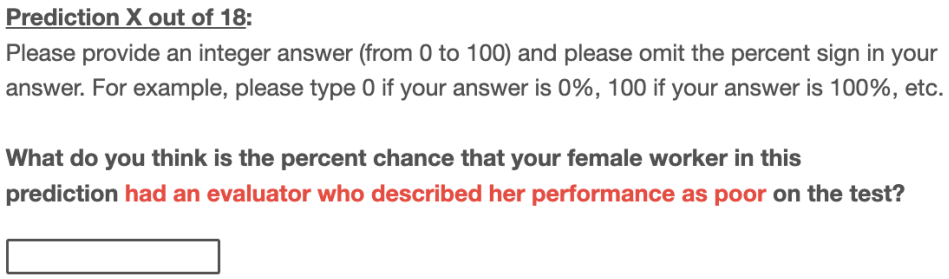


Figure I.1.10: Prior Belief (main self-evaluation) about Female Workers, *Evaluator (Alternative Questions) Study*

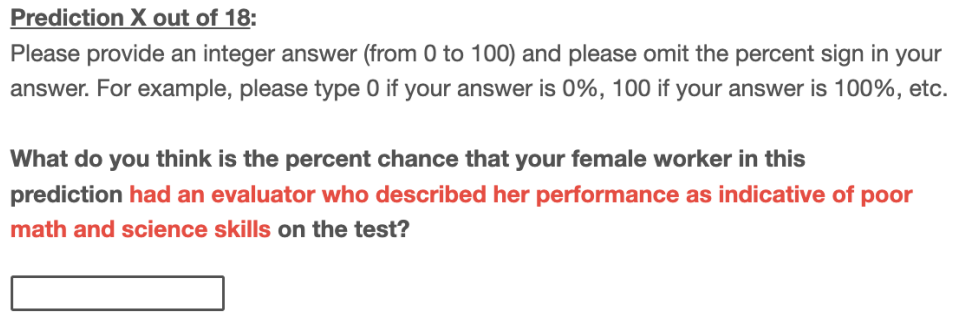


Figure I.1.11: Prior Belief (top-half) about Female Workers, *Evaluator (Alternative Questions) Study*

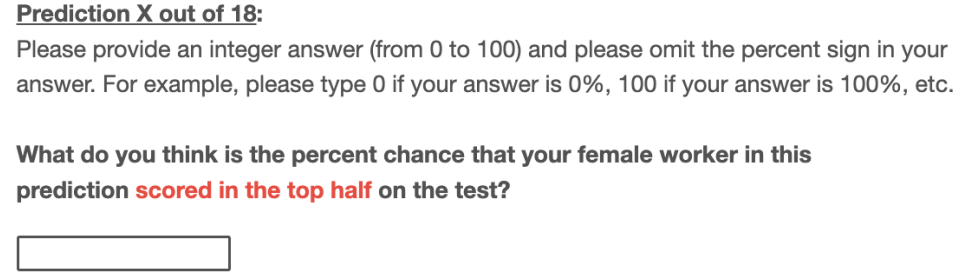


Figure I.1.12: Posterior Belief Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

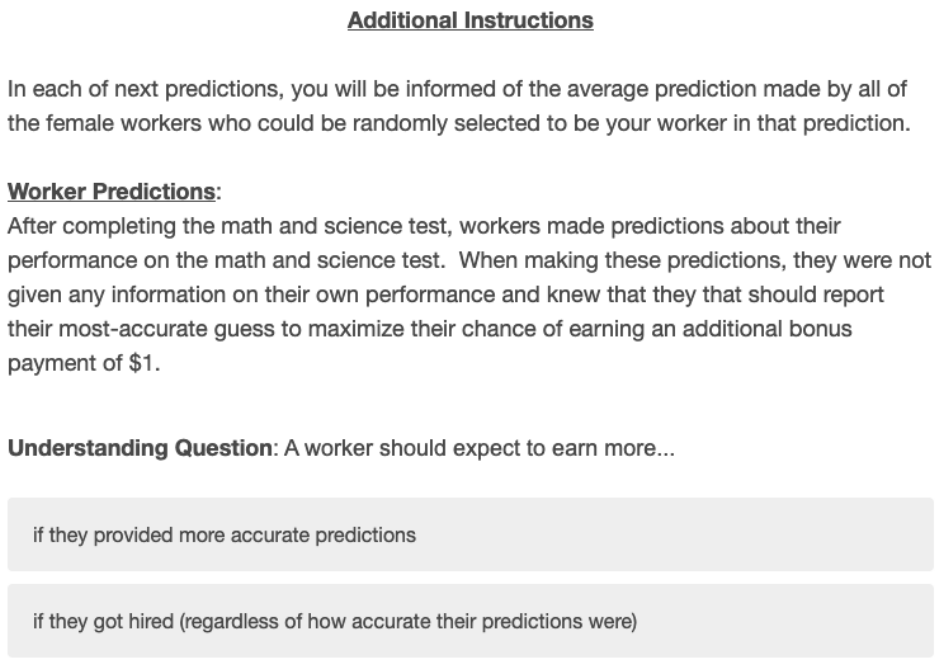


Figure I.1.13: Posterior Belief (3+) about Female Workers, *Evaluator (Alternative Questions) Study*

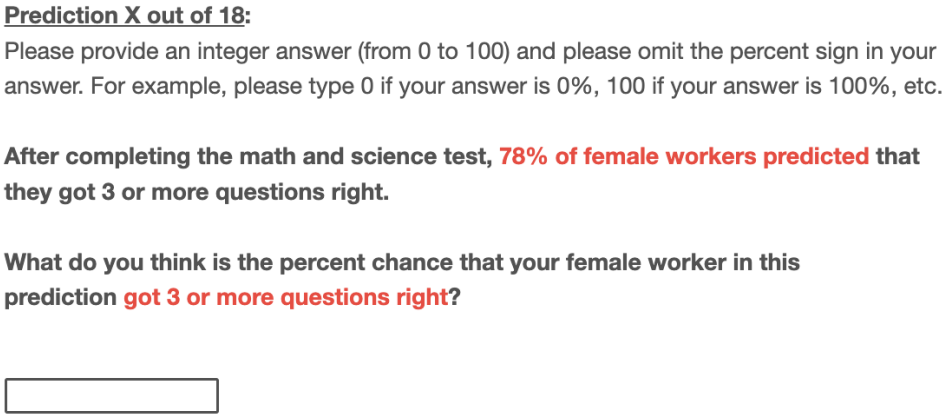


Figure I.1.14: Posterior Belief (5+) about Female Workers, *Evaluator (Alternative Questions) Study*

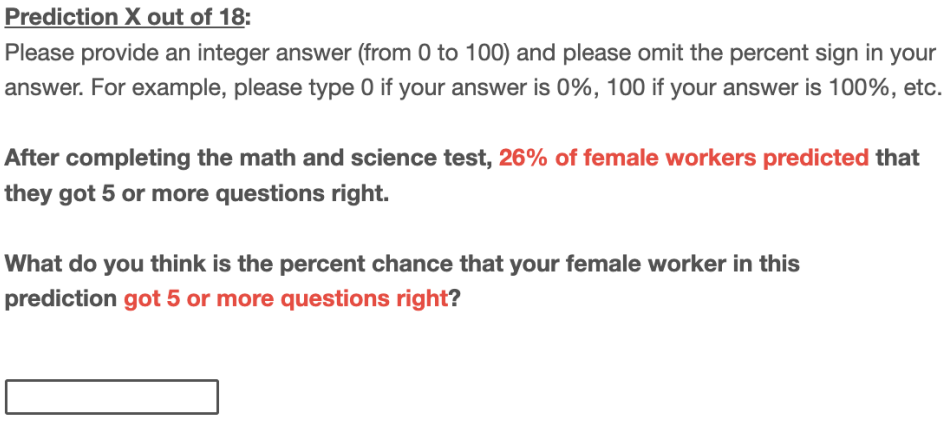


Figure I.1.15: Posterior Belief (7+) about Female Workers, *Evaluator (Alternative Questions) Study*

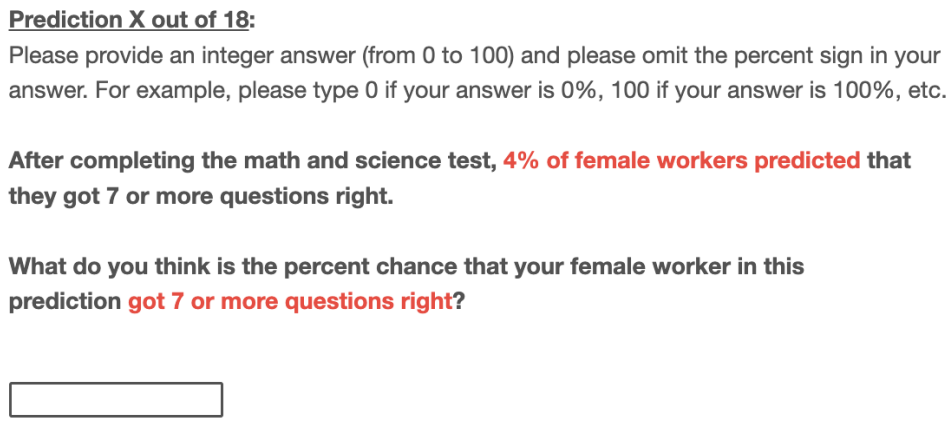


Figure I.1.16: Posterior Belief (poor-2) about Female Workers, *Evaluator (Alternative Questions) Study*

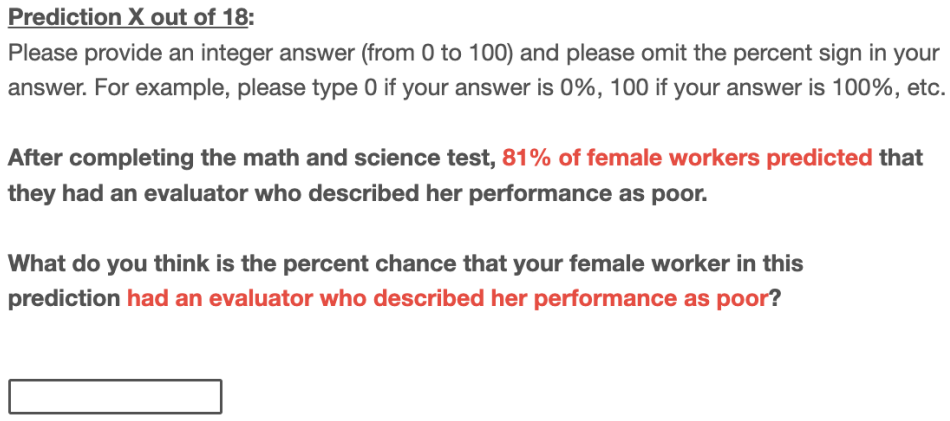


Figure I.1.17: Posterior Belief (main self-evaluation) about Female Workers, *Evaluator (Alternative Questions) Study*

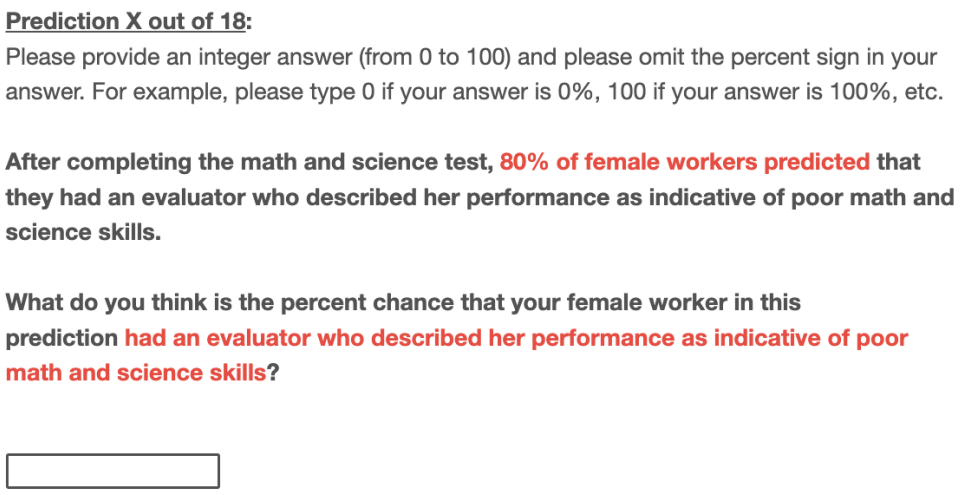


Figure I.1.18: Posterior Belief (top-half) about Female Workers, *Evaluator (Alternative Questions) Study*

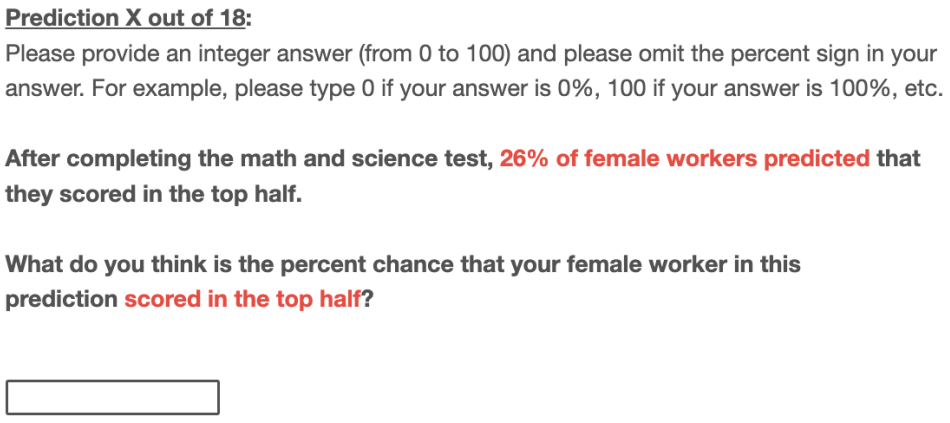


Figure I.1.19: Over/Underconfidence Beliefs Instructions about Female Workers, *Evaluator (Alternative Questions) Study*

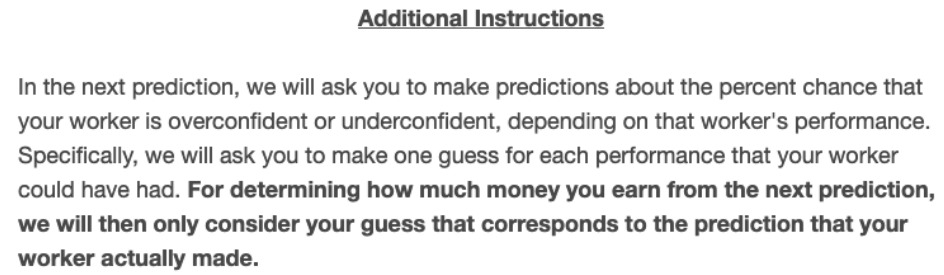


Figure I.1.20: Over/Underconfidence Beliefs (3+) about Female Workers, *Evaluator (Alternative Questions) Study*

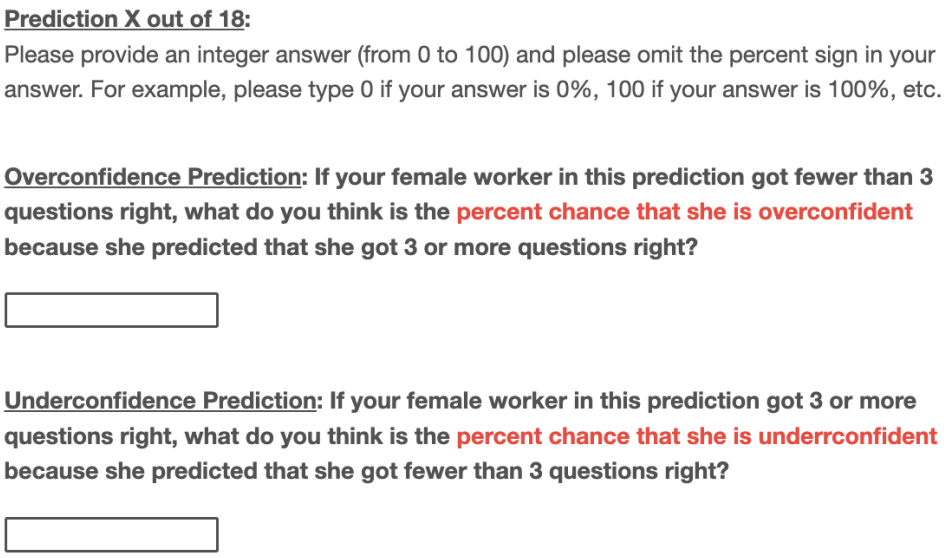


Figure I.1.21: Over/Underconfidence Beliefs (5+) about Female Workers, *Evaluator (Alternative Questions) Study*

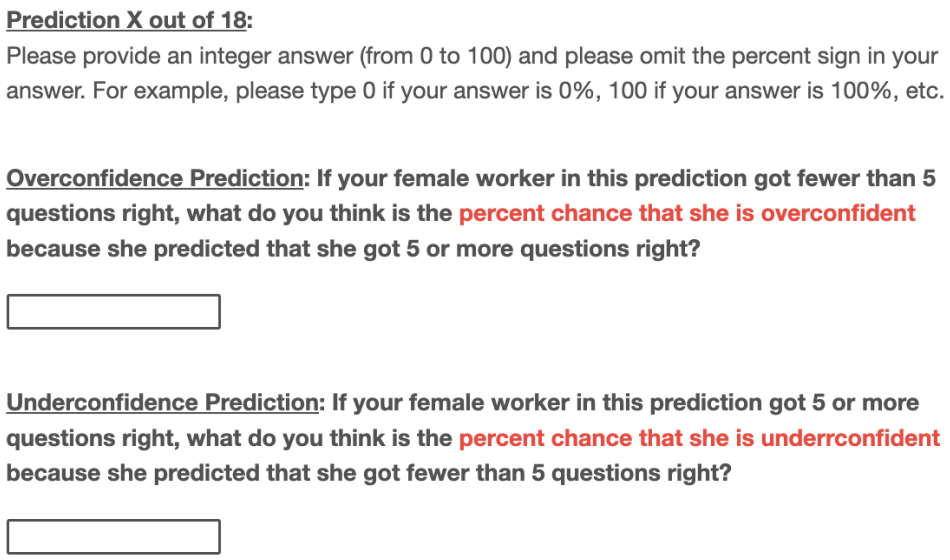


Figure I.1.22: Over/Underconfidence Beliefs (7+) about Female Workers, *Evaluator (Alternative Questions) Study*

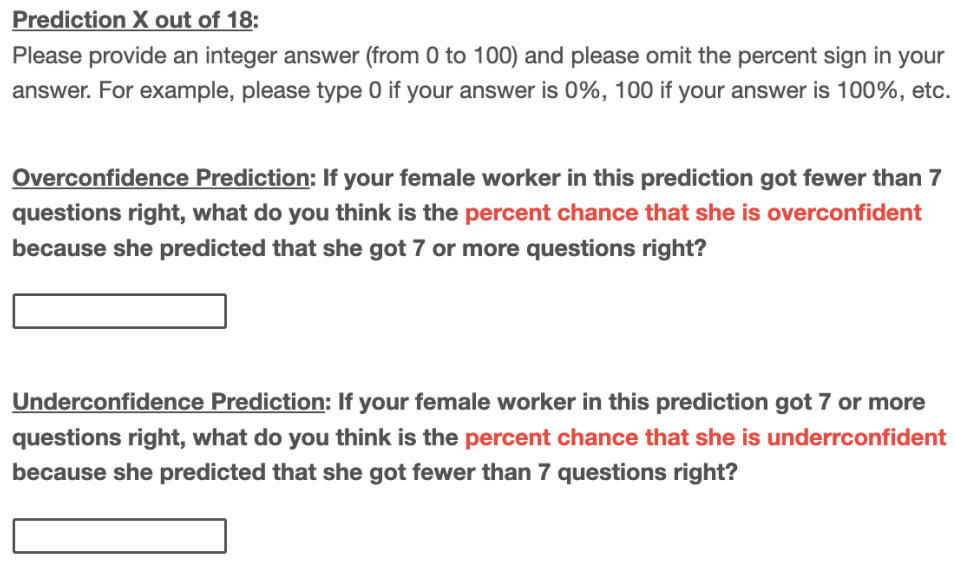


Figure I.1.23: Over/Underconfidence Beliefs (poor-2) about Female Workers, *Evaluator (Alternative Questions) Study*

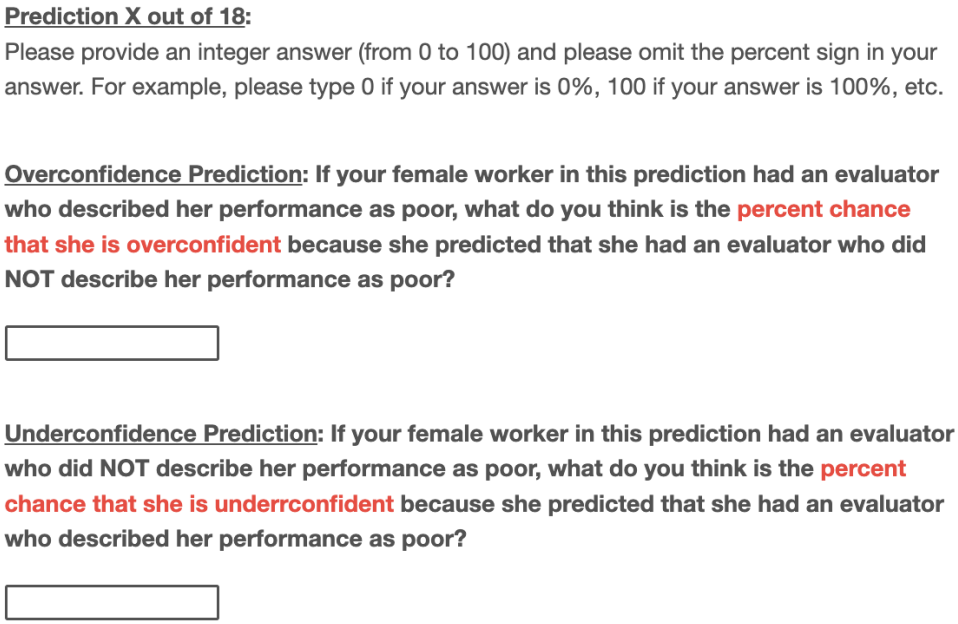


Figure I.1.24: Over/Underconfidence Beliefs (main self-evaluation) about Female Workers, *Evaluator (Alternative Questions) Study*

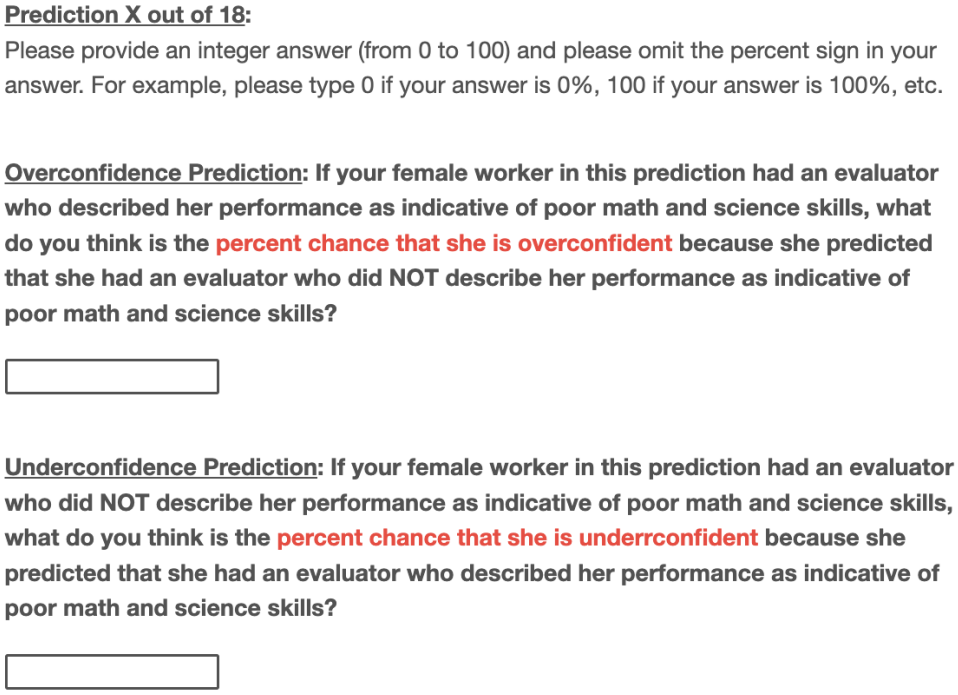


Figure I.1.25: Over/Underconfidence Beliefs (top-half) about Female Workers, *Evaluator (Alternative Questions) Study*

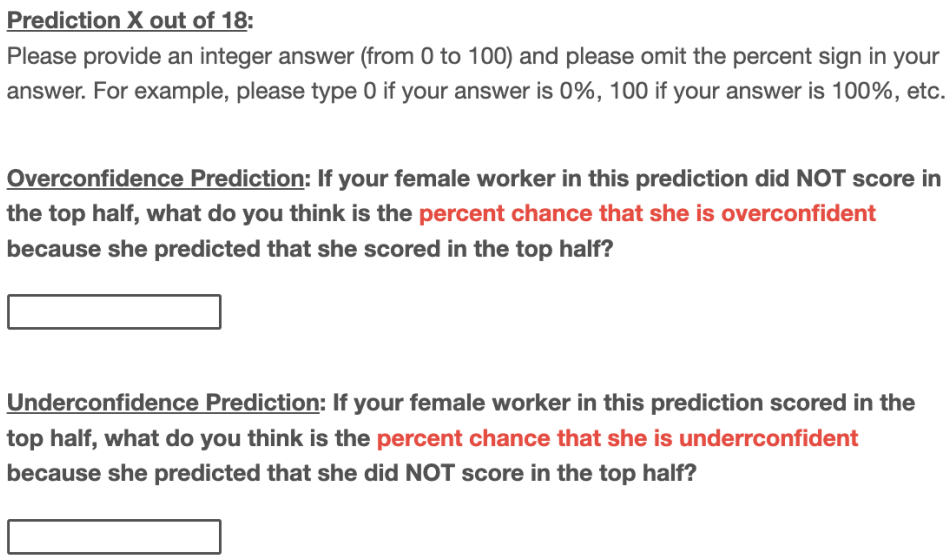


Figure I.1.26: Posterior Belief (3+) about Male Workers, *Evaluator (Alternative Questions) Study*

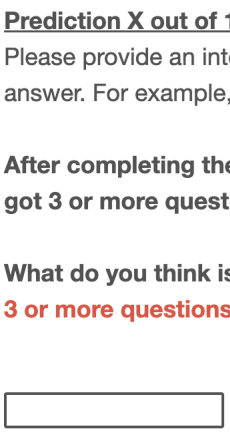


Figure I.1.27: Posterior Belief (5+) about Male Workers, *Evaluator (Alternative Questions) Study*

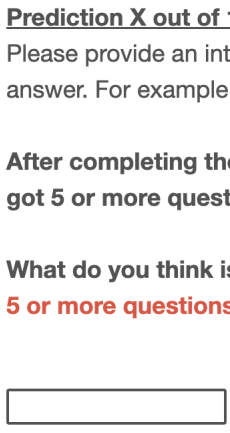


Figure I.1.28: Posterior Belief (7+) about Male Workers, *Evaluator (Alternative Questions) Study*

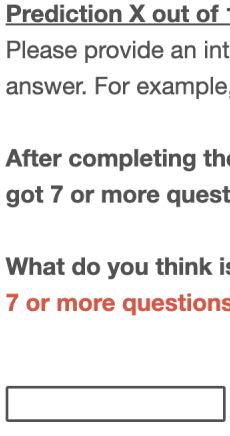


Figure I.1.29: Posterior Belief (poor-2) about Male Workers, *Evaluator (Alternative Questions) Study*

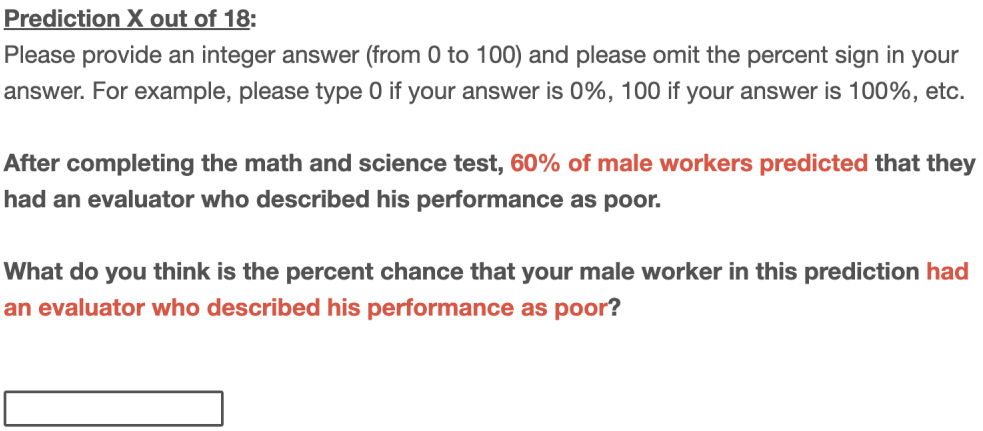


Figure I.1.30: Posterior Belief (main self-evaluation) about Male Workers, *Evaluator (Alternative Questions) Study*

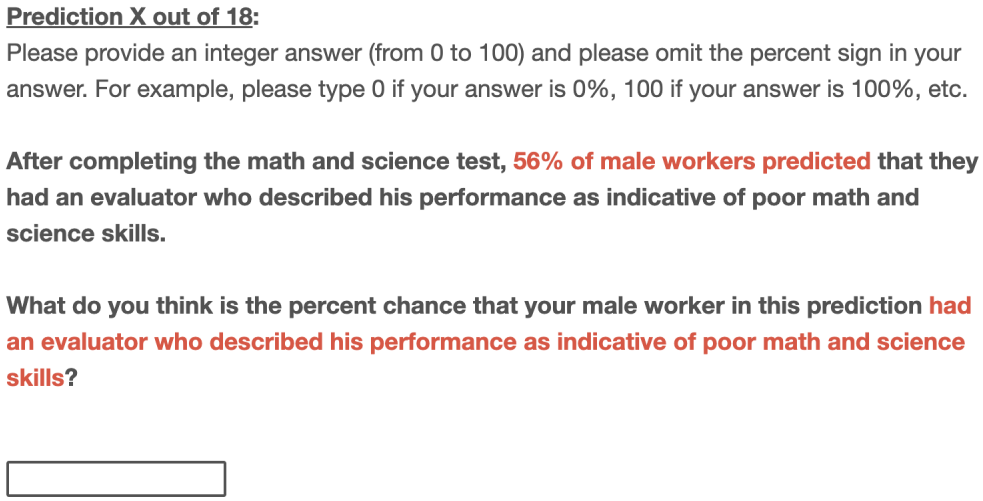
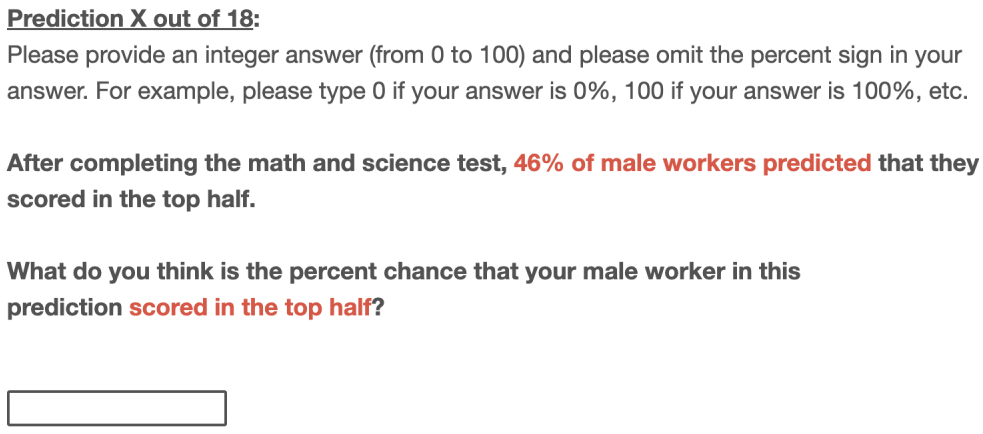


Figure I.1.31: Posterior Belief (top-half) about Male Workers, *Evaluator (Alternative Questions) Study*



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*

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 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

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*

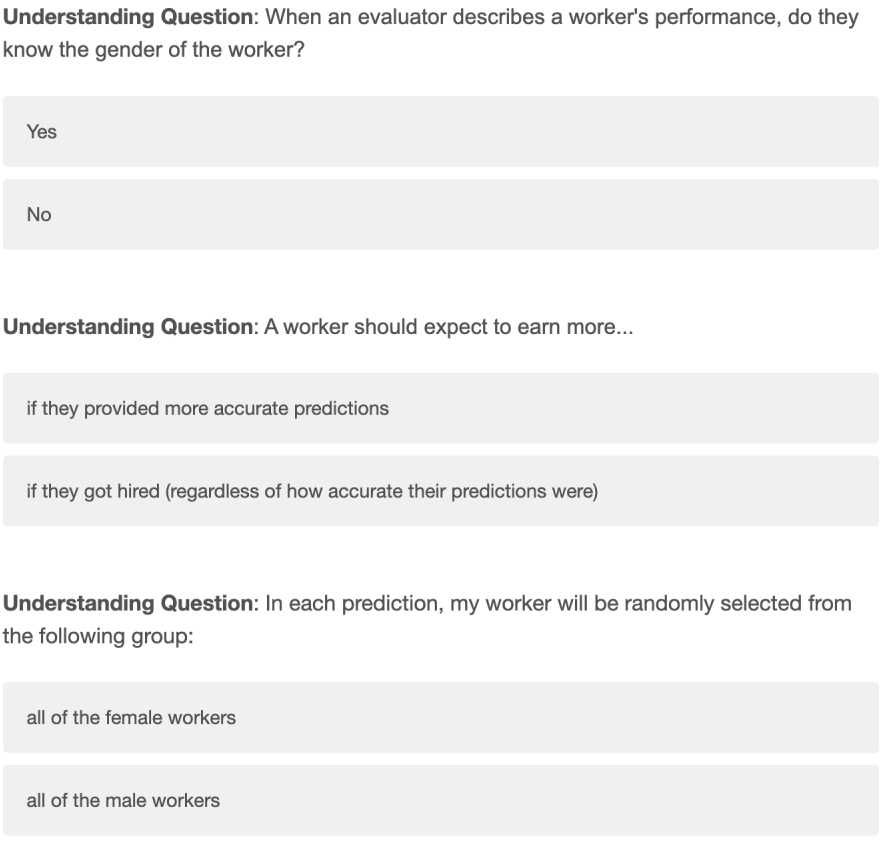


Figure I.3.3: Prior Belief about Female Workers, *Evaluator (Full Distribution) Study*

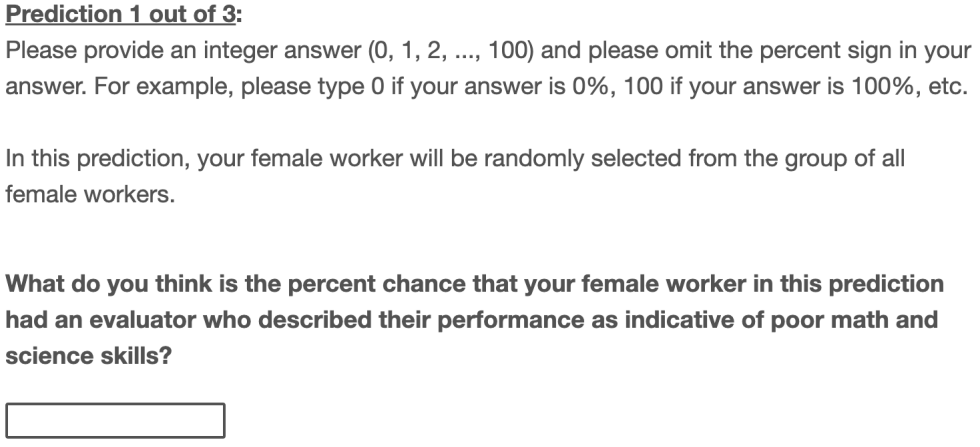


Figure I.3.4: Posterior Belief about Female Workers, *Evaluator (Full Distribution) Study*

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*

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 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.**

Overconfidence Prediction: 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?

Underconfidence Prediction: 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%**.

Figure I.4.1: Study Overview, *Baseline* Treatment of the *Evaluator (Professional Evaluators) 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, 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.

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

Female Workers

Figure I.4.2: Instructions about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

Your Workers

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.

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 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.

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 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*

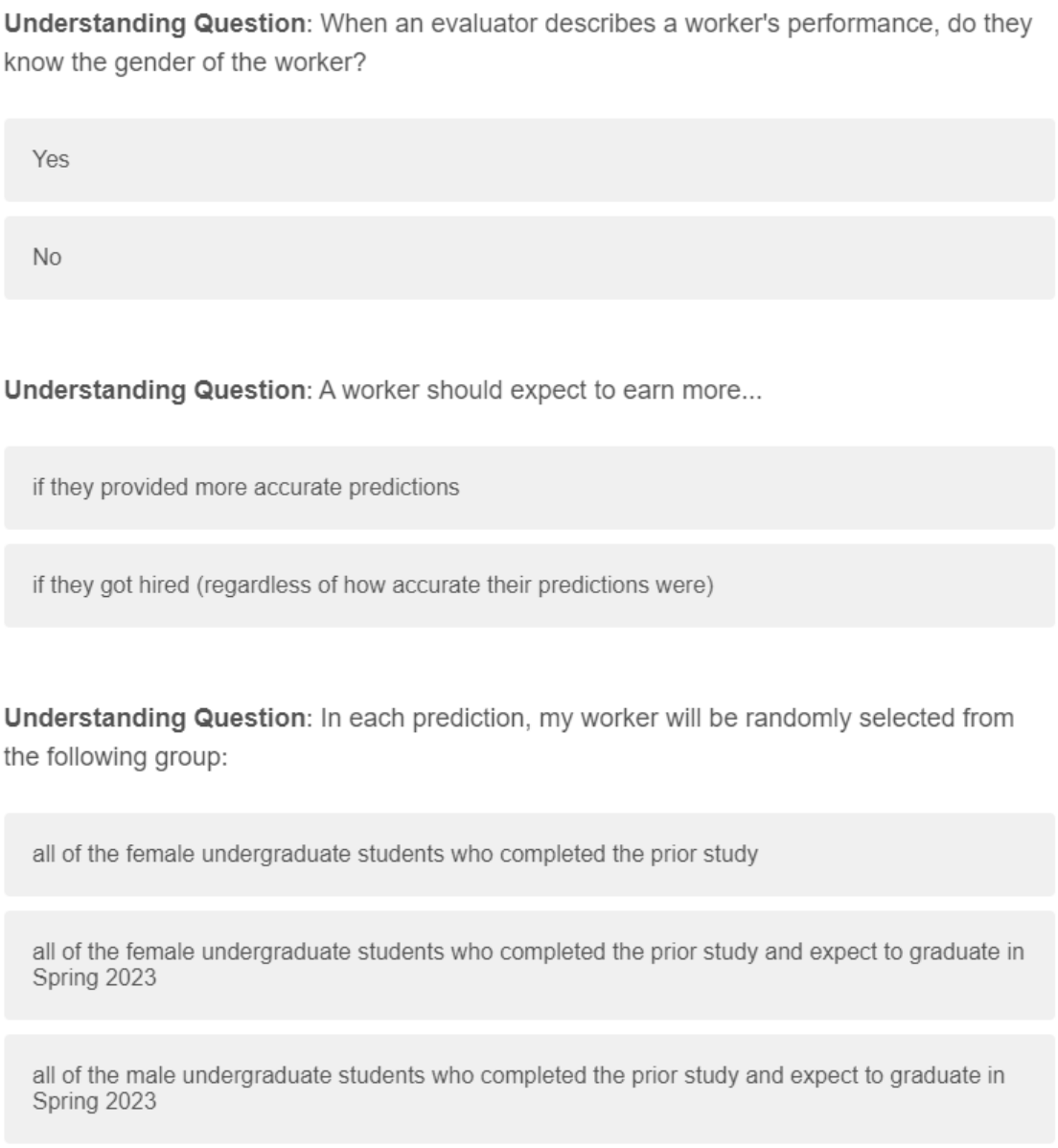


Figure I.4.4: Prior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

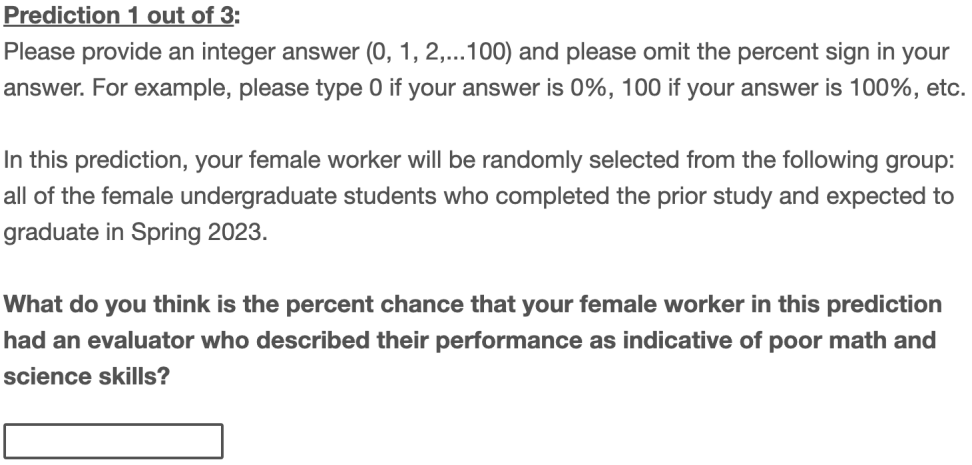


Figure I.4.5: Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

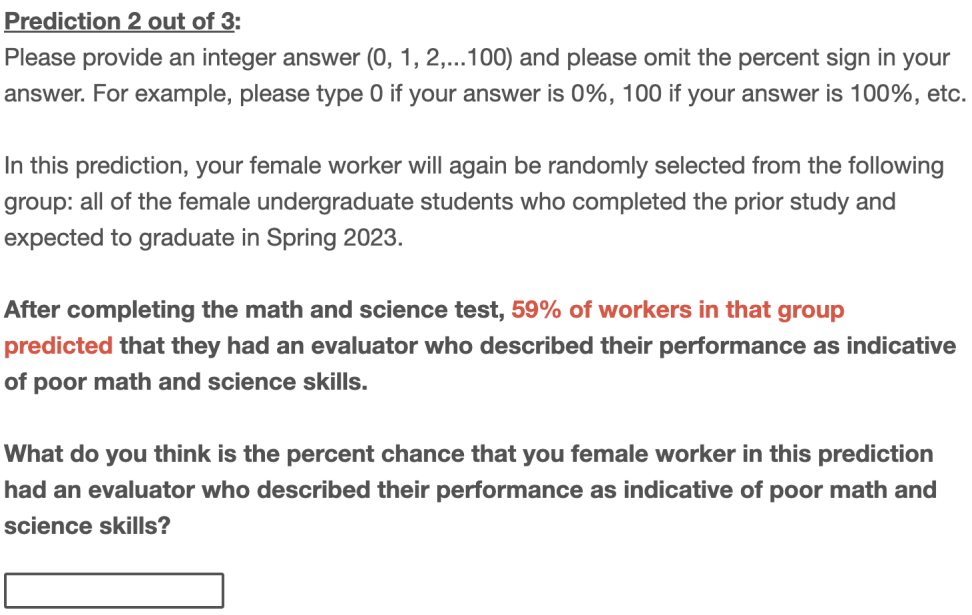


Figure I.4.6: Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

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 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.**

Overconfidence Prediction: 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?

Underconfidence Prediction: 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%**.

Group-2 (Female) Workers

Figure I.4.7: Instructions about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) 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 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.

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 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.

Worker Predictions:

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.

Figure I.4.8: Comprehension Questions about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

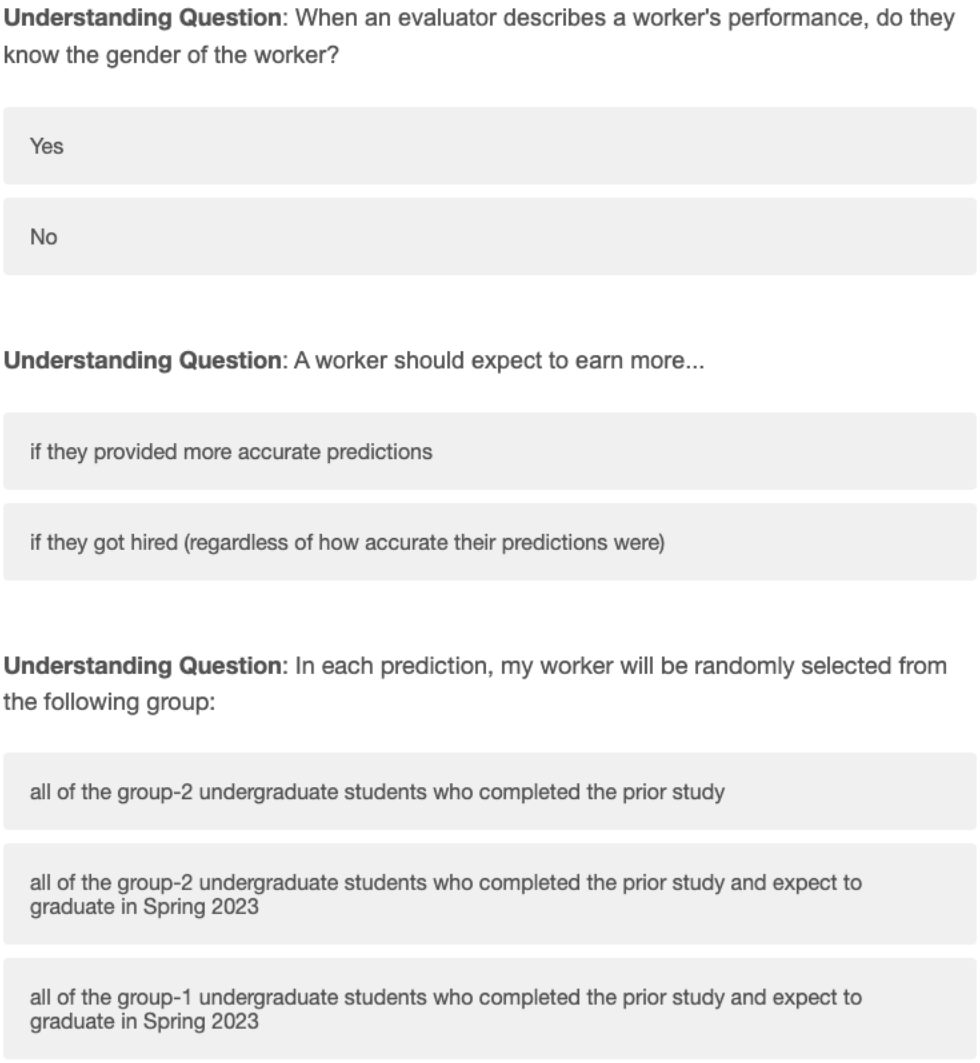


Figure I.4.9: Prior Belief about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

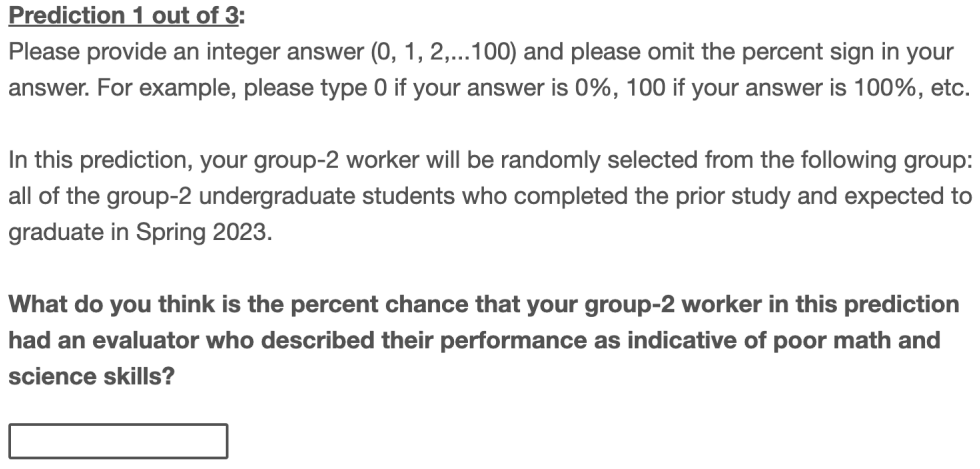


Figure I.4.10: Posterior Belief about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

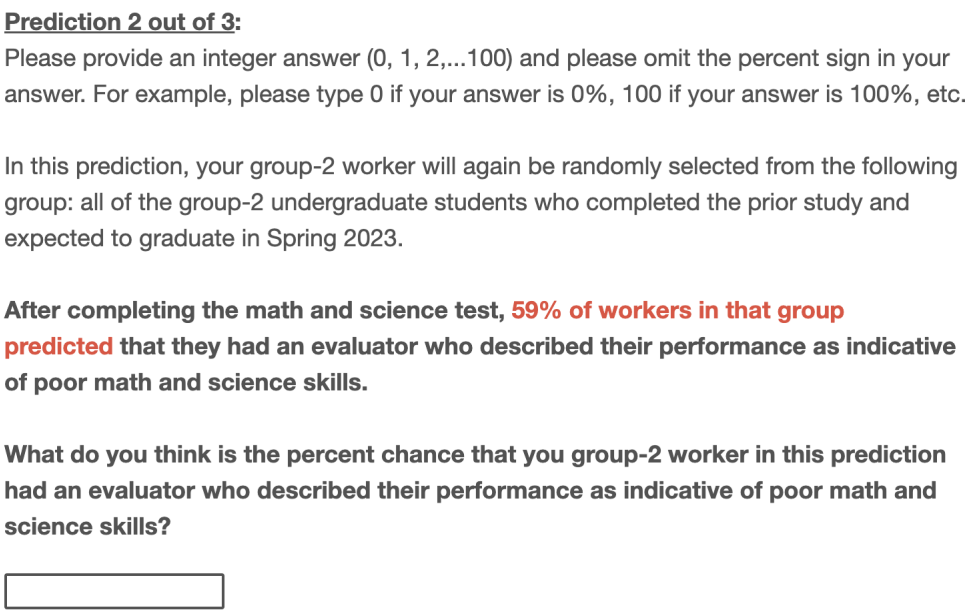


Figure I.4.11: Over/Underconfidence Beliefs about Group-2 Workers, *Baseline* Treatment of the *Evaluator (Professional Evaluators) Study*

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%**.

Figure I.5.1: Study Overview, *Baseline* Treatment of the *Evaluator (Extended) Study*

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*

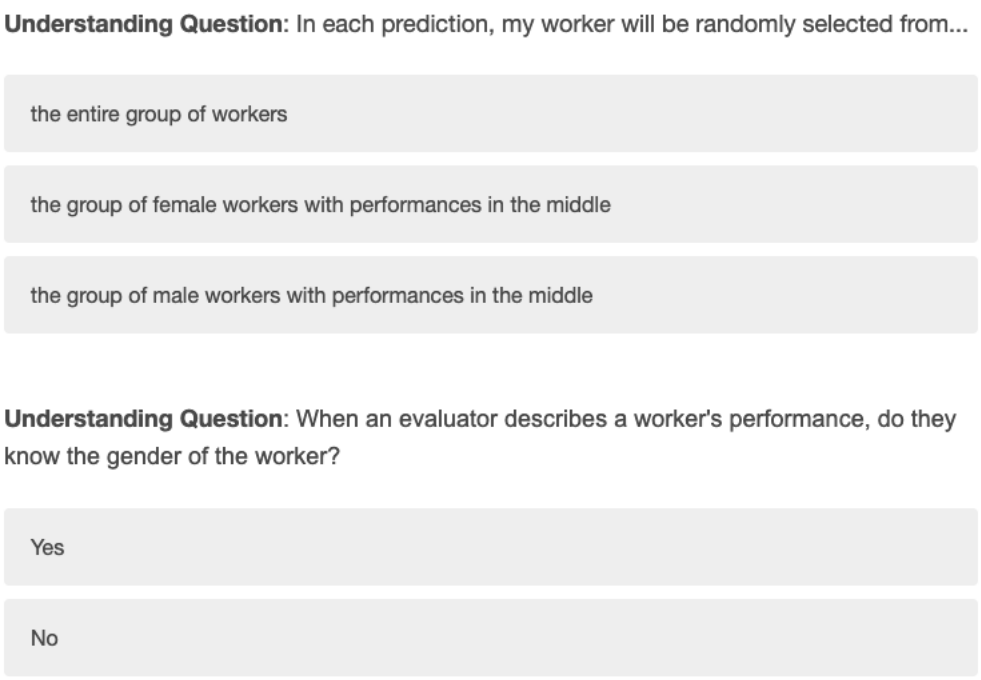


Figure I.5.4: Prior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

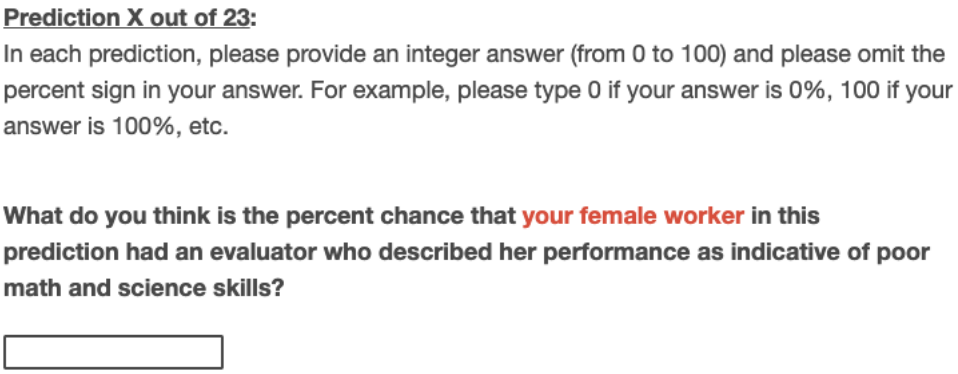


Figure I.5.5: Additional Instructions about Worker-Specific Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

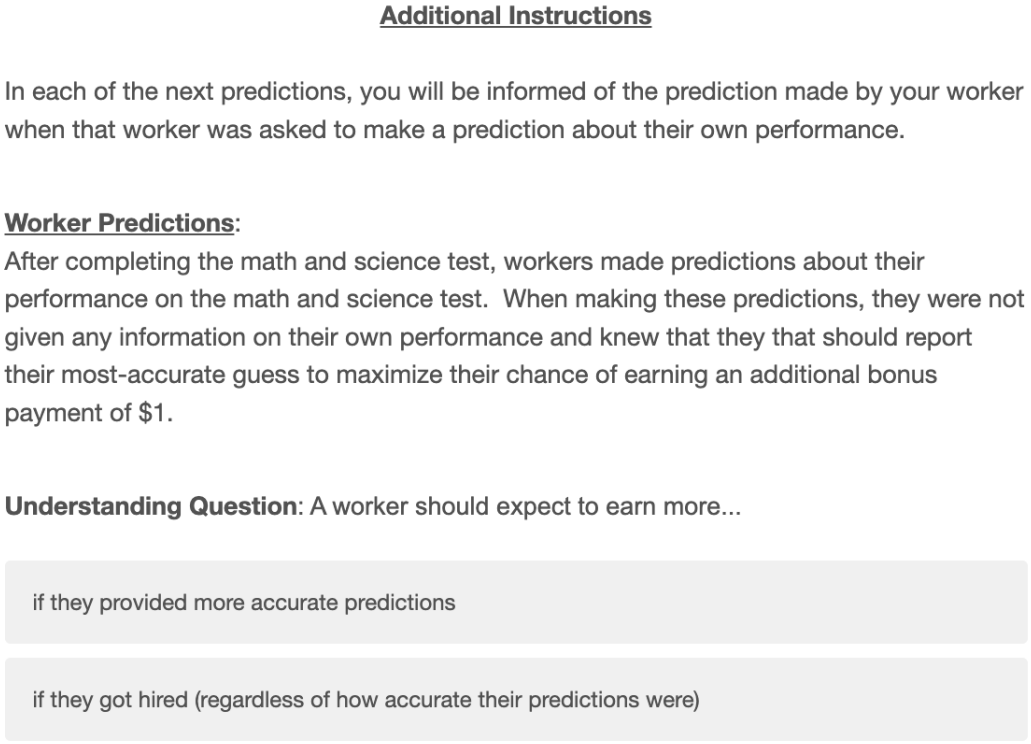


Figure I.5.6: Example of Worker-Specific Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

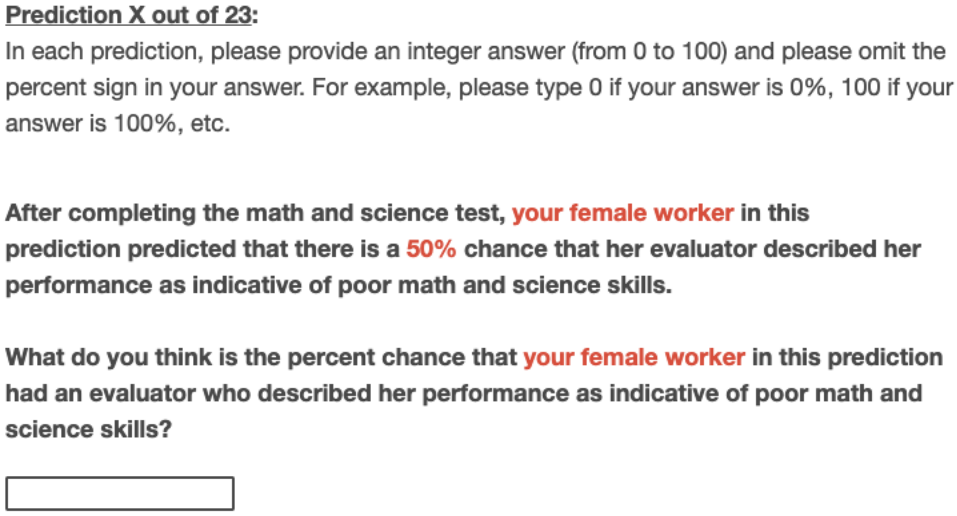


Figure I.5.7: Additional Instructions about Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

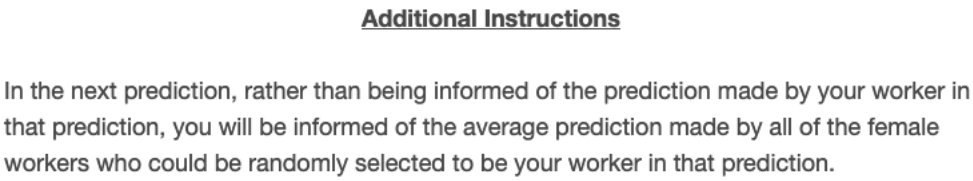


Figure I.5.8: Posterior Belief about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

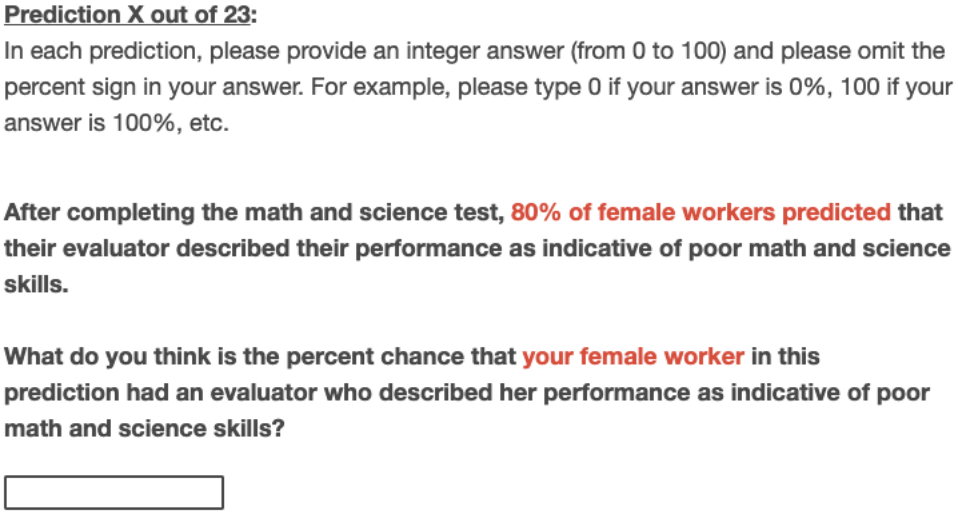


Figure I.5.9: Additional Instructions about Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*

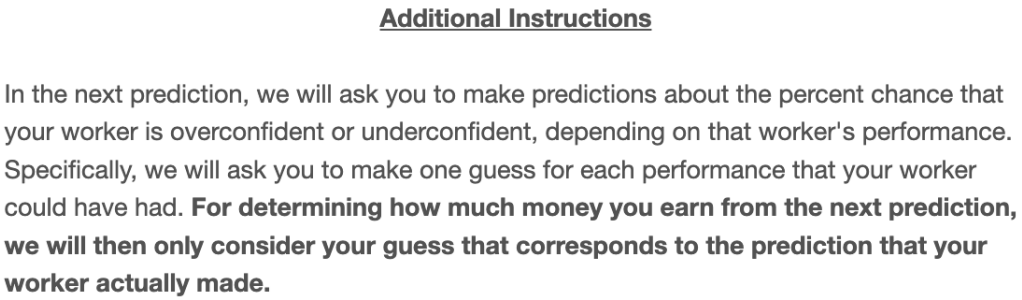
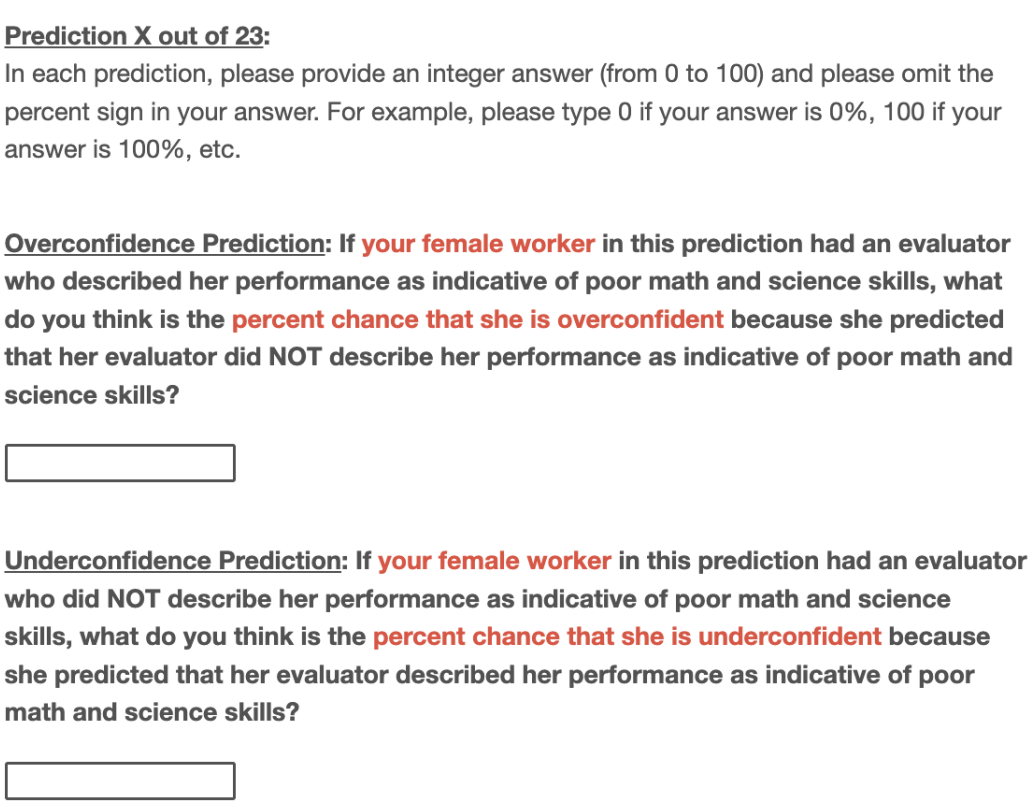


Figure I.5.10: Over/Underconfidence Beliefs about Female Workers, *Baseline* Treatment of the *Evaluator (Extended) Study*



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.

Figure I.5.11: Study Overview, *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

Main Instructions (Page 2 out of 2)

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.

Your workers

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.

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.

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*

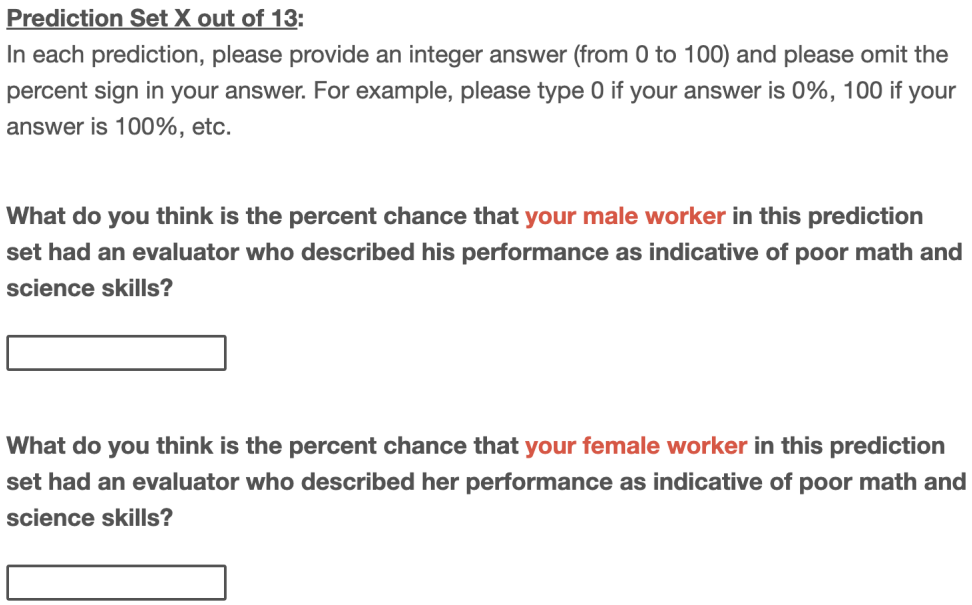


Figure I.5.13: Example of Worker-Specific Posterior Belief about Female and Male Workers, *Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

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:

- **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*

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:

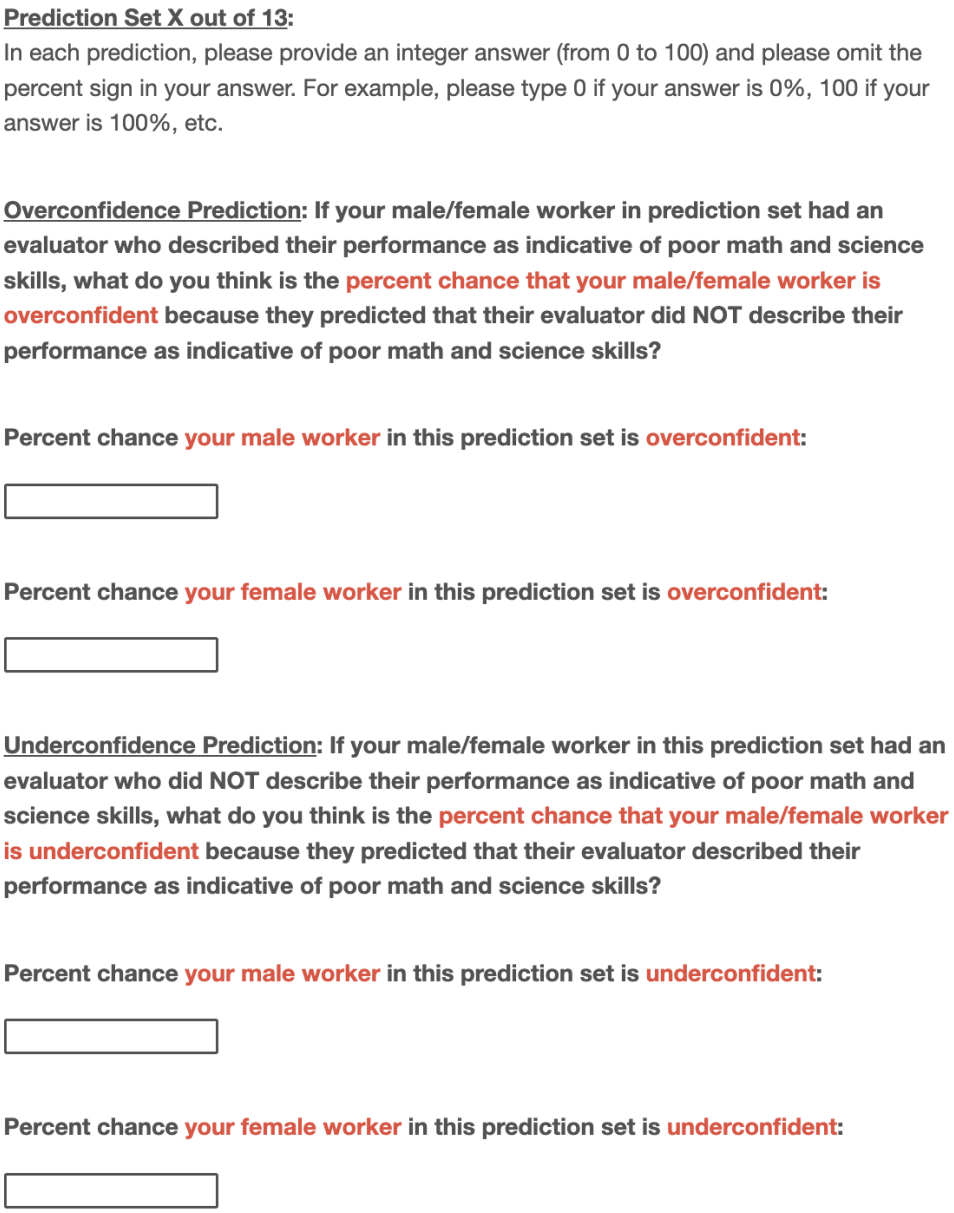
- **56% of male workers predicted that their evaluator described their performance as indicative of poor math and science skills, and**
- **80% of female workers predicted that their evaluator described their performance as indicative of poor math and science skills.**

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*



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*

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 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)

Female Workers

Figure I.5.17: Prior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

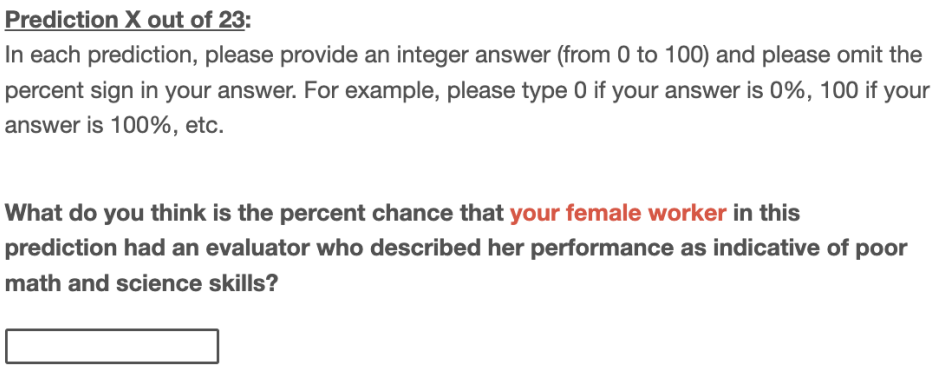


Figure I.5.18: Additional Instructions about Worker-Specific Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

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 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.5.19: Worker-Specific Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

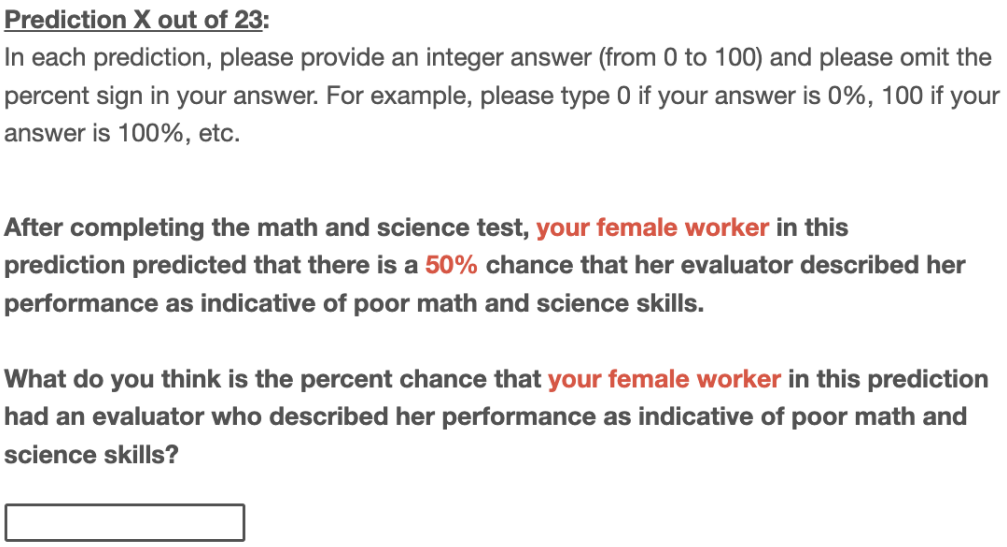


Figure I.5.20: Additional Instructions about Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

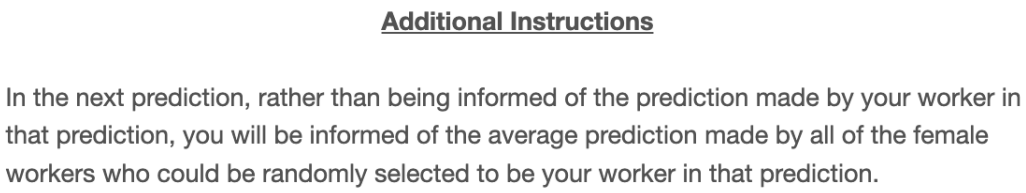


Figure I.5.21: Posterior Belief about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

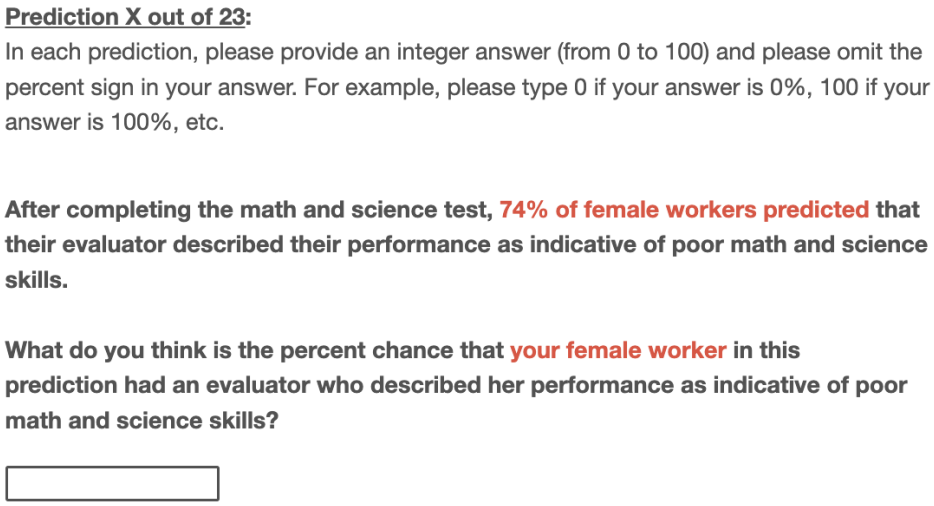


Figure I.5.22: Additional Information about Over/Underconfidence Beliefs about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*

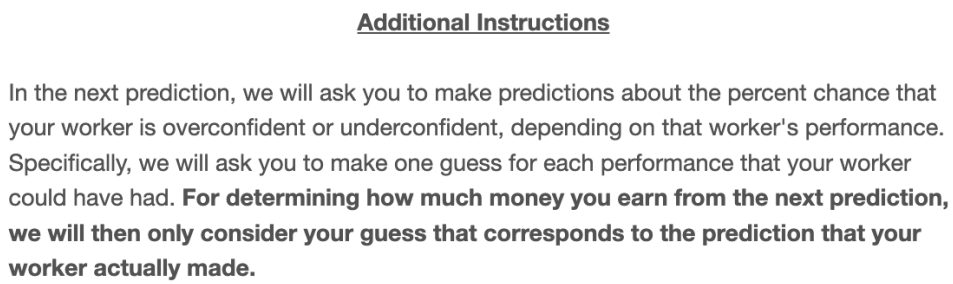
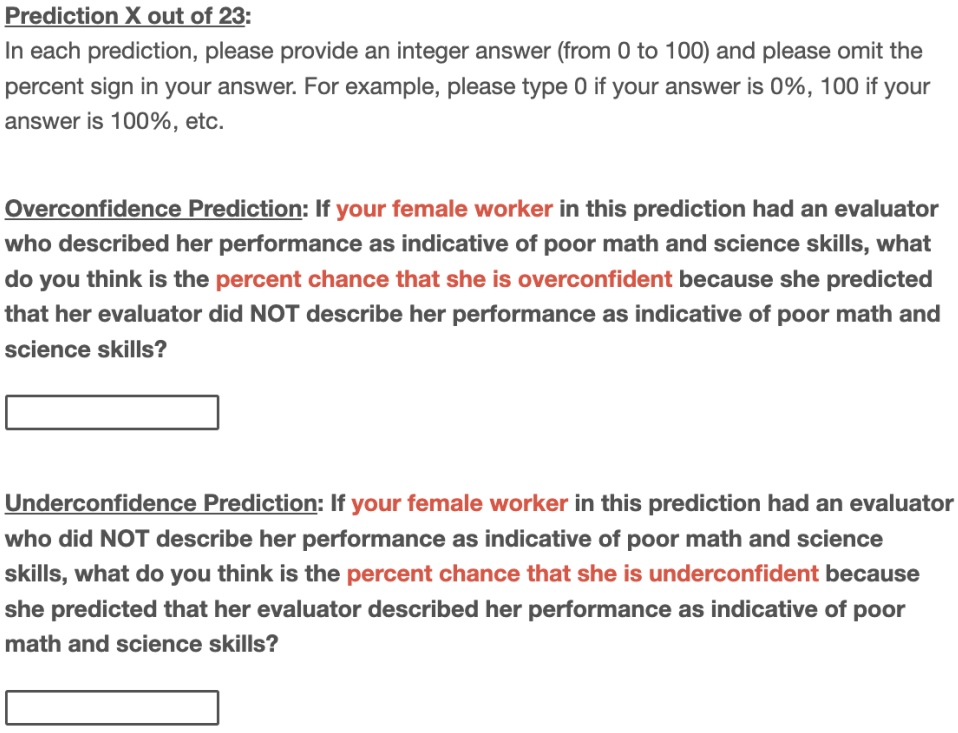


Figure I.5.23: Over/Underconfidence Beliefs about Female Workers, *Strategic Incentives* Treatment of the *Evaluator (Extended) Study*



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*

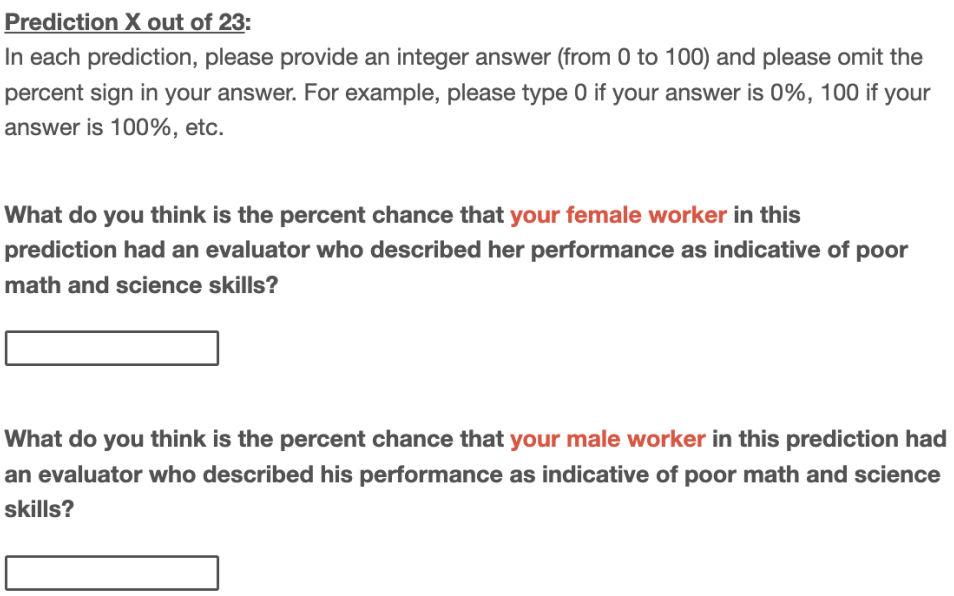


Figure I.5.25: Worker-Specific Posterior Belief about Male and Female Workers, *Strategic Incentives, Joint Evaluations* Treatment of the *Evaluator (Extended) Study*

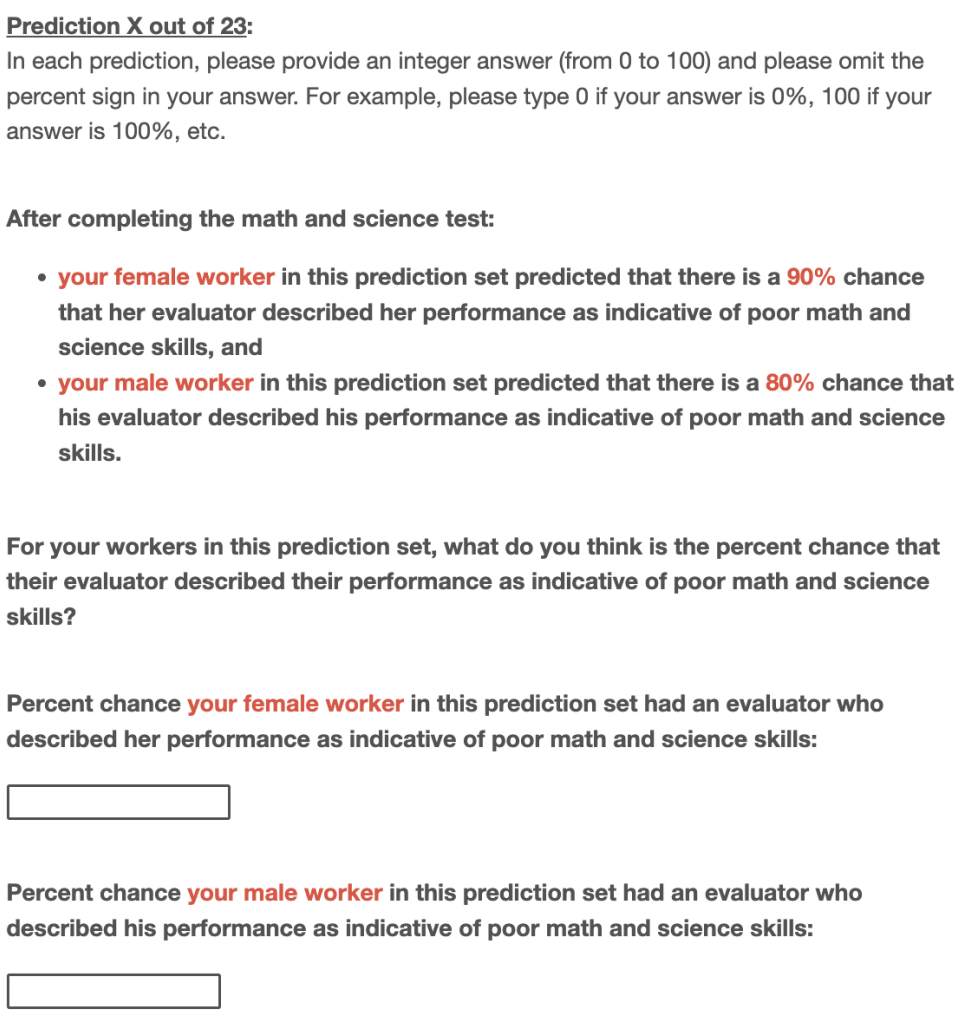
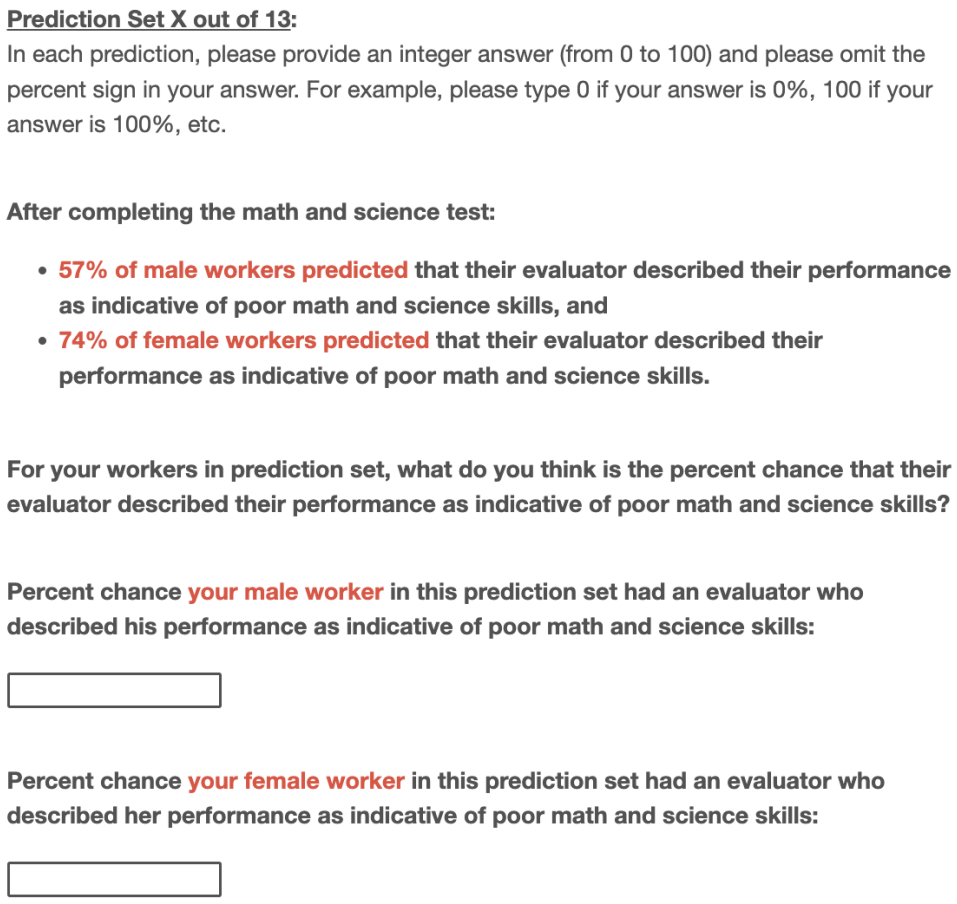


Figure I.5.26: Posterior Belief about Male and Female Workers, *Strategic Incentives, Joint Evaluations* Treatment of the *Evaluator (Extended) Study*



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:

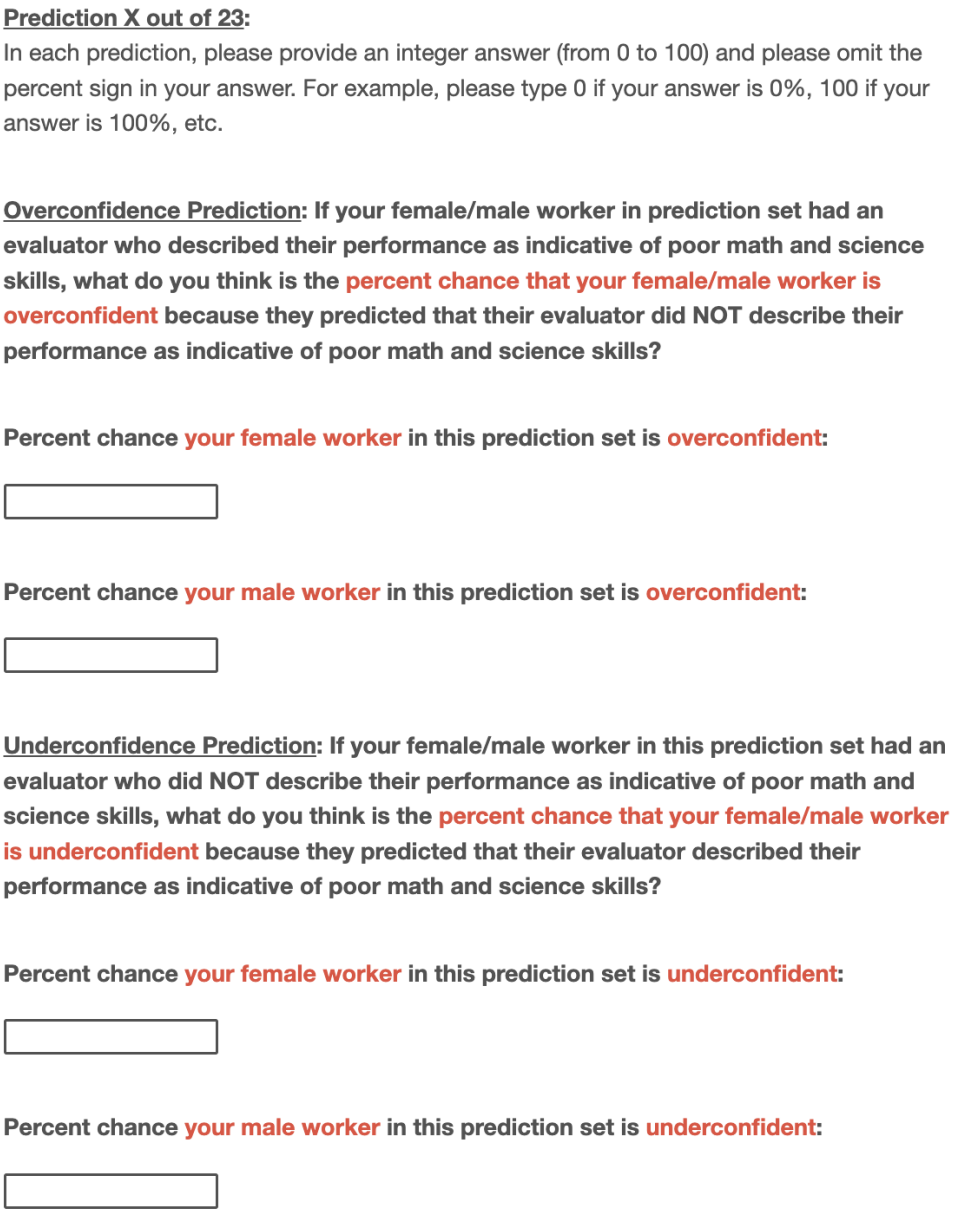
- **57% of male workers predicted that their evaluator described their performance as indicative of poor math and science skills, and**
- **74% of female workers predicted that their evaluator described their performance as indicative of poor math and science skills.**

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*



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%**.

Figure I.6.1: Study Overview, *Evaluator (Additional Demographics) 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 $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

Female Workers

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*

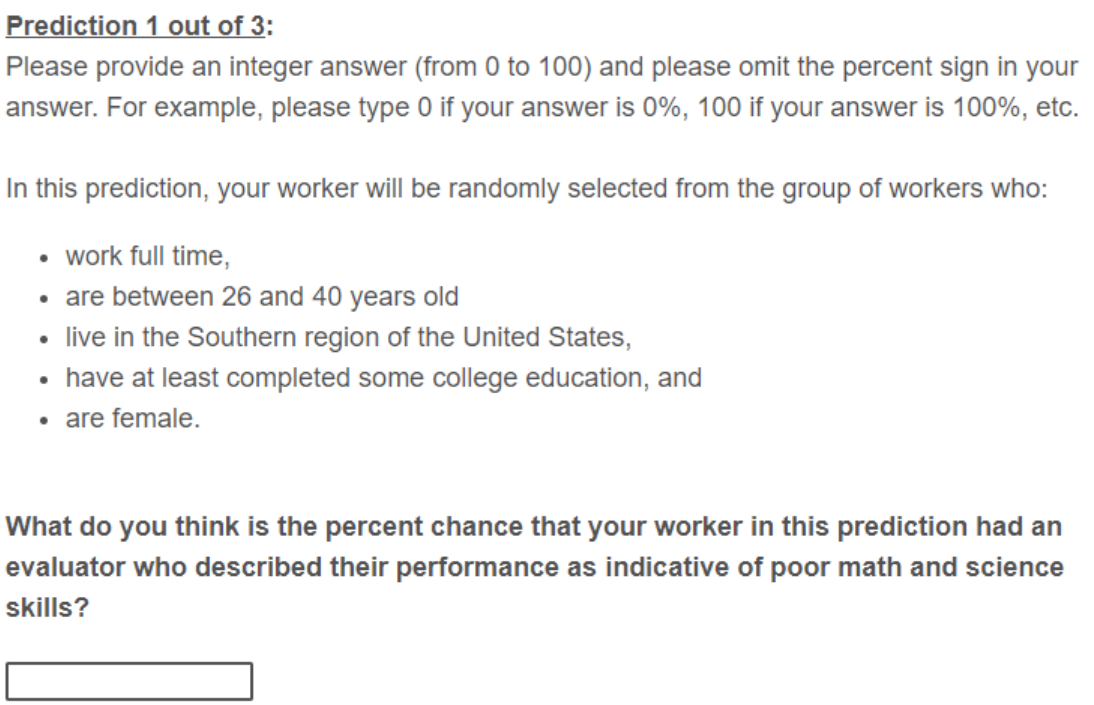


Figure I.6.5: Posterior Belief about Female Workers, *Evaluator (Additional Demographics) Study*

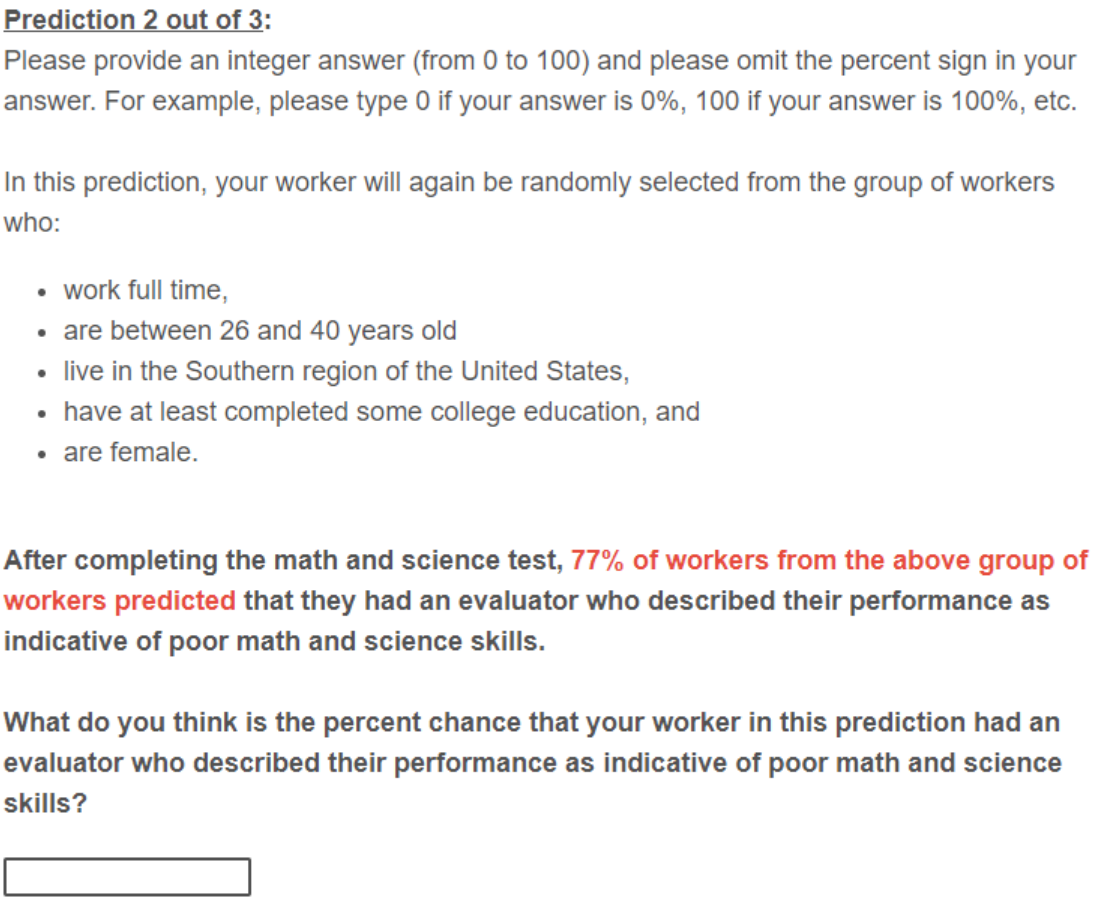


Figure I.6.6: Over/Underconfidence Beliefs about Female Workers, *Evaluator (Additional Demographics) Study*

Prediction 3 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.

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 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?

Underconfidence Prediction: 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*

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 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*

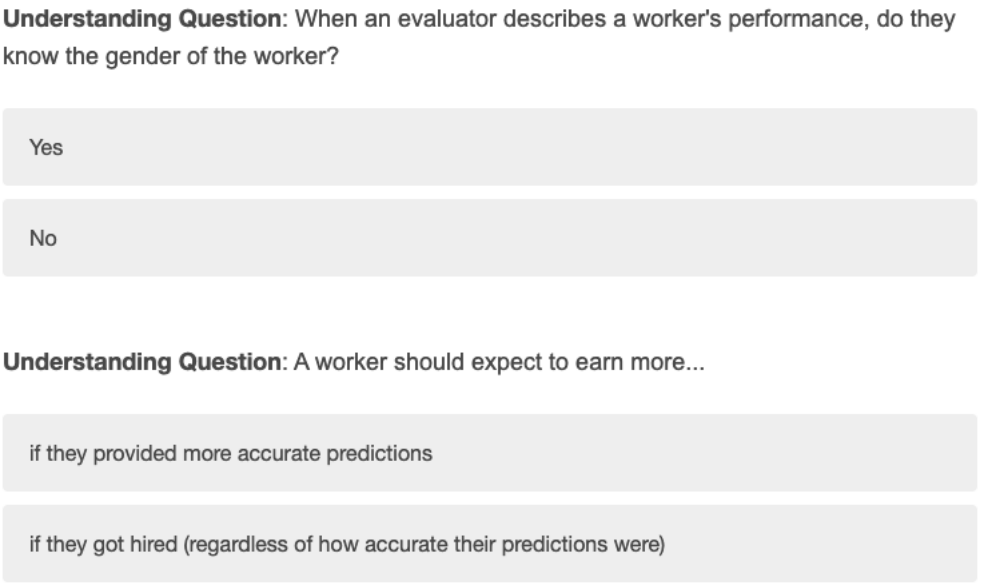


Figure I.7.4: Prior Belief about Female Workers, *Evaluator (Known Performance) Study*

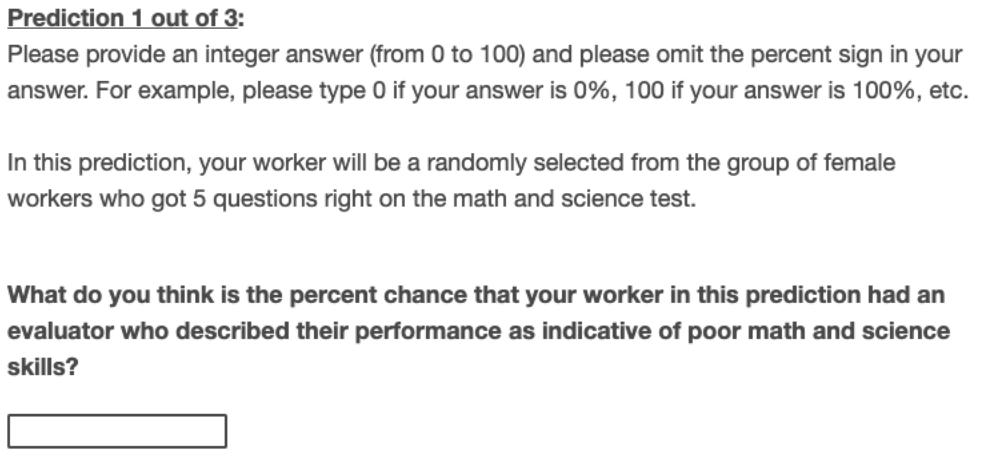


Figure I.7.5: Posterior Belief about Female Workers, *Evaluator (Known Performance) Study*

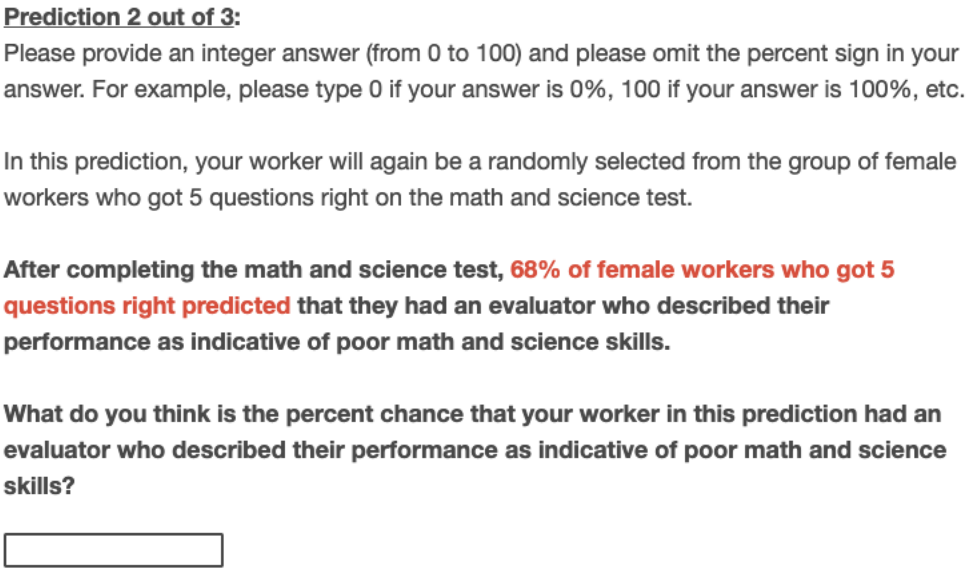


Figure I.7.6: Over/Underconfidence Beliefs about Female Workers, *Evaluator (Known Performance) Study*