NBER WORKING PAPER SERIES

THE GENDER GAP IN SELF-PROMOTION

Christine L. Exley
Judd B. Kessler

Working Paper 26345
http://www.nber.org/papers/w26345

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2019, Revised May 2021

The Gender Gap in Self-Promotion
Christine L. Exley and Judd B. Kessler
NBER Working Paper No. 26345
October 2019, Revised May 2021
JEL No. C91,D90,J16

ABSTRACT

In applications, interviews, performance reviews, and many other environments, individuals subjectively describe their ability and performance to others. We run a series of experiments, involving over 4,000 participants from online labor markets and over 10,000 school-aged youth. We find a large gender gap in self-promotion: Women subjectively describe their ability and performance to potential employers less favorably than equally performing men. Even when all incentives to promote are removed, however, the gender gap remains. The gender gap in self-promotion is reflective of an underlying gender gap in how individuals subjectively evaluate their own performance. This underlying gender gap proves persistent and arises as early as the sixth grade.

Christine L. Exley
Harvard Business School
25 Baker Way
Baker Library 449
Boston, MA 02163
clexley@gmail.com

Judd B. Kessler
The Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104
and NBER
judd.kessler@wharton.upenn.edu

1 Introduction

At various points in their educational and professional lives—in college and professional school applications, in job applications and interviews, in performance reviews—individuals are explicitly asked about their performance and ability. In many other contexts, individuals face implicit invitations or opportunities to discuss their performance and ability. How they respond can directly impact their educational and labor market outcomes, and so gender differences in responses could contribute to observed gender gaps in educational and labor market outcomes (Goldin, 2014; Blau and Kahn, 2017).

Consequently, one might be worried about a gender gap in "self-promotion," or how individuals communicate their performance and ability to others. Given that individuals typically communicate their performance and ability to others in *subjective* terms (e.g., asserting that they are "good" at math) rather than in more precise terms (e.g., asserting that they fall in the 90th percentile of math ability according to some observable metric), one may be particularly concerned about a gender gap in self-promotion that relates to such subjective descriptions. Indeed, prior work shows that women are less likely to report being "proficient" or "skilled" in programming languages on their resumes (Murciano-Goroff, 2021), are less likely to use "positive" words in their titles and abstracts for papers on clinical research (Lerchenmueller, Sorenson and Jena, 2019), and are more likely to use narrow topic-specific—rather than broad—words in their research grant proposals (Kolev, Fuentes-Medel and Murray, 2019).[1]

However, research on self-promotion that relates to how individuals subjectively communicate their performance and ability—the ubiquitous type of self-promotion that we focus on in this paper—faces three distinct challenges. First, measuring self-promotion can be difficult because subjective descriptions of performance and ability are often qualitative in nature. Second, establishing a gender gap in self-promotion among *equally* performing men and women is difficult because, in many settings where self-promotion is relevant, there are not precise and observable measures of true underlying performance. Third, examining the underlying drivers of self-promotion is difficult absent the ability to exogenously manipulate the environment.

The contributions of this paper stem from our ability—through a carefully controlled experimental setting—to document a gender gap in self-promotion among *equally* performing men and women and to narrow in on the underlying drivers of this gap. We find a large gender gap in self-promotion when participants are not informed about their performance. We also find a large gender gap in *informed* self-promotion when participants are perfectly informed about their absolute and relative performance. We find that these gaps reflect an underlying gender gap in how individuals subjectively evaluate their performance that arises even when incentives to "promote" are removed. This underlying gender gap proves notably persistent when participants are asked about their own performance on a math and science test, but gender gaps do not emerge if participants are instead

[1] For work on gender differences in communication and perceptions of that communication, see also Bohren, Imas and Rosenberg (2018), Grossman et al. (2019), and Manian and Sheth (2020).

asked to describe their performance on a more female-typed task or are instead asked to describe the performance of someone else. These results—along with additional heterogeneity analyses—point towards the role of gender norms and culture coloring how men and women subjectively view and evaluate their performances on male-typed tasks. Finally, by collecting data from more than 10,000 middle-school and high-school students, we show that this gap is robust to a broader population and that it arises at a young age.

In our main study version, the *Self-Promotion* version, participants first complete a task that requires them to answer a test of math and science questions. They then provide subjective answers—on quantitative scales that facilitate measurement—to self-evaluation questions about their performance on that task. Participants are aware that potential employers will use one of these subjective answers—and only that answer—to decide whether to hire them and how much to pay them.[2] Answers to these questions reveal a substantial and significant gender gap in self-promotion. For example, when asked to indicate agreement on a scale from 0 to 100 with a statement that reads "I performed well on the test," women provide answers that are 13 points lower than equally performing men. The average participant describes their performance as a 53 out of 100, so this 13-point gender gap represents 24% of the mean. We find similarly substantial and statistically significant gaps in response to the three other self-evaluation questions we ask, including two others on this 0-to-100 scale and one on a six-point Likert scale that defines 1 as "terrible" performance and 6 as "excellent" performance.

Motivated by the possibility that women describe their performance more negatively because they think that they had a lower performance—either in absolute or relative terms—we then explore whether we also observe a gender gap in *informed* self-promotion. Specifically, we investigate whether a gender gap persists when participants are provided with perfect information about their absolute and relative performance on the task (i.e., we tell them exactly how many questions they answered correctly on the test and where they fell in the performance distribution of other participants). Results suggest that the gender gap in informed self-promotion is somewhat smaller than the gender gap in self-promotion, but we still find a substantial and statistically significant gender gap in informed self-promotion.

In light of these robust gender gaps in self-promotion and informed self-promotion, we then narrow in on whether there is an underlying gender gap even absent any incentives to promote. Specifically, we turn to the *Private* version of our study, which is nearly identical to the *Self-Promotion* version, except that answers to the self-evaluation questions are not shared with potential employers. Consistent with individuals responding to the incentives to promote, both men and women subjectively evaluate their performance more negatively in the absence of employers. Nevertheless, there is a statistically significant gender gap in the *Private* version that is just as large as in the *Self-Promotion* version. This finding highlights that the gender gap in self-promotion reflects an

[2]Data from our experimental employers confirm that self-promotion pays: more favorable answers to the self-evaluation questions increase the chance that participants are hired and the subsequent earnings they receive.

underlying gender gap in subjective evaluations of performance.

Additional study versions document the robustness of this underlying gap. For example, in our *Private (Social Norms)* version, participants are informed about their absolute and relative performance and also informed about how others—with the same performance as them—answer self-evaluation questions. We find that the gender gap in this version remains just as large as in the *Private* version. More generally, when considering additional study versions and several replications of our results, we find a statistically significant gender gap 64 out of 64 times in responses to self-evaluation questions about own performance on the math and science task.

In only two of our study versions does the gender gap not persist. The gap does not persist when we ask individuals to evaluate the performances of others rather than themselves in the *Private (Other-Evaluations)* version, and the gap does not persist when we ask individuals to evaluate their performance on a more female-typed task in the *Private (Verbal)* version. These two findings highlight that the gap is not due to men and women having different views about how to subjectively evaluate performance in general (e.g., having different "standards" for performance in general). Instead, the results highlight that the gap arises because women subjectively evaluate their *own* performance on a *male-typed* task less favorably than equally performing men.

That the gender gap is so robust when men and women subjectively evaluate their own performance on a male-typed task—but is not present when they evaluate others or evaluate themselves on a more female-typed task—is consistent with the importance of gender stereotypes (Bordalo et al., 2019).[3] It is also consistent with the ways in which gender norms and culture may influence behavior given one's life experiences (Bowles, Babcock and McGinn, 2005; Gneezy, Leonard and List, 2009; Cárdenas et al., 2012). Even when asked about performance on a specific math and science test, and even when they know their absolute and relative performance on that test, experiences in a society with certain gender norms and culture may color how participants subjectively view and evaluate that performance. Indeed, additional results—highlighting correlations of subjective evaluations with other demographic characteristics such as age, education, and political affiliation—point towards the role of life experiences, even beyond those specifically tied to gender, in contributing to subjective evaluations of performance.

If the ways in which women and men subjectively view and evaluate their performance becomes colored over time due to their life experiences, an important question is how early these gender differences arise. We explore this question by recruiting more than 10,000 middle-school and high-school students to participate in a modified *Private* version of our study. We find large and statistically significant gender gaps in their answers to self-evaluation questions across all ages, including among sixth-graders, the youngest students we study.

Our work contributes to a robust prior literature on gender gaps in economic outcomes and the drivers of these gaps (Croson and Gneezy, 2009; Bertrand, 2011; Azmat and Petrongolo, 2014;

[3]Gender stereotypes are less relevant when we ask participants to evaluate other participants because we do not reveal the gender of those other participants.

Niederle, 2016). We complement this literature by documenting a new gender gap on an important behavior that—given the prevalence of explicit and implicit opportunities for individuals to communicate subjective descriptions about their performance—can impact educational and labor market outcomes. We find a large and robust gender gap in self-promotion among *equally* performing men and women and investigate the drivers of this gap. We find that the gender gap in self-promotion is reflective of an underlying gender gap in how men and women subjectively evaluate their own performance on male-typed tasks, even absent any promotion incentives. We find that this underlying gap arises at a young age and is difficult to mitigate.

We see a number of avenues for future work. First, we view our findings as informative and complementary to the broader gender literature. That the underlying gender gap we observe seems deeply ingrained and related to gender norms and culture suggests that it could be related to gender gaps previously observed in the economics literature. Gender differences in subjective views about performance could relate to—and perhaps contribute to—gender differences in other outcomes, such as those relating to negotiation (Hernandez-Arenaz and Iriberri, 2918) and competition (Niederle and Vesterlund, 2011). Future work could investigate the relationship of subjective views of performance with these other outcomes. Second, from a policy perspective, future work could explore how to mitigate this underlying gender gap. For example, if the goal is to treat equally performing men and women equally, future work could investigate the impact of deemphasizing subjective self-evaluations relative to more objective metrics to determine hiring and promotion decisions or investigate strategies to change perceptions of gender norms or the norms themselves. Given the persistence of the gender gap and the fact that it arises at a young age, more work is warranted.

2 Design, Data Collection, and Setting

In this Section, we first describe the structure of our data collection and what motivated it. We then present details about the seven primary versions of our study in which subjects were recruited from online labor markets: the *Self-Promotion* version, the *Self-Promotion (Risky)* version, the *Private* version, the *Private (Social Norms)* version, the *Private (Informed Immediately)* version, the *Private (Other-Evaluation)* version, and the *Private (Verbal)* version. Section 2.1 describes the *Self-Promotion* version in detail; each Section 2.2–2.7 describes a different study version by comparing it to prior versions and emphasizing what we can learn from it. We then provide results on participants' performance and beliefs on the math and science test, in Section 2.8, which highlights that our setting—like many of the settings in which gender gaps in pay and labor market representation are observed—is "male-typed."

We recruited 3,892 participants from online labor markets—Amazon's Mechanical Turk (MTurk) and Prolific—to participate in one of our seven versions across five waves of data collection, as shown in the first five rows of Table 1.[4] Each participant was guaranteed a completion fee plus a possible

[4]To be eligible for any study version, workers must have previously completed at least 100 tasks (on MTurk or Prolific, depending on the sample) with a 95% or better approval rating from prior employers, and workers must be working from a United States IP address. The median age is 33 years old, the median educational attainment is a

bonus payment from one randomly selected part of the study.[5] After participants completed all parts of the study, they took a short follow-up survey that collected demographic information, including gender. Gender was not mentioned prior to this follow-up survey, so participants were not primed to think about gender when answering the self-evaluation questions.

Why did five waves of data collection occur? We collected data over five waves because of the persistence of the gender gap across study versions and because of our desire to test the boundaries of this gap. In the first wave, we randomly assigned workers to either the *Self-Promotion* version, the *Self-Promotion (Risky)* version, or the *Private* version. Then, in subsequent waves, we ran additional versions, while always replicating one of our prior versions. New versions run in waves 2–5 were built off of the *Private* version. Our focus on the *Private* version in these waves was motivated by our desire to explore the potential drivers of the underlying gender gap in subjective evaluations of performance that wave 1 demonstrated was driving our gender gap in self-promotion.

As will be discussed in what follows, our data collection and continual replication of our earlier findings—across time and across labor market platforms—highlights the robustness of our results. When considering responses to self-evaluation questions relating to a participant's own math and science performance, we find a statistically significant gender gap in 64 out of 64 of specifications (as defined by the self-evaluation question, whether subjects are informed, the study version, and the wave).

As noted in the final row of Table 1, an additional 10,637 youth participated in a modified *Private* version of our study designed to explore the origins of the underlying gender gap in how individuals subjectively evaluate their performance. The specific design of this version, and the associated results, are discussed in Section 5.

In addition to the data described above, 300 participants completed a version of our study as "employers," who are relevant for the *Self-Promotion* and *Self-Promotion (Risky)* versions of our study.[6] Appendix B describes this study version and presents results from it. Those results demonstrate that self-promotion pays. Participants who report higher self-evaluations in the *Self-Promotion* and *Self-Promotion (Risky)* versions of our study are paid more by employers.

Bachelor's Degree, and the percentage of male participants is 59%. While participants were required to correctly answer understanding questions at various points to proceed in the study, no participants were excluded from our data analysis.

[5]In all of our studies run on MTurk (i.e., data collected in waves 1–4), participants received a $2 completion fee for a 20-minute study. In our studies run on Prolific (i.e., data collected in wave 5), participants received a $4 completion fee for a 25-minute study. The longer study time reflects additional questions asked on the follow-up survey in wave 5, detailed in Section 4. The higher completion payment reflects the longer study, that we collected wave 5 data in 2021 when norms were to pay more, and that we recruited from Prolific rather than MTurk.

[6]In addition to the participants described in the main text, we use performance data from 200 participants to create reference groups to provide participants with information on relative performance (100 participants who completed the math and science test and 100 participants who completed the verbal test). We also analyze data from 399 MTurk workers who evaluated free-response comments generated by participants as described below and discussed in Appendix C. Including these 599 participants and the 300 employers described in the main text, this paper involves a total of 4,791 study participants from online labor markets.

Table 1: Study Versions by Wave

	Self-Promotion	Private	Self-Promotion (Risky)	Private (Social Norms)	Private (Imm. Informed)	Private (Other-Evaluation)	Private (Verbal)
Wave 1	New (n=302)	New (n=304)	New (n=294)				
Wave 2		Replication (n=302)		New (n=298)			
Wave 3		Replication (n=300)			New (n=299)		
Wave 4					Replication (n=597)	New (n=597)	
Wave 5		Replication (n=294)					New (n=305)
Youth		Replication (n=10,637)					

Data was collected in October 2018 for wave 1, November 2019 for wave 2, April 2020 for wave 3 and wave 4, and January 2021 for wave 5. Participants came from MTurk in waves 1–4 and from Prolific in wave 5. Youth data was collected in October 2020 as part of a partnership with the Character Lab Research Network, as described in Section 5. In all but wave 4, we aimed to recruit 300 participants per study version. In wave 4, to generate more data from the *Private (Immediately Informed)* version, we aimed to recruit 600 participants per study version. Realized sample size for each study version appear in each cell.

2.1 The *Self-Promotion* Version

The *Self-Promotion* version of our study proceeds as follows: participants complete a math and science test, provide their beliefs about their absolute performance on that test, provide responses to self-evaluation questions about their test performance, are informed of their absolute and relative test performance, provide informed responses to self-evaluation questions about their test performance, and then answer questions that provide control and demographic information, including gender. More specifically, the *Self-Promotion* version has four parts, described in sequence below. See Appendix D.1 for screenshots and additional details.

Part 1: Performance and Performance Beliefs

In part 1 of the study, participants are asked to take a test comprised of 20 multiple choice questions from the Armed Services Vocational Aptitude Battery (ASVAB). They have up to 30 seconds to answer each question. There are four questions each from the following five categories: General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. Participants are informed that "In addition to being used by the military to determine

which jobs armed service members are qualified for, performance on the ASVAB is often used as a measure of cognitive ability by academic researchers." If part 1 is randomly selected for payment, a participant's bonus payment is equal to 5 cents times the number of ASVAB questions answered correctly.

As a measure of beliefs about their absolute performance, after participants complete the 20 ASVAB questions, they are asked: "Out of the 20 questions on the test you took in part 1, how many questions do you think you answered correctly?" This question is not incentivized, and participants can select any number from 0 to 20.[7]

Part 2: Self-Promotion

In part 2 of the study, participants are asked five questions about their performance on the test. Participants are told that if part 2 is randomly selected for payment, one of the responses to one of the questions will be shared with another study participant called their "employer." The employer will see the response to the randomly selected question—and only that response to that question (i.e., not any of the other responses or any information about actual performance)—and will determine whether to hire them and how much to pay them if hired.

If an employer chooses not to hire a participant, the participant will earn a bonus of 25 cents, and the employer will earn a bonus of 100 cents. If an employer chooses to hire a participant, the employer will choose a wage between 25 and 100 cents, which will be the bonus for the participant. The employer's bonus payment will then equal: 100 cents minus the wage paid to the participant plus 5 cents times the number of questions the participant answered correctly on the test.[8]

To encourage participants to reflect on their performance, the first question in part 2 is a free-response question that states: "Please describe how well you think you performed on the test that you took in part 1 and why." The remaining four are the quantitative self-evaluation questions that we analyze for the remainder of the paper.[9]

The first two self-evaluation questions focus solely on participants' past performance on the test. First, we ask participants to indicate how well they think they performed on the test by selecting an adjective from a six-point Likert scale ranging from "terrible" to "exceptional." We refer to this as the *performance-bucket* question. We then elicit a more continuous response, asking participants to indicate the extent to which they agree, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I performed well on the test I took in part 1." We refer to this response as the *performance* question.

[7] A main reason to provide incentives for such belief elicitation is to mitigate noise in responses. As discussed in what follows, however, we do not use responses to this question to control for beliefs statistically. Instead, we control for beliefs about absolute and relative performance by design, mitigating concerns about noise in this measure.

[8] The employer earnings are based on the number of correct answers that the participant completed previously. Participants do not have to complete additional tasks and the decision environment avoids any potential uncertainty that might arise about future performance.

[9] One could also imagine analyzing responses to the free-response question. Analyzing responses to this question is fraught, however, as the text is hard to evaluate and can convey additional information that makes measuring the "positivity" of the response difficult. Nevertheless, we attempt to learn what we can from this data by having 399 MTurk participants evaluate the free responses from wave 1, and we summarize those findings in Appendix C.

The latter two self-evaluation questions relate to participants' past performance but also allow room for participants to hold preferences and beliefs about a related, hypothetical job. Using the same 0-to-100 scale described above, participants are asked to indicate the extent to which they agree with the following statements: "I would apply for a job that required me to perform well on the test I took in part 1" and "I would succeed in a job that required me to perform well on the test I took in part 1." We refer to these as the *willingness-to-apply* question and the *success* question, respectively.

The answers to these four self-evaluation questions allow us to quantity—on a 1-to-6 scale for the performance-bucket question and on a 0-to-100 scale for the three other questions—how participants subjectively describe their performance to a potential employer. The answers to these self-evaluation questions serve as as our measures of self-promotion.

Part 3: Informed Self-Promotion

In part 3 of the study, participants are asked precisely the same questions about their performance on the test as in part 2, and participants are told that if part 3 is randomly selected for payment, one of their answers to one of the questions will be shared with their employer.

We refer to their answers on the self-evaluation questions in part 3 as our measure of *informed* self-promotion because, before answering these questions, participants learn precise information about their absolute and relative performance on the test. In particular, participants are told exactly how many of the 20 questions they answered correctly (i.e., their absolute performance) and they are compared to 100 other participants who were asked the same questions and told how many of those participants answered more questions correctly and how many answered fewer questions correctly (i.e., their relative performance). As an attention check, participants must correctly report how many of the 20 questions they answered correctly before proceeding to answer the self-evaluation questions in part 3.

Part 4: Financial-Deservingness Question and Demographics

In part 4, participants are first asked a question that measures perceptions of deservingness for earnings from our experiment: "Out of a maximum amount of 100 cents, what amount of bonus payment, in cents, do you think you deserve for your performance on the test you took in part 1?" If this part is randomly selected for payment, their bonus payment equals whatever amount they indicate from 0 to 100 cents. This question allows us to consider the potential gender difference in how much participants claim that they deserve to earn, elicited with a 1-to-1 correspondence with financial payoffs.[10] We then collect demographic information on participants, including gender.

2.2 The *Self-Promotion (Risky)* Version

To explore the robustness of the gender gap in self-promotion, we ran the *Self-Promotion (Risky)* version. The *Self-Promotion (Risky)* version proceeds exactly as the *Self-Promotion* version except

[10]Since this measure of deservingness occurs *after* participants respond to self-evaluation questions, and since it may theoretically be influenced by these responses, it is not an appropriate "control" variable in regressions with evaluations as the dependent variable. Rather, we consider it as an alternative dependent variable in Section 4.3.

that participants are told that there is *some chance* that their employers will learn their actual performance (i.e., be informed of how many questions they answered correctly on the test) along with one of their answers to a self-evaluation question.[11] See Appendix D.2 for screenshots and additional details.

If participants expect that employers may learn their actual performance, the *Self-Promotion (Risky)* version could cause workers to feel constrained to provide answers that are more likely to be viewed as appropriate by their employers. More generally, the *Self-Promotion (Risky)* version helps us to show robustness to a labor-market setting where individuals are aware that signals about true performance may be available to employers.

2.3 The *Private* Version

The *Private* version proceeds exactly as the *Self-Promotion* version except that participants provide their answers to part 2 and part 3 self-evaluation questions in a non-strategic, non-incentivized setting. In particular, there is no mention of any employer, and participants are told that if part 2 or part 3 is randomly selected for payment, their bonus will equal 25 cents regardless of how they answer the self-evaluation questions. See Appendix D.3 for screenshots and additional details.

Given the lack of employers, the *Private* version eliminates the relevance of strategic incentives to provide more favorable responses to self-evaluation questions in order to achieve higher financial returns. Put differently, it eliminates the incentives to promote that were present in the *Self-Promotion* version. Eliminating employers also eliminates other potential drivers of the gender gap. In the *Private* version, gender differences in response to self-evaluation questions cannot be driven by potential gender differences in risk aversion, gender differences arising from lack of control over payoffs, or gender differences in preferences towards employers.

Consequently, while the *Self-Promotion* version and *Self-Promotion (Risky)* version investigate whether there is a gender gap in self-promotion, the *Private* version investigates whether there is an underlying gender gap in how individuals subjectively evaluate their performance.

2.4 The *Private (Social Norms)* Version

The *Private (Social Norms)* proceeds exactly as the *Private* version except that participants are provided with additional information when providing responses in part 3 (i.e., after they receive performance information). In particular, each of the four self-evaluation questions now includes a message that reads: "Also note that, among participants in a prior study who scored the same as you on the test, the average answer to this question was: [insert relevant average answer]." See Appendix D.4 for screenshots and additional details.

This additional information in the *Private (Social Norms)* version may mitigate gender differences in beliefs about what responses to self-evaluation questions are typical or appropriate.

[11] This chance is left ambiguous in the experimental instructions. In practice, there was a 1% chance we would run a version in which employers received this additional information. This resulted in us not running such a version.

2.5 The *Private (Immediately Informed)* Version

The *Private (Immediately Informed)* version proceeds exactly as the *Private* version except that participants are immediately informed of their absolute and relative performance and then respond to the self-evaluation questions. That is, this study never asks participants to respond to self-evaluation questions before they are informed of their absolute and relative performance. See Appendix D.5 for screenshots and additional details.

By only asking self-evaluation questions when participants are informed, the *Private (Immediately Informed)* version eliminates the potential role of consistency motives or anchoring effects that could arise from first asking self-evaluation questions when participants are not informed of their performance and then asking self-evaluation questions when participants are informed of their performance.[12]

2.6 The *Private (Other-Evaluation)* Version

The *Private (Other-Evaluation)* version builds off of the *Private (Immediately Informed)* version but asks participants to answer evaluation questions about others rather than themselves. More specifically, the *Private (Other-Evaluation)* version proceeds exactly as the *Private (Immediately Informed)* version except that participants are informed of the absolute and relative performance of another MTurk worker and asked to evaluate the performance of that other MTurk worker.

Unbeknownst to participants, they are asked about an MTurk worker with the same test score as them. That is, a participant who answers X out of 20 questions correctly on the test is asked to provide informed evaluations about another participant who also answered X out of 20 questions correctly on the test (without being told that X out of 20 is also their score). See Appendix D.6 for screenshots and additional details.

Examining whether a gender gap persists in the *Private (Other-Evaluation)* speaks to whether there is a gender difference in standards or in evaluations of performance generally, or, instead, whether the gender difference in evaluations is specific to own performance.

2.7 The *Private (Verbal)* Version

The *Private (Verbal)* version proceeds exactly as the *Private* version except that participants complete a test that assesses their verbal skills rather than math and science skills. See Appendix D.7 for screenshots and additional details.

Given that verbal skills are typically considered "female-typed," the *Private (Verbal)* version allows us to explore responses to self-evaluation questions in a more "female-typed" setting. In addition, in the follow-up survey to this version (and the *Private* version we run in the same wave), we ask participants additional questions in part 4 of the study that we describe and analyze in Section 4.4.

[12] Of course, different types of consistency—such as how individuals view their performance both before and after they take the test (Beyer, 1990), or even how individuals view their performance in tasks like these before they complete the study—could still be relevant.

2.8 Our Study Environment

In this section, we present data on performance on the math and science test and on the beliefs that participants report about their absolute performance (i.e., how many questions they think they answered correctly on the test). These analyses are intended to provide a sense of our study environment. In particular, we find results consistent with our setting being "male-typed" in that women think they answered significantly fewer questions correctly than equally performing men.[13]

Panel A of Figure 1 shows CDFs of the number of test questions answered correctly by male participants and by female participants. On average, women answer 9.94 questions correctly and men answer 9.34 questions correctly. The mean difference is statistically significant ($p < 0.01$) and the distributions are statistically significantly different (a Kolmogorov–Smirnov test yields $p < 0.01$).

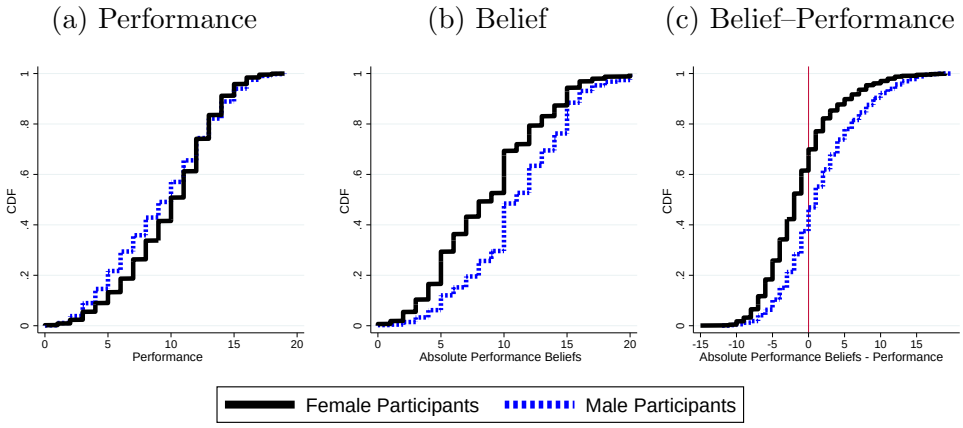
Despite women performing better than men, Panels B and C of Figure 1 show that women believe they performed worse on the test than men. Panel B shows raw beliefs about performance. On average, men believe they answered 11.05 questions correctly while women believe they answered only 8.77 questions correctly. The mean difference is statistically significant ($p < 0.01$), and the distributions are statistically significantly different (a Kolmogorov–Smirnov test yields $p < 0.01$). Panel C shows the difference between beliefs about performance and actual performance. Again, the mean difference is statistically significant ($p < 0.01$), and the distributions are statistically significantly different (a Kolmogorov–Smirnov test yields $p < 0.01$). Looking at where the CDFs cross 0, we see that the gender gap in beliefs about performance is driven both by the majority of women underestimating their performance and the majority of men overestimating their performance.

Appendix Table A.1 presents the corresponding regression results. Column 1 shows that women outperform men on the test (the coefficient on *Female* is positive and statistically significant), and the remaining columns confirm the statistically significant gender gaps in beliefs about performance, including when considering the raw data only (Column 2), when controlling for performance with dummies for each possible test score (Column 3), and when the outcome variable directly captures the difference between beliefs about performance and actual performance (Column 4). In the latter three columns, the coefficient on *Female* is negative, large, and statistically significant.

These results highlight that women believe they answered fewer questions correctly than equally performing men. We will consider the role of such beliefs in the self-promotion that we observe. As discussed in Sections 2.1 and 3.2, however, rather than using these reported beliefs as statistical controls, we will instead control for beliefs by design. We will explore how participants answer self-evaluation questions *after they are perfectly informed* of exactly how many questions they answered correctly (i.e., their absolute performance) and how their score compared to other participants (i.e., their relative performance).

[13] In this section, we pool across all study versions from waves 1–5 in which participants take the math and science test (i.e., all versions except the *Private (Verbal)* version) because participants always take the test and report beliefs about their absolute performance before the study versions vary by treatment. However, we note that the gender gap in beliefs about absolute performance that we report on below persists for each individual study version.

Figure 1: Performance and Absolute Performance Belief Distributions



Graphs show CDFs for the associated outcome. *Performance* is the number of questions a participant answered correctly out of the 20 questions on the test. *Belief* is the number of questions a participant believes he or she answered correctly. *Belief–Performance* is the difference between these two variables, calculated for each participant. Data are from all study versions from waves 1–5 involving the math and science test (i.e., all but the *Private (Verbal)* version).

3 Results

In this section, we report on our experimental results. Section 3.1 documents a large gender gap in self-promotion among equally performing men and women. In Sections 3.2–3.7, we report on our various study versions run across waves 1–5 to narrow in on the underlying sources of this gender gap.

3.1 The Gender Gap in Self-Promotion

The *Self-Promotion* version of the experiment allows us to assess whether there is a gender gap in self-promotion—that is, whether there is a gender gap in how participants subjectively answer self-evaluation questions when they know one of their answers will be shared with employers.

While our main analysis will use regressions to compare the self-promotion of equally performing men and women, Figure 2 shows raw responses to the four questions from the self-evaluation in part 2 of the *Self-Promotion* version. These responses are provided before participants learn their absolute and relative performance on the test. Women provide significantly lower responses to each question and hence engage in statistically significantly lower self-promotion ($p < 0.01$ for each corresponding t-test and for each Kolmogorov–Smirnov test).

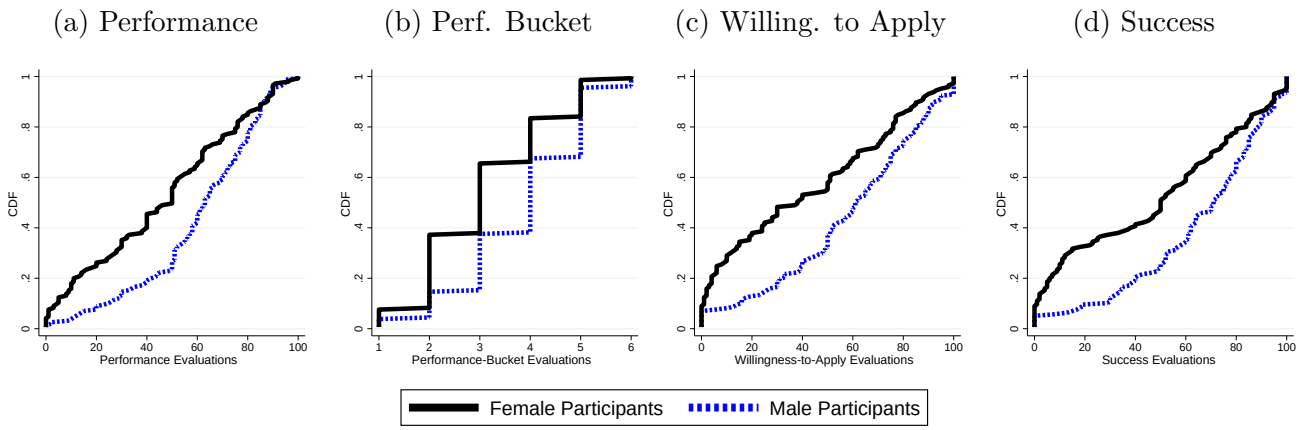
Panel 1 of Table 2 confirms the statistical significance of these gender gaps in self-promotion when controlling for performance with fixed effects for each possible test score (0 to 20) to allow us to compare equally performing men and women. The coefficient on *Female* is negative, large, and statistically significant for all four questions.

Column 1 presents results from the *performance* question that asks participants to respond to the statement "I performed well on the test I took in part 1" on a scale from 0 (entirely disagree) to 100 (entirely agree). The average responses provided by women are 12.68 points lower, which

represents a 24% decrease relative to the mean. Column 2 presents results from the *performance-bucket* question that asks participants to "Please indicate how well you think you performed on the test you took in part 1" on the six-point Likert scale question. The average responses provided by women are 0.59 points lower, which represents a 17% decrease relative to the mean. Columns 3 and 4 present results from the more "context rich" questions that may relate to participants' preferences and beliefs about a related, hypothetical job. Column 3 presents results from the *willingness-to-apply* question that asks participants to respond to the statement "I would apply for a job that required me to perform well on the test I took in part 1" on a scale from 0 (entirely disagree) to 100 (entirely agree). The average responses provided by women are 15.31 points lower, which represents a 31% decrease relative to the mean. Column 4 presents results from the *success* question that asks participants to respond to the statement "I would succeed in a job that required me to perform well on the test I took in part 1" on a scale from 0 (entirely disagree) to 100 (entirely agree). The average responses provided by women are 15.08 points lower, which represents a 27% decrease relative to the mean. Thus, across all four questions, there is a substantial and statistically significant gender gap in self-promotion among equally performing men and women.

The remaining panels in Table 2 and Table 3 present corresponding results from our other study versions. These additional results, which we will detail in the following subsections, allow us to narrow in on the underlying sources of the gender gap.

Figure 2: In the *Self-Promotion* version, CDFs showing the Gender Gap in Self-Promotion



Graphs show CDFs of responses to the question noted in each panel, as defined in the notes of Table 2, elicited before performance information is provided. Data are from the *Self-Promotion* version.

Table 2: Results from Evaluations (before performance information is provided)

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Self-Promotion Version, Wave 1 (N=302)				
Female	-12.68***	-0.59***	-15.31***	-15.09***
	(2.96)	(0.13)	(3.46)	(3.46)
Panel 2: Self-Promotion (Risky) Version, Wave 1 (N=294)				
Female	-9.15***	-0.47***	-12.82***	-9.24***
	(2.93)	(0.13)	(3.29)	(3.32)
Panel 3: Private Version, Wave 1 (N=304)				
Female	-13.46***	-0.56***	-17.57***	-16.46***
	(2.93)	(0.13)	(3.51)	(3.61)
Panel 4: Private Version, Wave 2 (N=302)				
Female	-12.21***	-0.55***	-17.25***	-14.39***
	(3.18)	(0.15)	(3.54)	(3.53)
Panel 5: Private (Social Norms) Version, Wave 2 (N=298)				
Female	-15.14***	-0.80***	-16.93***	-15.62***
	(3.28)	(0.16)	(3.71)	(3.71)
Panel 6: Private Version, Wave 3 (N=300)				
Female	-16.45***	-0.79***	-15.69***	-16.16***
	(3.18)	(0.15)	(3.92)	(3.87)
Panel 7: Private (Immediately Informed) Version, Wave 3: no evaluations				
Panel 8: Private (Immediately Informed) Version, Wave 4: no evaluations				
Panel 9: Private (Other-Evaluation) Version, Wave 4: no evaluations				
Panel 10: Private Version, Wave 5 (N=294)				
Female	-13.05***	-0.59***	-18.77***	-19.18***
	(2.61)	(0.11)	(3.30)	(3.17)
Panel 11: Private (Verbal) Version, Wave 5 (N=305)				
Female	1.15	-0.12	1.99	-0.36
	(2.40)	(0.11)	(3.19)	(3.02)
Panel 12: All Evaluations of Own Math and Science Performance (N=2094)				
Female	-13.83***	-0.67***	-17.28***	-16.12***
	(1.13)	(0.05)	(1.31)	(1.32)
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column before participants are informed of their absolute and relative performance. Responses to the *Performance* question indicate the extent of each participant's agreement (from 0–100) with the following statement: "I performed well on the test I took in part 1." Responses to the *Performance-Bucket* question indicate which Likert-scale response (coded from 1 for the lowest to 6 for the highest) a participant selects when asked to "indicate how well you think you performed on the test in part 1." Responses to the *Willingness-to-Apply* question indicates the extent of each participant's agreement (from 0–100) with the following statement: "I would apply for a job that required me to perform well on the test I took in part 1." Responses to the *Success* question indicates the extent of each participant's agreement (from 0–100) with the following statement: "I would succeed in a job that required me to perform well on the test I took in part 1." *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data in each panel are from the noted study version(s).

Table 3: Results from Informed Evaluations (after performance information is provided)

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
Panel 1: Self-Promotion Version, Wave 1 (N=302)				
Female	-7.01**	-0.40***	-10.73***	-11.73***
	(2.90)	(0.13)	(3.40)	(3.30)
Panel 2: Self-Promotion (Risky) Version, Wave 1 (N=294)				
Female	-7.24**	-0.36***	-9.11***	-8.07**
	(2.83)	(0.14)	(3.38)	(3.29)
Panel 3: Private Version, Wave 1 (N=304)				
Female	-8.01***	-0.33**	-13.25***	-13.15***
	(2.88)	(0.14)	(3.53)	(3.53)
Panel 4: Private Version, Wave 2 (N=302)				
Female	-7.58**	-0.42***	-14.15***	-14.37***
	(3.18)	(0.15)	(3.53)	(3.46)
Panel 5: Private (Social Norms) Version, Wave 2 (N=298)				
Female	-11.93***	-0.62***	-16.39***	-15.77***
	(3.15)	(0.16)	(3.42)	(3.58)
Panel 6: Private Version, Wave 3 (N=300)				
Female	-12.70***	-0.52***	-16.55***	-15.87***
	(3.04)	(0.14)	(3.73)	(3.76)
Panel 7: Private (Immediately Informed) Version, Wave 3 (N=299)				
Female	-7.61**	-0.47***	-11.42***	-12.48***
	(3.35)	(0.16)	(3.81)	(3.61)
Panel 8: Private (Immediately Informed) Version, Wave 4 (N=597)				
Female	-8.54***	-0.42***	-16.63***	-18.66***
	(2.22)	(0.10)	(2.42)	(2.30)
Panel 9: Private (Other-Evaluation) Version, Wave 4 (N=597)				
Female	0.29	-0.11	-3.54**	-3.17*
	(1.58)	(0.08)	(1.69)	(1.68)
Panel 10: Private Version, Wave 5 (N=294)				
Female	-7.74***	-0.24**	-12.91***	-14.24***
	(2.26)	(0.10)	(3.09)	(3.01)
Panel 11: Private (Verbal) Version, Wave 5 (N=305)				
Female	-0.93	-0.05	-1.34	-1.36
	(1.94)	(0.09)	(2.76)	(2.61)
Panel 12: All Evaluations of Own Math and Science Performance (N=2990)				
Female	-9.83***	-0.47***	-15.12***	-15.59***
	(0.94)	(0.04)	(1.08)	(1.07)
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2, after participants are informed of their absolute and relative performance (or the other participant's absolute and relative performance in Panel 9). *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data in each panel are from the noted study version(s).

3.2 The gap when informed about performance

Individuals are frequently asked to describe their performance—including in response to explicit self-evaluation questions—when they do not know how well they performed in absolute or relative terms. That we document a gender gap in self-promotion when participants are uncertain about their absolute and relative performance (i.e., the gap shown in Section 3.1) is thus important for considering the role of self-promotion in driving gender gaps educational and labor market outcomes.

In light of the large literature on gender gaps in beliefs about performance and how they contribute to gender gaps in behavior (Lundeberg, Fox and Punćcohaŕ, 1994; Niederle and Vesterlund, 2007; Coffman, 2014; Niederle, 2016; Apicella, Demiral and Mollerstrom, 2017; Bordalo et al., 2019; Born, Ranehill and Sandberg, 2018; Isaksson, 2018; Coffman, Collis and Kulkarni, 2020, 2019)—and since we indeed observe a gender gap in beliefs about absolute performance as detailed in Section 2.8—one may also wonder if the gender gap in self-promotion reflects women thinking they had a lower performance (in absolute or relative terms) than equally performing men.[14] Motivated by this question, we turn to the results from the self-evaluation questions in part 3 of the *Self-Promotion* version. Since these questions are asked after participants are informed of their absolute and relative performance on the test—and thus after we close any gender gap in beliefs about absolute and relative performance "by design," given the provision of this information—we refer to the corresponding results as our measures of *informed* self-promotion. In this way, we differ from much of the aforementioned literature because we are not interested in the beliefs participants hold about their absolute and relative performance; rather, we are interested in whether a gender gap in self-promotion persists even when we close any gender gap in these beliefs about absolute and relative performance.

As shown in Panel 1 of Table 3, which presents results from responses after participants have learned their absolute and relative performance, we observe substantial and statistically significant gender gaps in *informed* self-promotion. When considering the questions asked on the 0–100 scale, the gender gap in informed self-promotion is 7.01 percentage points for the performance question, 10.73 for the willingness-to-apply question, and 11.73 for the success question. When considering the question asked on the 1–6 scale, the gender gap in informed self-promotion is 0.40.

The gender gap in informed self-promotion makes clear that the gender gap in self-promotion is not just a result of women thinking they had a lower performance—either in terms of absolute or relative performance—than men. The gender gap in self-promotion also arises when participants are perfectly informed of their absolute and relative performance on the task (i.e., closing any gender gap in beliefs about absolute and relative performance on the task).[15]

That said, it is interesting to note that the gender gap in informed self-promotion appears smaller than the gender gap in self-promotion measured before participants are informed of their

[14] For work on biased beliefs more generally, see also Ertac (2011); Mobius et al. (2011); Buser, Gerhards and Van der Weele (2018); Coutts (2018).

[15] We fix beliefs about absolute and relative performance, but other beliefs could theoretically be relevant (e.g., in the willingness-to-apply and success questions, beliefs about the referenced hypothetical job may be relevant).

absolute and relative performance. While Panel 1 of Appendix Table A.2 shows that the difference between these gaps is not statistically significant, Panel 2—which pools across all study versions in which participants provide evaluations of their math and science performance before and after being informed—shows that information significantly reduces the gender gap by anywhere from 10–31%. Nevertheless, as shown in Table 3, in all of these study versions—and in a version that immediately informs participants of their absolute and relative performance, as further discussed in Section 3.5—the gender gap remains large and statistically significant even when participants are fully informed of their absolute and relative performance.

3.3 The gap when promotion incentives are removed

In Section 3.1 and Section 3.2, we document a gender gap in self-promotion and a gender gap in informed self-promotion when participants know that a potential employer, who is deciding whether to hire them and how much to pay them, will learn how they answer a self-evaluation question. In Section 4.3, we show the robustness of the gender gap in self-promotion and informed self-promotion to a different promotion incentive and discuss the value of future work considering other promotion incentives.

An important advantage of our context, however, is that we can remove any incentives to promote from the decision environment. This allows us to explore whether the gender gap in self-promotion is reflective of an underlying gender gap in how individuals subjectively evaluate their own performance (e.g., how they view or describe their performance) even in the absence of incentives to promote.

To examine whether the gender gap in self-promotion is reflective of an underlying gender gap absent promotion incentives, we turn to the *Private* version. In this version, there are no employers. Participants receive a fixed payment regardless of how they complete their self-evaluations, thus eliminating any promotion incentives. More broadly, this version allows us to ensure that any remaining underlying gap is not driven by gender differences in strategic incentives (Reuben, Sapienza and Zingales, 2014; Charness, Rustichini and Van de Ven, 2018; Soldà et al., 2019; Schwardmann and van der Weele, 2019), in risk aversion over payoffs (Dwyer, Gilkeson and List, 2002; Eckel and Grossman, 2008; Croson and Gneezy, 2009), in lack of control over payoffs (Cobb-Clark, 2015; Apicella, Demiral and Mollerstrom, 2020), or in preferences relating to the payoffs of others (Andreoni and Vesterlund, 2001; Croson and Gneezy, 2009; DellaVigna et al., 2013).

As shown in Panel 3 of Table 2, we observe substantial and statistically significant gender gaps across all four self-evaluation questions when incentives to promote are removed. As shown in Panel 3 of Table 3, we also observe substantial and statistically significant gender gaps when incentives to promote are removed *and* participants are fully informed of their absolute and relative performance.

One may speculate that removing promotion incentives had little impact on the gender gap in our setting because participants did not value or attend to the incentives to promote in the first place. This proves not to be the case. As shown in Appendix Table A.3, which compares responses to self-evaluations in the *Private* and *Self-Promotion* versions run in the same wave—

so participants were randomly assigned to these versions—both men and women respond to the promotion incentives by providing more favorable answers to the self-evaluation questions in the *Self-Promotion* version. However, the extent to which men and women respond to these promotion incentives is very similar, and thus a similar gender gap emerges in the *Private* and *Self-Promotion* versions.

3.4 The gap when informed of typical responses

That the gender gap persists—even when promotion incentives are removed and when participants are fully informed of their absolute and relative performance—makes clear that there is an underlying gender gap in how men and women subjectively evaluate their performances. Motivated by the possibility that this underlying gap relates to views about what is typical or socially appropriate, we next explore whether the underlying gap persists even when participants are informed about the average response to each self-evaluation question by others who had the same performance as them.

In our second wave of data collection, we replicate the gender gap in the *Private* version—both when participants are not informed about their performance (see Panel 4 of Table 2) and when participants are informed about their absolute and relative performance (see Panel 4 of Table 3). We also show that the gender gap arises in the *Private (Social Norms)* version, both when participants are not informed about their performance (see Panel 5 of Table 2, which is essentially another replication of the *Private* version, since subjects have not yet received additional information) and when participants are informed about their absolute and relative performance *as well as* the average answers to self-evaluation questions provided by others who had the same performance as them (see Panel 5 of Table 3). The gender gap in informed self-promotion is just as large in the *Private (Social Norms)* version as in the *Private* version.

3.5 The gap when immediately informed about their performance

In the study versions discussed above, consistency motives or anchoring effects could influence the persistence of the gender gaps we observe after participants are informed of their absolute and relative performance. This is because participants first answer self-evaluation questions in part 2 before they are informed of their performance and then answer self-evaluation questions again in part 3 after they are informed of their performance. Such consistency motives and anchoring effects are important to investigate, because subjective views about one's performance are often formed before individuals receive feedback about their performance.

In our third wave of data collection, we again replicate the gender gap in the *Private* version—both when participants are not informed about their performance (see Panel 6 of Table 2) and when participants are informed about their absolute and relative performance (see Panel 6 of Table 3). We also show that the gender gap arises in the *Private (Immediately Informed)* version when participants are immediately informed about their absolute and relative performance and then asked self-evaluation questions (see Panel 7 of Table 3). That is, even when participants are not asked

self-evaluation questions before being informed of their performance, we still observe a gender gap after they are informed of their performance.

3.6 The gap when individuals evaluate others

Motivated by the possibility that gender norms contribute to women subjectively viewing their own performance more negatively, and by prior work on how women are better advocates for others than themselves in negotiations (Bowles, Babcock and McGinn, 2005), we investigate whether gender gaps also emerge when participants are asked to evaluate the performance of others rather than themselves.

In our fourth wave of data collection, we replicate the gender gap in the *Private (Immediately Informed)* version when participants are informed about their absolute and relative performance (see Panel 8 of Table 3). However, we find small, often statistically insignificant, gender gaps in the *Private (Other-Evaluation)* version when participants are informed about another participant's absolute and relative performance and then asked the four evaluation questions about that other participant's performance (see Panel 9 of Table 3).

3.7 The gap when asked about a more female-typed task

Given the gender gaps in pay and in occupational and industry representation that motivate our study, the main task that we study is a "male-typed" task. In particular, we ask individuals to provide self-evaluations relating to their math and science skills. Inspired by prior work on gender stereotypes and how gender differences relate to the involved task (Coffman, 2014; Dreber, von Essen and Ranehill, 2014; Bordalo et al., 2019; Coffman, Collis and Kulkarni, 2019; Coffman, Flikkema and Shurchkov, 2019; Atwater and Saygin, 2020), one may expect that the gender gap we observe in the male-typed task might be mitigated, or even reversed, when we consider a more "female-typed" task.

In our fifth wave of data collection, we again replicate the gender gap in the *Private* version—both when participants are not informed about their performance on the math and science test (see Panel 10 of Table 2) and when participants are informed about their absolute and relative performance on the math and science test (see Panel 10 of Table 3). When considering data from the *Private (Verbal)* version, however, we find no statistically significant gender gaps, either when participants are not informed about their performance on the verbal test (see Panel 11 of Table 2) or when participants are informed about their performance on the verbal test (see Panel 11 of Table 3).

Together with the evidence in Section 3.6, these findings point towards the role of gender norms and culture in shaping views of one's performance, which may help to explain why the gender gap in subjective evaluations of own performance in math and science proves so persistent. These two sets of findings also make clear that the gender gap is not driven by women subjectively evaluating performance differently than men *in general* (e.g., having different "standards" in general), since it does not persist when they are asked about their own performance relating to verbal

skills or when they are asked about someone else's performance on the math and science test.

4 Discussion

In this section, we present additional analysis of the data collected in waves 1–5. Section 4.1 documents the robustness of our results. Section 4.2 presents several heterogeneity analyses to further explore the gender gap and to identify avenues for future work. Section 4.3 explores the robustness of the gender gap in self-promotion to different promotion incentives. Section 4.4 explores whether individuals anticipate the gender gap in self promotion; we find that they do not.

4.1 The robustness of the gender gap

How robust is the gender gap in subjective evaluations of performance? Consider all of the versions and replications that we ran with the math and science task. Separately considering each self-evaluation question, whether or not participants are informed, each study version, and each wave, we have 64 possible settings to look for a gender gap. Table 2 (Panels 1–6 and 10) and Table 3 (Panels 1–8 and 10) report these 64 tests. We find a statistically significant gender gap 64 out of 64 times.

Panel 12 of Table 2 analyzes data from the 28 tests when participants are not informed about their performance. The gender gap is 13.83 for the performance question, 0.67 for the performance-bucket question, 17.28 for the willingness-to-apply question, and 16.12 for the success question. Panel 12 of Table 3 analyzes data from the 36 tests when participants are informed about their performance. The gender gap is 9.83 for the performance question, 0.47 for performance-bucket question, 15.12 for the willingness-to-apply question, and 15.59 for the success question.

Appendix Tables A.4–A.7 provide further robustness tests of this pooled data. Appendix Table A.4 shows that the gender gaps are robust to excluding performance controls; indeed the gaps get directionally larger because women outperform men on the task. Appendix Table A.5 shows that the gender gaps remains statistically significant when controlling for other demographic information we collected from participants along with gender. Appendix Table A.6 shows that the results are robust to excluding "inattentive" participants who answered no better than chance on the test (i.e., answered five or fewer questions correctly out of 20). Appendix Table A.7 shows that gender gaps are statistically significant when considering quantile regressions estimated at the 25th, 50th, and 75th percentiles, highlighting that the gender gaps are not driven by a subset of "extreme" answers.[16]

4.2 Heterogeneity Analyses

The set of results presented in Section 3 highlight that we observe gender gaps in subjective evaluations whenever participants evaluate their performance on a male-typed task. In this section, we further explore the data from these study versions to identify heterogeneity in the gender gap

[16]These quantile regressions are only presented for the questions elicited on the 0-to-100 scale. We do not run these for the performance-bucket question elicited on six-point scale to avoid convergence issues given the discrete nature of this question and the inclusion of performance fixed effects.

and to explore what other factors correlate with subjective evaluations of performance. In doing so, we aim to better understand the underlying gender gap and to highlight possible avenues for future work.

Heterogeneity by performance level

To explore how the gender gap varies by performance, the specifications in Appendix Tables A.8 and A.9 drop the performance fixed effects and instead control linearly for performance and for an interaction of performance and gender. (Recall that our main regression specifications include fixed effects for each performance level—so that we can compare equally performing men and women—and do not assume linearity between performance and answers to self-evaluation questions.) Appendix Table A.9 additionally excludes the lowest performers who may be inattentive (following the criteria for Appendix Table A.6) and estimates coefficients on *Performance* that are uniformly positive—those who perform better provide more favorable subjective evaluations of performance.

At the average performance level, the gender gaps in Appendix Tables A.8 and A.9 are large and statistically significant. In addition, the gap is estimated to be larger at low performance levels and to be smaller at high performance levels. (Recall, however, that very high test performances are rare; Figure 1 shows that 75% of participants perform no more than 3 questions above average and 90% of participants perform no more than 5 questions above average.) Since we observe a smaller gender gap for the higher performers, future work might explore the relationship between performance and such gender gaps. It is worth noting, however, that we see the opposite pattern in our data on youth, presented in Section 5. In that data, the highest performers display the largest gaps.

Heterogeneity by beliefs about absolute performance

To explore how the gender gap varies by beliefs about absolute performance, the specifications in Appendix Table A.10 add a linear control for participants' beliefs about their absolute performance on the test. The results in Panel 1 show that—holding performance (i.e., the number of questions they answered correctly) constant—a more optimistic belief about their absolute performance (i.e., the number of questions they believe they answered correctly) is associated with more favorable subjective evaluations. This relationship is even stronger for women, suggesting that the gender gap is larger among those who were more pessimistic about their absolute performance and smaller among those who were more optimistic about it. We see similar results in Appendix Table A.11, which replaces the linear belief control with a linear control for the gap between a participant's belief and their actual performance.

Intriguingly, the results in Panel 2 of Appendix Tables A.10 and A.11 show that beliefs about absolute performance are *still* correlated with subjective evaluations of performance even after participants have been informed about their absolute and relative performance on the test. That is, individuals who initially thought they answered fewer questions correctly on the test *still* evaluate their performance less favorably *even after they learn how many questions they answered correctly*

on the test.[17] Why could this be? One explanation is that there are certain types of individuals who view their performance in math and science more positively than others or view their performance more negatively than others. Such positive types could subjectively evaluate their performance more positively and also overestimate their absolute performance. Such negative types could subjectively evaluate their performance more negatively and also underestimate their absolute performance. Because such a type is not caused by the belief about absolute performance (indeed the type could cause the belief), the subjective evaluations continue to be influenced by the type, even after individuals are perfectly informed of their absolute (and relative) performance.

The presence of types like those posited in the prior paragraph highlights a methodological advantage of controlling for beliefs "by design," by providing participants with precise information to fix their beliefs. When a research question asks what role beliefs play in driving some outcome (i.e., rather than how beliefs update in response to information), it may be preferable to control for beliefs "by design" than to measure beliefs and control for them statistically. Controlling for beliefs by design allows a researcher to avoid potential confounds related to measurement error, omitted variable bias (which could be caused by the positive and negative types discussed above), and reverse causality. Indeed, as shown in Appendix Table A.12, if we had controlled for beliefs statistically—by adding a fixed effect for each belief when analyzing the data of uninformed participants—we would have inferred that beliefs account for much more of the gender gap (compare Panel 1 to Panel 3) than we concluded when we controlled for beliefs by design (compare Panel 1 to Panel 2).

Heterogeneity by broader beliefs

To further explore the possibility that certain types of individuals systematically view their math and science performance less favorably, we added two questions to the follow-up survey in our fifth wave of data collection to measure broader beliefs about performance.

One question asked participants to indicate their agreement (on a seven-point Likert scale from "strongly disagree" to "strongly agree") with a statement that reads "In general, I perform well when asked questions that test my math and science skills." As shown in Appendix Table A.13, answers to this question are highly and positively predictive of subjective evaluations that relate to math and science skills in the *Private* version (and equally so for men and women).

The other question asks participants to indicate their agreement (on the same scale) with a statement that reads "In general, I perform well when asked questions that test my verbal skills." As shown in Appendix Table A.14, answers to this question are also highly and positively predictive of subjective evaluations that relate to verbal skills in the *Private (Verbal)* version (and, again, equally so for both men and women).[18]

[17] Panel 3 of Appendix Tables A.10 and A.11 show that beliefs about absolute performance are just as correlated with subjective evaluations of performance when we only look in the *Private (Immediately Informed)* version, emphasizing that this result is not driven by consistency motives related to providing evaluations before receiving information on performance.

[18] If we simultaneously include both performance beliefs and these broader beliefs in a regression, both measures of beliefs are positive and statistically significant.

These results further suggest the possibility of positive and negative types noted above and is consistent with individuals allowing their general perception of their math and science skills (or their verbal skills) to influence their perceptions of their specific performance on the math and science test (or verbal test) they take in our experiment.

Heterogeneity by demographics

Appendix Table A.5 controlled for the other demographics we collected in the experiment and demonstrated that the gender gap remained large and statistically significant. It also showed, however, that some of these other demographic traits predict subjective evaluations, both when participants are not informed and when they are informed of their absolute and relative performance.

Appendix Table A.15 further explores these other demographic characteristics and whether they interact with the size of the gender gap. Two sets of findings follow. First, we find statistically significantly more favorable subjective evaluations among younger participants, more educated participants, and more Republican-leaning participants. Second, gender gaps get statistically significantly larger among participants who are more Republican-leaning. To the extent that gender norms differ between Republicans and Democrats, a larger gender gap among Republicans could be a result of these differing gender norms. Future work may follow-up on these results to investigate the extent to which culture—which likely varies by demographic variables beyond gender—affects subjective views of performance.

4.3 Does a gender gap in self-promotion arise under different promotion incentives?

An important and interesting question for future work relates to the persistence of the gender gap in self-promotion across different promotion incentives—beyond those we explored in our *Self-Promotion* version. We take a first step in this direction by presenting results from the *Self-Promotion (Risky)* version, allowing us to explore the gender gap under slightly different promotion incentives.

Participants in the *Self-Promotion (Risky)* version are told that information about their actual performance could be communicated to employers along with their answer to a self-evaluation question. This could make participants feel more constrained to provide appropriate answers since there is some chance of "being caught" if they seem to inflate their subjective evaluations too much. Nonetheless, Panel 3 of Table 2 and Panel 3 of Table 3 show that the gender gap in self-promotion and the gender gap in informed self-promotion remains substantial and significant. Like the results from Section 3.3, we find that the gender gap is similarly sized in the *Self-Promotion (Risky)* version and the *Private* version, highlighting that men and women respond similarly to the promotion incentives in the *Self-Promotion (Risky)* version.

That men and women do not differentially respond to the incentives to promote in either the *Self-Promotion* version or the *Self-Promotion (Risky)* version is consistent with findings from our part 4 question on deservingness. That question asked subjects to claim an amount of money based

on what they thought they deserved from the study: "Out of a maximum amount of 100 cents, what amount of bonus payment, in cents, do you think you deserve for your performance on the test you took in part 1?" As shown in Appendix Table A.16, when pooling across all versions in which participants are asked about their performance on the math and science test—but do not have an opportunity to influence their payments through self-promotion—there is no evidence for a gender difference in how much money equally performing men and women claim. This finding is consistent with equally performing men and women reacting similarly to incentives to promote.

4.4 Do employers account for the gender gap in self-promotion?

In considering the consequences of the gender gap in self-promotion, an interesting question for future work is whether individuals anticipate this gender gap. If they do anticipate the gap, one might hypothesize smaller economic consequences of the gender gap in self-promotion because employers can account for women providing less favorable subjective evaluations than men. That is, employers who seek to identify the true underlying quality of a worker might infer that a less favorable evaluation provided by a woman is equivalent to a more favorable evaluation provided by a man. Prior work in Reuben, Sapienza and Zingales (2014) provides evidence against the empirical relevance of this possibility. That paper finds that employers do not fully account for the fact that men, more than women, tend to inflate their performance estimates (i.e., the number of problems they will correctly answer on a math task) relative to their actual performance.

To assess whether individuals anticipate the gender gap in self-promotion, we added eight incentivized questions to the end of the studies we ran in wave 5 of data collection (see screenshots in Appendix Figures D.24 and D.25). Each question asked participants to make predictions about the average performance of male or female participants in the *Self-Promotion* version after learning how they answered the self-evaluation questions. Participants were told that one of their predictions would be randomly selected and that they would receive an additional bonus payment of $0.50 if they answered that question correctly.

One of these questions asks participants to predict the average number of questions answered correctly on the test among a set of *female* workers whose average response to the performance question was 48 out of 100. Another of these questions asks predictors to predict the average number of questions answered correctly on the test among a set of *male* workers whose average response to the performance question was 59 out of 100. These averages correspond to the average responses provided by women and men on the performance question in the *Self-Promotion* version, after participants were informed of their performance. Since both men and women answer 10 out of 20 questions correctly on average in that version, the correct prediction for both of these questions is 10.

As shown in Appendix Table A.17, participants do not correctly predict the performance of the workers in the *Self-Promotion* version. For example, the coefficient estimate on the constant in Column 1 shows that participants predict the average performance of men to be around 12. The coefficient estimate on *Predictions about women* shows that they predict the average performance of

women to be 1.5 questions worse than men, despite the actual performance being equal. Participants fail to correct for women providing less favorable evaluations and thus predict a significantly lower average performance for women.

Three additional findings on these predictions are of note. First, the small and statistically insignificant coefficient estimate on *Predictions about women***Female* shows that female participants are no better than male participants at predicting the relative performance of men and women. Second, Columns 2–4 of Appendix Table A.17 show similar results when considering predictions based off of the average answer provided by men and by women in response to the other self-evaluation questions. Third, since the gender difference in self-promotion is even larger when participants are not informed of their performance, it is reasonable to expect that the predicted performance gap would be even greater if the predictors were instead provided with average responses to the self-evaluation questions that were asked before workers were informed of their absolute and relative performance.

That neither male nor female participants anticipate the gender gap in self-promotion suggests that we might not expect employers to correct for the gender gap in self-promotion when making assessments about workers. Future work—both in the laboratory and in the field—should investigate if this applies more broadly in other settings. Particularly important questions for this future work are whether experience helps employers get better at identifying the gender gap and correcting for it, and whether employers display a bias against women that leads them to be less likely to correct for the gap.

5 The Gender Gap Among Youth

One explanation for the gender gap in subjective evaluations of math and science performance relates to the role of gender norms and culture. Indeed, consistent with this possibility, we document a robust gender gap when participants are asked about their *own* performance on a *male-typed* task, but no gap when they are asked about the performance of others or when they are asked about their own performance on a more female-typed task. In Section 4.2, we also document several correlations with other demographic characteristics that may be picking up the role of gender norms or culture on evaluations more generally.

Another way to investigate the role of gender norms and culture, and to explore the robustness of gender gaps more generally, is to examine the age at which gender gaps emerge. Indeed, there is a rich literature—mostly related to gender gaps in competition—that does just this (Gneezy and Rustichini, 2004; Dreber, Von Essen and Ranehill, 2011; Cárdenas et al., 2012; Buser, Niederle and Oosterbeek, 2014; Dreber, von Essen and Ranehill, 2014; Sutter and Glätzle-Rützler, 2015; Khachatryan et al., 2015; Almås et al., 2018). The age at which the gender gap arises also speaks to when interventions to close the gap may need to be targeted.

In this section, we present results from an additional experiment involving 10,637 middle-school and high-school students recruited through the Character Lab Research Network, a network of

schools and researchers that partner to run studies that help "to advance scientific insights that help kids thrive." Our sample is nicely balanced by gender (48% of students are male) and skewed slightly towards middle-school students, giving us particular power at relatively younger ages.

These students completed a *Private* version of our study with four main modifications to accommodate this population and the recruitment process. First, rather than answering a test involving 20 questions of various difficulty levels, the test for youth only involved the 10 easiest questions from our math and science test. Second, in the willingness-to-apply question, rather than asking youth about their willingness to apply to a hypothetical job, we asked them about their willingness to take a class that involved topics like those covered on the test. Third, in the success question, rather than asking youth about their likelihood of success in a hypothetical job, we asked them about their likelihood of success in a hypothetical class that involved topics like those covered on the test. Fourth, when we provided information on performance, we only provided absolute performance information (we did not have prior performance data on youth to provide relative information). See Appendix D.8 for screenshots and additional details.

As seen in Table 4, the gender gap persists across all questions and across all grades. There is some evidence that the gender gap in willingness to take a class is smaller for older students, perhaps because what classes they have left to take in school is already determined. The clear takeaway, however, is that the gender gap in responses to self-evaluation questions is robust to this very different setting and that it appears as early as sixth grade, the youngest set of students that we study.

Following the heterogeneity analysis presented in Section 4.2, Appendix Tables A.18–A.22 present parallel results exploring heterogeneity based on performance, beliefs about absolute performance, and other demographics. In addition, given the population of this study, Appendix Table A.23 explores heterogeneity by grade point average in school.

On performance, Appendix Table A.18 drops the performance fixed effects to explore the heterogeneity by performance on the test. As with our prior results in Section 4.2, performance is significantly and positively correlated with more favorable subjective evaluations. Unlike our prior results, however, the gender gap is *larger* for the higher performers.

On beliefs about performance, Appendix Tables A.19 and A.20 consider heterogeneity by beliefs about absolute performance and beliefs about absolute performance relative to actual performance. As with our prior results, the beliefs are positively and significantly correlated with subjective evaluations, even when individuals have been informed about performance. Evidence for how these beliefs correlate with the size of the gap is mixed.

On demographics, Appendix Table A.21 considers heterogeneity by race. In our sample, 38% of participants are non-Hispanic Whites, the racial majority in the United States. We define *Racial Minority* as an indicator that a participant is not classified as a non-Hispanic White in the administrative data provided by Character Lab Research Network. We find that, relative to students who are non-Hispanic Whites, students from racial minority groups provide less positive responses to

the self-evaluation questions about performance and more positive responses about their willingness to take a class. The gender gap does not appear to systematically differ by racial minority status. Appendix Table A.22 considers heterogeneity according to whether a student qualifies for a free or reduced price lunch (FRPL). In our sample, 38% of students quality for FRPL. While mixed, evidence mostly points towards FRPL status being somewhat negatively correlated with answers to the self-evaluation questions and not being correlated with the gender gap.

Table 4: The Gender Gap in Evaluations Among Youth

	Among students in grade:						
	6th	7th	8th	9th	10th	11th	12th
Performance Question							
Female	-10.52***	-11.81***	-11.05***	-11.80***	-12.14***	-11.40***	-10.44***
	(1.26)	(1.04)	(0.79)	(1.45)	(1.41)	(1.49)	(1.74)
Performance-Bucket Question							
Female	-0.47***	-0.56***	-0.51***	-0.59***	-0.52***	-0.53***	-0.45***
	(0.07)	(0.05)	(0.04)	(0.07)	(0.07)	(0.08)	(0.09)
Willingness Question							
Female	-6.82***	-6.48***	-3.68***	-3.86**	-6.92***	-0.29	-5.77**
	(1.60)	(1.31)	(1.00)	(1.82)	(1.88)	(1.98)	(2.38)
Success Question							
Female	-9.42***	-9.85***	-7.19***	-7.41***	-8.40***	-4.58***	-7.29***
	(1.52)	(1.24)	(0.93)	(1.73)	(1.76)	(1.69)	(2.16)
Informed Performance Question							
Female	-4.00***	-7.10***	-6.98***	-6.51***	-9.55***	-6.75***	-6.24***
	(1.45)	(1.19)	(0.91)	(1.66)	(1.73)	(1.74)	(2.10)
Informed Performance-Bucket Question							
Female	-0.15**	-0.33***	-0.27***	-0.27***	-0.33***	-0.26***	-0.22**
	(0.07)	(0.06)	(0.05)	(0.09)	(0.09)	(0.09)	(0.11)
Informed Willingness Question							
Female	-4.54***	-4.02***	-2.35**	-3.43*	-6.65***	0.00	-5.62**
	(1.74)	(1.38)	(1.03)	(1.87)	(1.87)	(1.98)	(2.39)
Informed Success Question							
Female	-5.02***	-7.42***	-4.94***	-4.61**	-7.12***	-5.10***	-8.20***
	(1.68)	(1.36)	(1.01)	(1.83)	(1.93)	(1.88)	(2.32)
N	1521	2208	3367	1031	989	871	650
Perf. FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each row among students in the grade indicated by the column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. Performance FEs are dummies for each possible performance out of the 10 questions on the test.

Using data on students' grade point average (GPA), Appendix Tables A.23 considers heterogeneity by current marking period GPA while controlling for performance on the test with fixed effects. This measure of their grades is positively and significantly correlated with answers to the

self-evaluation questions, and the gap is—if anything—larger among those with a higher GPA, which is consistent with the higher performers on the test displaying a larger gap in this sample.

6 Conclusion

This paper documents a gender gap in self-promotion: women subjectively describe their performance less favorably to potential employers than equally performing men. This paper also documents a gender gap in *informed* self-promotion: the gender gap persists even when participants know their absolute and relative performance and are asked about that performance. The gender gap in self-promotion relates to an underlying gender gap in how men and women subjectively evaluate their performance absent any incentive to "promote." This underlying gap is robust to an environment that provides information about how others subjectively evaluate the same performance and an environment in which the potential for consistency motives or anchoring effects is reduced. We also replicate our results among more than 10,000 middle-school and high-school students. The gender gap only closes when we ask participants about the performance of others—rather than themselves—or when we ask participants about their performance on a more female-typed task.

There are many exciting and important avenues for future work. A first avenue relates to exploring other gaps, beyond gender, in how individuals subjectively evaluate their performance. Our results in Section 4.2 highlight gains to more work considering gaps that may relate to age, political affiliation, education, socio-economic status, race, nationality, and others. Exploring these additional individual characteristics may also help to provide further insight into role of culture and life experiences on subjective evaluations of performance and self-promotion.

A second avenue relates to how to mitigate the consequences of the gender gap in self-promotion. Given the potential difficulty of altering how men and women subjectively view their performance—particularly in the short run if such perceptions are deeply ingrained—promising approaches may require "changing the system" rather than "changing the women."[19] Such changes may involve implementing processes that rely less on subjective self-evaluations for promotion and hiring. For excellent recent work on change-the-system approaches, see Apicella, Demiral and Mollerstrom (2017), Coffman, Collis and Kulkarni (2020), He, Kang and Lacetera (2019), Apicella, Demiral and Mollerstrom (2020), and Carlana, La Ferrara and Pinotti (2020).

In light of the large literature on discrimination and gender-specific backlash (Riach and Rich, 2002; Bowles, Babcock and Lai, 2007; Rudman and Phelan, 2008; Blau and Kahn, 2017), a third avenue for future work is to explore how making gender known impacts the gender gap in self-promotion and its economic consequences.[20] To narrow in on whether the gender gap in self-

[19] There are other reasons to avoid a "changing the women" approach. For work on how forcing women to take actions that they would not choose themselves can backfire, in the context of choosing whether to negotiate, see Exley, Niederle and Vesterlund (2020).

[20] Bursztyn, Fujiwara and Pallais (2017) show how image concerns can cause women to downplay how they describe their career ambitions to others. For recent evidence on gender discrimination and how men and women are judged differently, see Reuben, Sapienza and Zingales (2014); Milkman, Akinola and Chugh (2015); Baert, De Pauw and

promotion reflects an underlying gap in how men and women subjectively evaluate their performance, we sought to strip away factors by constructing a *Private* version that was as comparable as possible to our *Self-Promotion* version but without incentives to promote. Future work, however, may take an alternative approach by investigating the impact of layering on important factors, such as gender being known. While results in Section 4.4 cast doubt on this possibility, the gender gap in self-promotion may be mitigated by employers accurately accounting for it when assessing the performance of their male and female workers. Alternatively, the gender gap in self-promotion could become larger if women fear backlash from self-promotion and hence self-promote less when their gender is known. Exploring these channels—along with how they may interact with other factors, such as discrimination, that become relevant when gender is known—are important avenues for future work.

Deschacht (2016); Bohnet and Bazerman (2016); Boring (2017); Sarsons (2017*a*,*b*); Kübler, Schmid and Stüber (2018); Alston (2019); Bohren, Imas and Rosenberg (2019); Bohren et al. (2019); Coffman, Exley and Niederle (Forthcoming); Kessler, Low and Sullivan (2019); Grossman et al. (2019); Sarsons et al. (Forthcoming). For how the gender composition of a group can influence outcomes, see also Eckel and Grossman (2001) and Hernandez-Arenaz and Iriberri (2018).

References

Almås, Ingvild, Alexander W Cappelen, Kjell G Salvanes, Erik Ø Sørensen, and Bertil Tungodden. 2018. "Competition, Subjective Feedback, and Gender Gaps in Performance." *Management Science*, 62(7).

Alston, Mackenzie Alston. 2019. "The (Perceived) Cost of Being Female: An Experimental Investigation of Strategic Responses to Discrimination." *Working Paper.*

Andreoni, James, and Lise Vesterlund. 2001. "Which is the fair sex? Gender differences in altruism." *Quarterly Journal of Economics*, 116(1): 293–312.

Apicella, Coren L., Elif E. Demiral, and Johanna Mollerstrom. 2017. "No Gender Difference in Willingness to Compete When Competing against Self." *American Economic Review: Papers & Proceedings*, 107(5): 136–140.

Apicella, Coren L, Elif E Demiral, and Johanna Mollerstrom. 2020. "Compete with others? No, thanks. With myself? Yes, please!" *Economics Letters*, 187.

Atwater, Ann, and Perihan O. Saygin. 2020. "Gender Differences in Leaving Questions Blank on High-Stakes Standardized Test." *Working Paper.*

Azmat, Ghazala, and Barbara Petrongolo. 2014. "Gender and the labor market: What have we learned from field and lab experiments?" *Labour Economics*, 30: 32–40.

Baert, Stijn, Ann-Sophie De Pauw, and Nick Deschacht. 2016. "Do employer preferences contribute to sticky floors?" *ILR Review*, 69(3): 714–736.

Bertrand, Marianne. 2011. "New perspectives on Gender." *Handbook of Labor Economics*, 4: 1543–1590.

Beyer, Sylvia. 1990. "Gender differences in the accuracy of self-evaluations of performance." *Journal of personality and social psychology*, 59(5).

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

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2016. "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation." *Management Science*, 62(5): 1225–1234.

Bohren, Aislinn, Alex Imas, and Michael Rosenberg. 2018. "The Language of Discrimination: Using Experimental versus Observational Data." *AEA Papers and Proceedings*, 108(169–74).

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

Bohren, J. Aislinn, Kareem Haggag, Alex Imas, and Devin G. Pope. 2019. "Inaccurate Statistical Discrimination." *University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2019-86.*

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review*.

Boring, Anne. 2017. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics*, 27–41.

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2018. "A man's world? – The impact of a male dominated environment on female leadership." *University of Gothenburg Working Paper in Economics No. 744*.

Bowles, Hannah Riley, Linda Babcock, and Kathleen L. McGinn. 2005. "Constraints and triggers: situational mechanics of gender in negotiation." *Journal of personality and social psychology*, 89(6): 951–965.

Bowles, Hannah Riley, Linda Babcock, and Lei Lai. 2007. "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask." *Organizational Behavior and Human Decision Processes*, 103(1): 84–103.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review*, 107(11): 3288–3319.

Buser, Thomas, Leonie Gerhards, and Joël J Van der Weele. 2018. "Measuring responsiveness to feedback as a personal trait." *Journal of Risk and Uncertainty*, 56(2): 165–192.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, competitiveness, and career choices." *The Quarterly Journal of Economics*, 129(3): 1409–1447.

Cárdenas, Juan-Camilo, Anna Dreber, Emma Von Essen, and Eva Ranehill. 2012. "Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden." *Journal of Economic Behavior & Organization*, 83(1): 11–22.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2020. "Goals and Gaps: Educational Careers of Immigrant Children." *Working Paper*.

Charness, Gary, Aldo Rustichini, and Jeroen Van de Ven. 2018. "Self-confidence and strategic behavior." *Experimental Economics*, 21(1): 72–98.

Cobb-Clark, Deborah A. 2015. "Locus of control and the labor market." *IZA Journal of Labor Economics*, 4(1).

Coffman, Katherine Baldiga. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *The Quarterly Journal of Economics*, 129(4): 1625–1660.

Coffman, Katherine B., Christine L. Exley, and Muriel Niederle. Forthcoming. "The Role of Beliefs in Driving Gender Discrimination." *Management Science*.

Coffman, Katherine B., Manuela R. Collis, and Leena Kulkarni. 2020. "When to Apply?" *Working Paper.*

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov. 2019. "Gender Stereotypes in Deliberation and Team Decisions." *Harvard Business School Working Paper.*

Coffman, Katherine, Manuela Collis, and Leena Kulkarni. 2019. "Stereotypes and Belief Updating." *Working Paper.*

Coutts, Alexander. 2018. "Good news and bad news are still news: Experimental evidence on belief updating." *Experimental Economics*, 1–27.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448–474.

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao. 2013. "The Importance of Being Marginal: Gender Differences in Generosity." *American Economic Review: Papers & Proceedings*, 103(3): 586–590.

Dreber, Anna, Emma Von Essen, and Eva Ranehill. 2011. "Outrunning the gender gap—boys and girls compete equally." *Experimental Economics*, 14(4).

Dreber, Anna, Emma von Essen, and Eva Ranehill. 2014. "Gender and competition in adolescence: task matters." *Experimental Economics*, 17(1): 154–172.

Dwyer, Peggy D, James H Gilkeson, and John A List. 2002. "Gender differences in revealed risk taking: evidence from mutual fund investors." *Economics Letters*, 76(2): 151–158.

Eckel, Catherine C., and Philip J. Grossman. 2001. "Chivalry and solidarity in ultimatum games." *Economic Inquiry*, 39(2): 171–188.

Eckel, Catherine C., and Philip J. Grossman. 2008. "Men, Women and Risk Aversion: Experimental Evidence." In *Handbook of Experimental Economics Results.* 1061–1073.

Ertac, Seda. 2011. "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback." *Journal of Economic Behavior & Organization*, 80(3): 532–545.

Exley, Christine L., Muriel Niederle, and Lise Vesterlund. 2020. "Knowing When to Ask: The Cost of Leaning-in." *Journal of Political Economy*, 128(3): 816–854.

Gneezy, Uri, and Aldo Rustichini. 2004. "Gender and competition at a young age." *American Economic Review*, 94(2): 377–381.

Gneezy, Uri, Kenneth L Leonard, and John A List. 2009. "Gender differences in competition: Evidence from a matrilineal and a patriarchal society." *Econometrica*, 77(5): 1637–1664.

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

Grossman, Philip J, Catherine Eckel, Mana Komai, and Wei Zhan. 2019. "It pays to be a man: Rewards for leaders in a coordination gam." *Journal of Economic Behavior & Organization*, 161: 197–215.

He, Joyce, Sonia Kang, and Nicola Lacetera. 2019. "Leaning In or Not Leaning Out? Opt-Out Choice Framing Attenuates Gender Differences in the Decision to Compete." *NBER Working Paper No. 26484.*

Hernandez-Arenaz, Iñigo, and Nagore Iriberri. 2018. "Women ask for less (only from men): Evidence from alternating-offer bargaining in the field." *Journal of Economic Behavior & Organization.*

Hernandez-Arenaz, Iñigo, and Nagore Iriberri. 2918. "A review of gender differences in negotiation." *Oxford Research Encyclopedia of Economics and Finance.*

Isaksson, Siri. 2018. "It Takes Two: Gender Differences in Group Work." *Working Paper.*

Kessler, Judd B, Corinne Low, and Colin D Sullivan. 2019. "Incentivized Resume Rating: Eliciting Employer Preferences without Deception." *American Economic Review*, 109(11): 3713–44.

Khachatryan, Karen, Anna Dreber, Emma Von Essen, and Eva Ranehill. 2015. "Gender and preferences at a young age: Evidence from Armenia." *Journal of Economic Behavior & Organization*, 118: 319=322.

Kolev, Julian, Yuly Fuentes-Medel, and Fiona Murray. 2019. "Is blinded review enough? How gendered outcomes arise even under anonymous evaluation." *NBER Working Paper No. 25759.*

Kübler, Dorothea, Julia Schmid, and Robert Stüber. 2018. "Gender discrimination in hiring across occupations: a nationally-representative vignette study." *Labour Economics*, 55: 215–229.

Lerchenmueller, Marc J, Olav Sorenson, and Anupam B Jena. 2019. "Gender differences in how scientists present the importance of their research: observational study." *British Medical Journal*, 367.

Lundeberg, Mary A, Paul W Fox, and Judith Punćcohaŕ. 1994. "Highly confident but wrong: Gender differences and similarities in confidence judgments." *Journal of educational psychology*, 86(1).

Manian, Shanthi, and Keith Sheth. 2020. "Follow my Lead: Assertive Cheap Talk and the Gender Gap." *Working Paper.*

Milkman, Katherine L, Modupe Akinola, and Dolly Chugh. 2015. "What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations." *Journal of Applied Psychology*, 100(6).

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2011. "Managing Self-Confidence: Theory and Experimental Evidence." *NBER Working Paper No. 17014.*

Murciano-Goroff, Raviv. 2021. "Missing women in tech: The labor market for highly skilled software engineers." *Management Science*.

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

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics*, 122(3): 1067–1101.

Niederle, Muriel, and Lise Vesterlund. 2011. "Gender and Competition." *Annual Review of Economics*, 3: 601–630.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How stereotypes impair women's careers in science." *Proceedings of the National Academy of Sciences*, 111(12): 4403–4408.

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

Rudman, Laurie A, and Julie E Phelan. 2008. "Backlash effects for disconfirming gender stereotypes in organizations." *Research in organizational behavior*, 28(6-79).

Sarsons, Heather. 2017*a*. "Interpreting Signals in the Labor Market: Evidence from Medical Referrals." *Working Paper*.

Sarsons, Heather. 2017*b*. "Recognition for Group Work: Gender Differences in Academia." *American Economic Review: Papers & Proceedings*, 107(5): 141–145.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. Forthcoming. "Gender Differences in Recognition for Group Work." *Journal of Political Economy*.

Schwardmann, Peter, and Joël van der Weele. 2019. "Deception and Self-Deception." *Nature Human Behavior*, 3: 1055–1061.

Soldà, Alice, Changxia Ke, Lionel Page, and William von Hippel. 2019. "Strategically Delusional." *Experimental Economics*.

Sutter, Matthias, and Daniela Glätzle-Rützler. 2015. "Gender differences in the willingness to compete emerge early in life and persist." *Management Science*, 61(10).

A Appendix

A.1 Additional Tables

Table A.1: Performance and Absolute Performance Beliefs

DV:	Performance	Belief		Belief–Performance
	(1)	(2)	(3)	(4)
Female	0.60***	-2.29***	-2.19***	-2.88***
	(0.13)	(0.14)	(0.14)	(0.17)
Constant	9.34***	11.05***		1.71***
	(0.09)	(0.09)		(0.12)
N	3587	3587	3587	3587
Performance FEs	No	No	Yes	No

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The SEs are robust. Results are from OLS regressions of the noted dependent variable (DV). *Performance* is the number of questions a participant answered correctly out of the 20 questions on the test. *Belief* is the number of questions a participant believes he or she answered correctly. *Belief–Performance* is the difference between these two variables, calculated for each participant. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all study versions from waves 1–5 involving the math and science test (i.e., all but the *Private (Verbal)* version).

Table A.2: Regression results on the role of providing information on absolute and relative performance

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: *Self-Promotion* Version				
Female	-11.75***	-0.55***	-14.09***	-14.29***
	(2.95)	(0.13)	(3.44)	(3.43)
Informed	-1.10	0.04	1.67	-0.04
	(1.36)	(0.07)	(1.50)	(1.51)
*Informed*Female*	3.80	0.11	2.15	1.76
	(2.37)	(0.11)	(2.44)	(2.39)
N	604	604	604	604
Panel 2: All Versions with Evaluations Before and After Being Informed				
Female	-13.89***	-0.67***	-17.17***	-16.15***
	(1.14)	(0.05)	(1.31)	(1.32)
Informed	-1.49***	0.00	0.32	-0.84
	(0.56)	(0.03)	(0.55)	(0.52)
*Informed*Female*	4.10***	0.21***	2.30***	1.59**
	(0.88)	(0.04)	(0.81)	(0.80)
N	4188	4188	4188	4188
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered at subject-level. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Informed* is an indicator for the evaluation being provided after the participant is informed of their absolute and relative performance. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data in Panel 1 are from the *Self-Promotion* version. Data in Panel 2 are from all versions that elicit evaluations of math and science performance before and after participants are informed of their absolute and relative performance (i.e., all but the *Private (Immediately Informed)* version, *Private (Other-Evaluation)* version, and *Private (Verbal)* version). Each participant in these versions is in the data twice for each specification, once providing an evaluation before being informed and once providing an evaluation after being informed.

Table A.3: Regression results on the impact of promotion incentives from the *Self-Promotion* and *Private* versions

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-13.86***	-0.59***	-17.85***	-16.52***
	(2.82)	(0.13)	(3.36)	(3.45)
Self-Promotion	6.25**	0.26**	4.27	6.93**
	(2.72)	(0.13)	(3.35)	(3.30)
*Self-Promotion*Female*	1.66	-0.00	2.30	1.07
	(4.04)	(0.18)	(4.77)	(4.84)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-8.55***	-0.33**	-13.81***	-13.88***
	(2.79)	(0.14)	(3.40)	(3.41)
Self-Promotion	7.79***	0.34**	6.72**	9.00***
	(2.85)	(0.14)	(3.34)	(3.24)
*Self-Promotion*Female*	1.41	-0.09	2.08	1.31
	(3.93)	(0.18)	(4.74)	(4.70)
N	606	606	606	606
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Self-Promotion* is an indicator for the evaluation being from the *Self-Promotion* version. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from the *Self-Promotion* version and *Private* version run in wave 1, so participants were randomly assigned between these study versions.

Table A.4: Robustness to excluding performance fixed effects

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-15.76***	-0.78***	-19.25***	-18.07***
	(1.14)	(0.05)	(1.30)	(1.32)
Constant	58.50***	3.76***	57.36***	61.39***
	(0.72)	(0.04)	(0.84)	(0.81)
N	2094	2094	2094	2094
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-11.16***	-0.56***	-16.78***	-16.92***
	(1.01)	(0.05)	(1.12)	(1.11)
Constant	57.86***	3.77***	58.51***	61.92***
	(0.62)	(0.03)	(0.68)	(0.66)
N	2990	2990	2990	2990
Performance FEs	No	No	No	No

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. Performance FEs are not included. Data are from all study versions involving evaluations of the participant's own math and science performance. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.5: Robustness to controlling for other demographic variables

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
Panel 1: Evaluations (before performance information is provided)				
Female	-12.70***	-0.61***	-15.95***	-14.82***
	(1.09)	(0.05)	(1.28)	(1.29)
Age	-0.30***	-0.01***	-0.32***	-0.24***
	(0.05)	(0.00)	(0.06)	(0.06)
Education (demeaned)	4.08***	0.21***	4.44***	4.90***
	(0.39)	(0.02)	(0.45)	(0.46)
Republican Leaning (demeaned)	0.12***	0.01***	0.10***	0.10***
	(0.02)	(0.00)	(0.02)	(0.02)
N	2092	2092	2092	2092
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-8.67***	-0.41***	-13.76***	-14.20***
	(0.90)	(0.04)	(1.05)	(1.04)
Age	-0.29***	-0.01***	-0.25***	-0.20***
	(0.04)	(0.00)	(0.05)	(0.05)
Education (demeaned)	3.38***	0.16***	4.22***	4.47***
	(0.33)	(0.02)	(0.38)	(0.37)
Republican Leaning (demeaned)	0.15***	0.01***	0.13***	0.11***
	(0.02)	(0.00)	(0.02)	(0.02)
N	2986	2986	2986	2986
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Education (demeaned)* is a number from 1 to 9 that corresponds with education level (where the least education is 1 and the most education is 9), demeaned by the average. *Republican Leaning (demeaned)* is a number from 0 to 100 that is the extent to which the participant indicated feeling favorably about the Republican party, demeaned by the average. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all study versions involving evaluations of the participant's own math and science performance but excludes the participants who selected "other" as their educational attainment. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.6: Robustness to excluding very low performers

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-13.50***	-0.62***	-17.38***	-16.38***
	(1.18)	(0.05)	(1.40)	(1.41)
N	1771	1771	1771	1771
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-8.72***	-0.38***	-14.63***	-15.04***
	(0.96)	(0.04)	(1.15)	(1.14)
N	2456	2456	2456	2456
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all study versions involving evaluations of the participant's own math and science performance, restricted to the set of participants who answered 6 or more questions correctly out of 20. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.7: Robustness to quantile regressions

Question:	Performance	Willingness-to-Apply	Success
	(1)	(2)	(3)
Panel 1: Evaluations (before information), 25th percentile			
Female	-18.00***	-25.00***	-30.00***
	(1.87)	(2.73)	(2.91)
N	2094	2094	2094
Panel 2: Informed Evaluations (after information), 25th percentile			
Female	-10.00***	-20.00***	-24.00***
	(1.50)	(2.13)	(2.09)
N	2990	2990	2990
Panel 3: Evaluations (before information), 50th percentile			
Female	-14.00***	-24.00***	-19.00***
	(2.18)	(2.61)	(2.40)
Constant	75.00***	65.00***	82.00***
N	2094	2094	2094
Panel 4: Informed Evaluations (after information), 50th percentile			
Female	-9.00***	-18.00***	-17.00***
	(1.13)	(1.94)	(1.81)
N	2990	2990	2990
Panel 5: Evaluations (before information), 75th percentile			
Female	-11.00***	-13.00***	-11.00***
	(0.99)	(1.47)	(1.65)
N	2094	2094	2094
Panel 6: Informed Evaluations (after information), 75th percentile			
Female	-6.00***	-11.00***	-10.00***
	(0.96)	(1.31)	(1.08)
N	2990	2990	2990
Performance FEs	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from quantile regressions, estimated at the percentile noted in each panel, of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all study versions involving evaluations of the participant's own math and science performance. Panels 1, 3, and 5 analyze evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panels 2, 4, and 6 analyze evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.8: Considering the relationship between performance and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-16.37***	-0.80***	-19.64***	-18.57***
	(1.18)	(0.06)	(1.33)	(1.35)
Performance (demeaned)	-0.54***	-0.06***	-0.42**	-0.27
	(0.17)	(0.01)	(0.19)	(0.18)
Performance (demeaned)	1.59***	0.09***	1.10***	1.13***
**Female*	(0.33)	(0.02)	(0.37)	(0.37)
N	2094	2094	2094	2094
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-12.26***	-0.59***	-17.44***	-17.81***
	(1.00)	(0.05)	(1.12)	(1.11)
Performance (demeaned)	0.61***	-0.01	0.18	0.58***
	(0.14)	(0.01)	(0.16)	(0.15)
Performance (demeaned)	2.05***	0.10***	1.55***	1.48***
**Female*	(0.28)	(0.01)	(0.31)	(0.30)
N	2990	2990	2990	2990
Performance FEs	No	No	No	No

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Performance (demeaned)* is the number of questions a participant answered correctly out of the 20 questions on the test, demeaned by the average performance. Data are from all study versions involving evaluations of the participant's own math and science performance. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.9: Considering the relationship between performance and evaluations when excluding very low performers

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-17.68***	-0.82***	-20.81***	-20.15***
	(1.40)	(0.06)	(1.57)	(1.61)
Performance (demeaned)	0.55**	0.01	0.69**	0.78***
	(0.25)	(0.01)	(0.29)	(0.29)
Performance (demeaned)	2.34***	0.11***	1.80***	1.98***
* *Female*	(0.41)	(0.02)	(0.48)	(0.48)
N	1771	1771	1771	1771
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-13.15***	-0.59***	-18.52***	-18.65***
	(1.23)	(0.06)	(1.34)	(1.35)
Performance (demeaned)	2.53***	0.10***	1.93***	2.33***
	(0.22)	(0.01)	(0.24)	(0.23)
Performance (demeaned)	2.71***	0.12***	2.27***	2.10***
**Female*	(0.34)	(0.02)	(0.40)	(0.39)
N	2456	2456	2456	2456
Performance FEs	No	No	No	No

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Performance (demeaned)* is the number of questions a participant answered correctly out of the 20 questions on the test, demeaned by the average performance. Data are from all study versions involving evaluations of the participant's own math and science performance, restricted to the set of participants who answered 6 or more questions correctly out of 20. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.10: Considering the relationship between beliefs and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-4.49***	-0.25***	-8.46***	-7.07***
	(0.84)	(0.04)	(1.14)	(1.11)
Belief (demeaned)	3.49***	0.17***	3.55***	3.52***
	(0.17)	(0.01)	(0.19)	(0.20)
*Belief (demeaned)*Female*	1.41***	0.05***	0.95***	1.15***
	(0.21)	(0.01)	(0.26)	(0.26)
N	2094	2094	2094	2094
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-4.01***	-0.21***	-8.49***	-9.18***
	(0.86)	(0.04)	(1.00)	(0.99)
Belief (demeaned)	2.25***	0.11***	2.55***	2.43***
	(0.14)	(0.01)	(0.15)	(0.16)
*Belief (demeaned)*Female*	0.73***	0.02	0.87***	0.89***
	(0.20)	(0.01)	(0.22)	(0.22)
N	2990	2990	2990	2990
Panel 3: Informed Evaluations from *Private (Immediately Informed)* version				
Female	-2.04	-0.16**	-8.16***	-10.38***
	(1.58)	(0.08)	(1.80)	(1.73)
Belief (demeaned)	2.26***	0.11***	2.17***	2.12***
	(0.24)	(0.01)	(0.25)	(0.24)
*Belief (demeaned)*Female*	1.09***	0.03*	1.41***	1.25***
	(0.36)	(0.02)	(0.41)	(0.39)
N	896	896	896	896
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Belief (demeaned)* is the number of questions a participant believes he or she answered correctly, demeaned by the average belief. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all study versions involving evaluations of the participant's own math and science performance. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3). Panel 3 restricts to evaluations from the *Private (Immediately Informed)* version in which subjects are informed of their absolute and relative performance before responding to any self-evaluation questions.

Table A.11: Considering the relationship between beliefs relative to performance and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-4.49***	-0.26***	-8.36***	-6.91***
	(0.88)	(0.04)	(1.15)	(1.12)
Belief–Performance (demeaned)	3.87***	0.19***	3.75***	3.75***
	(0.15)	(0.01)	(0.18)	(0.18)
Belief–Performance (demeaned)	0.65***	0.01	0.56***	0.72***
**Female*	(0.18)	(0.01)	(0.21)	(0.22)
N	2094	2094	2094	2094
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-4.33***	-0.22***	-8.67***	-9.34***
	(0.89)	(0.04)	(1.01)	(1.01)
Belief–Performance (demeaned)	2.69***	0.13***	2.89***	2.76***
	(0.13)	(0.01)	(0.14)	(0.14)
Belief–Performance (demeaned)	-0.25	-0.03***	0.12	0.17
**Female*	(0.17)	(0.01)	(0.18)	(0.18)
N	2990	2990	2990	2990
Panel 3: Informed Evaluations from *Private (Immediately Informed)* version				
Female	-2.66	-0.19**	-8.69***	-10.84***
	(1.64)	(0.08)	(1.83)	(1.75)
Belief–Performance (demeaned)	2.73***	0.13***	2.64***	2.53***
	(0.22)	(0.01)	(0.23)	(0.22)
Belief–Performance (demeaned)	-0.05	-0.02	0.30	0.27
**Female*	(0.27)	(0.01)	(0.29)	(0.28)
N	896	896	896	896
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Belief–Performance (demeaned)* is the number of questions a participant believes he or she answered correctly minus the number of questions the participant actually answered correctly, demeaned by the average difference. Data are from all study versions involving evaluations of the participant's own math and science performance. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3). Panel 3 restricts to evaluations from the *Private (Immediately Informed)* version in which subjects are informed of their absolute and relative performance before responding to any self-evaluation questions.

Table A.12: Statistically controlling for beliefs versus controlling for beliefs by design

Question:	Performance	Performance-Bucket	Willingness to Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before information)				
Female	-13.83***	-0.67***	-17.28***	-16.12***
	(1.13)	(0.05)	(1.31)	(1.32)
Belief FEs	No	No	No	No
Panel 2: Informed Evaluations (after information)				
Female	-9.84***	-0.46***	-14.75***	-14.60***
	(1.09)	(0.05)	(1.29)	(1.29)
Belief FEs	No	No	No	No
Panel 3: Evaluations (before information) with belief controls				
Female	-4.45***	-0.24***	-8.39***	-6.88***
	(0.88)	(0.04)	(1.16)	(1.14)
Belief FEs	Yes	Yes	Yes	Yes
N	2094	2094	2094	2094
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Belief FEs are dummies for each possible belief about how many questions the participant answered correctly out of the 20 questions on the test. Data are from all versions that elicit evaluations of math and science performance before and after participants are informed of their absolute and relative performance (i.e., all but the *Private (Immediately Informed)* version, *Private (Other-Evaluation)* version, and *Private (Verbal)* version). Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance, reproducing Panel 12 of Table 2. Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance from the same participants presented in Panel 1. Panel 3 analyzes evaluations from before participants are informed of their absolute and relative performance but adds Belief FEs to control for beliefs statistically.

Table A.13: Considering the relationship between general math and science beliefs and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-7.89***	-0.36***	-11.68***	-11.60***
	(2.44)	(0.10)	(3.06)	(2.82)
General Math Belief (demeaned)	6.80***	0.33***	9.61***	10.42***
	(1.17)	(0.05)	(1.42)	(1.17)
General Math Belief (demeaned)	0.76	-0.01	0.56	0.33
**Female*	(1.44)	(0.06)	(1.59)	(1.50)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-3.45*	-0.05	-5.97**	-7.30***
	(2.07)	(0.09)	(2.81)	(2.69)
General Math Belief (demeaned)	6.24***	0.30***	9.46***	9.65***
	(0.99)	(0.04)	(1.31)	(1.14)
General Math Belief (demeaned)	-0.46	-0.06	0.47	0.09
**Female*	(1.20)	(0.05)	(1.50)	(1.49)
N	294	294	294	294
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. *General Math Belief (demeaned)* is a participant's answer on a seven-point scale (where 1 is "strongly disagree" and 7 is "strongly agree" with the statement "In general, I perform well when asked questions that test my math and science skills"), demeaned by the average response. Data are from the *Private* version that was conducted in wave 5 when we added the general belief questions to the follow-up survey. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance. Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance.

Table A.14: Considering the relationship between general verbal beliefs and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-0.20	-0.18*	0.14	-2.23
	(2.13)	(0.10)	(2.76)	(2.57)
General Verbal Belief (demeaned)	8.40***	0.38***	11.37***	11.58***
	(1.24)	(0.05)	(1.28)	(1.20)
General Verbal Belief (demeaned)	-1.39	-0.09	-1.53	-1.91
**Female*	(1.50)	(0.07)	(1.52)	(1.46)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-1.92	-0.09	-2.96	-3.00
	(1.74)	(0.08)	(2.38)	(2.23)
General Verbal Belief (demeaned)	5.41***	0.29***	9.33***	9.54***
	(1.13)	(0.04)	(1.18)	(1.12)
General Verbal Belief (demeaned)	1.14	-0.05	0.24	0.06
**Female*	(1.37)	(0.06)	(1.39)	(1.38)
N	305	305	305	305
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. *General Verbal Belief (demeaned)* is a participant's answer on a seven-point scale (where 1 is "strongly disagree" and 7 is "strongly agree" with the statement "In general, I perform well when asked questions that test my verbal skills"), demeaned by the average response. Data are from the *Private (Verbal)* version that was conducted in wave 5 when we added the general belief questions to the follow-up survey. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance. Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance.

Table A.15: Considering the relationship between other demographics and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
Panel 1: Evaluations (before performance information is provided)				
Female	-12.77***	-0.61***	-15.95***	-14.76***
	(1.09)	(0.05)	(1.29)	(1.29)
Age	-0.24***	-0.01***	-0.22**	-0.09
	(0.07)	(0.00)	(0.09)	(0.08)
Education (demeaned)	4.21***	0.21***	4.17***	4.44***
	(0.51)	(0.02)	(0.62)	(0.60)
Republican (demeaned)	0.16***	0.01***	0.16***	0.15***
	(0.03)	(0.00)	(0.03)	(0.03)
*Age*Female*	-0.11	-0.00	-0.21*	-0.29**
	(0.11)	(0.00)	(0.12)	(0.12)
*Education (demeaned)*Female*	-0.28	-0.02	0.57	1.01
	(0.78)	(0.03)	(0.91)	(0.92)
*Republican (demeaned)*Female*	-0.08**	-0.01***	-0.12***	-0.11**
	(0.04)	(0.00)	(0.05)	(0.05)
N	2092	2092	2092	2092
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-8.67***	-0.41***	-13.72***	-14.13***
	(0.90)	(0.04)	(1.05)	(1.04)
Age	-0.24***	-0.01***	-0.18***	-0.10
	(0.06)	(0.00)	(0.07)	(0.06)
Education (demeaned)	3.42***	0.17***	4.15***	4.40***
	(0.44)	(0.02)	(0.49)	(0.47)
Republican (demeaned)	0.22***	0.01***	0.18***	0.15***
	(0.02)	(0.00)	(0.02)	(0.02)
*Age*Female*	-0.11	-0.00	-0.16*	-0.23**
	(0.08)	(0.00)	(0.10)	(0.10)
*Education (demeaned)*Female*	-0.11	-0.04	0.13	0.15
	(0.66)	(0.03)	(0.77)	(0.76)
*Republican (demeaned)*Female*	-0.15***	-0.01***	-0.10***	-0.09**
	(0.03)	(0.00)	(0.04)	(0.04)
N	2986	2986	2986	2986
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Education (demeaned)* is a number from 1 to 9 that corresponds with education level (where the least education is 1 and the most education is 9), demeaned by the average. *Republican Leaning (demeaned)* is a number from 0 to 100 that is the extent to which the participant indicated feeling favorably about the Republican party, demeaned by the average. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all study versions involving evaluations of the participant's own math and science performance but excludes the participants who selected "other" as their educational attainment. Panel 1 analyzes evaluations from before participants are informed of their absolute and relative performance (as in Panel 12 of Table 2). Panel 2 analyzes evaluations from after participants are informed of their absolute and relative performance (as in Panel 12 of Table 3).

Table A.16: Deservingness Measure

Female	-0.88
	(1.23)
N	2394
Performance FEs	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the deservingness measure, which ranges from 0–100, in response to the following question: "Out of a maximum amount of 100 cents, what amount of bonus payment, in cents, do you think you deserve for your performance on the test you took in part 1." *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data are from all versions in which participants are asked about their own performance on the math and science test but do not have an opportunity to influence their payments through self-promotion (i.e., all but the *Self-Promotion* version, *Self-Promotion (Risky)* version, *Private (Other-Evaluation)* version, and *Private (Verbal)* version).

Table A.17: Predicted Performance

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Predictions about women	-1.54***	-1.47***	-2.25***	-2.24***
	(0.17)	(0.17)	(0.18)	(0.19)
Female predictor	0.18	0.25	0.22	0.36
	(0.20)	(0.21)	(0.20)	(0.23)
Predictions about women	0.21	0.08	0.01	-0.21
**Female predictor*	(0.24)	(0.24)	(0.27)	(0.29)
Constant	11.98***	12.49***	12.40***	13.04***
	(0.14)	(0.15)	(0.14)	(0.16)
N	1198	1198	1198	1198

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered at the participant level. Results are from OLS regressions of the predicted average performance (i.e., the average number of questions answered correctly by a set of female participants or a set of male participants) based on the gender's average response to the question noted in the column. (Average responses are from the *Self-Promotion* version after information about absolute and relative performance on the test has been provided.) *Predictions about women* is an indicator that the question elicited a prediction for the average performance of female workers. *Female predictor* is an indicator for the predictor being female. Data are from the study versions conducted in wave 5 when we added the prediction questions to the follow-up survey.

Table A.18: Among our youth sample: considering the relationship between performance and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-11.63***	-0.54***	-4.98***	-8.19***
	(0.45)	(0.02)	(0.58)	(0.54)
Performance (demeaned)	4.55***	0.24***	2.15***	3.97***
	(0.14)	(0.01)	(0.18)	(0.17)
*Performance (demeaned)*Female*	-0.61***	-0.05***	-0.24	-0.56**
	(0.21)	(0.01)	(0.26)	(0.24)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-7.20***	-0.29***	-3.73***	-6.11***
	(0.53)	(0.03)	(0.60)	(0.59)
Performance (demeaned)	4.99***	0.27***	2.66***	4.17***
	(0.18)	(0.01)	(0.19)	(0.18)
*Performance (demeaned)*Female*	-0.50**	-0.01	-0.63**	-0.76***
	(0.25)	(0.01)	(0.27)	(0.26)
N	10637	10637	10637	10637
Performance FEs	No	No	No	No

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. *Performance (demeaned)* is the number of questions a participant answered correctly out of the 10 questions on the test, demeaned by the average performance. Data are from the study among youth (i.e., middle-school and high-school students). Panel 1 analyzes evaluations from before participants are informed of their absolute performance. Panel 2 analyzes evaluations from after participants are informed of their absolute performance.

Table A.19: Among our youth sample: considering the relationship between beliefs and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-3.83***	-0.14***	-0.93	-2.48***
	(0.36)	(0.02)	(0.57)	(0.51)
Belief (demeaned)	7.22***	0.37***	4.21***	5.41***
	(0.15)	(0.01)	(0.20)	(0.18)
*Belief (demeaned)*Female*	-0.06	-0.00	-0.70***	-0.32
	(0.18)	(0.01)	(0.25)	(0.23)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-3.15***	-0.10***	-0.64	-2.48***
	(0.51)	(0.03)	(0.60)	(0.59)
Belief (demeaned)	3.51***	0.16***	3.11***	3.34***
	(0.18)	(0.01)	(0.20)	(0.20)
*Belief (demeaned)*Female*	-0.26	-0.00	-0.57**	-0.39
	(0.22)	(0.01)	(0.26)	(0.26)
N	10637	10637	10637	10637
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. *Belief (demeaned)* is the number of questions a participant believes he or she answered correctly out of the 10 questions on the test, demeaned by the average belief. Performance FEs are dummies for each possible performance out of the 10 questions on the test. Data are from the study among youth (i.e., middle-school and high-school students). Panel 1 analyzes evaluations from before participants are informed of their absolute performance. Panel 2 analyzes evaluations from after participants are informed of their absolute performance.

Table A.20: Among our youth sample: considering the relationship between beliefs relative to performance and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-3.85***	-0.14***	-0.94	-2.50***
	(0.35)	(0.02)	(0.57)	(0.51)
Belief–Performance (demeaned)	7.00***	0.35***	4.11***	5.21***
	(0.15)	(0.01)	(0.19)	(0.18)
Belief–Performance (demeaned)	0.34**	0.03***	-0.51**	0.07
**Female*	(0.17)	(0.01)	(0.24)	(0.22)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-3.16***	-0.10***	-0.67	-2.51***
	(0.51)	(0.03)	(0.60)	(0.59)
Belief–Performance (demeaned)	3.39***	0.16***	2.86***	3.08***
	(0.18)	(0.01)	(0.20)	(0.20)
Belief–Performance (demeaned)	-0.04	-0.01	-0.10	0.10
**Female*	(0.22)	(0.01)	(0.25)	(0.25)
N	10637	10637	10637	10637
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. *Belief–Performance (demeaned)* is the number of questions a participant believes he or she answered correctly minus the number of questions the participant actually answered correctly, demeaned by the average difference. Performance FEs are dummies for each possible performance out of the 10 questions on the test. Data are from the study among youth (i.e., middle-school and high-school students). Panel 1 analyzes evaluations from before participants are informed of their absolute performance. Panel 2 analyzes evaluations from after participants are informed of their absolute performance.

Table A.21: Among our youth sample: considering the relationship between racial minority status and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-9.71***	-0.47***	-4.70***	-7.90***
	(0.72)	(0.04)	(0.97)	(0.87)
Racial Minority	-1.11*	-0.09***	3.74***	-0.20
	(0.65)	(0.03)	(0.89)	(0.78)
*Racial Minority*Female*	-2.39***	-0.08*	-0.38	-0.05
	(0.91)	(0.05)	(1.20)	(1.10)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-5.67***	-0.23***	-3.73***	-6.81***
	(0.87)	(0.05)	(1.01)	(0.97)
Racial Minority	-1.80**	-0.13***	2.50***	-0.48
	(0.81)	(0.04)	(0.93)	(0.89)
*Racial Minority*Female*	-1.48	-0.04	0.23	1.60
	(1.08)	(0.06)	(1.24)	(1.21)
N	10637	10637	10637	10637
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. *Racial Minority* is an indicator that the participant is not classified as a non-Hispanic White in the administrative data. Performance FEs are dummies for each possible performance out of the 10 questions on the test. Data are from the study among youth (i.e., middle-school and high-school students). Panel 1 analyzes evaluations from before participants are informed of their absolute performance. Panel 2 analyzes evaluations from after participants are informed of their absolute performance.

Table A.22: Among our youth sample: considering the relationship between FRPL status and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-10.64***	-0.51***	-4.93***	-7.80***
	(0.56)	(0.03)	(0.74)	(0.67)
FRPL	-1.20*	-0.09***	1.16	-1.69**
	(0.67)	(0.03)	(0.88)	(0.80)
*FRPL*Female*	-1.68*	-0.03	-0.02	-0.38
	(0.93)	(0.05)	(1.19)	(1.11)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-6.91***	-0.27***	-3.94***	-5.97***
	(0.67)	(0.04)	(0.77)	(0.74)
FRPL	-1.02	-0.07	0.02	-1.37
	(0.81)	(0.04)	(0.91)	(0.89)
*FRPL*Female*	0.64	0.02	0.92	0.55
	(1.06)	(0.06)	(1.22)	(1.20)
N	10637	10637	10637	10637
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. *FRPL* is an indicator for the participant qualifying for free and reduced-price lunch according to the administrative data. Performance FEs are dummies for each possible performance out of the 10 questions on the test. Data are from the study among youth (i.e., middle-school and high-school students). Panel 1 analyzes evaluations from before participants are informed of their absolute performance. Panel 2 analyzes evaluations from after participants are informed of their absolute performance.

Table A.23: Among our youth sample: considering the relationship between GPA and evaluations

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Evaluations (before performance information is provided)				
Female	-11.87***	-0.54***	-5.36***	-8.92***
	(0.46)	(0.02)	(0.59)	(0.54)
GPA (demeaned)	0.16***	0.01***	0.11***	0.27***
	(0.03)	(0.00)	(0.04)	(0.04)
*GPA (demeaned)*Female*	0.02	-0.00	-0.00	0.04
	(0.04)	(0.00)	(0.05)	(0.05)
Panel 2: Informed Evaluations (after performance information is provided)				
Female	-6.80***	-0.27***	-4.19***	-6.63***
	(0.53)	(0.03)	(0.61)	(0.59)
GPA (demeaned)	0.09**	0.00*	0.20***	0.29***
	(0.04)	(0.00)	(0.04)	(0.04)
*GPA (demeaned)*Female*	-0.11**	-0.00*	-0.06	-0.07
	(0.05)	(0.00)	(0.05)	(0.05)
N	10618	10618	10618	10618
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column (additional details on the question wording can be found in Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. *GPA (demeaned)* is administrative data on participants' "overall marking period GPA" that ranges from 35 to 102, demeaned by the average. Data are from the study among youth (i.e., middle-school and high-school students) excluding the youth for whom we do not have a GPA recorded. Panel 1 analyzes evaluations from before participants are informed of their absolute performance. Panel 2 analyzes evaluations from after participants are informed of their absolute performance.

B The *Employer* Version

We recruited 300 workers on MTurk to complete the *Employer* version of our study using the same criteria as in the main study versions (see footnote 4). Each employer received a guaranteed $1.50 completion fee for the 15-minute study. In addition, two of their decisions, out of the 21 decisions in the study, were selected to determine a possible bonus payment for them and for associated "workers," participants in the *Self-Promotion* and *Self-Promotion (Risky)* study versions.

For each decision, employers are informed that they must decide whether to hire a worker, and, if so, how much to pay that worker. If an employer chooses not to hire a worker, the employer earns a bonus of 100 cents and the worker earns a bonus of 25 cents. If an employer chooses to hire a worker, the employer must also choose a wage between 25 and 100 cents. The worker will receive that wage, and that employer's bonus payment will equal 100 cents minus the wage paid to the worker plus 5 cents times the number of questions the worker answered correctly on the math and science test. The only information an employer receives about a worker before hiring them is how the worker answered one of the four self-evaluation questions.

Employers make hiring and wage decisions via the strategy method. That is, their decisions involve each of the six possible answers selected from the performance-bucket evaluation question and five randomly selected answers (i.e., numbers from 0 to 100) from each of the three other evaluation questions.[21] Two decisions that are randomly selected then result in corresponding payments for the employer and for the two workers who provided answers that corresponded with those decisions. See Appendix D.9 for screenshots and additional details.

As expected, self-promotion pays. Employers respond to more positive answers to self-evaluation questions by being more likely to hire workers and by paying them more. Table B.1 shows how higher answers to self-evaluation questions affect wages given to workers. In all specifications, the coefficient on *Answer* is positive and significant. Columns (1), (3), and (4) show that the wage given to workers increases by an average of 0.21 or 0.22 cents for every point on the 0-to-100 scale in response to the performance question, the willingness-to-apply question, and the success question. Column (2) shows that the wage given to workers increases by an average of 4.26 cents for each increase on the six-point Likert scale in the performance-bucket question. Table B.2 shows that these results do not vary according to the gender of the employer. These results highlight that participants have an incentive to inflate their answers to self-evaluation questions to increase their expected study earnings in the *Self-Promotion* version.

[21] As noted above, the three other evaluation questions ask participants to state their agreement with the following statements: "I performed well on the test I took in part 1," "I would apply for a job that required me to perform well on the test I took in part 1," and "I would succeed in a job that required me to perform well on the test I took in part 1." Employers face all hiring decisions related to one question before moving on to the next question, but the order in which they face answers to each question is randomized.

Table B.1: *Employer* Version, Wage Regressions

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
Answer	0.21***	4.26***	0.22***	0.21***
	(0.02)	(0.27)	(0.02)	(0.02)
Constant	22.70***	18.95***	21.94***	22.76***
	(0.75)	(0.70)	(0.61)	(0.78)
N	1490	1788	1490	1490

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered by employer. Results are from OLS regressions of the wage received by the participant (25 cents if not hired and a chosen wage from 25–100 cents if hired). *Answer* is the answer provided by each participant in the question noted in each column, as defined in the notes of Table 2. Data are from the hiring decisions in the *Employer* version.

Table B.2: *Employer Version*, Wage Regressions

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Answer	0.21***	4.23***	0.23***	0.22***
	(0.03)	(0.41)	(0.02)	(0.02)
Female Employer	-1.30	-2.23	-1.37	-1.21
	(1.51)	(1.40)	(1.24)	(1.56)
*Answer*Female Employer*	-0.01	0.06	-0.01	-0.02
	(0.03)	(0.55)	(0.03)	(0.03)
Constant	23.37***	20.11***	22.66***	23.39***
	(1.28)	(1.08)	(1.03)	(1.22)
N	1490	1788	1490	1490

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.SEs are clustered by employer. Results are from OLS regressions of the wage received by the participant (25 cents if not hired and a chosen wage from 25–100 cents if hired). *Answer* is the answer provided by each participant in the question noted in each column, as defined in the notes of Table 2. *Female Employer* is an indicator for the employer being female. Data are from the hiring decisions in the *Employer* version.

C The *Free-Response* Versions

In February 2019, we recruited 399 participants on MTurk to complete either the *Free-Response Employer* version (n=198) or the *Free-Response Predictor* version (n=201) of our study using the same criteria as in the main study versions (see footnote 4). Each participant received a guaranteed $1.50 completion fee for the 15-minute study. In addition, one of their decisions, out of the 21 decisions in the study, was selected to determine a possible bonus payment for them and, in the *Free-Response Employer* version, for an associated "worker."[22] After participants completed all decisions of the study, they took a short follow-up survey that collected demographic information.

In the *Free-Response Employer* version, participants made 21 hiring decisions. In the *Free-Response Predictor* version, participants made 21 sets of predictions. Before making each decision or set of predictions, participants were provided with the text entered by a wave 1 participant to the free-response question: "Please describe how well you think you performed on the test that you took in part 1 and why." The free response either came from part 2 or part 3. Participants were randomly assigned these 21 free responses from the set of eligible free responses written by the participants from wave 1.[23]

Participants assigned to the *Free-Response Employer* version were asked whether they would like to hire the participant who provided that free response and, if so, how much to pay them. The payoffs for these employers are the same as described in the *Employer* version. While similar to the *Employer* version, there are many more possible free responses than answers to the quantitative self-evaluation questions, which means our analysis on hiring decisions is underpowered relative to the *Employer* version, since we only have at most a few employers reacting to each free response.

Participants assigned to the *Free-Response Predictor* version were instead asked to predict whether the participant who wrote the free response was male or female and how many questions, out of 20, that participant answered correctly on the math and science test. The payoffs for predictors were determined as follows. One of the two predictions from one of the 21 sets was randomly selected. If the prediction was correct, the predictor received a bonus payment of 50 cents.

Relative to the *Employer* version, there are three important differences when considering results from the *Free-Response* versions. First, since there is no objective way to rank free-response answers, we cannot examine how hiring decisions or predictions vary as the responses improve (as we did when examining, e.g.,

[22] Each participant who completed the *Self-Promotion* or *Self-Promotion (Risky)* versions of our study was matched with an employer from the *Employer* version of our study and received corresponding payoffs from their employers' hiring decisions. By contrast, only select participants from the *Self-Promotion* and *Self-Promotion (Risky)* versions were matched with a participant from a *Free-Response Employer* version, and received corresponding payoffs, rather than everyone. Since we also wanted to collect data on the free responses from the *Private* version, participants in the *Free-Response Employer* version were (accurately) told that one of their decisions would be selected to count but *not* that one of their decisions would be randomly selected to count (as this would have required putting 0% weight on free responses from the *Private* version in the randomization).

[23] Not all of the free responses collected in the study were evaluated. First, the *Free-Response* versions were run after wave 1 but before wave 2, so free responses from waves 2–5 did not yet exist. We consequently consider the 1800 free responses from the *Self-Promotion* version, the *Self-Promotion (Risky)* version, and the *Private* version run in wave 1. Second, a research assistant—blinded to participant gender and study version—deemed 130 of the 1800 potentially eligible free responses "ineligible" due to the answer not relating to the question asked or due to severe grammar and/or spelling issues that made an answer incomprehensible. Consequently, the participants were each randomly shown 21 free-responses from the set of 1670 eligible free responses. Finally, note that some eligible free-responses were never randomly selected to be shown to a participant.

the impact of a one unit increase on a 0-to-100 scale in the *Employer* version). Second, while participants are not informed of the gender of the individual who answered the free-response question, they may be able to infer gender—to some degree—given how the free responses are written. Below, we test this hypothesis using data from the predictors. Third, as noted above, given the large number of possible free responses, we are underpowered to consider the effect of specific free responses.

For these reasons, we favor the analysis of the quantitative responses to the self-evaluation questions presented in the main text to examine the gender gap in self-promotion. Here, however, we investigate the hiring decisions and predictions from the *Free-Response* versions to present several interesting (but inherently secondary) results. Given our power issues, we combine data from all three study versions run in wave 1.

Table C.1 presents results from regressions testing whether the gender of the free response author affects predictions and hiring decisions. Columns (1) and (2) are estimated from predictions from the *Free-Response Predictor* version. The negative coefficients on *Female* in column (1) show that participants predict (at least directionally) lower scores when reading free responses authored by female participants. This evidence is consistent with our findings from the quantitative self-evaluation questions discussed in the main text—women appear to provide less favorable subjective evaluations of their performance, even in the free responses. The positive coefficients on *Female* in column (2) show that, even though predictors are not informed of the gender of the participant who authored the free response, evaluators can infer gender—to some degree—when viewing the responses. Predictors are significantly more likely to predict that a response was written by a female participant when it was indeed written by a female participant. Column (3) is estimated from hiring decisions from the *Free-Response Employer* version. Based on the free response answers, employers pay at least directionally less to female workers.

An important caveat to the analysis in the prior paragraph, however, is that since evaluators can infer the gender of the associated worker based off of the free responses, the predictions of performance and hiring decisions may be influenced by the perception of the gender of the free response author (e.g., predictors might expect women to perform worse than men; employers may want to pay women more than men based on social preferences, etc.), which makes it difficult to isolate the effect of the language used in the free response (i.e., the self-promotion). As mentioned in footnote 9 and in the main text of the paper, difficulties with using free responses, and other qualitative data, contributes to our decision to focus our analysis on the quantitative self-evaluation questions we explore in the main text of the paper.

Table C.1: Free Response Regressions

	Predicted Performance	Predicted Probability Female	Wage
	(1)	(2)	(3)
Panel 1: Free responses (before performance information is provided)			
Female	-0.67***	0.08***	-1.44*
	(0.22)	(0.03)	(0.81)
N	749	749	743
Panel 2: Free responses (after performance information is provided)			
Female	-0.35	0.09***	-0.66
	(0.23)	(0.03)	(1.04)
N	773	773	755
Performance FEs	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the noted dependent variable (DV). *Predicted Performance* equals the predictor's guess of the number of questions the participant answered correctly based on the free response. *Predicted Probability Female* equals the probability that the predictor placed on the participant being female. *Wage* equals the wage given to the participant by an employer. In cases where multiple participants responded to the same free response, we use the average decision (e.g., if a free response is predicted to be written by a female participant once and a male participant once, that participant is recorded as being predicted to be female with a 0.50 probability). *Female* is an indicator for the participant who wrote the free response being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test of the participant who wrote the free response. Data in Panel 1 are from free responses elicited before performance information is provided, and data in Panel 2 are from free responses elicited after performance information is provided. Data are from all three study versions run in wave 1: the *Self-Promotion* version, the *Self-Promotion (Risky)* version, and the *Private* version.

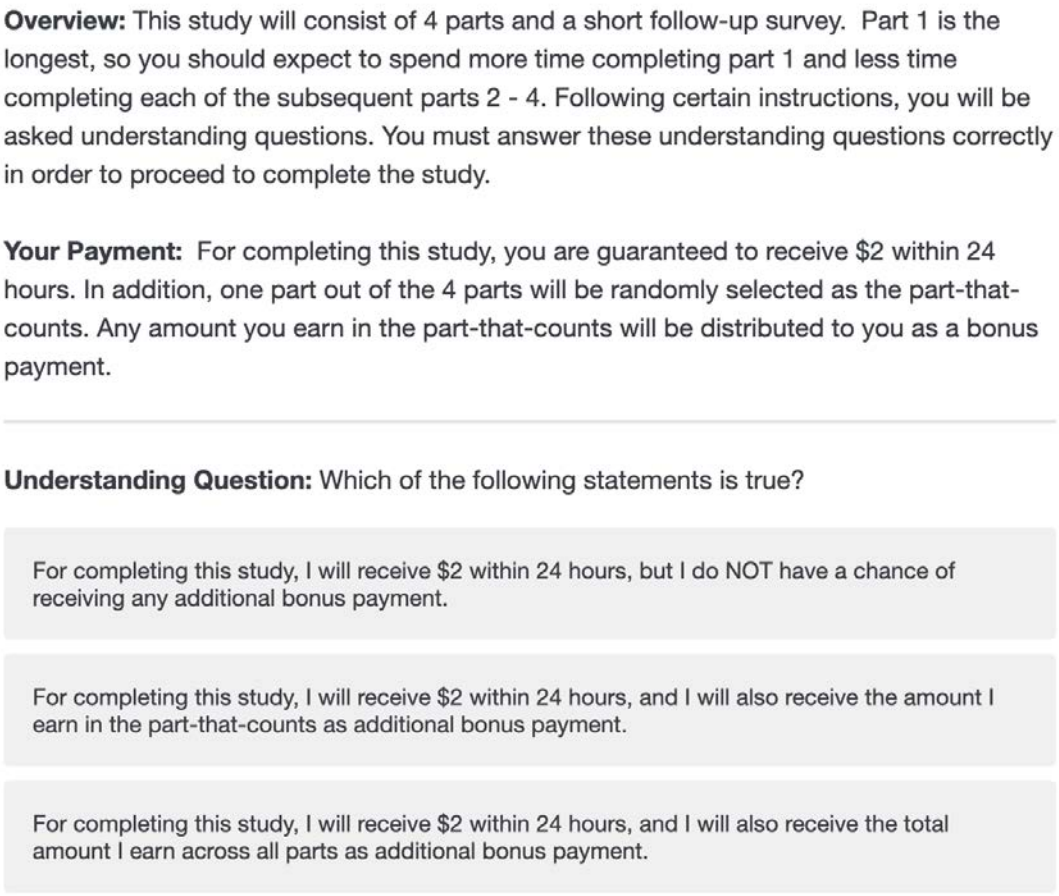
D Experimental Instructions

D.1 Instructions for *Self-Promotion* version

Prior to participating in the study, participants must correctly answer a captcha and consent to participate. At the end of the study, participants must complete a short follow-up survey to gather demographic information.

The study begins by informing each participant of the $2 study completion fee and of the opportunity to earn additional payment for themselves. Figure D.1 shows how this payment information is explained along with the understanding question that the participant must answer correctly to proceed.

Figure D.1: Payment Information



The instructions for part 1 are displayed in Figures D.2 and an example of an ASVAB question is displayed in Figure D.3 (note that the timer in that screenshot indicates the participant has 23 seconds left to answer the question although the timer starts at 30 seconds).

Figure D.2: Instructions for Part 1

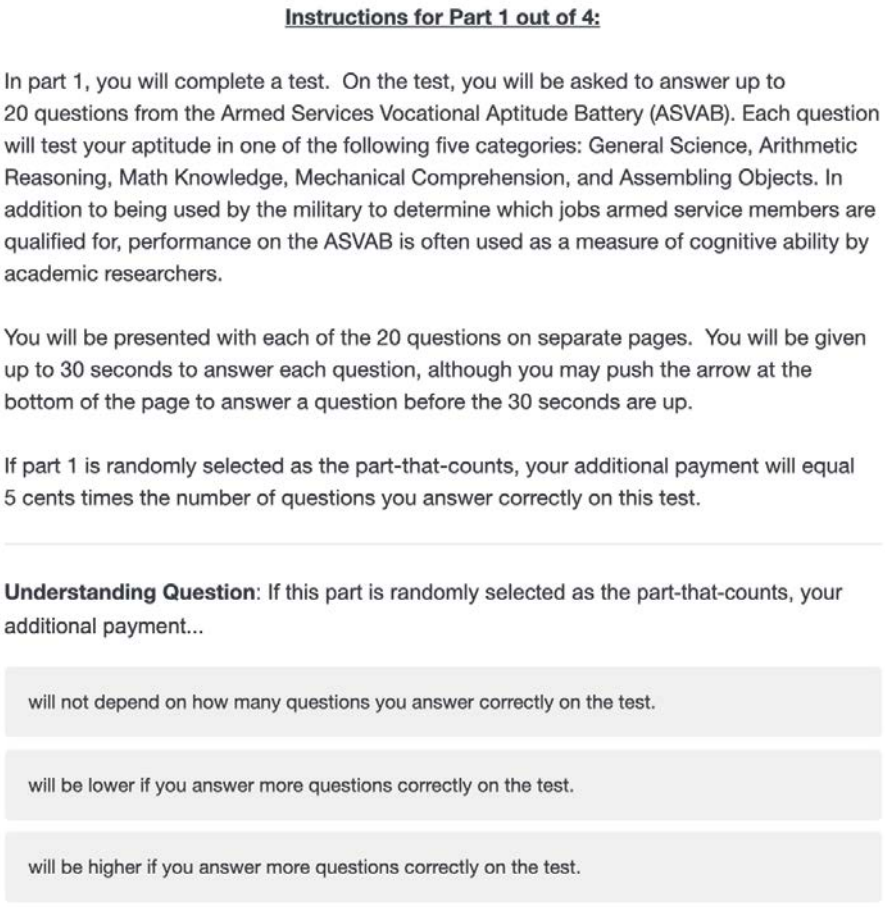
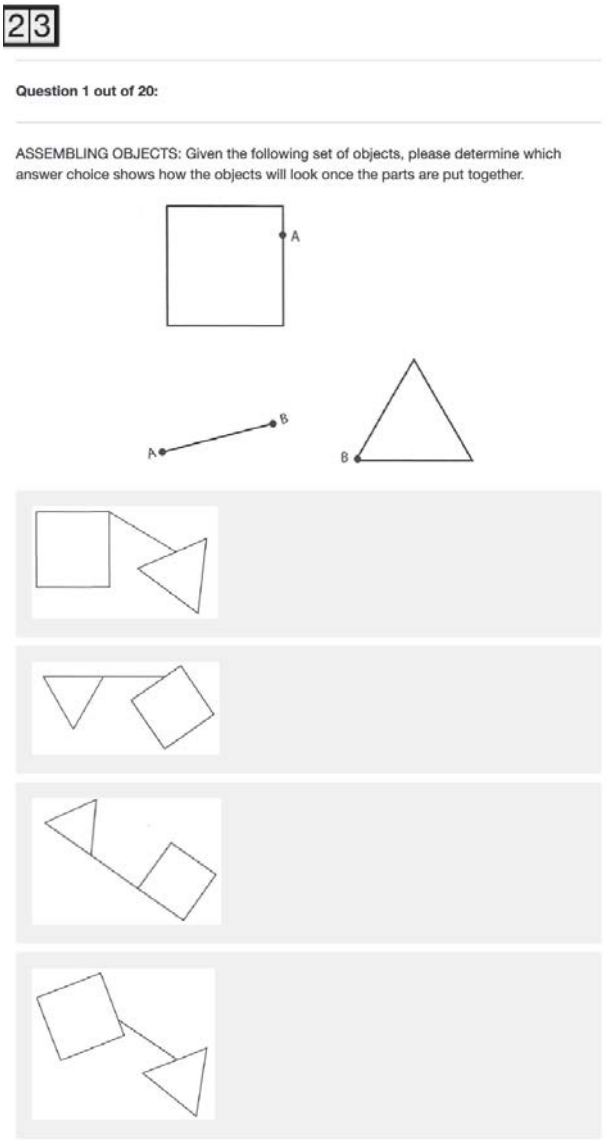
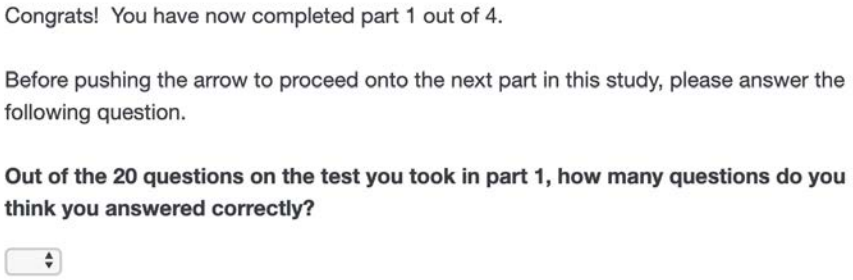


Figure D.3: Part 1: Example ASVAB question



After completing the ASVAB questions in part 1 but before proceeding to part 2, participants are asked about their absolute performance belief, as shown in Figure D.4.

Figure D.4: Absolute Performance Belief Question



Participants then receive instructions for part 2 (see Figure D.5), must correctly answer understanding questions about those instructions (see Figure D.6), and then are asked the self-evaluation questions (see Figure D.7).

Figure D.5: Part 2 Instructions

Instructions for Part 2 out of 4:

In part 2, you will be asked several questions -- on the next page -- related to your performance on the test you completed in part 1.

One of your answers to these questions will be shown to "your part 2 employer," who will be another MTurk worker who completes a different version of this study. Your part 2 employer can decide whether to hire you and, if so, how much to pay you.

Prior to deciding whether to hire you and, if so, how much to pay you, your part 2 employer will NOT be informed of how many questions you answered correctly on the test in part 1.

If this part is randomly selected as the part-that-counts, the additional payment given to your part 2 employer and to you will be determined as follows:

- If your part 2 employer chooses NOT to hire you, your additional payment will equal 25 cents and your part 2 employer's additional payment will equal 100 cents.

- If your part 2 employer chooses to hire you, your additional payment will equal how much they choose to pay you, and your part 2 employer's additional payment will equal (i) 100 cents minus how much they choose to pay you, plus (ii) 5 cents times the number of questions you answered correctly on the test in part 1. Your part 2 employer can choose to pay you any amount between 25 cents and 100 cents.

Figure D.6: Part 2 Understanding Questions

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

will equal 25 cents for sure.

will equal 5 cents times the number of questions you answered correctly on the test in part 1.

will equal 25 cents if you are not hired and the amount your part 2 employer chooses to pay you if you are hired.

Understanding Question: When deciding how much to pay you, your part 2 employer will only know...

how many questions you answered correctly on the test you took in part 1.

how you answer one of the questions -- on the next page -- about your performance on the test you took in part 1.

how you answer all of the questions -- on the next page -- about your performance on the test you took in part 1.

Figure D.7: Part 2 Self-Evaluation Questions

Now, please answer the five questions below to complete part 2. Note that, although the final three questions appear in the same block, they are three separate questions.

Please describe how well you think you performed on the test that you took in part 1 and why.

Please indicate how well you think you performed on the test you took in part 1.

Terrible	Very Poor	Neutral	Good	Very Good	Exceptional

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with each of the following statements:

Entirely Disagree	Strongly Disagree	Disagree	Somewhat Disagree	Neither Disagree Nor Agree	Somewhat Agree	Agree	Strongly Agree	Entirely Agree
0	10	20	30	40	50	60	70	80	90	100

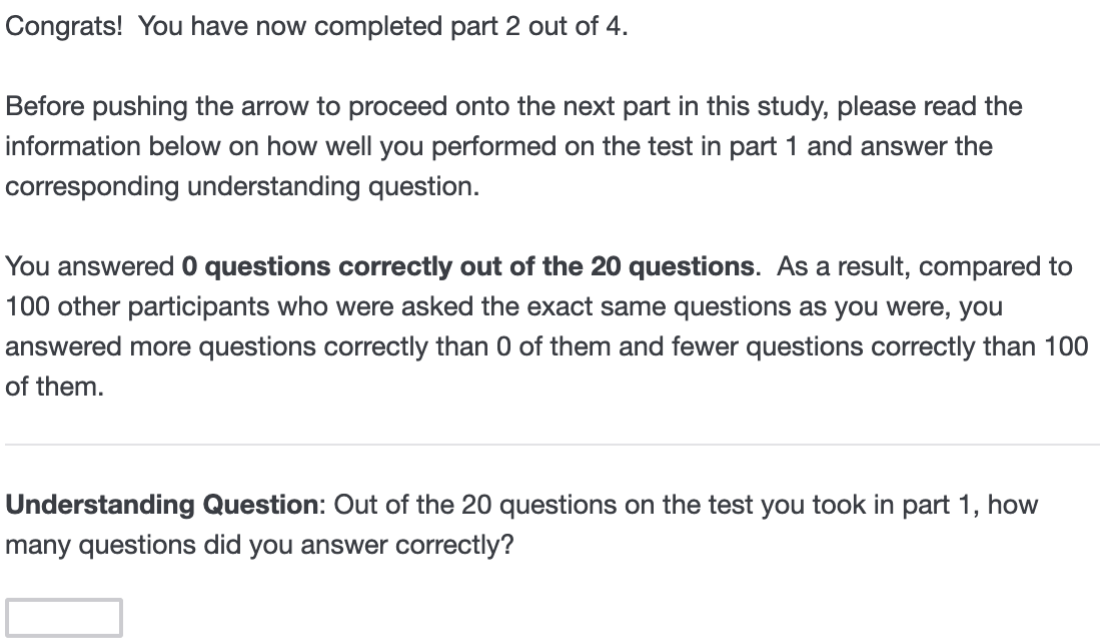
I performed well on the test I took in part 1.

I would apply for a job that required me to perform well on the test I took in part 1.

I would succeed in a job that required me to perform well on the test I took in part 1.

After completing part 2, participants are provided with perfect information on their absolute and relative performance and are required to correctly report back their absolute performance as shown in Figure D.8.

Figure D.8: Absolute and Relative Performance Information



In part 3, participants are provided with the same instructions (see Figure D.9), understanding questions (see Figure D.10), and self-evaluation questions (see Figure D.11) as they were in part 2.

Figure D.9: Part 3 Instructions

Instructions for Part 3 out of 4:

In part 3, you will be asked several questions -- on the next page -- related to your performance on the test you completed in part 1.

One of your answers to these questions will be shown to "your part 3 employer," who will be another MTurk worker who completes a different version of this study. Your part 3 employer can decide whether to hire you and, if so, how much to pay you.

Prior to deciding whether to hire you and, if so, how much to pay you, your part 3 employer will NOT be informed of how many questions you answered correctly on the test in part 1 (even though you were informed of this information on the previous page).

If this part is randomly selected as the part-that-counts, the additional payment given to your part 3 employer and to you will be determined as follows:

- If your part 3 employer chooses NOT to hire you, your additional payment will equal 25 cents and your part 3 employer's additional payment will equal 100 cents.

- If your part 3 employer chooses to hire you, your additional payment will equal how much they choose to pay you, and your part 3 employer's additional payment will equal (i) 100 cents minus how much they choose to pay you, plus (ii) 5 cents times the number of questions you answered correctly on the test in part 1. Your part 3 employer can choose to pay you any amount between 25 cents and 100 cents.

Figure D.10: Part 3 Understanding Questions

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

will equal 25 cents for sure.

will equal 5 cents times the number of questions you answered correctly on the test in part 1.

will equal 25 cents if you are not hired and the amount your part 3 employer chooses to pay you if you are hired.

Understanding Question: When deciding how much to pay you, your part 3 employer will only know...

how many questions you answered correctly on the test you took in part 1.

how you answer one of the questions -- on the next page -- about your performance on the test you took in part 1.

how you answer all of the questions -- on the next page -- about your performance on the test you took in part 1.

Figure D.11: Part 3 Self-Evaluation Questions

Now, please answer the five questions below to complete part 3. Note that, although the final three questions appear in the same block, they are three separate questions.

Please describe how well you think you performed on the test that you took in part 1 and why.

Please indicate how well you think you performed on the test you took in part 1.

Terrible	Very Poor	Neutral	Good	Very Good	Exceptional

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement:

Entirely Disagree		Strongly Disagree		Disagree		Somewhat Disagree		Neither Disagree Nor Agree		Somewhat Agree		Agree		Strongly Agree		Entirely Agree
0	10		20		30		40		50		60		70		80		90	100

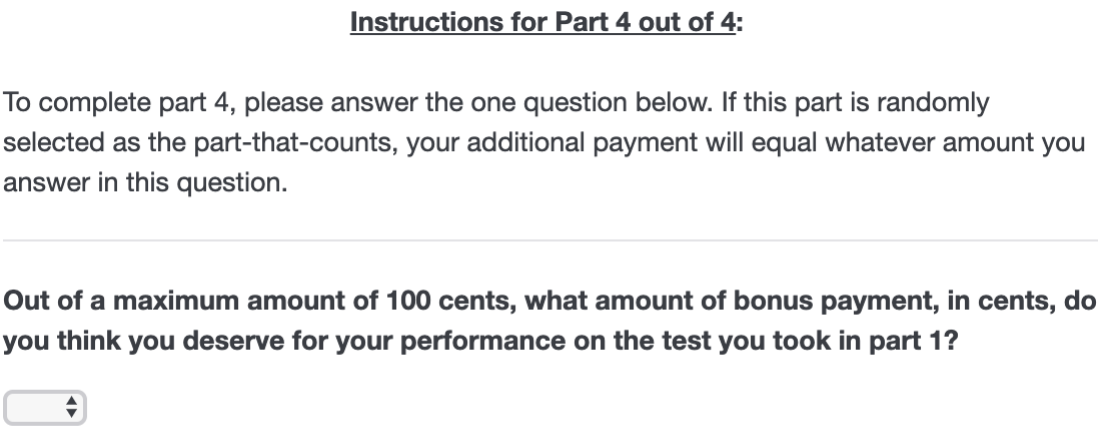
I performed well on the test I took in part 1.

I would apply for a job that required me to perform well on the test I took in part 1.

I would succeed in a job that required me to perform well on the test I took in part 1.

Finally, participants receive instructions about and are asked to answer the deservingness question in Part 4 (see Figure D.12). They then answer demographic questions, including the one that asks about their gender.

Figure D.12: Part 4 Instructions and Deservingness Question



D.2 Instructions for the *Self-Promotion (Risky)* version

The *Self-Promotion (Risky)* version of the study proceeds in the same manner as the *Self-Promotion* version of the study, except for the instructions about part 2 and part 3. Participants are informed that there is some chance that their employer will learn their actual performance. See Figures D.13 and D.14 for these instructions and the corresponding understanding questions, respectively.

Figure D.13: The *Self-Promotion (Risky)* version: Part 2 Instructions

Instructions for Part 2 out of 4:

In part 2, you will be asked several questions -- on the next page -- related to your performance on the test you completed in part 1.

There is some chance that one of your answers to these questions will be shown to "your part 2 employer," who will be another MTurk worker who completes a different version of this study. Your part 2 employer can decide whether to hire you and, if so, how much to pay you.

Prior to deciding whether to hire you and, if so, how much to pay you, there is also some chance that your part 2 employer will be informed of how many questions you answered correctly on the test in part 1.

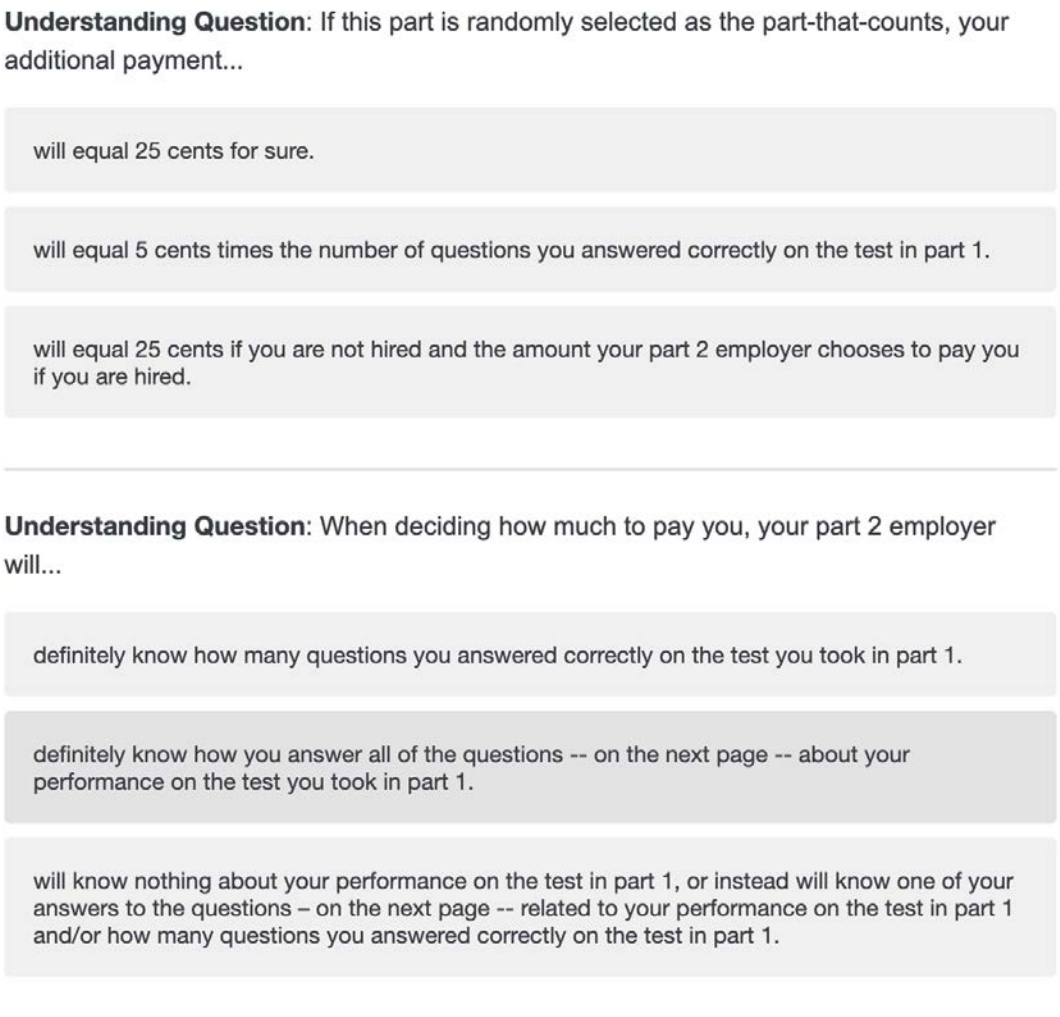
However, while your part 2 employer may learn one of your answers to the questions -- on the next page -- related to your performance on the test in part 1 and/or how many questions you answered correctly on the test in part 1, it is also possible that your part 2 employer will not learn any information related to your performance prior to deciding whether to hire you and, if so, how much to pay you.

If this part is randomly selected as the part-that-counts, the additional payment given to your part 2 employer and to you will be determined as follows:

- If your part 2 employer chooses NOT to hire you, your additional payment will equal 25 cents and your part 2 employer's additional payment will equal 100 cents.

- If your part 2 employer chooses to hire you, your additional payment will equal how much they choose to pay you, and your part 2 employer's additional payment will equal (i) 100 cents minus how much they choose to pay you, plus (ii) 5 cents times the number of questions you answered correctly on the test in part 1. Your part 2 employer can choose to pay you any amount between 25 cents and 100 cents.

Figure D.14: The *Self-Promotion (Risky)* version: Part 2 Understanding Questions



D.3 Instructions for the *Private* version

The *Private* version run in wave 1 proceeds in the same manner as the *Self-Promotion* version, except for the instructions about part 2 and part 3. Participants are simply informed that they will receive 25 cents regardless of how they answer the self-evaluation questions. See Figure D.15 for these instructions and the corresponding understanding question. The *Private* versions run in waves 2, 3, and 5 are identical to the *Private* version in the first wave, except for a slight formatting change in the part 2 and part 3 questions to allow for room to introduce the additional information in the *Private (Social Norms)* version. See Figure D.16 for the corresponding screenshot of the part 3 self-evaluation questions (and note that this is identical to how they appear in part 2).

Figure D.15: The *Private* version: Part 2 Instructions and Understanding Question

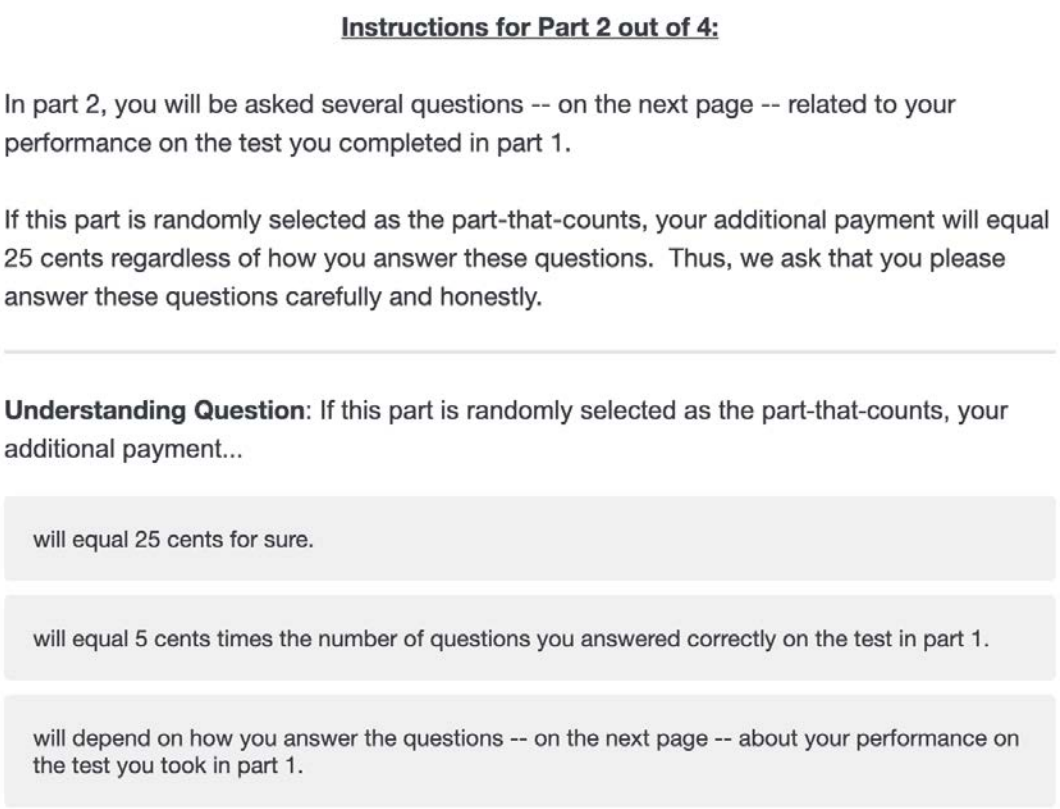
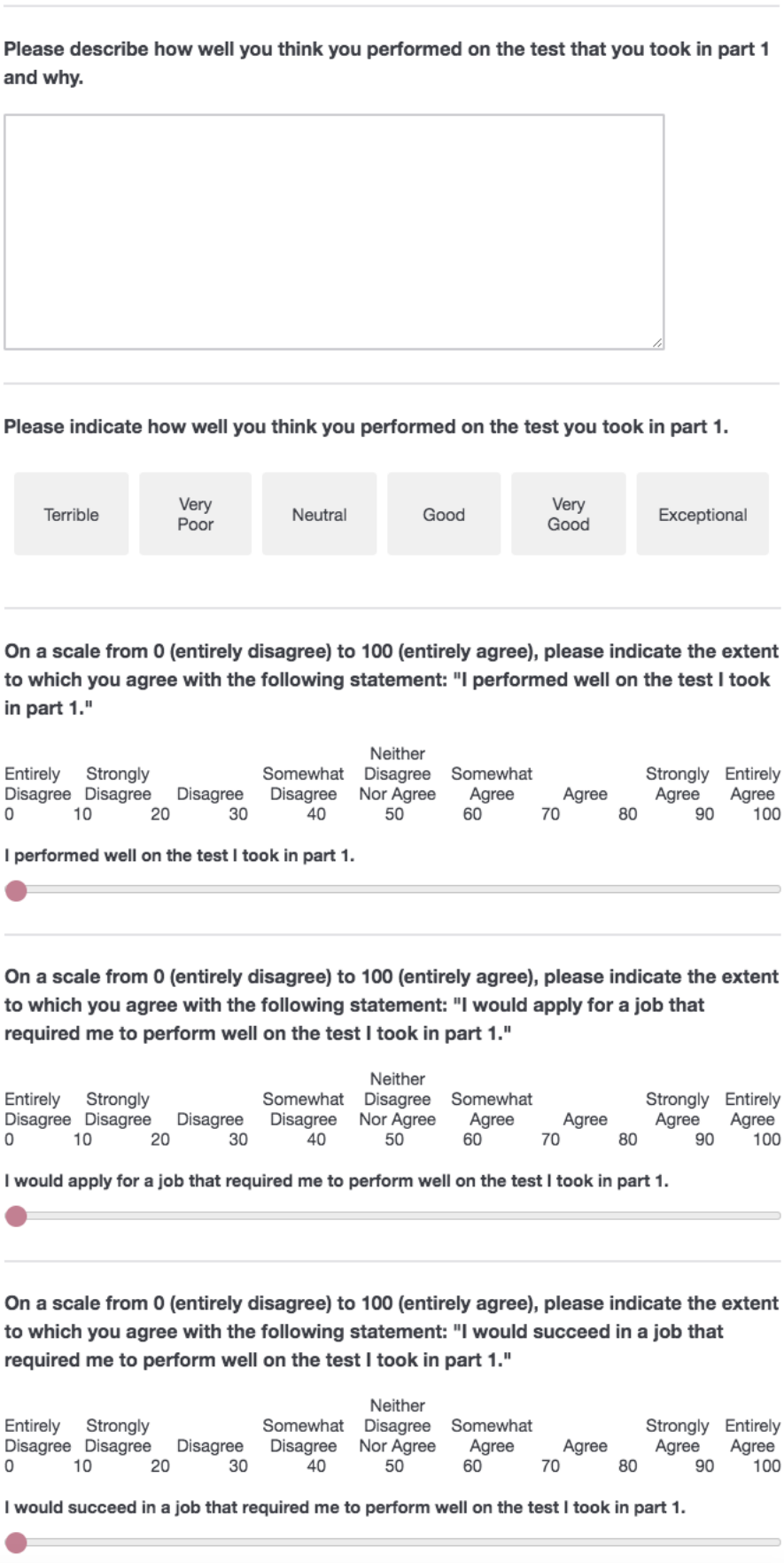


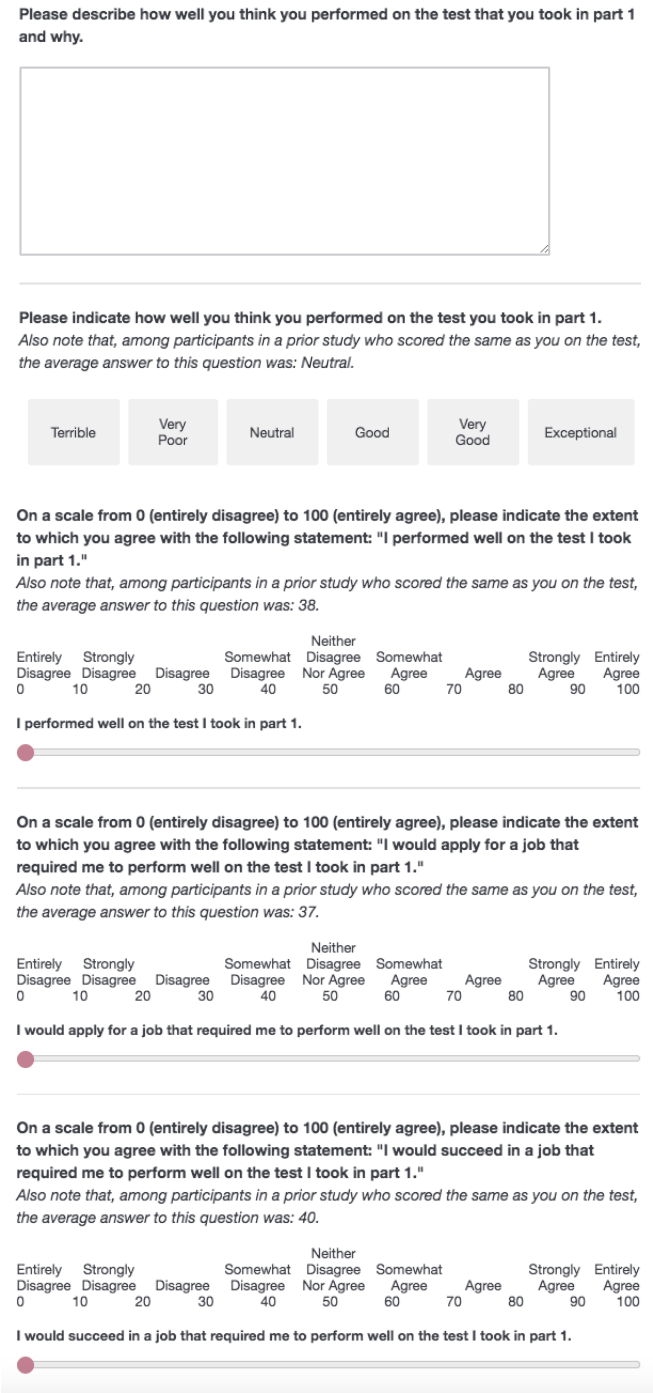
Figure D.16: The *Private* version: Part 3 Self-Evaluation Questions With a Slight Formatting Change



D.4 Instructions for the *Private (Social Norms)* version

The *Private (Social Norms)* version of the study proceeds in the same manner as the *Private* version of the study, except that, in part 3, additional information is provided on the average answer to each of the self-evaluation questions from prior participants with the same score as the participant. See Figure D.17 for the corresponding screenshot of the part 3 questions.

Figure D.17: The *Private (Social Norms)* version: Part 3 Self-Evaluation Questions for a Participant who Correctly Answered 10 out of 20 Questions



D.5 Instructions for the *Private (Immediately Informed)* version

The *Private (Immediately Informed)* version of the study proceeds in the same manner as the *Private* version of the study, except that participants learn their absolute and relative performance before answering any self-evaluation questions. That is, parts 3 and 4 in the *Private* version become parts 2 and 3 in this version so that the study proceeds as follows: participants complete the test in part 1, report their beliefs about their absolute performance on that test, are informed of their absolute and relative performance on that test, answer self-evaluation questions about that test in part 2, and answer the deservingness question in part 3.

D.6 Instructions for the *Private (Other-Evaluation)* version

The *Private (Other-Evaluation)* version proceeds in the same manner as the *Private (Immediately Informed)* version, except that participants are informed of the absolute and relative performance of another MTurk participant (see Figure D.18) and then are asked to provide informed other-evaluations about this other MTurk participant rather than themselves (see Figures D.19 and D.20).

Figure D.18: The *Private (Other-Evaluation)* version: Absolute and Relative Performance Information on Another MTurk Participant

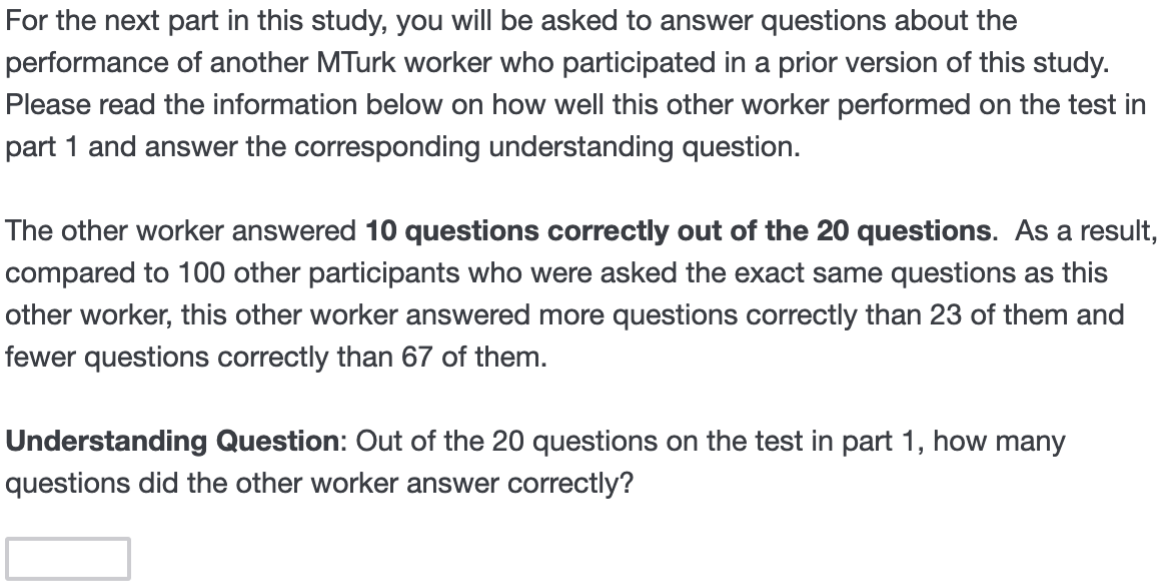


Figure D.19: The *Private (Other-Evaluation)* version: Part 2 Instructions and Understanding Questions

Instructions for Part 2 out of 3:

In part 2, you will be asked several questions -- on the next page -- related to the performance of the other worker, described on the previous page, on the test in part 1.

If this part is randomly selected as the part-that-counts, your additional payment will equal 25 cents regardless of how you answer these questions. Thus, we ask that you please answer these questions carefully and honestly.

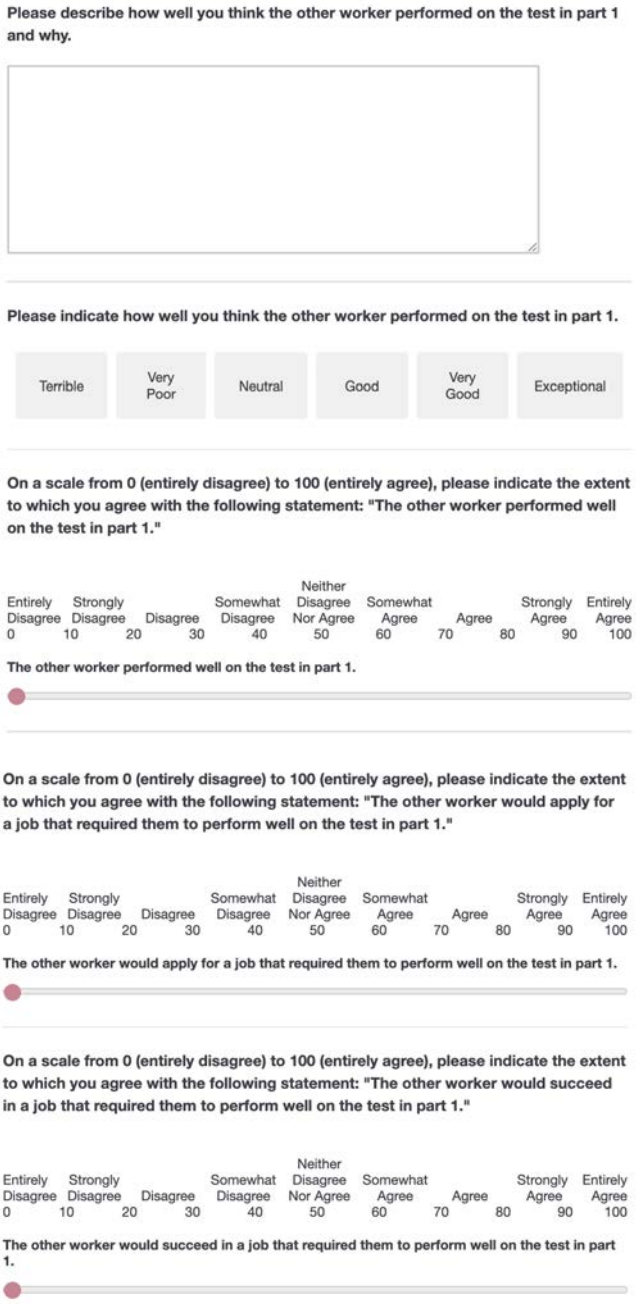
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

will equal 25 cents for sure.

will equal 5 cents times the number of questions you answered correctly on the test in part 1.

will depend on how you answer the questions -- on the next page -- about the performance of the other worker on the test in part 1.

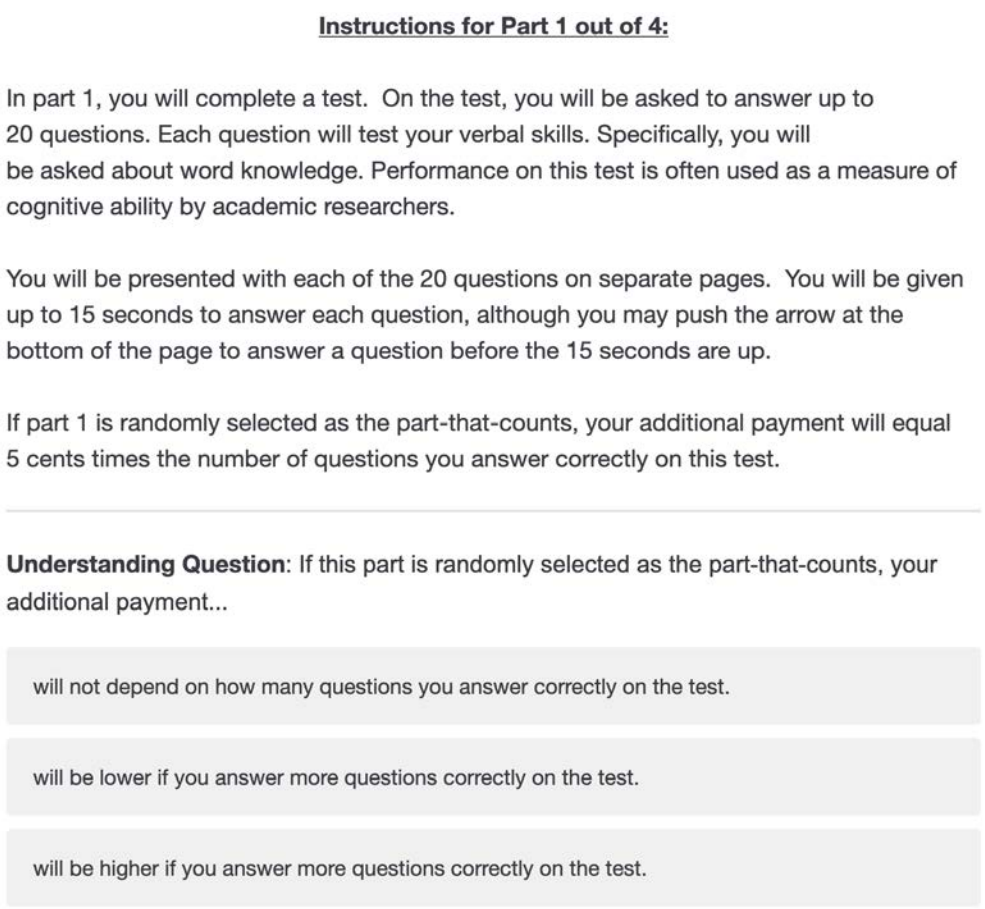
Figure D.20: The *Private (Other-Evaluation)* version: Part 2 Other-Evaluation Questions for Another Participant who Correctly Answered 10 out of 20 Questions



D.7 Instructions for the *Private (Verbal)* version

The *Private (Verbal)* version proceeds in the same manner as the *Private* version, except that the test that participants complete in part 1 asks them to answer 20 word knowledge questions rather than 20 math and science questions (see Figure D.21 for the instructions and Figure D.22 for an example question). In addition, there are two pages added to their follow-up survey that participants complete after they complete the other parts of the study.[24] As shown in Figure D.23, they learn (as a surprise) of the opportunity to earn additional bonus payment if they answer one of the eight prediction questions on the next two pages correctly. The order of the next two pages is randomly determined. On one of the pages, they are asked to answer four prediction questions about women (see Figure D.24). On the other page, they are asked to answer four prediction questions about men (see Figure D.25).

Figure D.21: Instructions for Part 1



[24] These same questions are also added to the *Private* version we ran in wave 5.

Figure D.22: Part 1: Example Verbal Question

Figure D.23: Instructions for Predictions

On the remaining two pages of the follow-up survey, you will be asked 8 questions. One of these questions will be randomly selected as the question-that-counts. If your answer to the question-that-counts is correct, you will receive an additional bonus payment of $0.50. This bonus payment will be in addition to any bonus payment you earned from the part-that-counts. **Thus, please answer these questions carefully and honestly to maximize your chance of earning more bonus payment.**

In each question, you will be asked to guess the performance of other participants in a prior study. These other participants answered 20 questions on a test from from the Armed Services Vocational Aptitude Battery (ASVAB). Each question tested their aptitude in one of the following five categories: General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects.

After completing this test, these participants were asked questions about their performance on the test. They were told that their response to one of the questions would be shared with an "employer," who would be another participant who completed a different version of this study. They were also told that their employer would decide whether to hire them, and if so, how much to pay them. They were also told that if they were hired, their employer would earn more money if they answered more questions correctly on the test.

Figure D.24: Predictions about Women

Workers were asked to to indicate whether they thought their performance on the test was terrible, very poor, neutral, good, very good, or exceptional.

Among the set of **female workers** who indicated to their employers that their performance on the test was **neutral**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I performed well on the test I took."

Among a set of **female workers** whose average response to the question above was **48 out of 100, somewhat disagreeing with the statement that they performed well**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: I would apply for a job that required me to perform well on the test I took."

Among a set of **female workers** whose average response to the question above was **45 out of 100, somewhat disagreeing with the statement that they would apply for such a job**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: I would succeed in a job that required me to perform well on the test I took."

Among a set of **female workers** whose average response to the question above was **49 out of 100, somewhat disagreeing with the statement that they would succeed in such a job**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Figure D.25: Predictions about Men

Workers were asked to to indicate whether they thought their performance on the test was terrible, very poor, neutral, good, very good, or exceptional.

Among the set of **male workers** who indicated to their employers that their performance on the test was **good**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I performed well on the test I took."

Among a set of **male workers** whose average response to the question above was **59 out of 100, somewhat agreeing with the statement that they performed well**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: I would apply for a job that required me to perform well on the test I took."

Among a set of **male workers** whose average response to the question above was **60 out of 100, somewhat agreeing with the statement that they would apply for such a job**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: I would succeed in a job that required me to perform well on the test I took."

Among a set of **male workers** whose average response to the question above was **65 out of 100, agreeing with the statement that they would succeed in such a job**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

D.8 Instructions for *Private* version run among youth

Prior to participating in the study, participants must correctly answer a captcha and consent to participate. At the end of the study, participants must complete a short follow-up survey to gather demographic information. Participants are recruited via the Character Lab Research Network and complete this study as part of the curriculum at school. There are no payments associated with this study.

The study begins by informing each participant about the test that they will take. The instructions for the test are displayed in Figure D.26 and an example of a question on the test is displayed as Figure D.27 (note that the timer in that screenshot indicates the participant has 24 seconds left to answer the question although the timer starts at 30 seconds).

Figure D.26: Instructions for the test

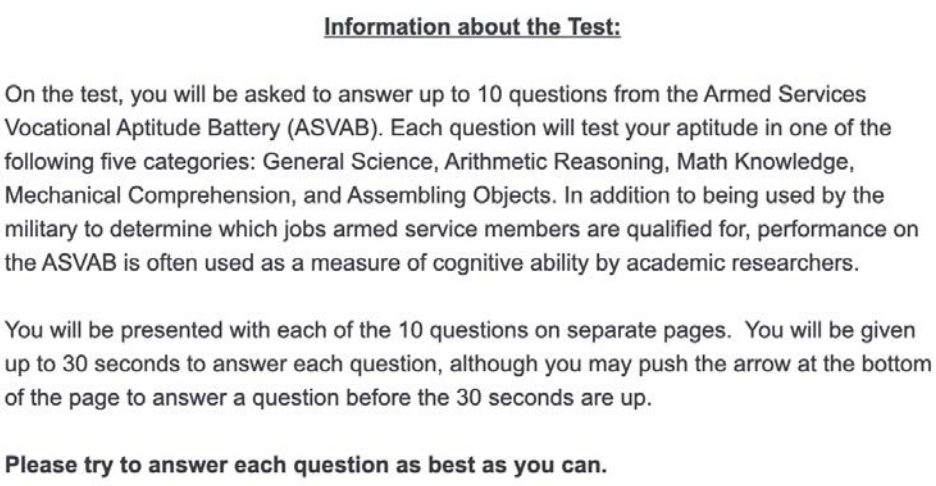
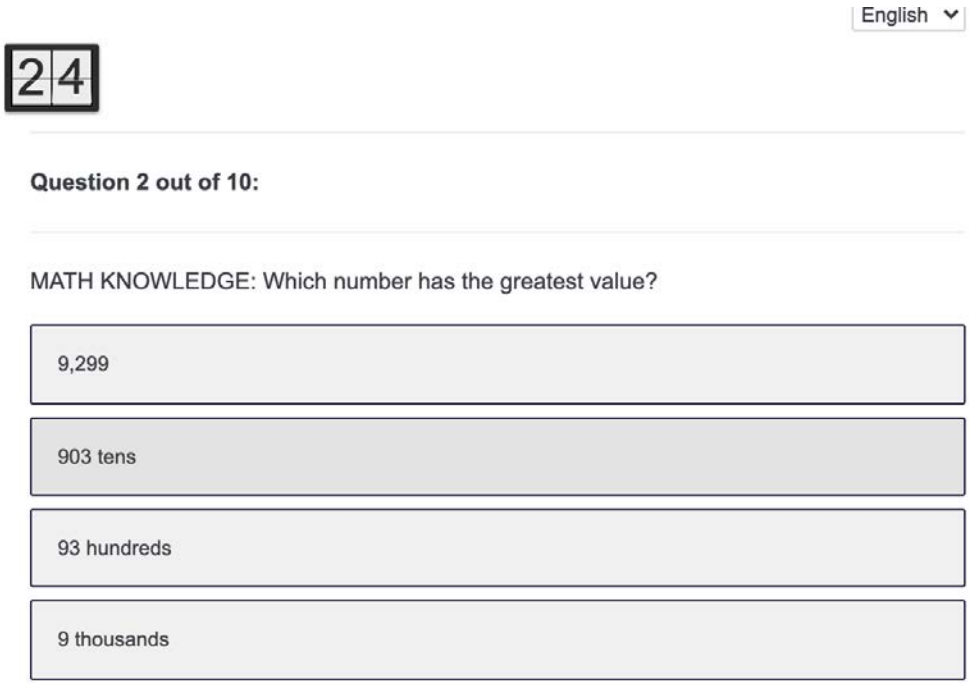
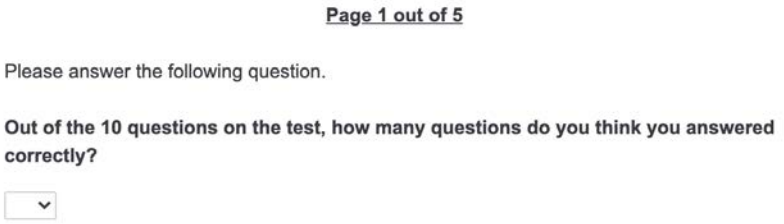


Figure D.27: Example question on the test



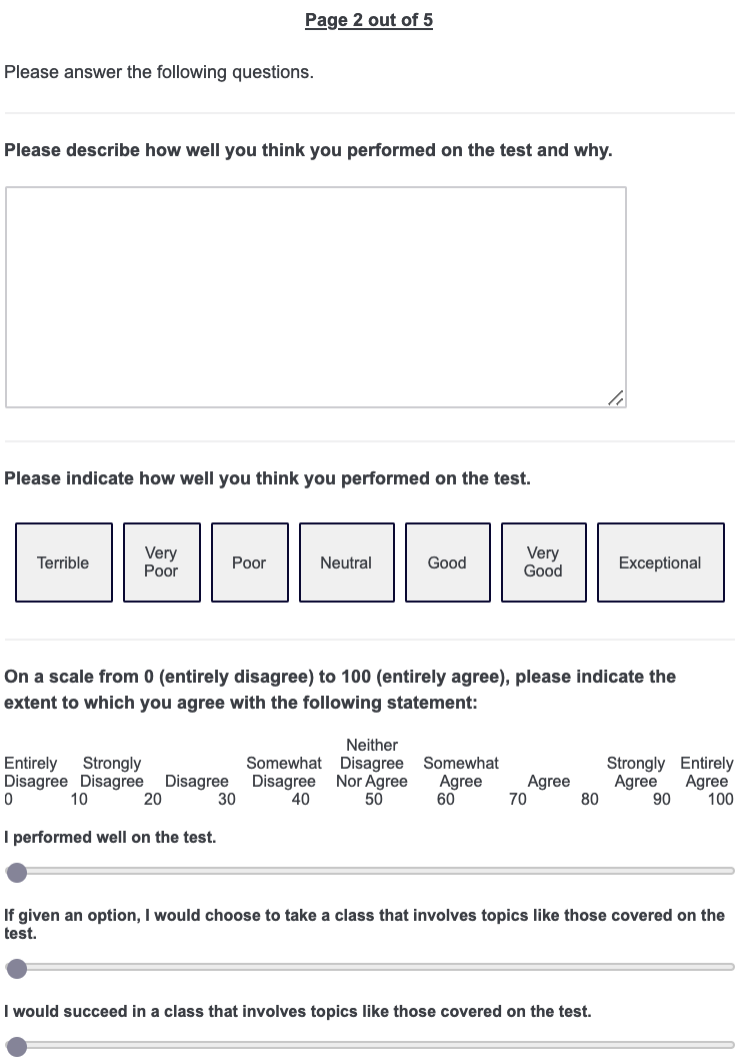
After completing the test, participants are asked to complete five additional pages of the study. On the first page, they are asked about their absolute performance belief, as shown in Figure D.28.

Figure D.28: Absolute Performance Belief Question



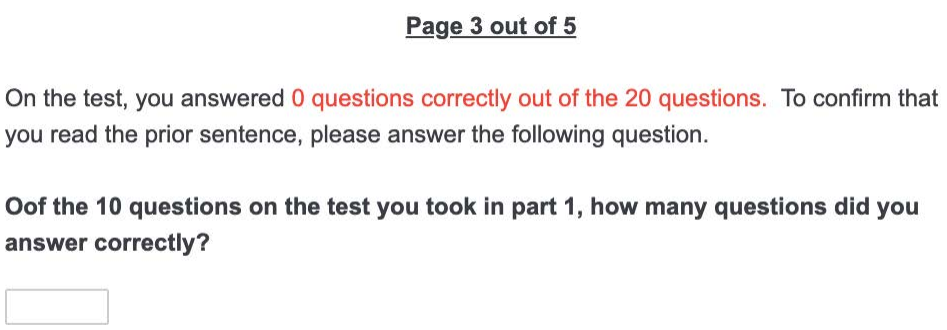
On the second page, they are asked the self-evaluation questions (see Figure D.29).

Figure D.29: Self-Evaluation Questions



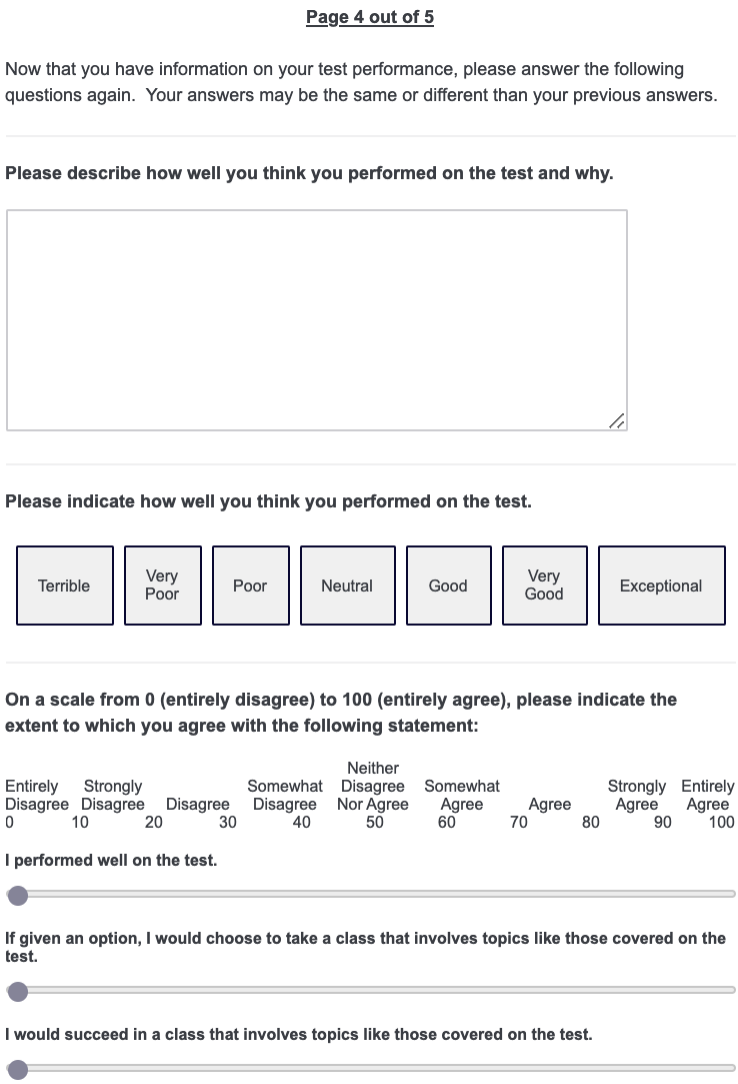
On the third page, participants are provided with perfect information on their absolute performance and are required to correctly report back their absolute performance as shown in Figure D.30.

Figure D.30: Absolute Performance Information



On the fourth page, they are asked the self-evaluation questions again (see Figure D.31). On the fifth page, they are asked for demographic information.

Figure D.31: Informed Self-Evaluation Questions



D.9 Instructions for *Employer* version

Prior to participating in the study, participants must correctly answer a captcha and consent to participate in the study. At the end of the study, participants must complete a short follow-up survey to gather demographic information.

The study begins by informing each participant of the $1.50 study completion fee and of the opportunity to earn additional payment. Figure D.32 shows how this payment information is explained. Figure D.33 shows the understanding questions that the participant must answer correctly to proceed.

Figure D.32: Payment Information

Overview:
This study will consist of 21 decisions 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, MTurk workers completed a test. On the test, they were asked to answer up to 20 questions from the Armed Services Vocational Aptitude Battery (ASVAB). Each question tested their aptitude in one of the following five categories: General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. In addition to being used by the military to determine which jobs armed service members are qualified for, performance on the ASVAB is often used as a measure of cognitive ability by academic researchers.

Your Decisions:
For each of the 21 decisions, you will be matched with one worker from the piror study. You then must decide whether to hire that worker, and if so, how much to pay that worker.

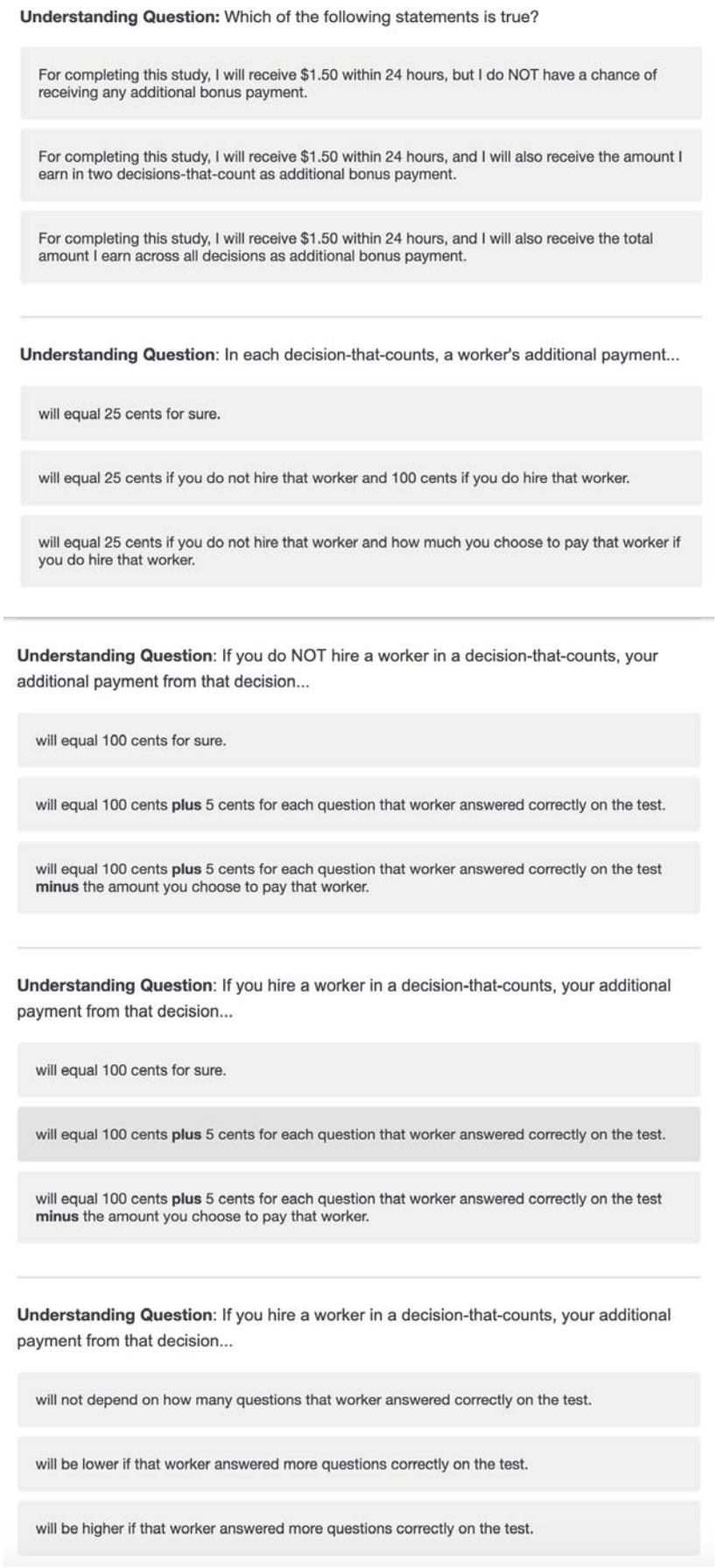
After you make all of your 21 decisions, two decisions will be selected as a decision-that-counts.

In each decision-that-counts, the additional payment given to the worker in that decision and to you will be determined as follows:

- If you choose NOT to hire the worker, that worker's additional payment will equal 25 cents and your additional payment will equal 100 cents.

- If you choose to hire the worker, that worker's additional payment will equal how much you choose to pay them, and your additional payment will equal (i) 100 cents minus how much you choose to pay them, plus (ii) 5 cents times the number of questions that worker answered correctly on the test. Your can choose to pay that worker any amount between 25 cents and 100 cents.

Figure D.33: Understanding Questions of Payment Information



The 21 decisions that employers face involve four blocks. Three blocks relate to the three evaluation questions that involve the 0-to-100 scale (i.e., the performance question, the willingness-to-apply question and the success question), and each of these blocks involves five decisions that correspond to five randomly selected evaluations (i.e., numbers from 0 to 100). Another block relates to the evaluation question involving a six point Likert-scale (i.e., the performance-bucket question), and this block involves six decisions that correspond to each of the six possible evaluations in that question. The order of these four blocks is randomized on the participant-level.

The instructions for, and examples of, decisions relating to the *performance* evaluations are displayed in Figures D.34 and D.35, respectively.

Figure D.34: Instructions for *Performance Evaluation* Decisions

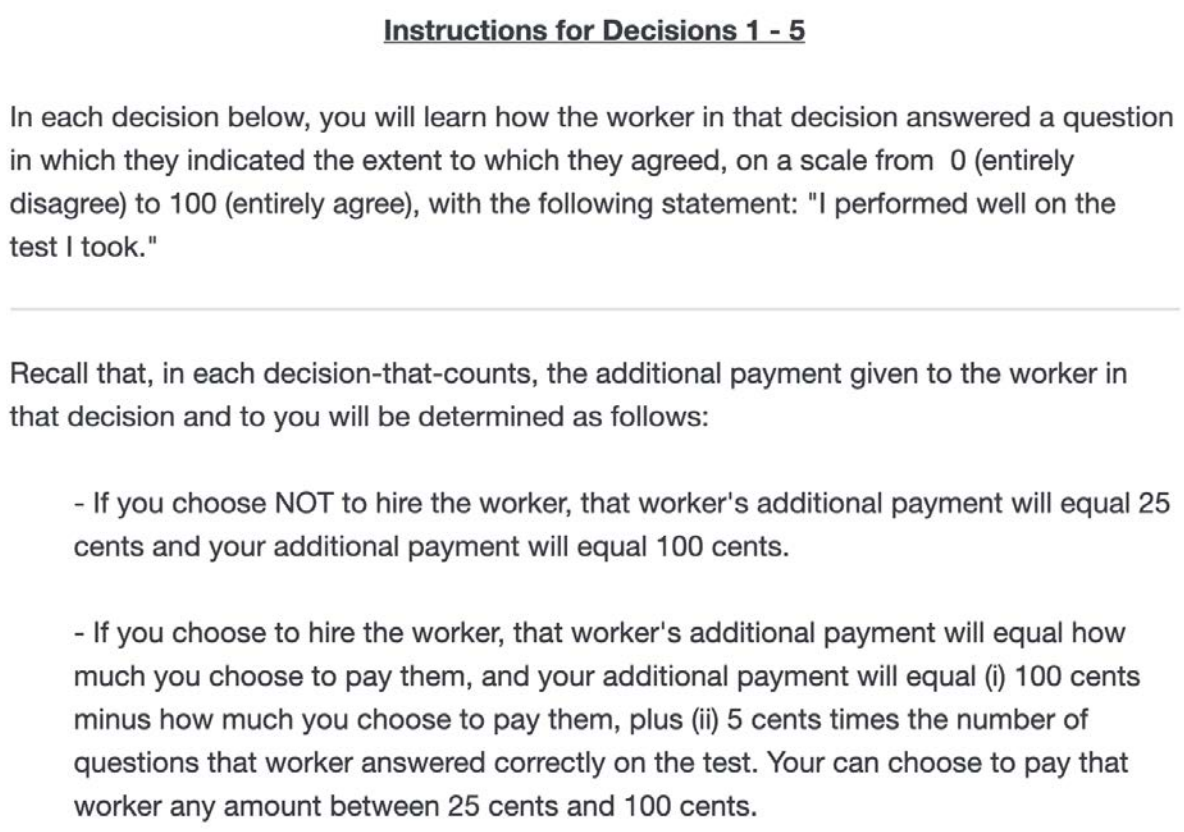
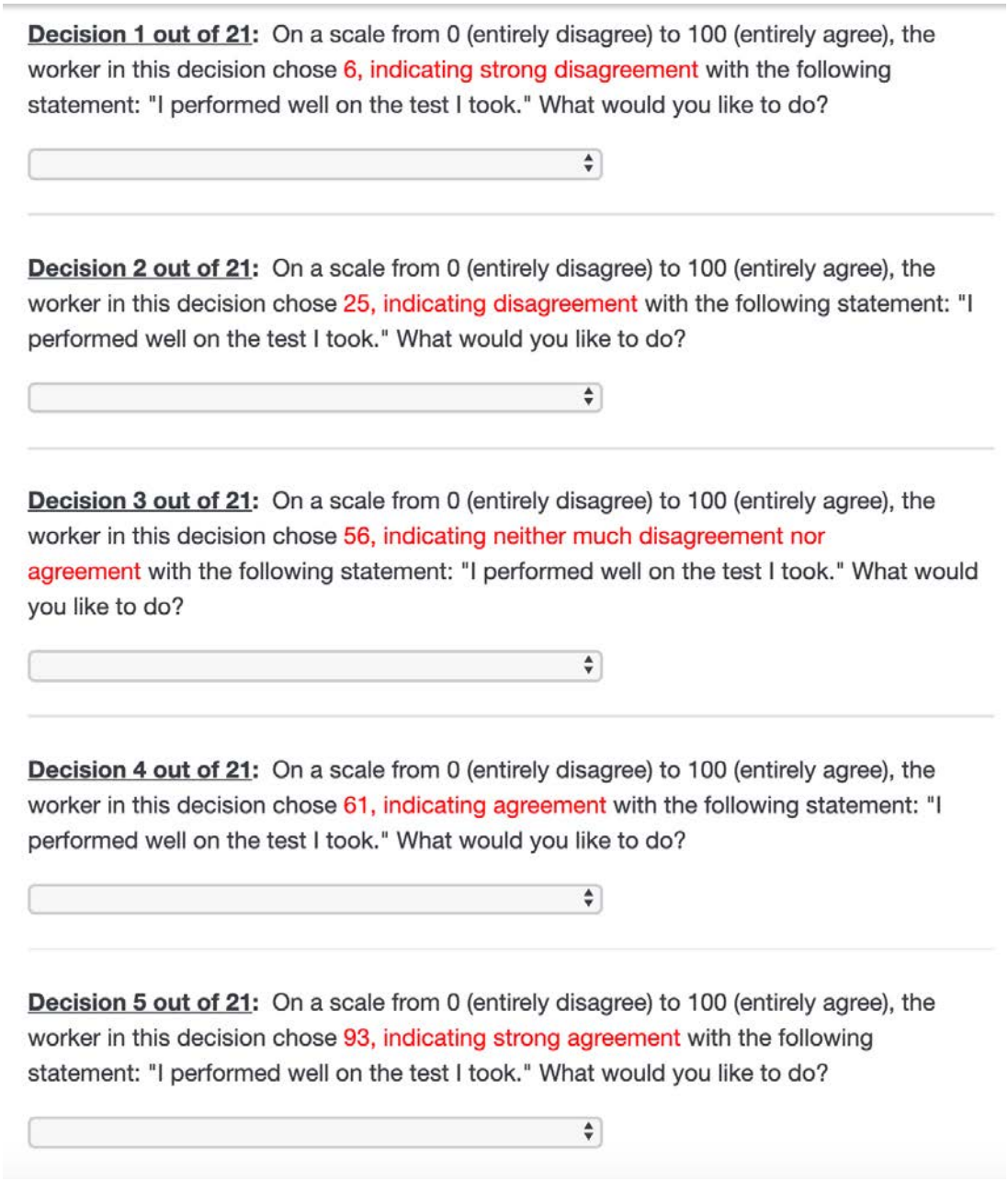


Figure D.35: *Performance Evaluation* Decisions



The instructions for, and examples of, decisions relating to the *performance-bucket* evaluations are displayed in Figures D.36 and D.37, respectively.

Figure D.36: Instructions for *Performance-Bucket Evaluation* Decisions

Instructions for Decisions 6 - 11

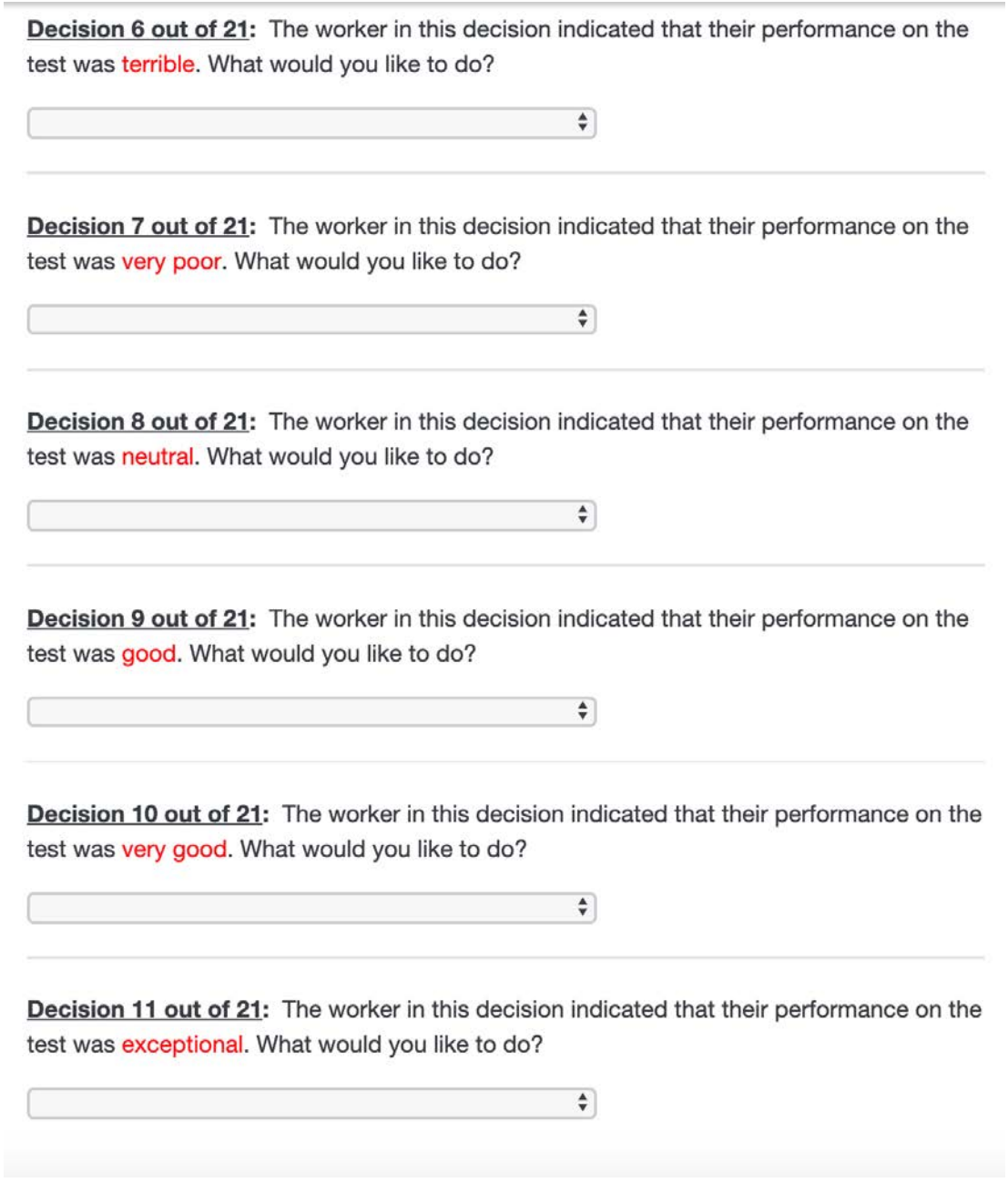
In each decision below, you will learn how the worker in that decision answered a question in which they indicated whether they thought their performance on the test was terrible, very poor, neutral, good, very good, or exceptional.

Recall that, in each decision-that-counts, the additional payment given to the worker in that decision and to you will be determined as follows:

- If you choose NOT to hire the worker, that worker's additional payment will equal 25 cents and your additional payment will equal 100 cents.

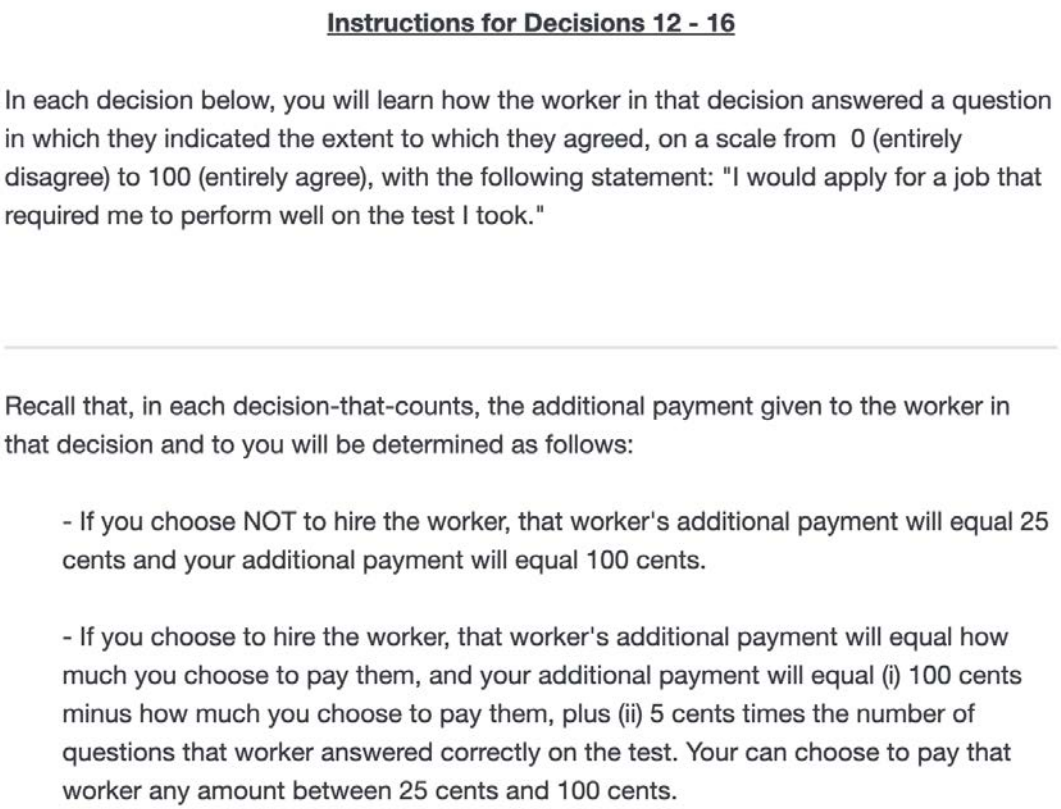
- If you choose to hire the worker, that worker's additional payment will equal how much you choose to pay them, and your additional payment will equal (i) 100 cents minus how much you choose to pay them, plus (ii) 5 cents times the number of questions that worker answered correctly on the test. Your can choose to pay that worker any amount between 25 cents and 100 cents.

Figure D.37: *Performance-Bucket Evaluation* Decisions



The instructions for, and examples of, decisions relating to the *willingness-to-apply* evaluations are displayed in Figures D.38 and D.39, respectively.

Figure D.38: Instructions for *Willingness To Apply Evaluation* Decisions



Instructions for Decisions 12 - 16

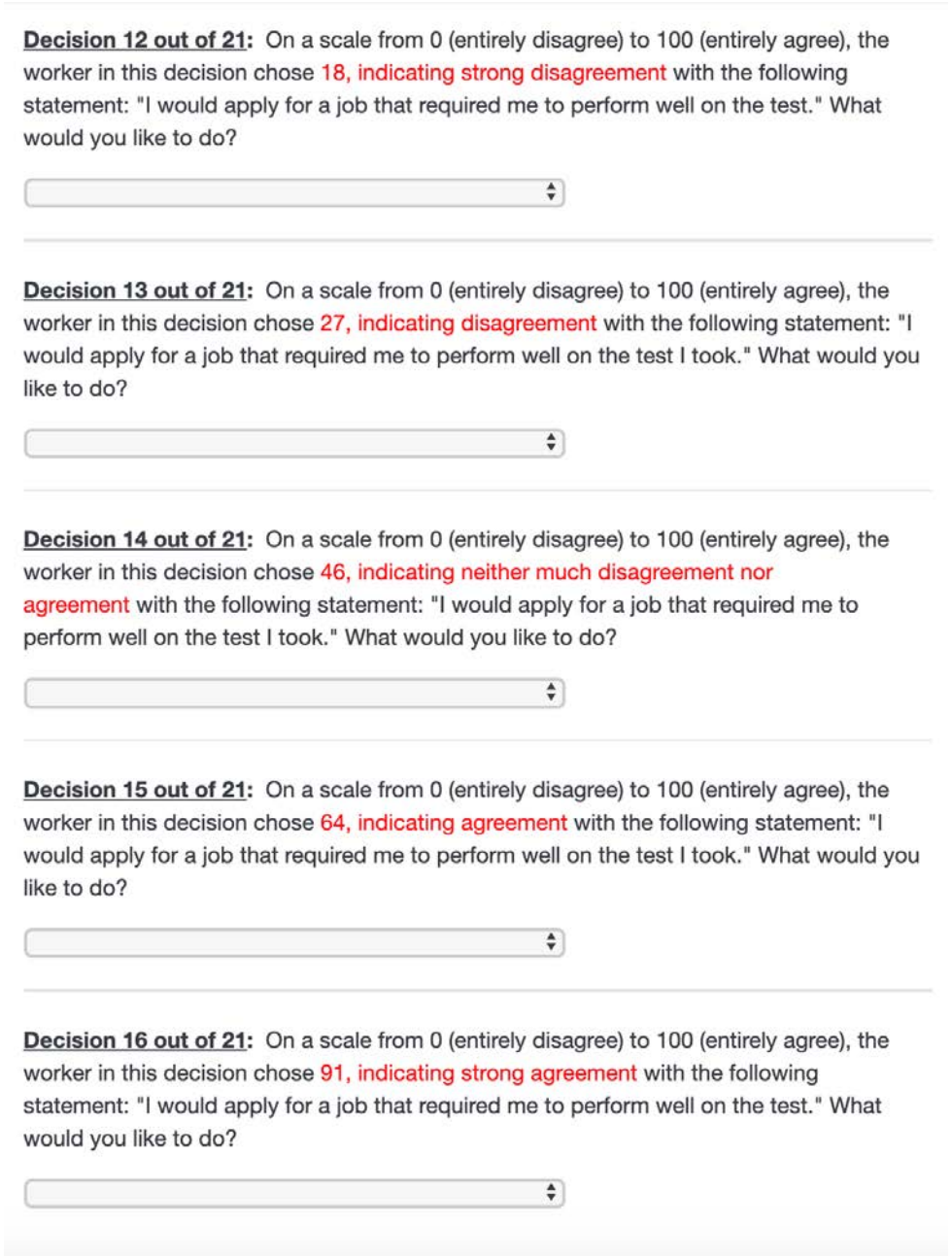
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I would apply for a job that required me to perform well on the test I took."

Recall that, in each decision-that-counts, the additional payment given to the worker in that decision and to you will be determined as follows:

- If you choose NOT to hire the worker, that worker's additional payment will equal 25 cents and your additional payment will equal 100 cents.

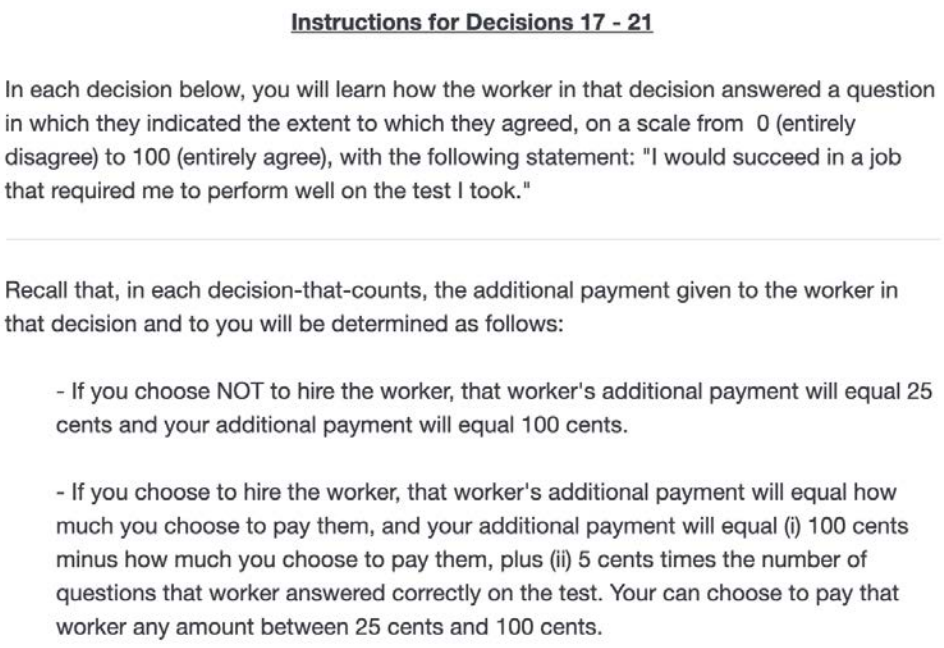
- If you choose to hire the worker, that worker's additional payment will equal how much you choose to pay them, and your additional payment will equal (i) 100 cents minus how much you choose to pay them, plus (ii) 5 cents times the number of questions that worker answered correctly on the test. Your can choose to pay that worker any amount between 25 cents and 100 cents.

Figure D.39: *Willingness To Apply Evaluation* Decisions



The instructions for, and examples of, decisions relating to the *success* evaluations are displayed in Figures D.40 and D.41, respectively.

Figure D.40: Instructions for *Success Evaluation* Decisions



Instructions for Decisions 17 - 21

In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I would succeed in a job that required me to perform well on the test I took."

Recall that, in each decision-that-counts, the additional payment given to the worker in that decision and to you will be determined as follows:

- If you choose NOT to hire the worker, that worker's additional payment will equal 25 cents and your additional payment will equal 100 cents.

- If you choose to hire the worker, that worker's additional payment will equal how much you choose to pay them, and your additional payment will equal (i) 100 cents minus how much you choose to pay them, plus (ii) 5 cents times the number of questions that worker answered correctly on the test. Your can choose to pay that worker any amount between 25 cents and 100 cents.

Figure D.41: *Success Evaluation* Decisions