The Gender Gap in Self-Promotion

Christine L. Exley and Judd B. Kessler*

October 20, 2021

Abstract

We run a series of experiments, involving over 4,000 online participants and over 10,000 school-aged youth. When individuals are asked to subjectively describe their performance on a male-typed task relating to math and science, we find a large gender gap in self-evaluations. This gap arises both when self-evaluations are provided to potential employers, and thus measure self-promotion, and when self-evaluations are not driven by incentives to promote. The gender gap in self-evaluations proves persistent and arises as early as the sixth grade. No gender gap arises, however, if individuals are instead asked about their performance on a more female-typed task.

*Exley: clexley@hbs.edu, Harvard Business School; Kessler: judd.kessler@wharton.upenn.edu, The Wharton School, University of Pennsylvania.

1 Introduction

Despite gender gaps in pay shrinking over the past few decades, women continue to earn less than men. These gender gaps can be partially explained by women being underrepresented in the highest paying industries and occupations, but gaps persist even when accounting for factors such as education and occupational selection (Goldin, 2014; Blau and Kahn, 2017). Gender differences in representation and pay are particularly pronounced in stereotypically male spaces. As evidence of prevalent gender gaps in the financial and corporate sectors, Bertrand, Goldin and Katz (2010) finds that the gender gap in annual earnings among elite MBA graduates expands over time to nearly 60 log points. Looking within STEM fields, Michelmore and Sassler (2016) finds that the largest pay gaps arise in the most male-dominated fields: engineering and computer science. The persistence of these gender gaps has inspired a rich literature on factors that can help to explain them.

Aiming to contribute to this literature, this paper is motivated by the observation that individuals regularly evaluate their own performance and often communicate their self-evaluations to others. Sometimes (e.g., in applications, interviews, and performance reviews), individuals are explicitly asked to evaluate their performance. Other times (e.g., when writing reports about their work, during presentations and meetings, and when discussing their work with colleagues), individuals face implicit invitations or opportunities to convey information about their performance. How individuals evaluate their performance may influence their future decisions, and how they communicate these self-evaluations may influence whether they are hired for a job, whether they are promoted, and how much they are paid.

When individuals convey evaluations of their performance, they frequently use *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 according to some observable metric). Thus, it is important to understand how individuals subjectively describe their performance and whether there is a gender gap in 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 how individuals subjectively describe their performance faces three distinct challenges. First, subjective descriptions are often qualitative in nature and hence difficult to measure. Second, comparing the subjective descriptions of *equally* performing men and women requires observing subjective descriptions about a well-defined performance that can be precisely measured. Third, the ability to examine the underlying drivers of subjective descriptions is limited in settings in which one cannot exogenously manipulate the environment.

[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).

The contributions of this paper stem from our ability, through a carefully controlled experimental setting, to document a gender gap in subjective descriptions of performance—elicited using self-evaluation questions—among *equally* performing men and women and to narrow in on the drivers of this gap. Motivated by gender gaps in the labor market, we focus on self-evaluations about performance on a stereotypically male-typed task.

In our first study version, participants complete a math and science task. 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. Answers to these questions reveal a substantial and significant gender gap in self-evaluations. 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 (Lundeberg, Fox and Punćcohaŕ, 1994; Niederle and Vesterlund, 2007; Bordalo et al., 2019), we then explore whether we also observe a gender gap in *informed* self-evaluations. Specifically, we investigate whether a gender gap persists when participants are provided with perfect information about their absolute and relative performance on the task. Results suggest that the gender gap in informed self-evaluations is somewhat smaller than the gender gap in (uninformed) self-evaluations, but we still find a substantial and statistically significant gender gap in informed self-evaluations.

Since these self-evaluations are conveyed to potential employers, they capture how individuals describe their performance in the presence of incentives to assess themselves favorably. We thus interpret these gender gaps in self-evaluations as gender gaps in "self-promotion." Indeed, we find that gender gaps in self-promotion make women significantly less likely to be hired—and make them earn significantly less—than equally performing men. A natural question is therefore whether the gender gaps in self-promotion reflect men, more so than women, strategically inflating their self-evaluations in response to incentives to promote.

To provide insight into this question, we investigate whether a gender gap in self-evaluations persists even absent any incentives to promote. In particular, we investigate whether a gender gap persists when self-evaluations are elicited privately and not shared with potential employers. Removing promotion incentives causes men to provide lower self-evaluations, but it also causes women to provide lower self-evaluations by a nearly identical amount. We thus observe statistically significant gender gaps in privately elicited self-evaluations that are just as large as the gender gaps when self-evaluations are provided to employers, implying that the gender gap in self-promotion reflects an underlying gender gap in self-evaluations even absent any incentives to promote.

Several additional study versions reveal the robustness of this underlying gap in self-evaluations, including when participants are informed about how self-evaluation questions are typically answered. In only two of our study versions are there no gender differences in subjective descriptions of performance. First, we observe no gender differences when we ask individuals to privately evaluate the performances of others, rather than themselves. Second, consistent with the importance of gender stereotypes (Bordalo et al., 2019), we observe no gender differences when we ask individuals to privately evaluate their performance on a more female-typed task relating to verbal skills. These two findings highlight that men and women do not have different views about how to subjectively describe performance in general. Instead, we only observe evidence for women subjectively describing their *own* performance on a *male-typed* task less favorably than equally performing men.

Given the robustness of the gender gap in self-evaluations on a male-typed task, an important question is how early these differences arise, particularly when considering the age at which to target potential interventions to counter this gap and given some prior work that finds gender differences emerge in later adolescence (Andersen et al., 2013). To investigate this question, we recruited more than 10,000 middle-school and high-school students to provide privately-elicited self-evaluations on a male-typed task. We find large and statistically significant gender gaps in self-evaluations across all ages, including among sixth-graders, the youngest students that we study. This suggests that—to the extent that the gender gap in self-evaluations arises because of formative experiences—some of these experiences occur quite early in children's lives.

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; Niederle, 2016). We complement this literature by documenting a gender gap in how individuals subjectively describe their performance on a male-typed task and investigate the drivers of this gap. Future work may investigate whether gender differences in subjective views about performance could relate to—and perhaps contribute to—gender differences in other outcomes. Akin to the role of confidence—as measured by absolute or relative performance—in helping to explain the gender gaps in the the willingness to compete (Niederle and Vesterlund, 2007; van Veldhuizen, 2017), speak up (Coffman, 2014), be a leader (Born, Ranehill and Sandberg, 2018), and claim credit (Isaksson, 2018), subjective assessments of one's own performance may cause women to not feel "good enough" to enter a competition or negotiation, to apply for a job, or to assert their expertise in stereotypically male domains. Indeed, such an explanation would correspond with prior work finding that female engineers ask for lower salaries, unless provided with information on the median salary requested (Roussille, 2021), and that women are deterred from applying to jobs that subjectively describe the requisite management, analytical, computer, or technology skills (Coffman, Collis and Kulkarni, 2020; Abraham and Stein, 2020).

2 Design, Data Collection, and Setting

We recruited 3,892 participants from online labor markets—Amazon's Mechanical Turk (MTurk) and Prolific—to participate in one of seven versions of our study across five waves of data collection, as shown in the first five rows of Table 1.[2] Each participant was guaranteed a completion fee plus a possible bonus payment from one randomly selected part of the study.[3] 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 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. These study versions allowed us to test two potential drivers of gender differences in self-evaluations that we expected, given prior literature. First, motivated by the vast literature on gender gaps in beliefs about performance, and how such gaps contribute to gender gaps in behavior, we hypothesized that differences in beliefs about performance could contribute to gender differences in self-evaluations. As further explained below, each of the three study versions in wave 1 allows us to examine the role of performance beliefs by comparing self-evaluations before and after participants are provided with perfect information about their absolute and relative performance on the task that their self-evaluations describe. Second, motivated by the gender gaps in the labor market and prior literature on gender differences in reported beliefs about performance that arise in strategic contexts (Reuben, Sapienza and Zingales, 2014; Charness, Rustichini and Van de Ven, 2018), we hypothesized that incentives to inflate self-evaluations that would be shared with potential employers could contribute to gender differences in self-evaluations. The study versions in wave 1 allow us to test whether this is the case because the *Self-Promotion* and *Self-Promotion (Risky)* version involve differing incentives to inflate self-evaluations, while the *Private* version removes all incentives to promote.

After observing a substantial gender gap in self-evaluations in the *Private* version in wave 1—even after participants are provided with perfect information about their absolute and relative performance—we explored the underlying drivers of this gender gap by investigating what changes to the decision environment could close it. To limit the potential drivers of gender differences in self-evaluations, we built off of the *Private* version for this exploration. Consequently, in our subsequent waves of data collection, we replicated the *Private* version and introduced new study

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

[3] 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.

versions built off of the *Private* version.

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. In addition, 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 gender gap in self-evaluations. The design of this version, and the associated results, are discussed in Section 5.

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.

In addition to the data described above, 298 participants completed a version of our study as "employers," who are relevant for the *Self-Promotion* and *Self-Promotion (Risky)* versions of our study.[4] As discussed in Section 4.3, results from the employers demonstrate that self-promotion pays. Participants who report higher self-evaluations in the *Self-Promotion* and *Self-Promotion*

[4] 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 600 MTurk workers who evaluated free-response comments generated by participants as described below and discussed in Appendix B. Including these 800 participants and the 298 employers described in the main text, this paper involves a total of 4,990 study participants from online labor markets.

(Risky) versions of our study are paid more by employers.

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. They have up to 30 seconds to answer each question. Given the gender gaps that motivate our study and the fact that gender gaps are often more prevalent in stereotypical male-typed tasks, we chose to select questions that related to math and science. Specifically, we selected four questions each from the following five categories on the Armed Services Vocational Aptitude Battery (ASVAB): General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. By selecting questions from the ASVAB, we are also able to follow prior literature that uses performance on the ASVAB as a measure of cognitive ability (Frey and Detterman, 2004) and to convey to participants why performance on questions like the ones they are answering are often informative. Specifically, 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."[5] 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?" Participants can select any number from 0 to 20, and their answers are not incentivized. Their answers are not incentivized because we control for beliefs about absolute (and relative) performance by design, as described later, mitigating concerns about noise in this measure.

Part 2: Self-Evaluations

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

[5] Our description of the Armed Services Vocational Aptitude Battery mentions that it is a test used by the military. One may wonder if this framing matters. While we do not vary the wording of this description to exclude the reference to the military, we note that—among participants in our fifth wave of data collection—we asked participants to indicate their agreement, on a 7-point Likert scale, with the following statement that does not mention the military: "In general, I perform well when asked questions that test my math and science skills." Results related to this follow-up question, which does not mention the military, have a remarkably similar pattern with respect to gender.

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 math and science test they took in part 1. Payment is determined by the participant's prior performance on the math and science test—rather than any future performance—to avoid any potential uncertainty that might arise around future performance. Thus, even if they are hired, participants do not have to complete any additional tasks.

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.[6]

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 call this 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 call this the *performance* question.

The latter two self-evaluation questions relate to participants' past performance but also allow 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 quantify—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.

[6] 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 a total of 600 MTurk participants evaluate the free responses from wave 1. We summarize those findings in Appendix B.

Part 3: Informed Self-Evaluations

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-evaluation 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). To ensure participants pay attention to this information, 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. 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 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.[7] 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.

[7] 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.

2.3 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. 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. In addition, 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 (e.g., caring about employers' earnings).

2.4 The *Private (Social Norms)* Version

The *Private (Social Norms)* version 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.

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. This version 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.

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. 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)* version 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 one's 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 their math and science skills. See Appendix D.7 for screenshots and additional details.

Given that verbal skills, relative to math and science skills, are more stereotypically 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 that we describe and analyze in Section 4.3.

2.8 Our Study Environment

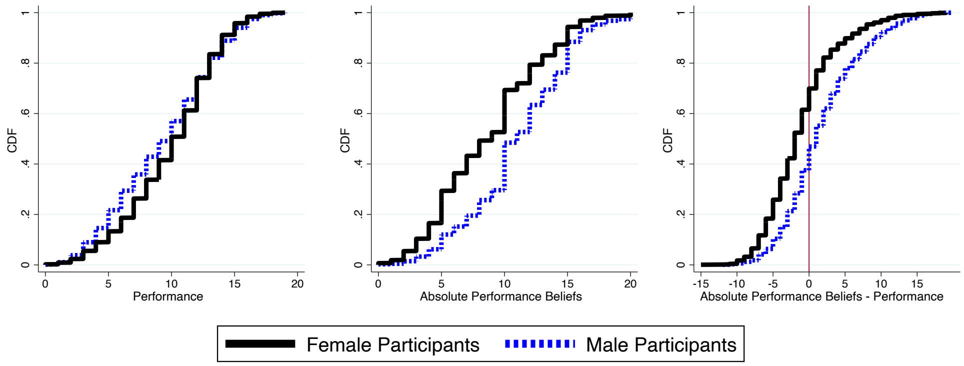
In this section, we present data on performance on the math and science test and on the beliefs that participants report about their absolute performance (i.e., how many questions they think they answered correctly on the test). Since our results are very similar across study versions, and since participants always take the test and report beliefs about their absolute performance *before* encountering any version-specific variation, 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). We find results consistent with our setting being "male-typed" in that women think they answered significantly fewer questions correctly than equally performing men.

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.

Figure 1: Performance and Absolute Performance Belief Distributions

(a) Performance (b) Belief (c) Belief–Performance



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).

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 gender gap in self-evaluations that we observe. As noted in Section 2.1, however, rather than using these reported beliefs as statistical controls, we will instead control for beliefs by design.

3 Results

Tables 2 and 3 present our experimental results from each study version in a separate panel. The following subsections discuss these results.

The first two subsections document persistent gender gaps in self-evaluations when participants are asked about their performance on the math and science test. Focusing on results from the

Self-Promotion and *Self-Promotion (Risky)* study versions, Section 3.1 documents a large gender gap in self-evaluations that are provided to potential employers, which we refer to as the gender gap in self-promotion. Focusing on results from the *Private*, *Private (Social Norms)* and *Private (Immediately Informed)* study versions, Section 3.2 documents a large gender gap in self-evaluations even absent any incentives to promote.

The last two subsections, by contrast, show that these gender gaps do not extend to all contexts. Focusing on results from the *Private (Other-Evaluation)* version, Section 3.3 finds little-to-no gender gap in how participants subjectively evaluate the performance of *others*. Focusing on the results from the *Private(Verbal)* version, Section 3.4 documents no gender gap in self-evaluations related to a *verbal* task.

3.1 The gender gap in self-promotion on a math and science task

The *Self-Promotion* version of the experiment allows us to examine how participants complete self-evaluations when they know one of their answers will be shared with employers. We thus consider any gender gap in self-evaluations in the *Self-Promotion* version as indicative of a gender gap in "self-promotion," that is, a gender gap in how individuals promote or describe their performance to others.

Figure 2 shows raw responses to the four self-evaluation questions from 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 ($p < 0.01$ 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-evaluations 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 than those provided by men, 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 a six-point Likert scale. 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

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).

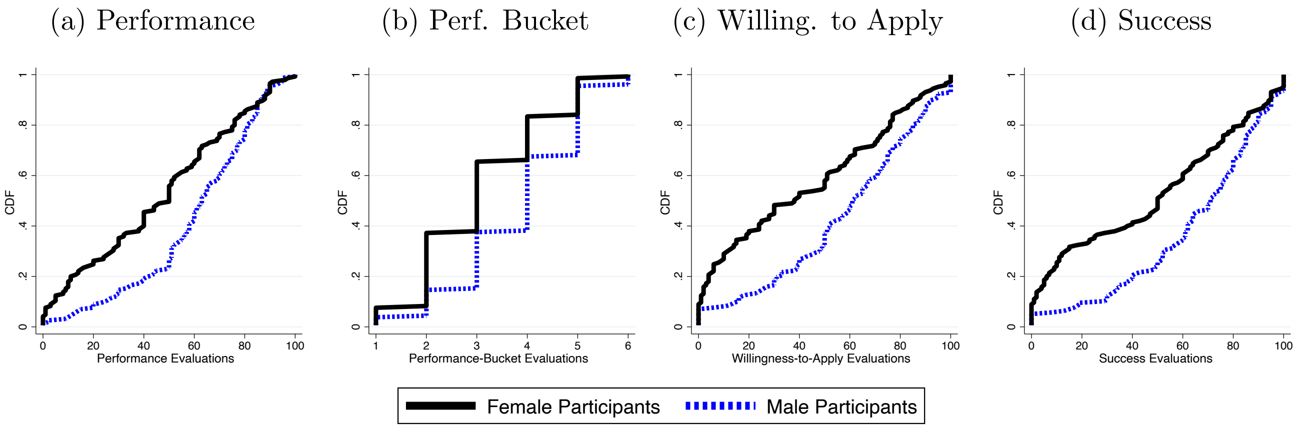
Table 3: Results from Informed Evaluations (after performance information is provided)

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).

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-evaluations among equally performing men and women.

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



(a) Performance (b) Perf. Bucket (c) Willing. to Apply (d) Success

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.

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-evaluations when participants are uncertain about their absolute and relative performance is thus important for considering the role of self-evaluations in driving gender gaps in educational and labor market outcomes.

To explore whether this gender gap in self-evaluations reflects women thinking they had a lower performance (in absolute or relative terms) than equally performing men—particularly in light of the gender gap in beliefs about absolute performance as detailed in Section 2.8—we turn to 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"—we refer to the responses to these questions as *informed* self-evaluations.

Panel 1 of Table 3 presents results from responses after participants have learned their absolute and relative performance. These results reveal substantial and statistically significant gender gaps in *informed* self-evaluations. When considering the questions asked on the 0–100 scale, the gender gap in informed self-evaluations is 7.01 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-evaluations is 0.40.

The gender gap in informed self-evaluations makes clear that the gap 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 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). Put differently, we document a gender gap in self-evaluations that cannot be attributed to gender differences in "confidence," if confidence is modeled as individuals' beliefs about their absolute and relative performance, an implicit definition often adopted in prior literature. That said, one may naturally wish to consider confidence more broadly, particularly in the case of the willingness-to-apply question and success question, and hence still consider our results as potentially relating to a gender gap in confidence. Indeed, one could even consider self-evaluations to directly measure a subjective form of confidence.

While the gender gap persists when participants are informed of their absolute and relative performance, the gender gap in *informed* self-evaluations appears smaller than the gender gap in (uninformed) self-evaluations that are elicited before participants are informed of their absolute and relative performance. As shown in Panel 1 of Appendix Table A.2, this is a result of men and women responding somewhat differently to information on their performance. While men inconsistently respond to this information (see the coefficient estimates on *Informed*), women directionally increase their self-evaluations in response to this information (the sum of the coefficient estimates on *Informed* and *Informed***Female* is directionally positive for all four questions). In addition, women directionally increase their self-evaluations in response to this information more so than men (the coefficient estimates on *Informed***Female* are always directionally positive). While none of these effects are statistically significant in the *Self-Promotion* version on its own, similar and statistically significant patterns follow when we pool across all study versions in which we elicit self-evaluations both before and after performance information is provided (see Panel 2 of Appendix Table A.2).

An important and interesting question for future work relates to the persistence of the gender gap in self-evaluations 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. Panel 3 of Table 2 and Panel 3 of Table 3 show that the gender gap in self-evaluations and the gender gap in informed self-evaluations remain substantial and significant under the slightly different promotion incentives in the *Self-Promotion (Risky)* version.

3.2 The gender gap in self-evaluations on a math and science task

The gender gaps in self-evaluations that are provided to potential employers in the *Self-Promotion* and *Self-Promotion (Risky)* versions—i.e., the gender gaps in self-promotion—could arise due to the incentives to promote one's performance to potential employers or could instead be reflective of an underlying gender gap in self-evaluations even absent any promotion incentives.

To examine the relevance of promotion incentives—and to assess whether there is an underlying gender gap in self-evaluations even absent any incentives to promote—we turn to the *Private* version. In the *Private* version, self-evaluations no longer serve as a measure of self-promotion because they are not shared with potential employers. Participants receive a fixed payment regardless of their self-evaluations, eliminating any incentives to promote. More broadly, this version allows us to measure any underlying gender gap in self-evaluations that cannot be driven by gender differences relating to strategic incentives (Reuben, Sapienza and Zingales, 2014; Charness, Rustichini and Van de Ven, 2018), risk aversion over payoffs (Dwyer, Gilkeson and List, 2002; Eckel and Grossman, 2008), lack of control over payoffs (Cobb-Clark, 2015; Apicella, Demiral and Mollerstrom, 2020), or preferences over others' payoffs (Andreoni and Vesterlund, 2001; DellaVigna et al., 2013).

Appendix Table A.3 compares answers to self-evaluation questions in the *Private* and *Self-Promotion* versions run in the same wave (i.e., wave 1). The positive and statistically significant coefficient estimates on *Self-Promotion*—in response to 7 out of the 8 self-evaluation questions—make clear that men respond to promotion incentives by providing more favorable responses in the *Self-Promotion* version than in the *Private* version. But, this pattern is not unique to men. In response to all 8 self-evaluation questions, the sum of the coefficient estimates on *Self-Promotion* and *Female*Self-Promotion* are positive and statistically significant, revealing that women also respond to promotion incentives by providing more favorable self-evaluations in the *Self-Promotion* version. Indeed, the insignificant and largely positive coefficient estimates on *Female*Self-Promotion* reveal that the gender gaps in the *Self-Promotion* version are *not* reflective of men responding more favorably to promotion incentives than women.[8] The gender gaps in the *Self-Promotion* version are instead reflective of an underlying gender gap in self-evaluations absent any incentives to promote.

Results from the *Private* version show that this underlying gender gap is large. When participants are not informed about their performance (see Panel 3 of Table 2), there is a statistically significant gender gap in self-evaluations in response to each of the four questions. When considering the questions asked on the 0–100 scale, the gender gap in self-evaluations is 13.46 points for the performance question, 17.57 for the willingness-to-apply question, and 16.46 for the success question. When considering the question asked on the 1–6 scale, the gender gap in self-evaluations is 0.56. When participants are informed about their absolute and relative performance (see Panel 3 of Table 3), the gender gap in self-evaluations is smaller but still quite large and statistically significant. When considering the questions asked on the 0–100 scale, the gender gap in informed self-evaluations is 8.01 points for the performance question, 13.25 for the willingness-to-apply question, and 13.15 for the success question. When considering the question asked on the 1–6 scale, the gender gap in informed self-evaluations is 0.33. Like the gender gap in self-promotion, the gender gap in self-evaluations—absent any promotion incentives—is not just a result of women thinking they had a lower performance. Even when participants know their absolute or relative performance, women subjectively evaluate their performance less favorably than equally performing men.

To further investigate the robustness and drivers of the gender gap in self-evaluations, we consider

[8] That men and women seem to care similarly about the incentives to promote is consistent with findings from our part 4 question that asks subjects to claim an amount of money based on what they think they deserve to earn from the study. As shown in Appendix Table A.4, when pooling across all versions in which participants are privately asked about their performance on the math and science test, there is no evidence for a gender difference in how much money equally performing men and women claim. This finding also suggests that the gender gap in self-evaluations may be specific to situations where individuals evaluate their performance by assigning subjective descriptions to their performance (rather than by assigning monetary values to their performance), a hypothesis that could be explored in future work.

results from our additional study versions that also do not involve any promotion incentives. In our second wave of data collection, we replicated 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-evaluations is just as large in the *Private (Social Norms)* version as in the *Private* version. Thus, the gender gap in self-evaluation persists even when information on what may be typical or socially appropriate is provided.

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). Even when participants are not asked self-evaluation questions before being informed of their performance—and, thus, when we remove any related consistency or anchoring effects—we still observe a gender gap after participants are informed about their absolute and relative performance.

3.3 No gender gap in other-evaluations on a math and science task

Given the robust gender gaps in self-evaluations on a math and science task, one may wonder whether similar gender differences emerge when participants are asked to evaluate the performance of others on the same task or whether, like in prior findings related to negotiation and competition (Bowles, Babcock and McGinn, 2005; Cassar, Wordofa and Zhang, 2016), this change in focus mitigates gender differences. To investigate this, 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).

3.4 No gender gap in self-evaluations on a verbal task

Given the gender gaps in pay and in occupational and industry representation that motivate our study, the main task that participants face is a stereotypical male-typed task relating to math and science skills. Given prior work on gender stereotypes and how the type of task can influence gender differences in beliefs (Bordalo et al., 2019; Coffman, Collis and Kulkarni, 2019), competitions

(Günther et al., 2010; Shurchkov, 2012; Dreber, von Essen and Ranehill, 2014), group decision-making (Coffman, 2014; Coffman, Flikkema and Shurchkov, 2019), and test-taking (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 stereotypical female-typed task. In our fifth wave of data collection, we again replicate the gender gap in self-evaluations 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 in self-evaluations, 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).

These findings suggest that the gender gap in self-evaluations may be more prevalent in male-typed tasks than in female-typed tasks and highlights the value of future work exploring whether gender gaps in self-evaluations arise across a wider range of tasks. Together with the evidence in Section 3.3, these findings also make clear that the gender gap in self-evaluations arising in response to the math and science task 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 participants 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 task.

4 Discussion

In this section, we present additional analysis of the data collected in waves 1–5, related to robustness (Section 4.1), heterogeneity (Section 4.2), and the consequences of the gender gap in self-evaluations (Section 4.3).

4.1 The robustness of the gender gap

We examine the gender gap in self-evaluations—on a math and science task—across a range of settings. 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. Not surprisingly, when we pool across all self-evaluations relating to the math and science task, the gender gaps in self-evaluations persist, regardless of whether participants are uninformed about their performance (see Panel 12 of Table 2) or informed about their performance (see Panel 12 of Table 3).

Further robustness tests of this pooled data reveal that the gender gaps in self-evaluations are robust to excluding performance controls (Appendix Table A.5), controlling for other demographic information (Appendix Table A.6), excluding "inattentive" participants who answered no better than chance on the math and science test (Appendix Table A.7), quantile regressions estimated at the 25th, 50th, and 75th percentiles (Appendix Table A.8), and ordered Probit specifications for answers to the performance-bucket question elicited on the six-point scale (Appendix Table A.9).

4.2 Heterogeneity Analyses

Given the robustness of the gender gap in self-evaluations relating to the math and science task, we conduct three sets of heterogeneity analyses on this pooled data.

First, Appendix Tables A.10 and A.11 show that—while gaps are large and statistically significant at the average performance level—the gap is estimated to be somewhat smaller at high performance levels. Future work might explore the relationship between performance and such gender gaps.

Second, Appendix Table A.12 shows statistically significantly more favorable self-evaluations among younger participants, more educated participants, and more Republican-leaning participants. Appendix Table A.12 also shows that the gender gaps are larger among more Republican-leaning participants. To garner additional insights about what drives differences in self-evaluations across groups—and to shed light on the potential role of culture—future work might investigate the relationship between self-evaluations and these demographics, as well as other factors such as socio-economic status, race, where someone lives, and where someone grew up. We hope that future work also gathers data from countries beyond the United States.

Third, as detailed in Appendix C, statistically controlling for participants' reported beliefs about their absolute performance introduces potential confounds related to measurement error, omitted variable bias, and reverse causality. These potential confounds are why we control for beliefs "by design" by examining informed self-evaluations that are elicited after participants are perfectly informed of their absolute and relative performance. Nevertheless, it is interesting to note that Appendix Tables A.13 and A.14 show that absolute beliefs are positively correlated with self-evaluations and that the gender gap in self-evaluations is generally smaller among individuals who believe they had a higher absolute performance. Appendix Tables A.15 and A.16 find similar results with a broader measure of views about ability. As also discussed in Appendix C, future work may investigate whether these findings reflect the existence of "types" of individuals who generally view their math and science performances more negatively or more positively.

4.3 Consequences of the gender gap

An important direction for future work is to explore how the gender gap in self-evaluations contributes to the various gender differences in economic outcomes. We provide two additional sets of results from our study to help inform this future work. One set relates to how employers respond to self-evaluations and one set relates to whether study participants predict the gender gap in self-evaluations that we observe.

The Employer Results

In order to determine bonus payments for the "workers" in the the *Self-Promotion* and *Self-Promotion (Risky)* study versions, we recruited 298 "employers" from MTurk in the *Employer* version.[9] These employers make 21 hiring decisions. In each decision, they must decide whether to hire a worker, and, if so, how much to pay that worker (recall payment details in Section 2.1 under the "Part 2" subheader). The only information an employer receives about a worker before hiring them is how the worker answered one of the four self-evaluation questions. Out of these 21 hiring decisions, two decisions are implemented to determine the bonus payments for the employer and for two corresponding workers. See Appendix D.9 for screenshots and additional details.

As shown in Panel 1 of Appendix Table A.17, employers are more willing to hire workers who provide more positive self-evaluations. Columns (1), (3), and (4) show that this willingness increases by 1 percentage point 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 this willingness increases by an average of 18 percentage points for each increase on the six-point Likert scale in the performance-bucket question. Panel 1 of Appendix Table A.18 shows similar results when we instead consider employers' wage decisions. We do not observe any significant differences by the gender of the employer.

As shown in Panel 2 of Appendix Table A.17, these hiring decisions imply that female workers are less likely to be hired than equally performing male workers in the *Self-Promotion* and *Self-Promotion (Risky)* versions.[10] Female workers are anywhere from 9 to 12 percentage points less likely to be hired than equally performing men. Panel 2 of Appendix Table A.18 shows that women also have significantly lower expected wages than equally performing men. Thus, the results from the *Employer* version confirm that the gender gap in self-evaluations can result in equally performing women receiving worse economic outcomes than equally performing men.

The Predictor Results

If employers anticipate the gender gap in self-evaluations, one might hypothesize smaller economic consequences from the gap because employers can account for women providing less favorable subjective evaluations than men. To assess whether the gender gap in self-evaluations is anticipated, we added eight incentivized questions to the end of the study versions we ran in wave 5 of data collection (see screenshots in Appendix Figures D.24 and D.25). Each question asked participants to predict the average performance of male and female workers in the *Self-Promotion* version after learning the average self-evaluation responses provided by those male and female workers.

As shown in Appendix Table A.19, participants do not correctly predict that male and female workers have a similar average performance (equal to about 10) in the *Self-Promotion* version. Rather, when considering predictions based off of answers to each self-evaluation question, both male and female participants predict that the average performance of men is significantly higher than the average performance of women. Thus, we find no evidence of predictors correcting the gap when making assessments about workers. Future work—both in the laboratory and in the field—

[9] Each employer received a guaranteed $1.50 completion fee for the 15-minute study and were recruited using the same criteria as noted in footnote 2.

[10] While we pool workers from both versions in Appendix Table A.17, these results are similar and remain statistically significant when separately considering each study version.

should investigate whether this applies more broadly in other settings, such as when experience helps employers get better at identifying the gender gap and, perhaps, correcting for it. 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), future work should also investigate the impact of making gender known on self-evaluations.

5 The Gender Gap in Self-Evaluations Among Youth

A growing literature investigates whether gender differences in competition arise among children (Gneezy and Rustichini, 2004; Dreber, Von Essen and Ranehill, 2011; Cárdenas et al., 2012). Work that considers a wide range of ages finds mixed evidence—some finds no gender differences among young children and that gaps emerge in later adolescence (Andersen et al., 2013) while some finds gender differences arising as early kindergarten (Sutter and Glätzle-Rützler, 2015). The age at which gender differences arise is informative, both in terms of the potential role of formative life experiences and in terms of determining the ideal ages at which to target policy interventions to potentially mitigate such gender gaps.

To gain insight into the age at which gender gaps in self-evaluations emerge, we ran an additional experiment involving 10,637 middle-school and high-school students. These students were 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 balanced by gender (48% of students are male) and skewed 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, the test for youth only involved the 10 easiest questions from our math and science test. Second, in the willingness-to-apply question, we asked them about their willingness to take a class that involved topics like those covered on the test. Third, in the success question, 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 self-evaluations is robust to this very different setting and that it appears as early as sixth grade, among the youngest students that we study.

Following much of the heterogeneity analysis presented in Section 4.2, Appendix Tables A.20–A.24 present parallel results exploring heterogeneity based on performance, beliefs about absolute performance, other demographics, and GPA. Appendix Table A.20 reveals that, unlike our prior

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.

results, the gender gaps for youth are *larger* among higher performers. Appendix Tables A.21 and A.22 reveal that—while beliefs are positively and significantly correlated with self-evaluations—evidence on how they correlate with the size of the gap is mixed. Appendix Table A.23 shows that, relative to the 34% of 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. But, the gender gap does not appear to systematically differ by race. Appendix Table A.24 reveals that the 39% of students who qualify for a free or reduced price lunch (FRPL) provide somewhat less favorable self-evaluations but that FRPL status does not correlate with the gender gap. Finally, Appendix Table A.25 shows that

GPA 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.

These findings leave many interesting questions for future work, such as investigating the self-evaluations among even younger children—such as elementary school students—to try to pinpoint the age at which this gender gap emerges and exploring interventions that close the gender gap among youth to see if youth display the same patterns as workers in our online labor markets.

6 Conclusion

This paper documents a large gender gap in self-evaluations on a male-typed task relating to math and science: women subjectively describe their performance less favorably than equally performing men. We first show a substantial and robust gender gap in self-evaluations that will be shared with potential employers, which we take as evidence of a gender gap in self-promotion. We then show that this gap is not specific to settings with promotion incentives. When self-evaluations are elicited privately, the gender gap remains just as large. Finally, by focusing on settings in which self-evaluations are elicited privately, we further show that the gender gap in self-evaluations is robust to a variety of environments and arises early (as evident from our results with over 10,000 middle-school and high-school students). A notable exception to the robustness of these results is that we do not observe a gender gap in self-evaluations when we ask participants about their performance on a test assessing verbal ability.

We end the paper by highlighting the many exciting and important avenues for future work. A first avenue relates to further exploring settings beyond those relating to math and science. That we do not observe a gender gap in self-evaluations when participants are asked about their performance on a verbal test suggests that a gender gap in self-evaluations is less likely in female-typed domains. But, since our paper only privately elicits self-evaluations in this female-typed domain, future work is needed to assess the impact of communicating self-evaluations to employers in female-typed domains. For example, focusing on a female-dominated profession, Biasi and Sarsons (Forthcoming) finds that female public school teachers are less willing to negotiate than male public school teachers.

A second avenue relates to considering the impact of extensive margin decisions. We document a gender gap in self-evaluations when individuals are required to answer self-evaluation questions that will be shared with potential employers. Given that women are often reluctant to enter negotiations (Hernandez-Arenaz and Iriberri, 2019), to enter competitions (Niederle and Vesterlund, 2011), and to speak up (Coffman, 2014), a natural question is whether gender gaps in self-evaluations—and corresponding gender gaps in what employers infer about the performance of workers—are exacerbated in settings where women may avoid communicating about their performance altogether.

A third avenue relates to investigating the impact of the information structure on self-evaluations and how employers respond to them. We find that women are less likely to be hired than equally performing men when a potential employer only learns their answer to one self-evaluation question, in the *Self-Promotion* version of our study, and when information on performance might be shared with employers, in the *Self-Promotion (Risky)* version. Future work may investigate the impact on self-evaluations when employers have additional information on performance—or signals of performance—or when gender is known (see discussion in Section 4.3).

A fourth avenue relates to examining the potential consequences of the gender gap in self-evaluations, and specifically, whether it contributes to other gender gaps observed in the labor market. On one hand, labor market decisions may involve higher stakes than those considered in our studies, which may impact performance evaluations. On the other hand, the cumulative effect of the many potential gender gaps in self-evaluations that can arise in labor market settings—such as self-evaluations conveyed in job interviews and applications, in performance and promotion reviews, in meetings and presentations, and in everyday communications—could have a substantial impact over time.

A fifth avenue relates to examining policy interventions to mitigate any consequences of the gender gap in self-evaluations. Akin to the findings in Kessel, Mollerstrom and van Veldhuizen (2021), future work may investigate the effectiveness of informing individuals of the gender gap in self-evaluations and the associated financial consequences when self-evaluations are communicated to employers.[11] In addition, 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."[12] Future work should investigate the impact of relying less on subjective self-evaluations for hiring and promotion.

[11] Indeed, this approach seems promising in light of the results from our part 4 question, discussed in footnote 8.

[12] For work on potential downsides to a "changing the women" approach, Exley, Niederle and Vesterlund (2020) show that focusing women to take actions they would not choose themselves backfires in the context of choosing when to negotiate. For excellent recent work on change-the-system approaches, see Apicella, Demiral and Mollerstrom (2017), He, Kang and Lacetera (2019), and Carlana, La Ferrara and Pinotti (2020).

References

Abraham, Lisa, and Alison Stein. 2020. "Words Matter: Experimental Evidence from Job Applications." *Wokring Paper.*

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. 2013. "Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society." *Review of Economics and Statistics*, 95(4): 1438–1443.

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.

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

Bertrand, Marianne, Claudia Goldin, and Lawrence F Katz. 2010. "Dynamics of the gender gap for young professionals in the financial and corporate sectors." *American Economic Journal: Applied Economic*, 2(3): 228–55.

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

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

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

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

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.

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.*

Cassar, Alessandra, Feven Wordofa, and Y Jane Zhang. 2016. “Competing for the benefit of offspring eliminates the gender gap in competitiveness.” *Proceedings of the National Academy of Sciences*, 113(19): 5201–5205.

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., 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.*

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. 2008. "Men, Women and Risk Aversion: Experimental Evidence." In *Handbook of Experimental Economics Results.* 1061–1073.

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.

Frey, Meredith C, and Douglas K Detterman. 2004. "Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability." *Psychological Science*, 15(5): 373–378.

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

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.

Günther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel. 2010. "Women can't jump?—An experiment on competitive attitudes and stereotype threat." *Journal of Economic Behavior & Organization*, 75(3): 395–401.

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. 2019. "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.*

Kessel, Dany, Johanna Mollerstrom, and Roel van Veldhuizen. 2021. "Can Simple Advice Eliminate the Gender Gap in Willingness to Compete?" *European Economic Review.*

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.*

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.*

Michelmore, Katherine, and Sharon Sassler. 2016. “Explaining the gender wage gap in STEM: Does field sex composition matter?” *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(4): 194–215.

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).

Roussille, Nina. 2021. “The central role of the ask gap in gender pay inequality.” *Working Paper.*

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

Shurchkov, Olga. 2012. “Under pressure: gender differences in output quality and quantity under competition and time constraints.” *Journal of the European Economic Association*, 10(5): 1189–1213.

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

van Veldhuizen, Roel. 2017. “Gender differences in tournament choices: Risk preferences, overconfidence or competitiveness?” *Working Paper.*

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: 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.5: 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.6: 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.7: 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.8: 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. We do not run quantile regressions 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. *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.9: Robustness to ordered probit regressions

Question:	Performance-Bucket	
	(1)	(2)
Panel 1: Evaluations (before performance information is provided)		
Female	-0.66***	-0.59***
	(0.05)	(0.05)
N	2094	2094
Panel 2: Evaluations (after performance information is provided)		
Female	-0.45***	-0.41***
	(0.04)	(0.04)
N	2990	2990
Performance FEs	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from ordered probit specifications of the responses provided to the performance-bucket question, 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. We show results both with and without performance FEs due to concerns related to the inclusion of fixed effects in order probit specifications. 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.10: 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.11: 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.12: 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.13: 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
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).

Table A.14: 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
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).

Table A.15: 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.16: 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.17: Probability of being hired

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
Panel 1: Employers' hiring decisions				
Answer	0.01***	0.18***	0.01***	0.01***
	(0.00)	(0.01)	(0.00)	(0.00)
Constant	0.07**	-0.02	0.08***	0.08***
	(0.03)	(0.03)	(0.03)	(0.03)
N	1490	1788	1490	1490
Panel 2: Workers' expected probability of being hired				
Female	-0.09***	-0.11***	-0.12***	-0.11***
	(0.02)	(0.02)	(0.02)	(0.02)
N	1192	1192	1192	1192
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Panel 1 presents results on the decisions made by the employers in the *Employer* version. The results are from a linear probability model of the likelihood that an employer indicates they will hire a worker in a decision, with SEs clustered by employer. *Answer* is the answer to the self-evaluation question they were asked to consider in that decision. Panel 2 presents results on the expected probability of a worker being hired in the *Self-Promotion* or *Self-Promotion (Risky)* version. For each worker, their expected probability of being hired was calculated as the average probability of being hired when considering all employers who made hiring decisions in response to the answer on the self-evaluation they provided. *Female* is an indicator for the worker being female. Performance FEs are dummies for each possible worker performance out of the 20 questions on the test. The columns restrict to the data associated with the noted question, as defined in the notes of Table 2.

Table A.18: Wages

Question:	Performance	Performance-Bucket	Willingness-to-Apply	Success
Panel 1: Employers' wage decisions				
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
Panel 2: Workers' expected wage				
Female	-1.77***	-2.13***	-2.16***	-1.89***
	(0.49)	(0.37)	(0.56)	(0.53)
N	1192	1192	1192	1192
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Panel 1 presents results on the decisions made by the employers in the *Employer* version. The results are from OLS regressions of the wage an employer chose in a decision with SEs clustered by employer. *Answer* is the answer to the self-evaluation question they were asked to consider in that decision. Panel 2 presents results on the expected wage of a worker in the *Self-Promotion* or *Self-Promotion (Risky)* version. For each worker, their expected wage was calculated as the average wage when considering all employers who made hiring decisions in response to the answer on the self-evaluation they provided. *Female* is an indicator for the worker being female. Performance FEs are dummies for each possible worker performance out of the 20 questions on the test. The columns restrict to the data associated with the noted question, as defined in the notes of Table 2.

Table A.19: Predictions about 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.20: 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.21: 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.22: 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.23: 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.24: 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.25: 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 *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. In July 2021, we recruited 201 participants on MTurk to complete the *Free-Response Coding* version. Each participant received a guaranteed completion fee, which equaled $1.50 for the study versions run in 2019 and $3 for the study version run in 2021. 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. In the *Free-Response Coding* version, participants made 21 coding decisions. Before making each decision or set of predictions, participants were provided with the text—but no other information—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.[13]

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. One of their decisions—out of the 21 decisions in the study—was selected to determine a possible bonus payment for them and for an associated "worker."[14] The payoffs resulting from the one randomly selected decision for these employers are the same as described in the *Employer* version.

Participants assigned to the *Free-Response Predictor* version were 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.

Participants assigned to the *Free-Response Coding* version were asked to indicate either "yes" or "no" to whether the participant who wrote the free response was engaging in self-promotion.

Relative to the *Employer* version discussed in the main text, 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., the impact of a one unit increase on a 0-to-100 scale in the *Employer*

[13] Not all of the free responses collected in wave 1 of the study were evaluated. A research assistant—blinded to participant gender and study version—deemed 130 of the 1800 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.

[14] 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). .

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, 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 jointly analyze free responses from all three study versions run in wave 1.

Table B.1 presents results from regressions testing whether the gender of the free response author affects how responses are coded, predictions, and hiring decisions. Column (1) is estimated from ratings in the *Free-Response Coding* version. The negative coefficients on *Female* in column (1) show that participants are less likely (at least directionally) to indicate that female participants engage in self-promotion given their free responses. Columns (2) and (3) are estimated from predictions from the *Free-Response Predictor* version. The negative coefficients on *Female* in column (2) 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 (3) 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 (4) 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 6 and in the main text of the paper, difficulties with using free responses, and other qualitative data, contribute to our decision to focus our analysis on the quantitative self-evaluation questions we explore in the main text of the paper.

Table B.1: Free Response Regressions

	Coded as Self-Promotion	Predicted Performance	Predicted Probability Female	Wage
	(1)	(2)	(3)	(4)
Panel 1: Free responses (before performance information is provided)				
Female	-0.07***	-0.67***	0.08***	-1.44*
	(0.03)	(0.22)	(0.03)	(0.81)
N	764	749	749	743
Panel 2: Free responses (after performance information is provided)				
Female	-0.03	-0.35	0.09***	-0.66
	(0.03)	(0.23)	(0.03)	(1.04)
N	757	773	773	755
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 noted dependent variable (DV). *Coded as Self-Promotion* equals 1 if the predictor indicated that the the participant was engaging in self-promotion and 0 otherwise. *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 to participants, and data in Panel 2 are from free responses elicited after performance information is provided to participants. Neither predictors nor employers were provided with any information on participants aside from these free responses. Data are from all three study versions run in wave 1: the *Self-Promotion* version, the *Self-Promotion (Risky)* version, and the *Private* version.

C Methodological Note: The Role of Beliefs

To explore how the gender gap varies by beliefs about absolute performance, the specifications in Appendix Table A.13 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 self-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.14, 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.13 and A.14 show that beliefs about absolute performance are *still* correlated with self-evaluations 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.* 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 in self-evaluations *and* overestimate their absolute performance. Such negative types could subjectively evaluate their performance less positively in self-evaluations *and* 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.

To further explore the possibility that certain types of individuals systematically view their math and science performance less favorably than others, 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.15, 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 asked 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.16, 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).[15]

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.

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

The presence of types like those posited above highlights why caution is warranted when trying to assess the role of beliefs about absolute performance in contributing to self-evaluations by statistically controlling for reported beliefs about absolute performance. Such results may be confounded by measurement error, omitted variable bias (which could be caused by the positive and negative types discussed above), or reverse causality. Indeed, absent the relevance of such confounds, one cannot explain why the reported beliefs about absolute performance remain statistically significant even after participants are perfectly informed of their absolute and relative performance.

Controlling for beliefs by design—by providing participants with precise information on their absolute and relative performance prior to eliciting their informed self-evaluations—allows us to avoid these potential confounds. Thus, it is interesting to note that the apparent relevance of beliefs about performance in explaining the gender gap in self-evaluations is dependent on whether we control for beliefs by design or instead control for beliefs statistically. This is shown most clearly by the results in Appendix Table C.1. Panel 1 presents results the gender gap in self-evaluations before performance information is provided. Panel 2 shows the gender gap in self-evaluations after participants are perfectly informed of their absolute and relative performance and thus after controlling for these beliefs by design. Panel 3 returns to analyzing the data before performance information is provided but now adds in a fixed effect for each reported belief about absolute performance and hence controls for beliefs statistically. While a comparison between Panels 1 and 2 makes clear that beliefs about absolute and relative performance explain the minority of the gender gap in self-evaluations, a comparison between Panels 1 and 3 would have instead suggested that beliefs about absolute performance alone explain the majority of the gender gap in self-evaluations. Thus, when a research question asks what role beliefs play in driving some outcome (i.e., rather than how beliefs update in response to information in which case focusing on measured beliefs is essential), it may be preferable to control for beliefs "by design" than to measure beliefs and control for them statistically.

Table C.1: 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.

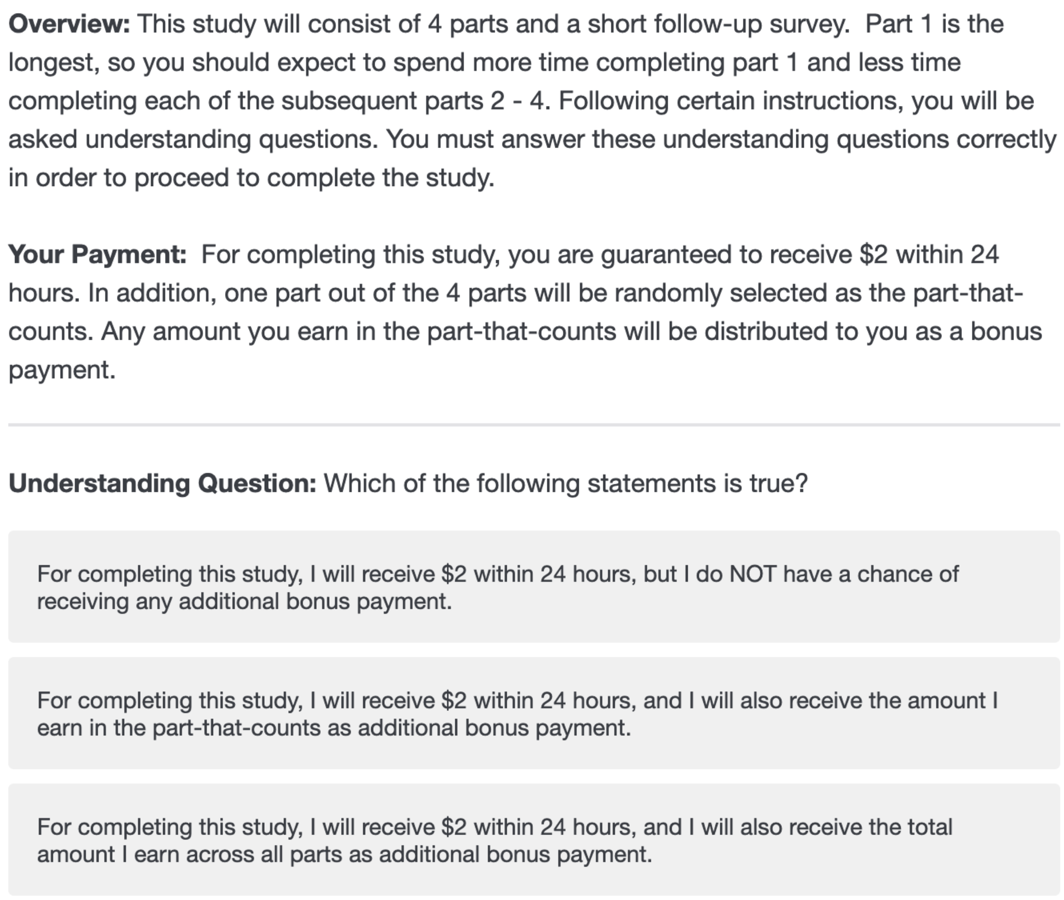
D Experimental Instructions

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



Overview: This study will consist of 4 parts and a short follow-up survey. Part 1 is the longest, so you should expect to spend more time completing part 1 and less time completing each of the subsequent parts 2 - 4. Following certain instructions, you will be asked understanding questions. You must answer these understanding questions correctly in order to proceed to complete the study.

Your Payment: For completing this study, you are guaranteed to receive $2 within 24 hours. In addition, one part out of the 4 parts will be randomly selected as the part-that-counts. Any amount you earn in the part-that-counts will be distributed to you as a bonus payment.

Understanding Question: Which of the following statements is true?

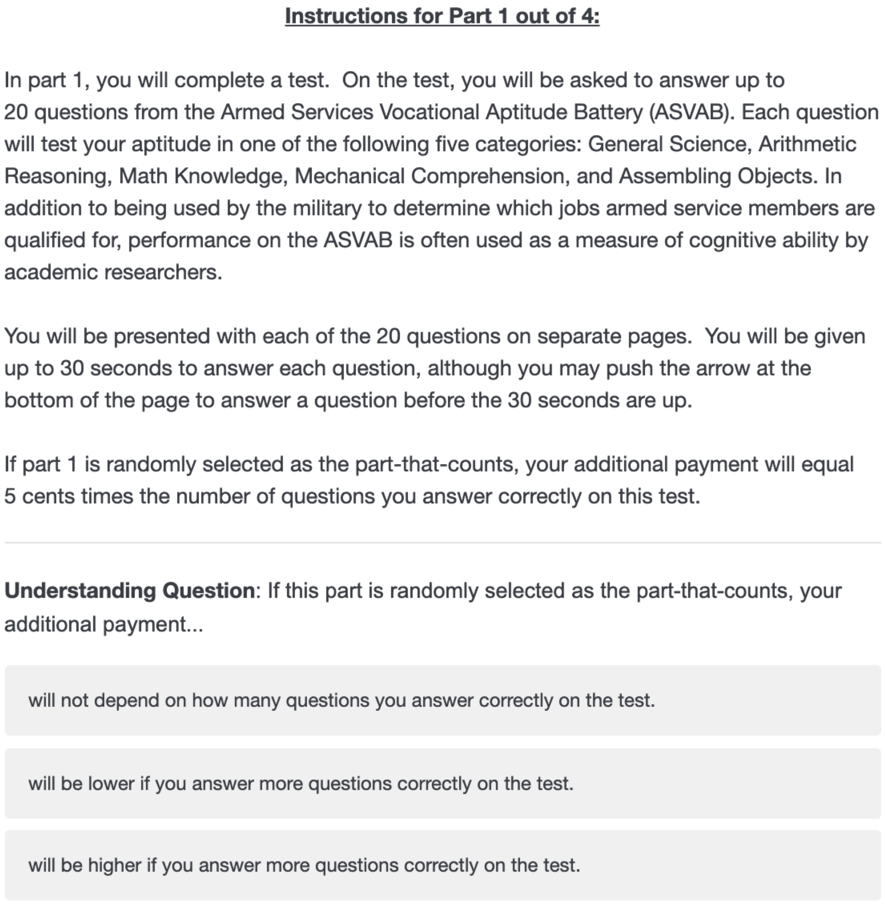
For completing this study, I will receive $2 within 24 hours, but I do NOT have a chance of receiving any additional bonus payment.

For completing this study, I will receive $2 within 24 hours, and I will also receive the amount I earn in the part-that-counts as additional bonus payment.

For completing this study, I will receive $2 within 24 hours, and I will also receive the total amount I earn across all parts as additional bonus payment.

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



Instructions for Part 1 out of 4:

In part 1, you will complete a test. On the test, you will be asked to answer up to 20 questions from the Armed Services Vocational Aptitude Battery (ASVAB). Each question will test your 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.

You will be presented with each of the 20 questions on separate pages. You will be given up to 30 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 30 seconds are up.

If part 1 is randomly selected as the part-that-counts, your additional payment will equal 5 cents times the number of questions you answer correctly on this test.

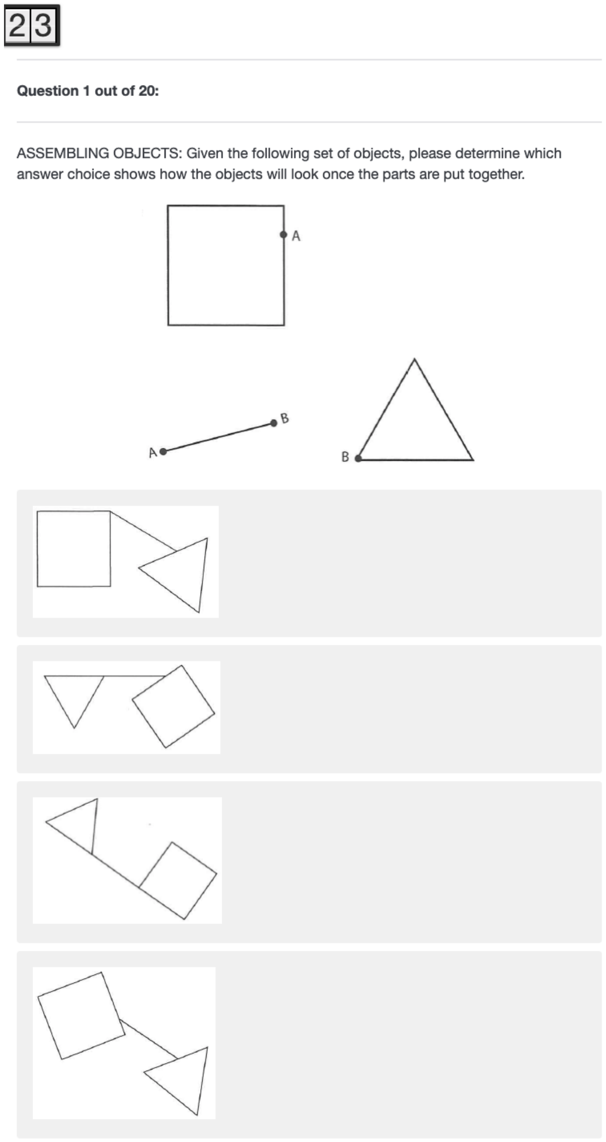
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

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

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

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

Figure D.3: Part 1: Example ASVAB question



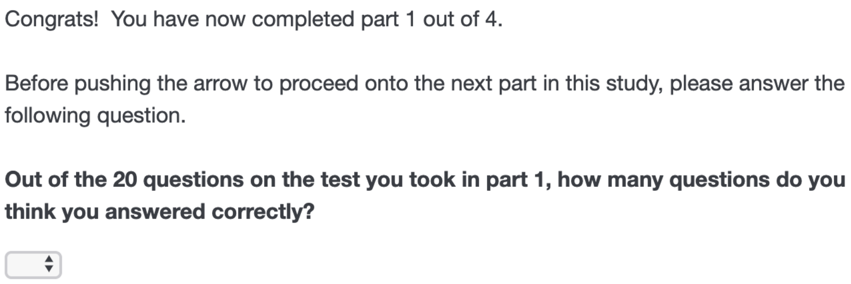
23

Question 1 out of 20:

ASSEMBLING OBJECTS: Given the following set of objects, please determine which answer choice shows how the objects will look once the parts are put together.

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



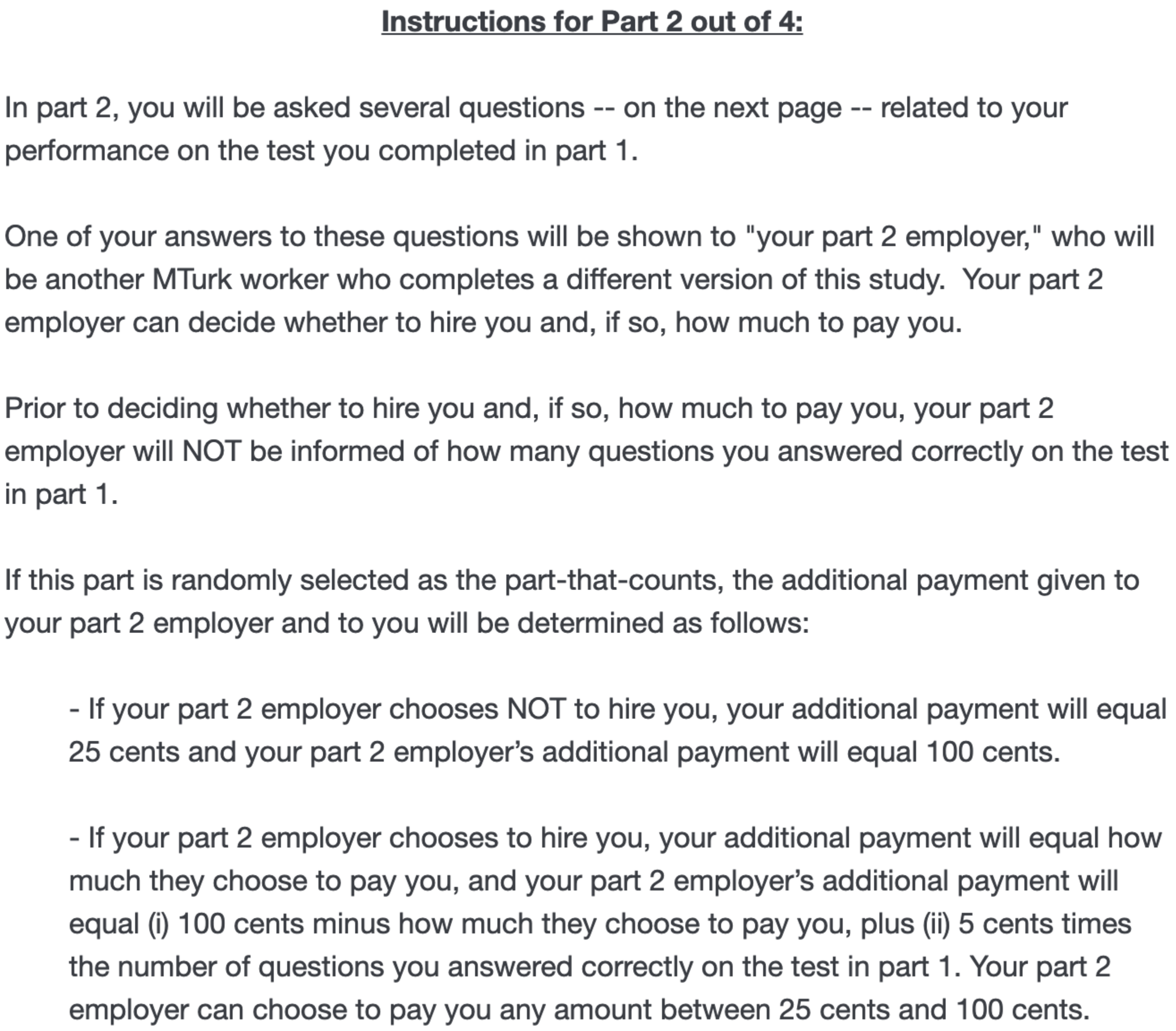
Congrats! You have now completed part 1 out of 4.

Before pushing the arrow to proceed onto the next part in this study, please answer the following question.

Out of the 20 questions on the test you took in part 1, how many questions do you think you answered correctly?

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.

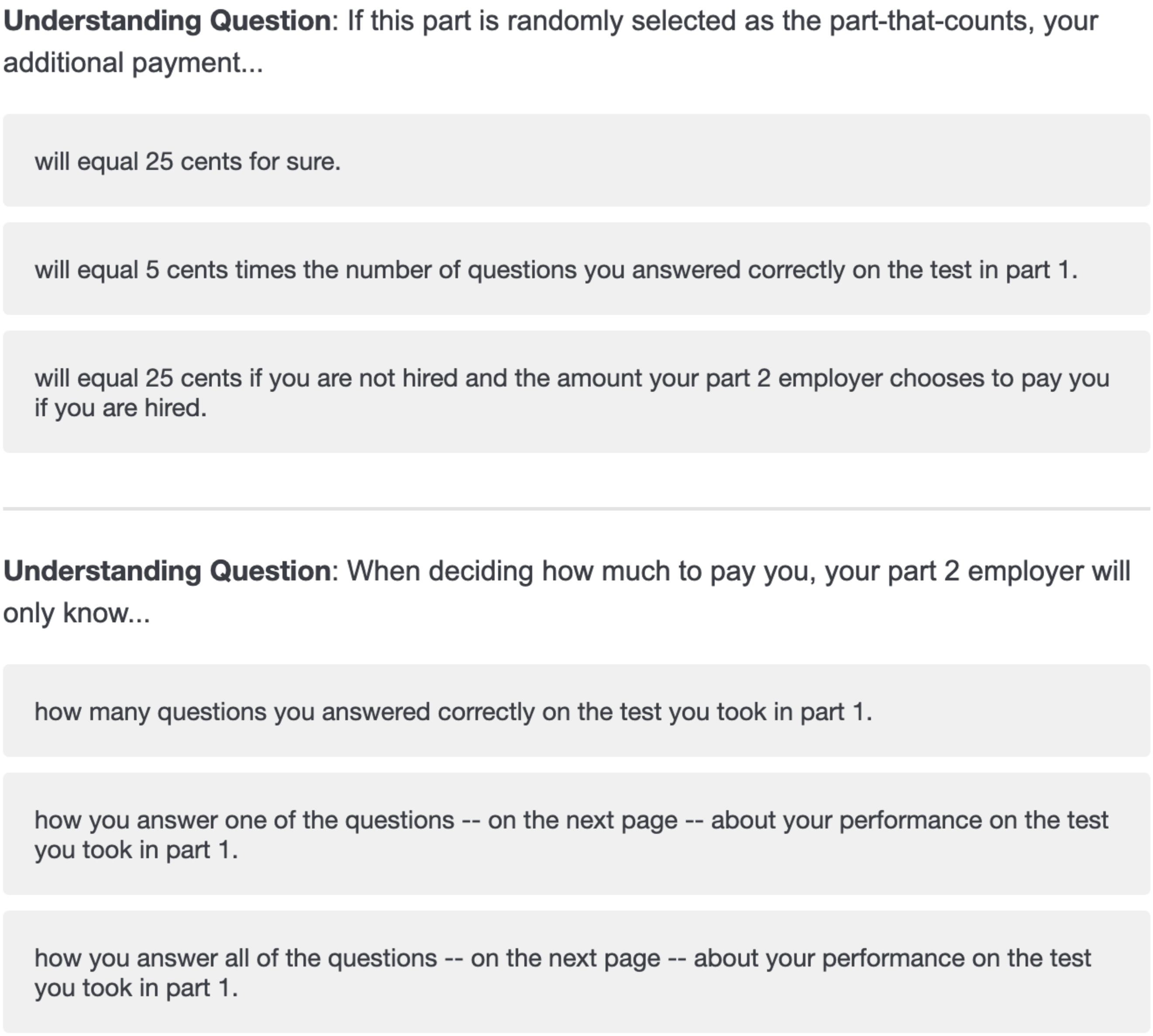
Prior to deciding whether to hire you and, if so, how much to pay you, your part 2 employer will NOT be informed of how many questions you answered correctly on the test in part 1.

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.

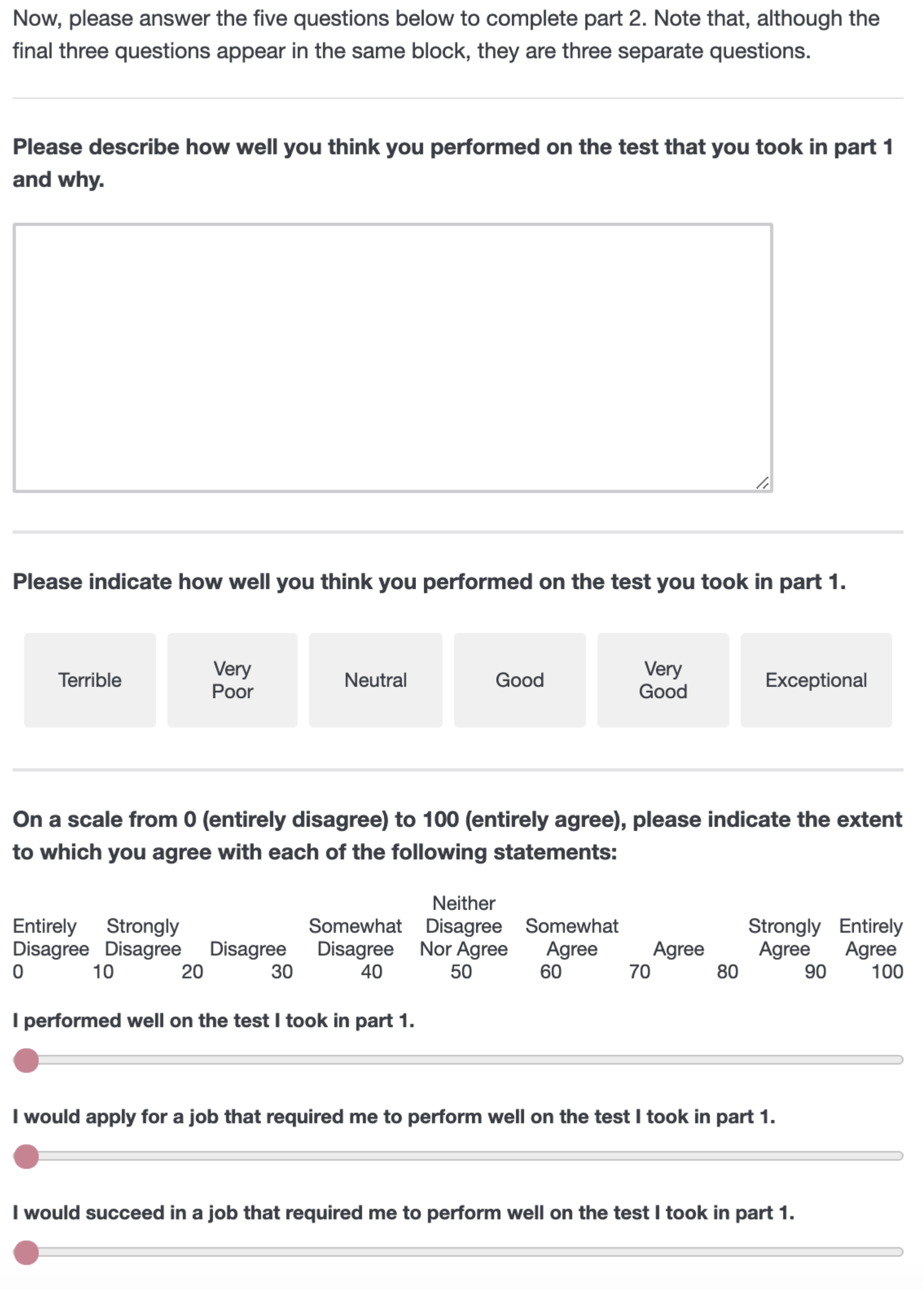
Understanding Question: When deciding how much to pay you, your part 2 employer will only know...

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.

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:

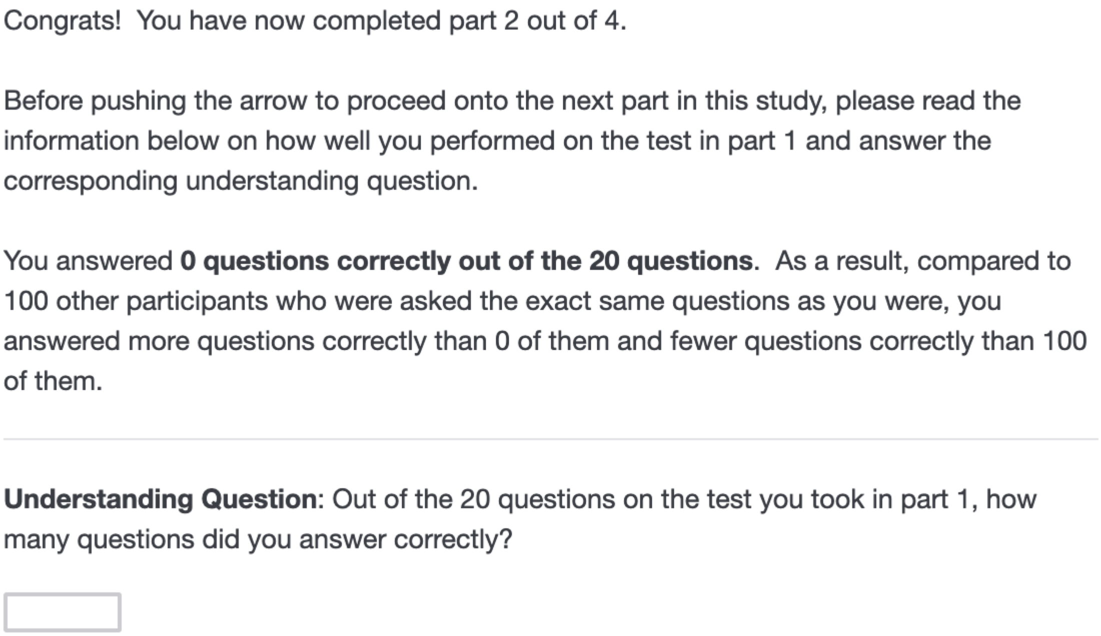
I performed well on the test I took in part 1.

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



Congrats! You have now completed part 2 out of 4.

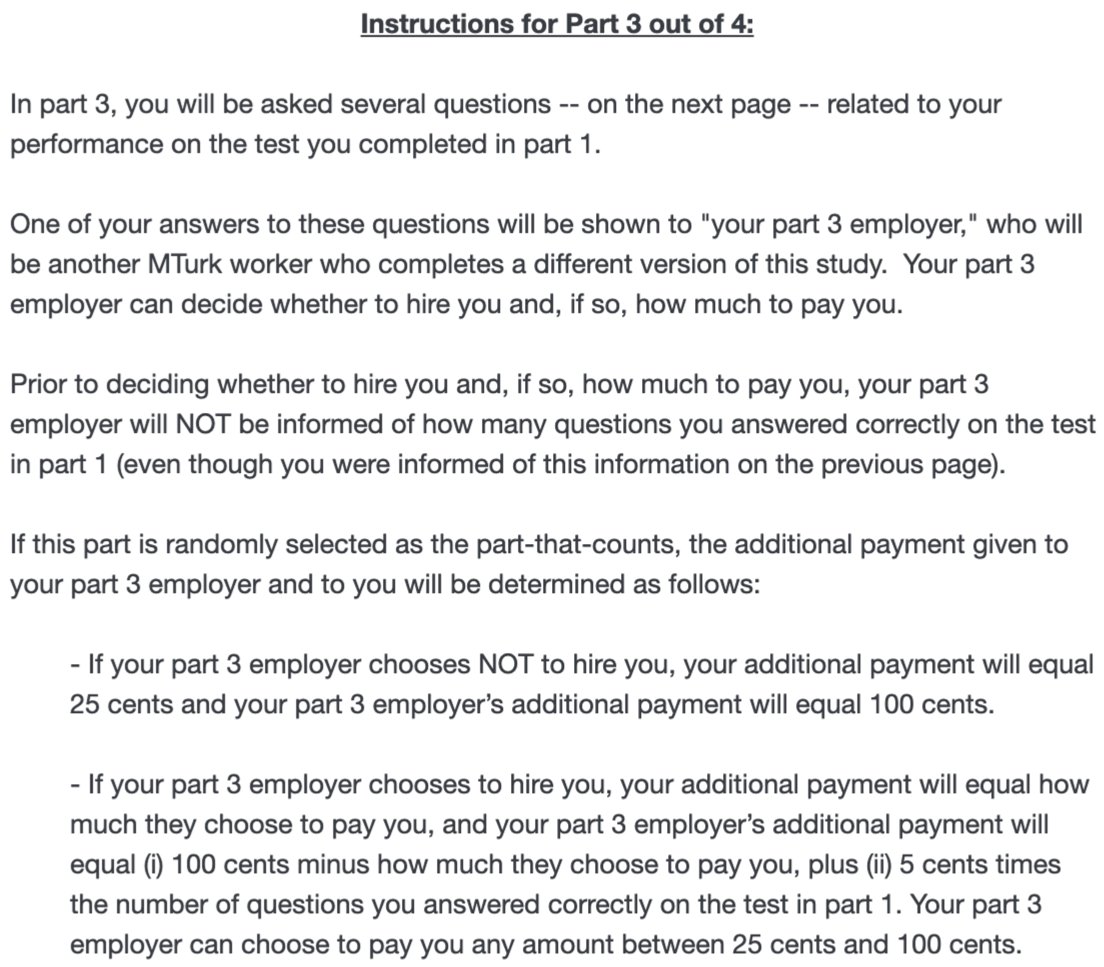
Before pushing the arrow to proceed onto the next part in this study, please read the information below on how well you performed on the test in part 1 and answer the corresponding understanding question.

You answered **0 questions correctly out of the 20 questions**. As a result, compared to 100 other participants who were asked the exact same questions as you were, you answered more questions correctly than 0 of them and fewer questions correctly than 100 of them.

Understanding Question: Out of the 20 questions on the test you took in part 1, how many questions did you answer correctly?

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.

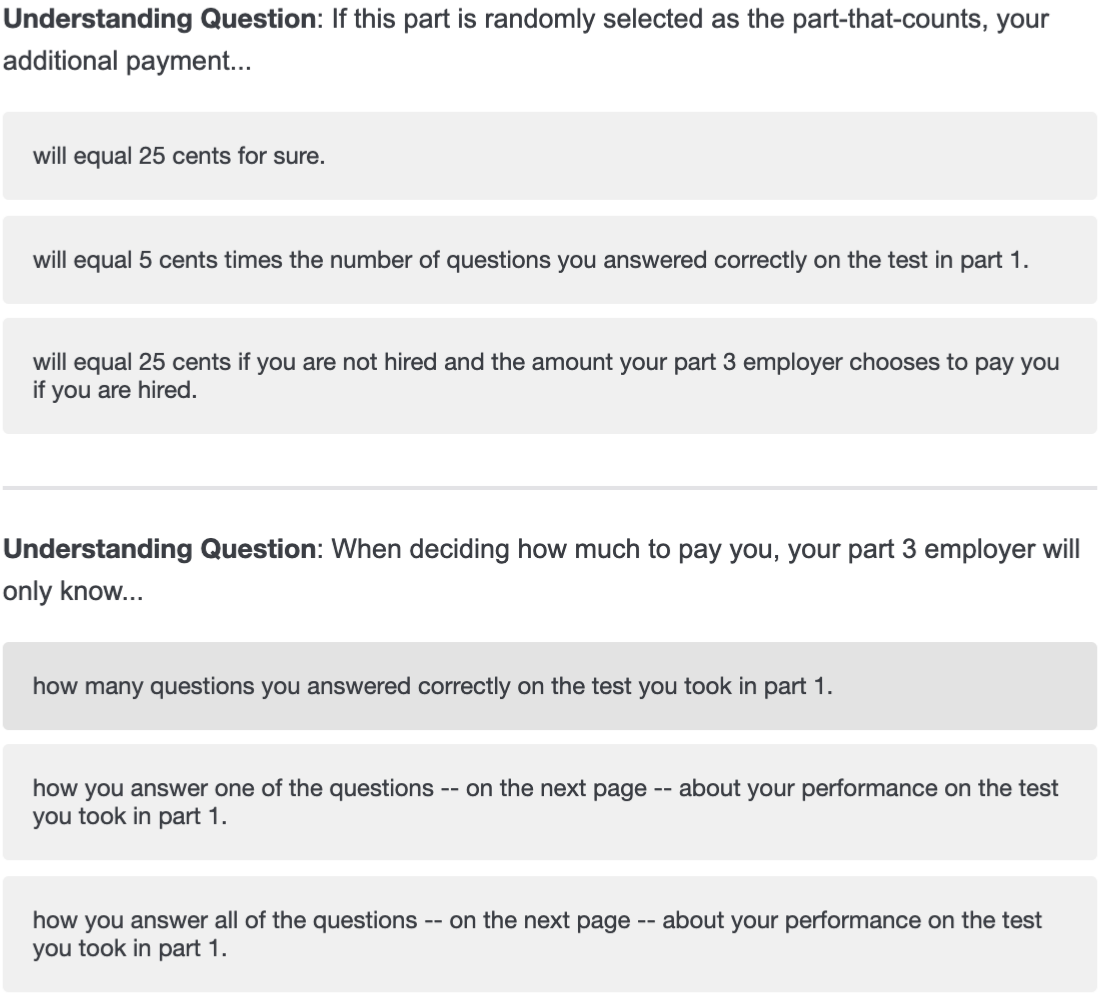
Prior to deciding whether to hire you and, if so, how much to pay you, your part 3 employer will NOT be informed of how many questions you answered correctly on the test in part 1 (even though you were informed of this information on the previous page).

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.

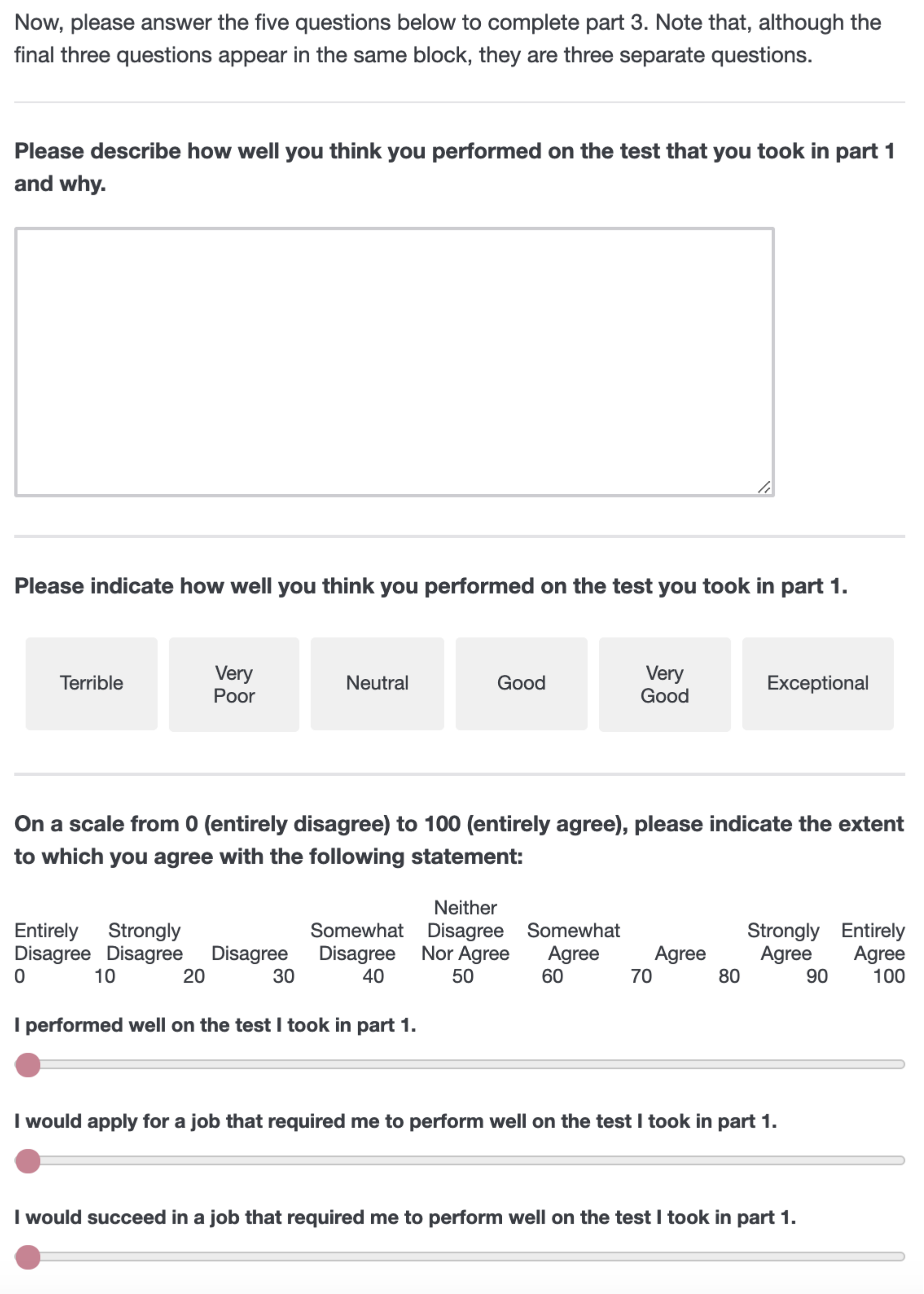
Understanding Question: When deciding how much to pay you, your part 3 employer will only know...

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.

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

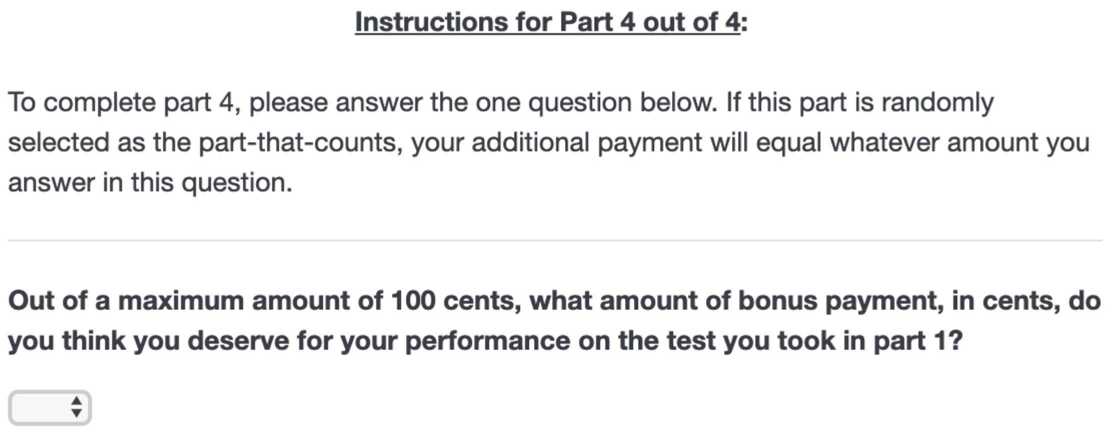
I performed well on the test I took in part 1.

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



Instructions for Part 4 out of 4:

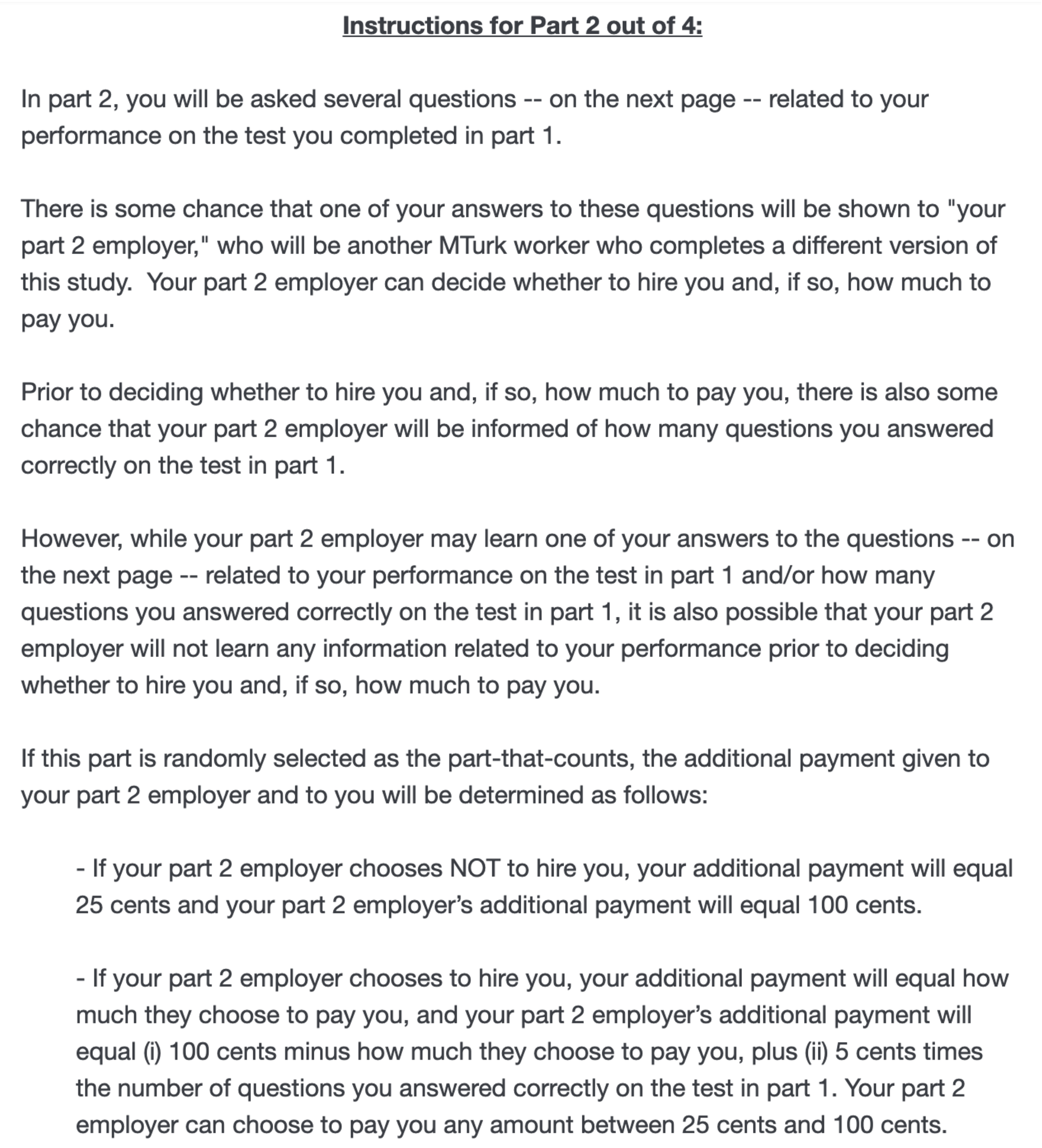
To complete part 4, please answer the one question below. If this part is randomly selected as the part-that-counts, your additional payment will equal whatever amount you answer in this question.

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?

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.

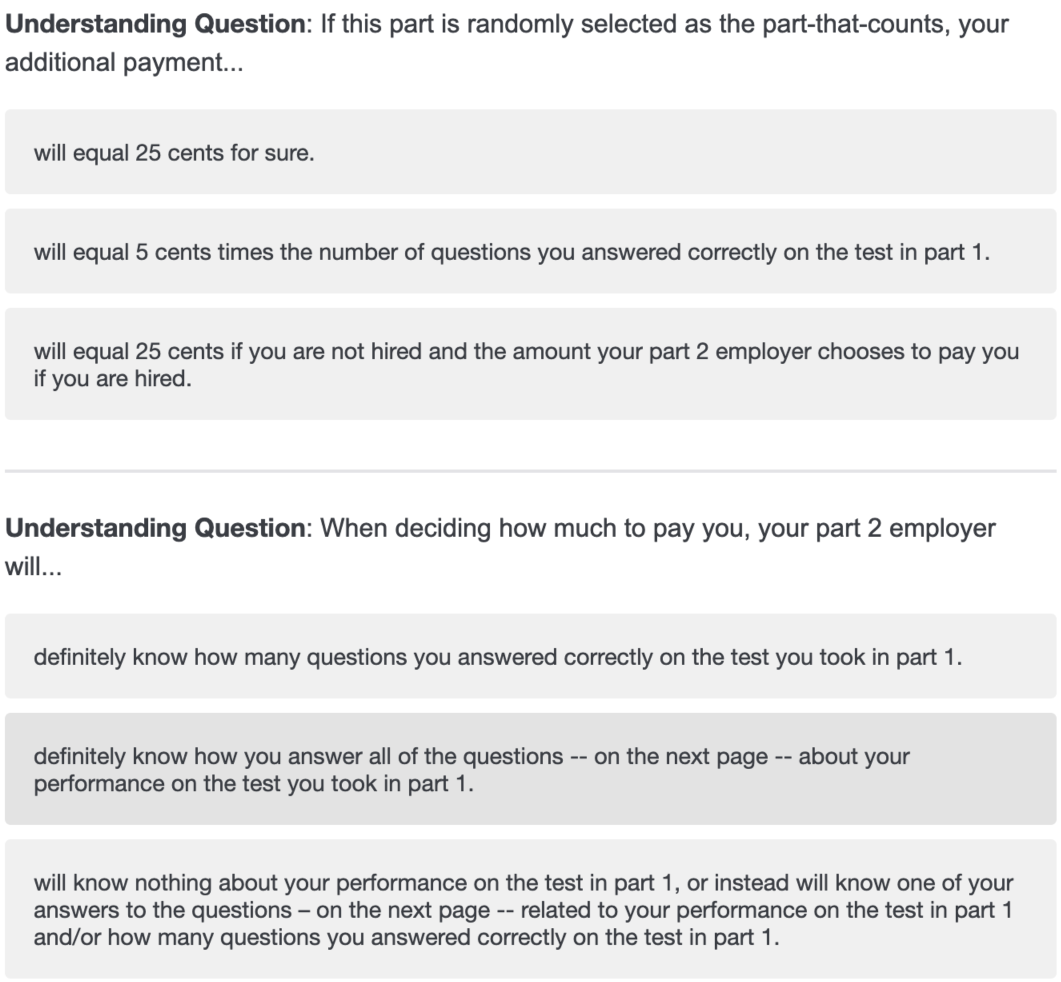
However, while your part 2 employer may learn one of your answers to the questions -- on the next page -- related to your performance on the test in part 1 and/or how many questions you answered correctly on the test in part 1, it is also possible that your part 2 employer will not learn any information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



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...

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

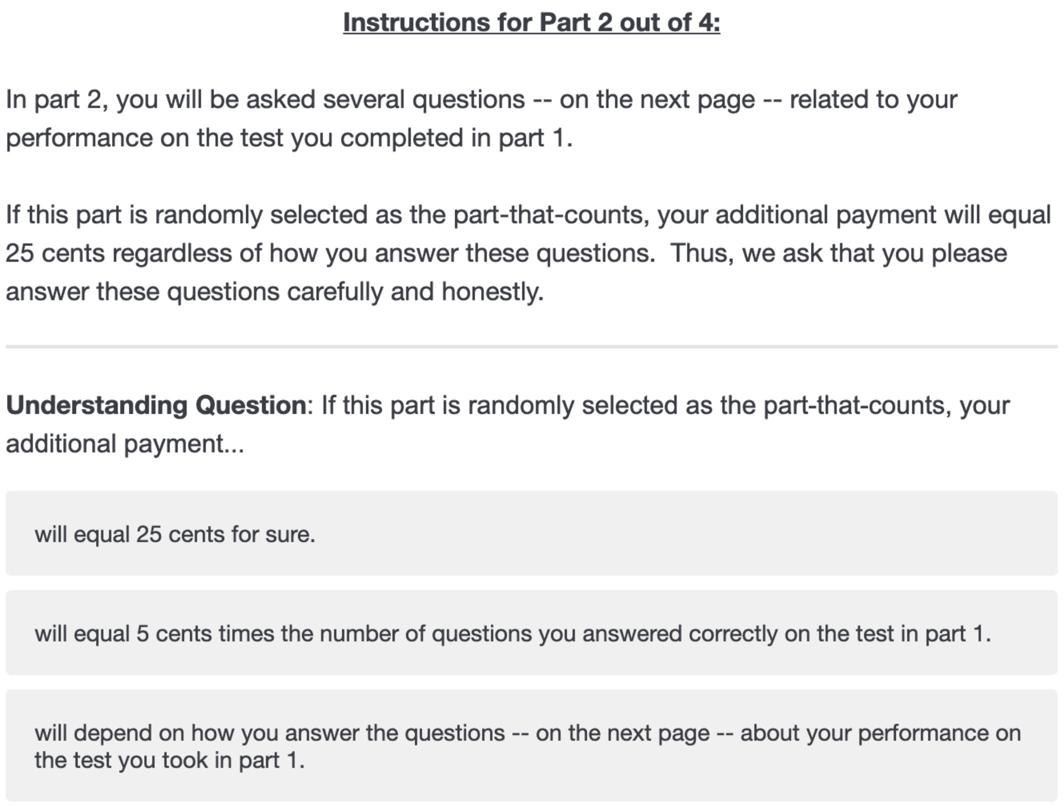
definitely know how you answer all of the questions -- on the next page -- about your performance on the test you took in part 1.

will know nothing about your performance on the test in part 1, or instead will know 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.

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



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.

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.

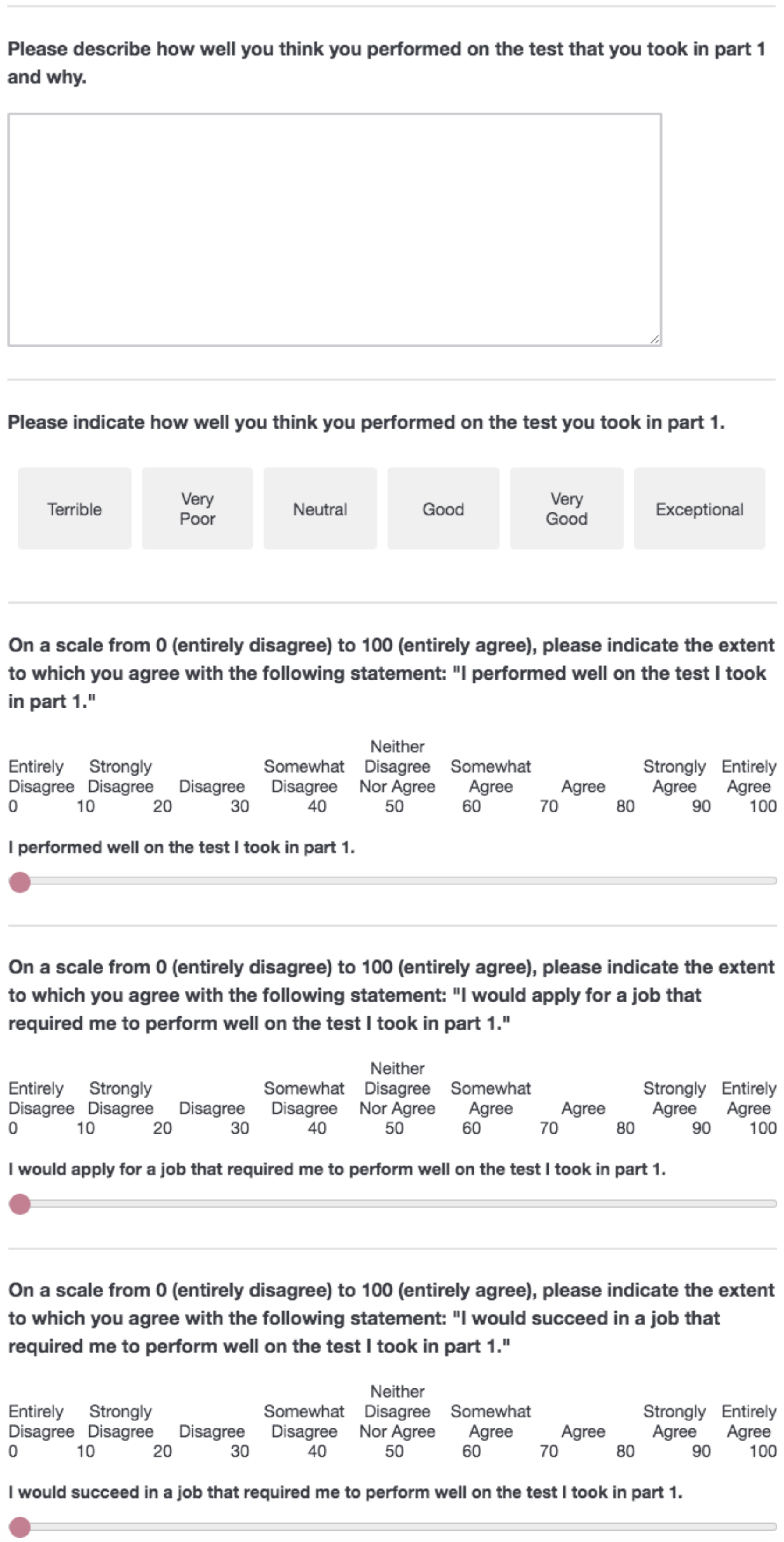
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

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 your performance on the test you took in part 1.

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



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

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.

●

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "I would apply for a job that required me to perform well on the test I took in part 1."

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 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), please indicate the extent to which you agree with the following statement: "I would succeed in a job that required me to perform well on the test I took in part 1."

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

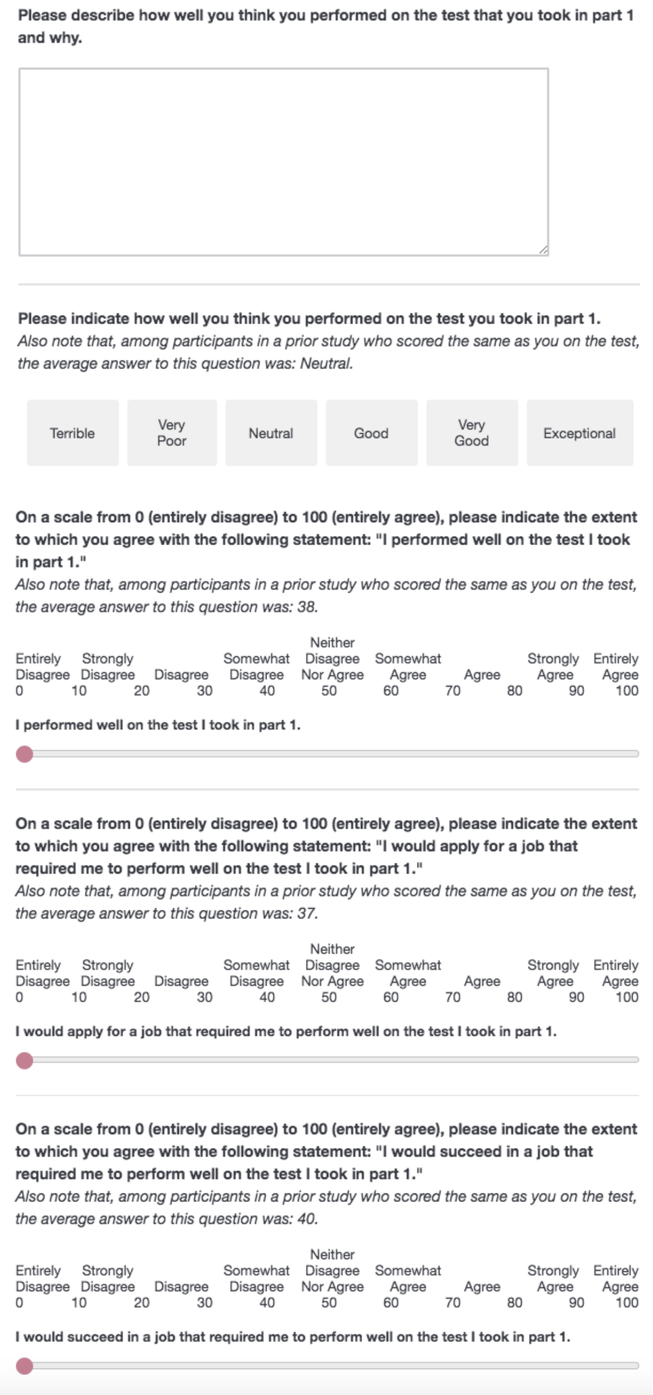
I would succeed in a job that required me to perform well on the test I took in part 1.

●

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



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.
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: Neutral.

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "I performed well on the test I took in part 1."
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: 38.

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

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "I would apply for a job that required me to perform well on the test I took in part 1."
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: 37.

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), please indicate the extent to which you agree with the following statement: "I would succeed in a job that required me to perform well on the test I took in part 1."
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: 40.

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

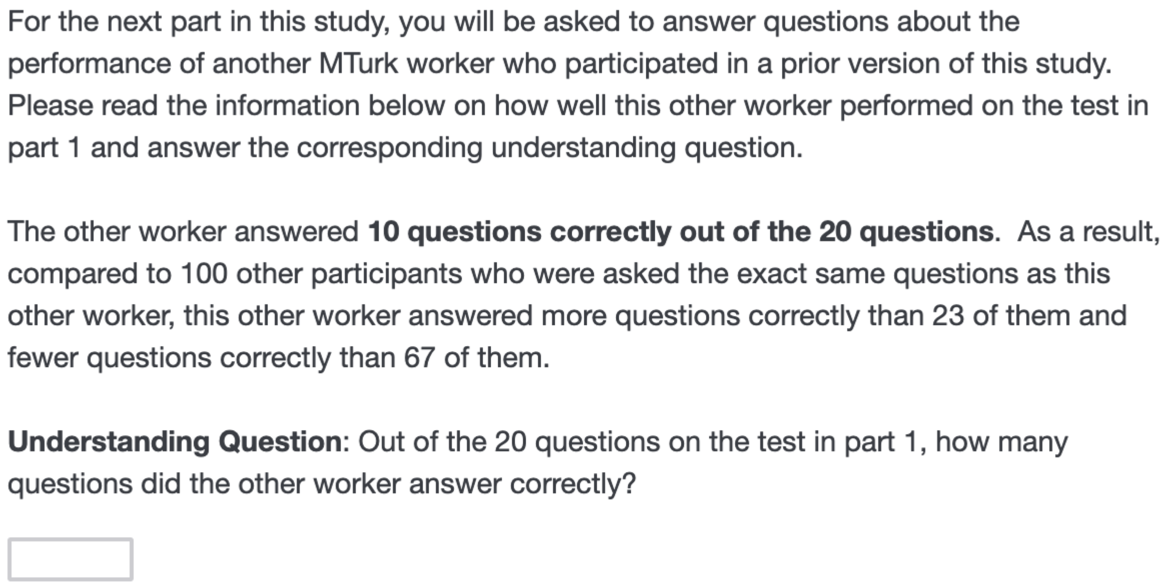
D.5 Instructions for the *Private (Immediately Informed)* version

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

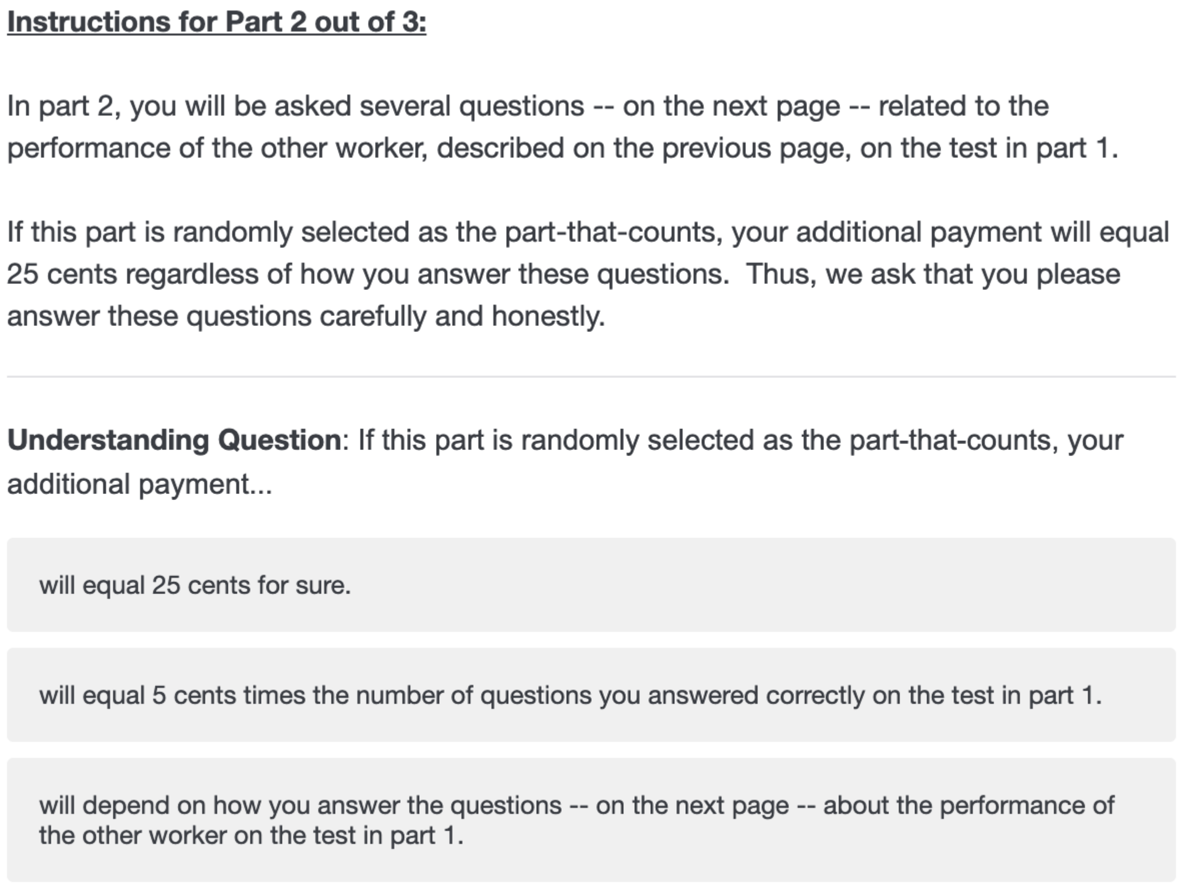


For the next part in this study, you will be asked to answer questions about the performance of another MTurk worker who participated in a prior version of this study. Please read the information below on how well this other worker performed on the test in part 1 and answer the corresponding understanding question.

The other worker answered **10 questions correctly out of the 20 questions**. As a result, compared to 100 other participants who were asked the exact same questions as this other worker, this other worker answered more questions correctly than 23 of them and fewer questions correctly than 67 of them.

Understanding Question: Out of the 20 questions on the test in part 1, how many questions did the other worker answer correctly?

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.

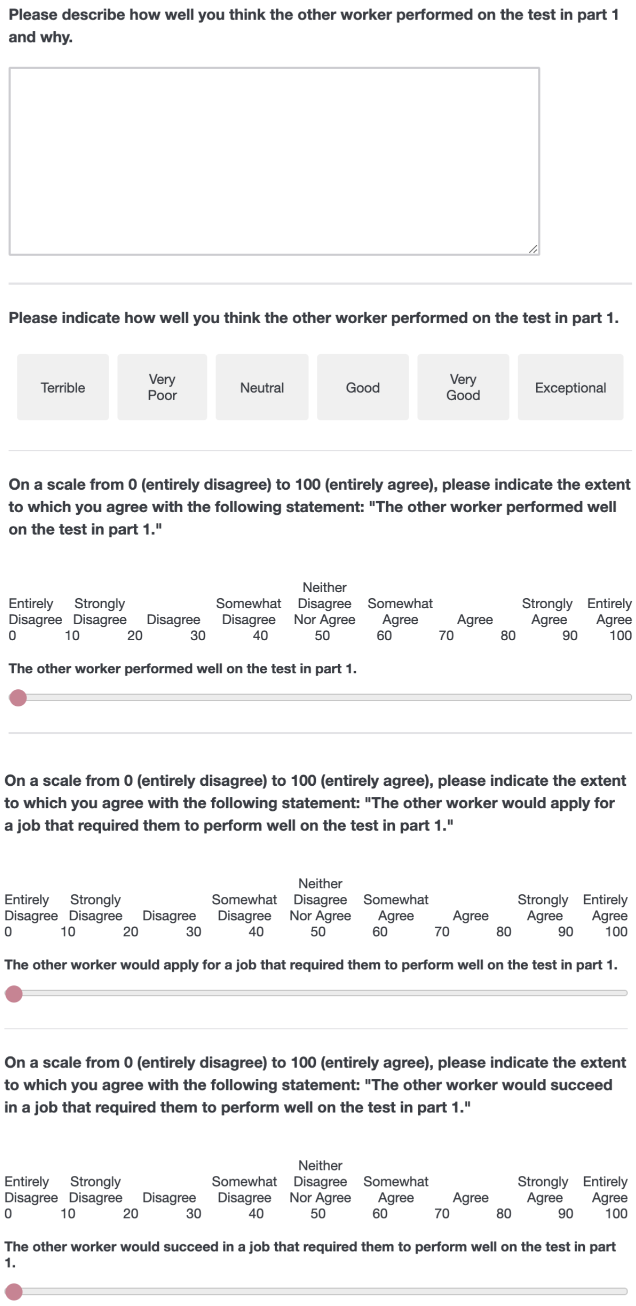
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

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



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

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

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "The other worker performed well on the test in part 1."

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

The other worker performed well on the test in part 1.

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "The other worker would apply for a job that required them to perform well on the test in part 1."

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

The other worker would apply for a job that required them to perform well on the test in part 1.

On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "The other worker would succeed in a job that required them to perform well on the test in part 1."

Entirely Disagree · Strongly Disagree · Disagree · Somewhat Disagree · Neither Disagree Nor Agree · Somewhat Agree · Agree · Strongly Agree · Entirely Agree

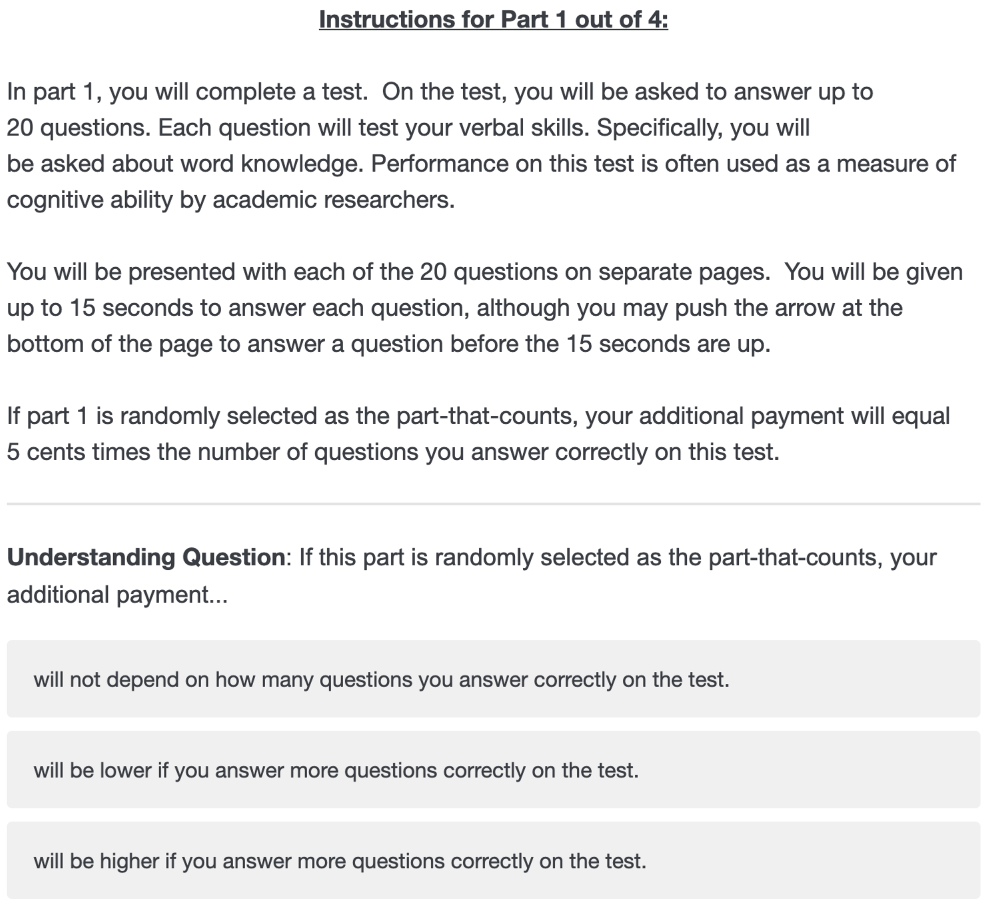
0 10 20 30 40 50 60 70 80 90 100

The other worker would succeed in a job that required them to perform well on the test in part 1.

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.[16] 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



Instructions for Part 1 out of 4:

In part 1, you will complete a test. On the test, you will be asked to answer up to 20 questions. Each question will test your verbal skills. Specifically, you will be asked about word knowledge. Performance on this test is often used as a measure of cognitive ability by academic researchers.

You will be presented with each of the 20 questions on separate pages. You will be given up to 15 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 15 seconds are up.

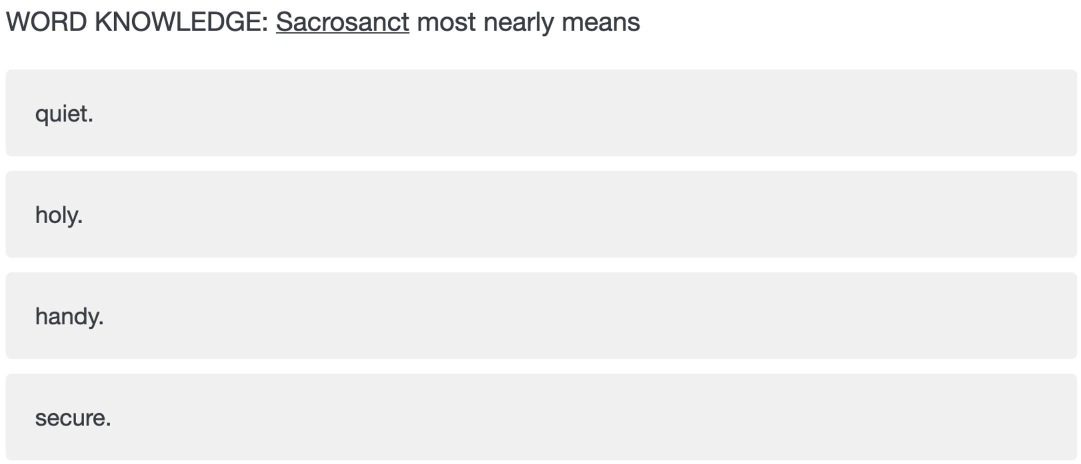
If part 1 is randomly selected as the part-that-counts, your additional payment will equal 5 cents times the number of questions you answer correctly on this test.

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

- will not depend on how many questions you answer correctly on the test.
- will be lower if you answer more questions correctly on the test.
- will be higher if you answer more questions correctly on the test.

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

Figure D.22: Part 1: Example Verbal Question



WORD KNOWLEDGE: Sacrosanct most nearly means

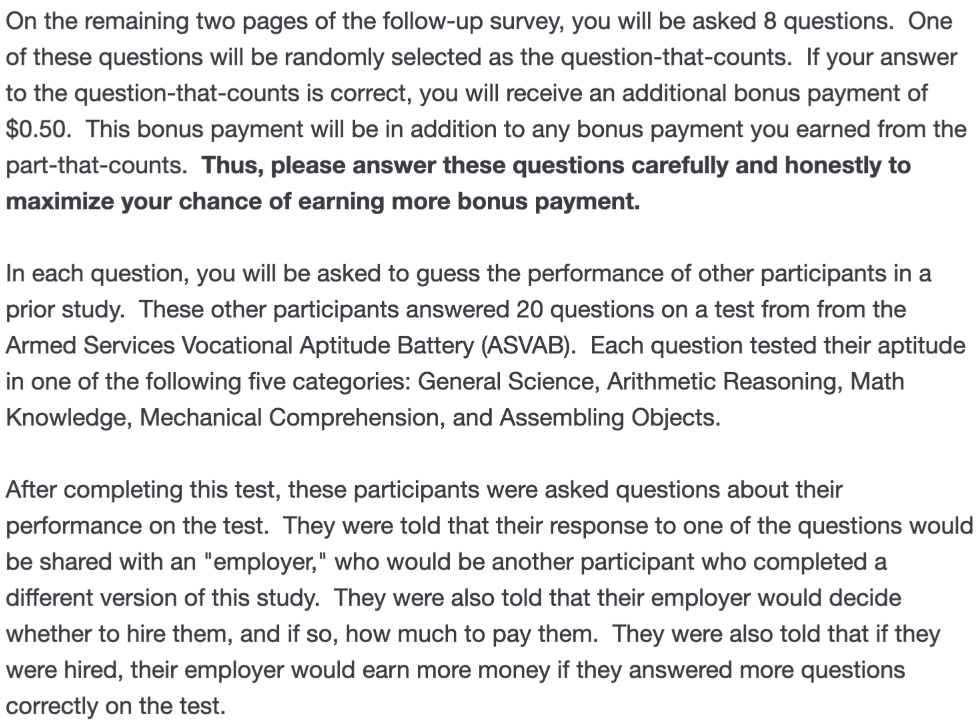
quiet.

holy.

handy.

secure.

Figure D.23: Instructions for Predictions

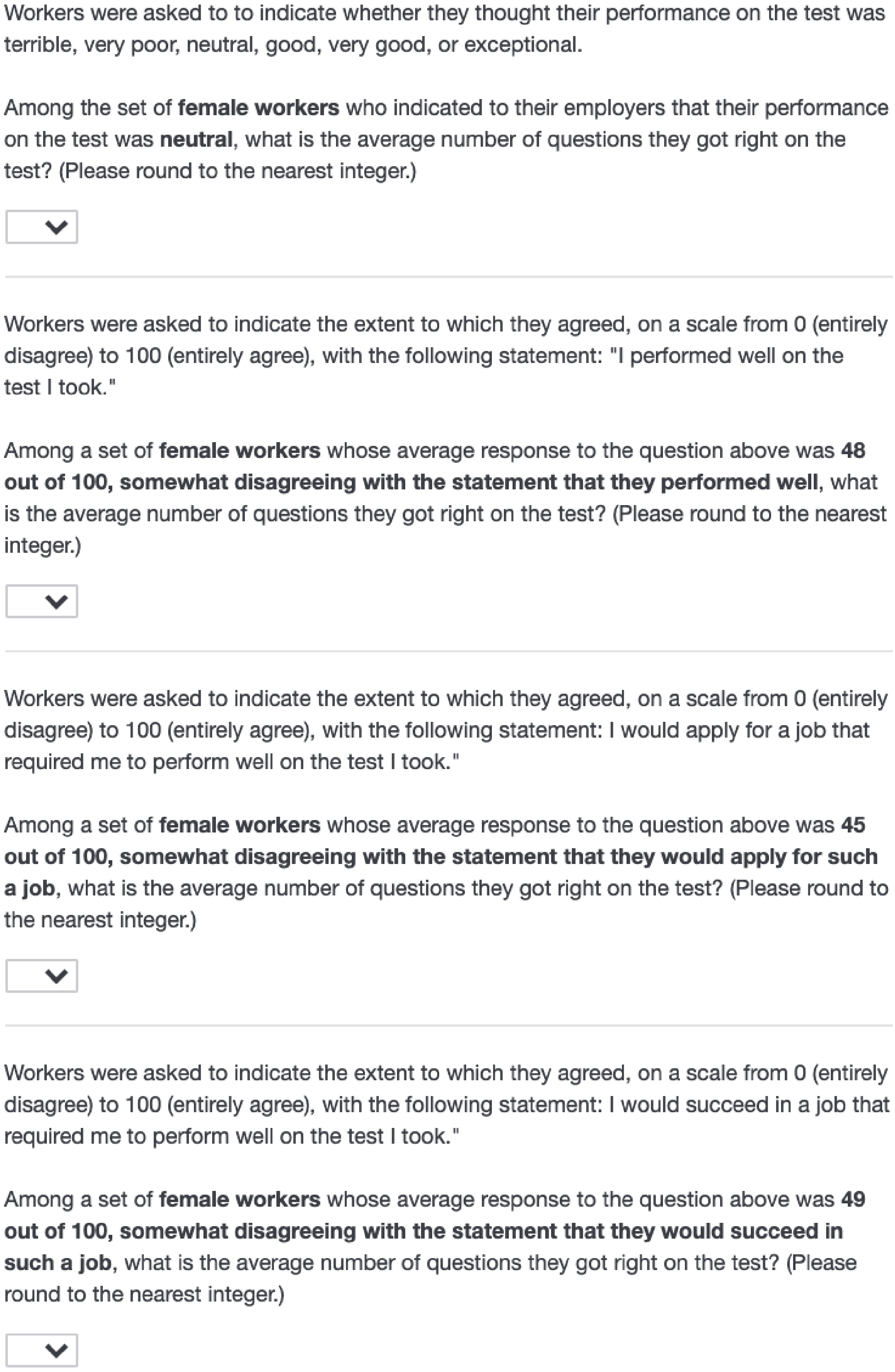


On the remaining two pages of the follow-up survey, you will be asked 8 questions. One of these questions will be randomly selected as the question-that-counts. If your answer to the question-that-counts is correct, you will receive an additional bonus payment of $0.50. This bonus payment will be in 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.)

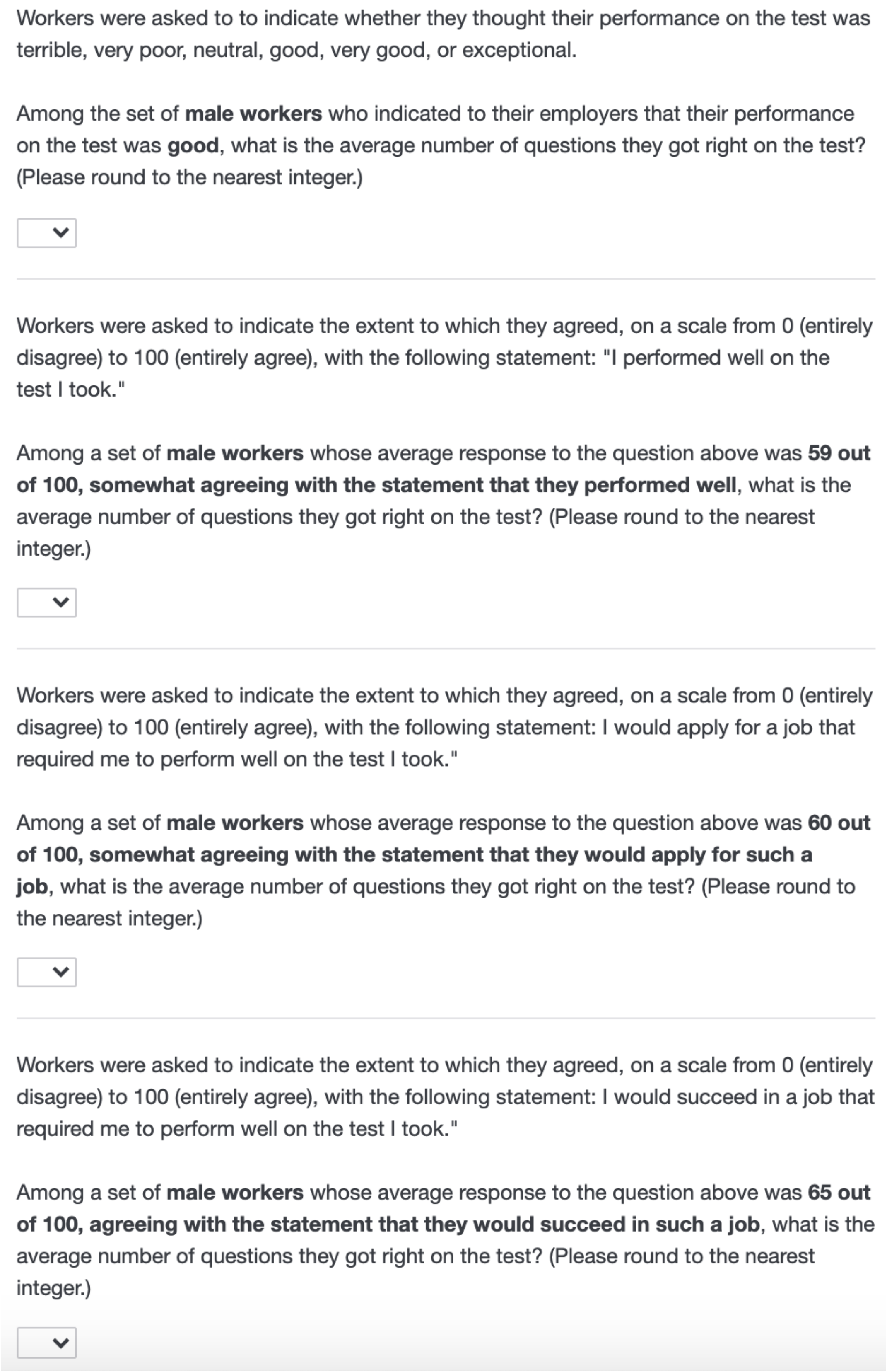
Workers were asked to indicate the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: I would apply for a job that required me to perform well on the test I took."

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."

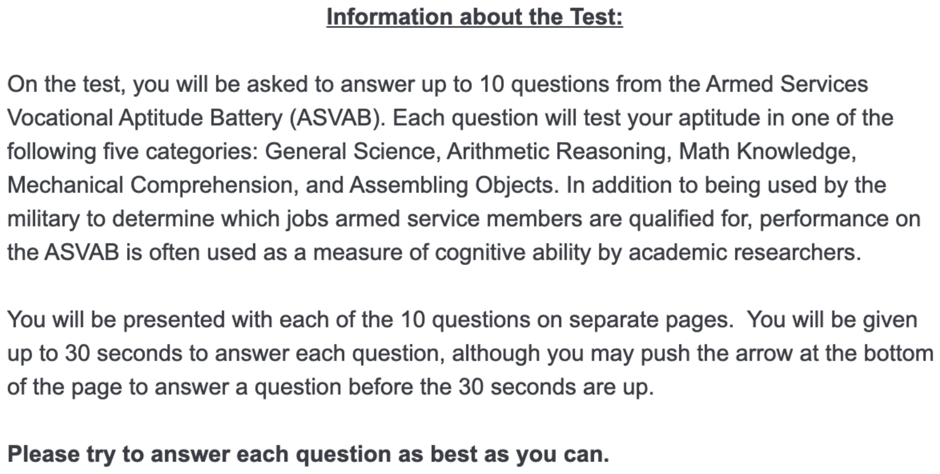
Among a set of **male workers** whose average response to the question above was **65 out of 100, agreeing with the statement that they would succeed in such a job**, what is the average number of questions they got right on the test? (Please round to the nearest integer.)

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



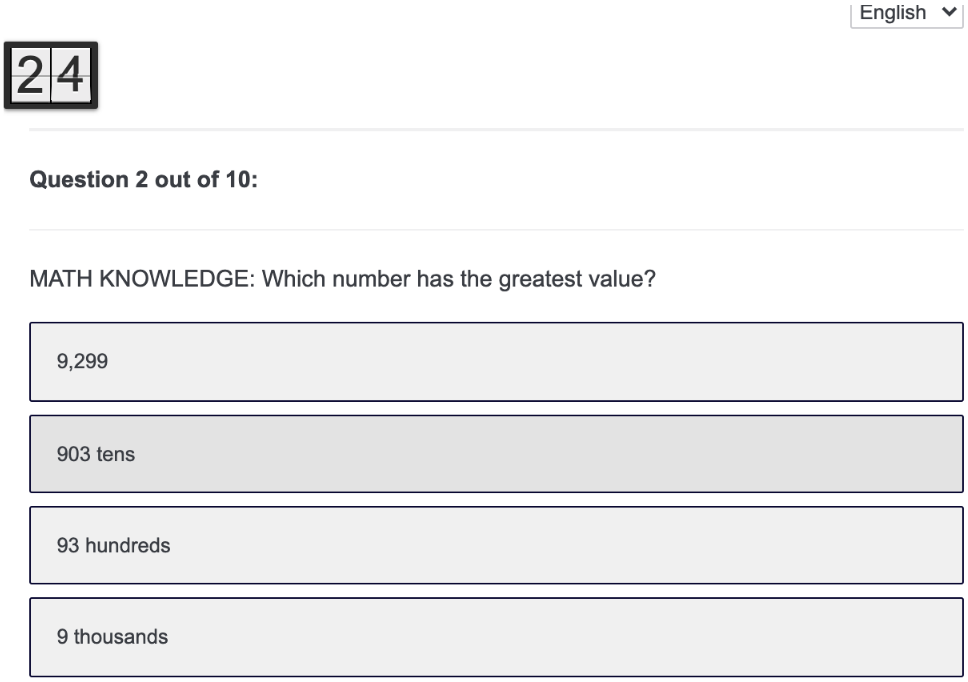
Information about the Test:

On the test, you will be asked to answer up to 10 questions from the Armed Services Vocational Aptitude Battery (ASVAB). Each question will test your 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.

You will be presented with each of the 10 questions on separate pages. You will be given up to 30 seconds to answer each question, although you may push the arrow at the bottom of the page to answer a question before the 30 seconds are up.

Please try to answer each question as best as you can.

Figure D.27: Example question on the test



Question 2 out of 10:

MATH KNOWLEDGE: Which number has the greatest value?

9,299

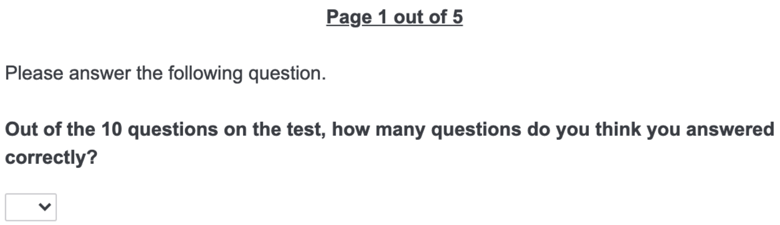
903 tens

93 hundreds

9 thousands

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



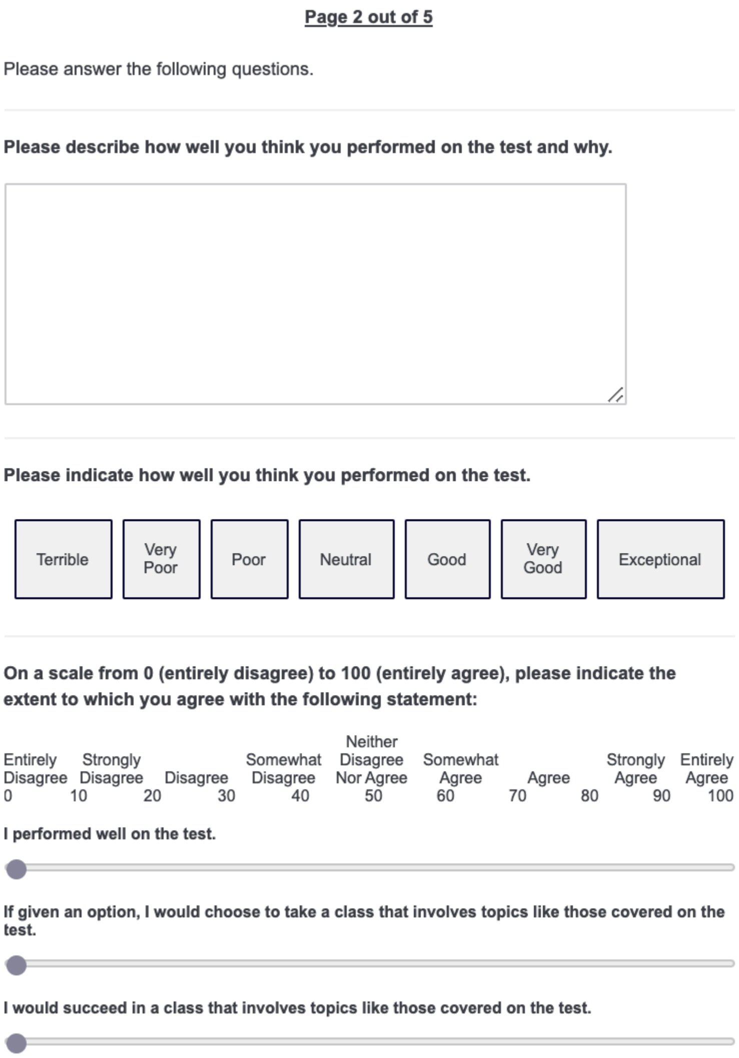
Page 1 out of 5

Please answer the following question.

Out of the 10 questions on the test, how many questions do you think you answered correctly?

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

Figure D.29: Self-Evaluation Questions



Page 2 out of 5

Please answer the following questions.

Please describe how well you think you performed on the test and why.

Please indicate how well you think you performed on the test.

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

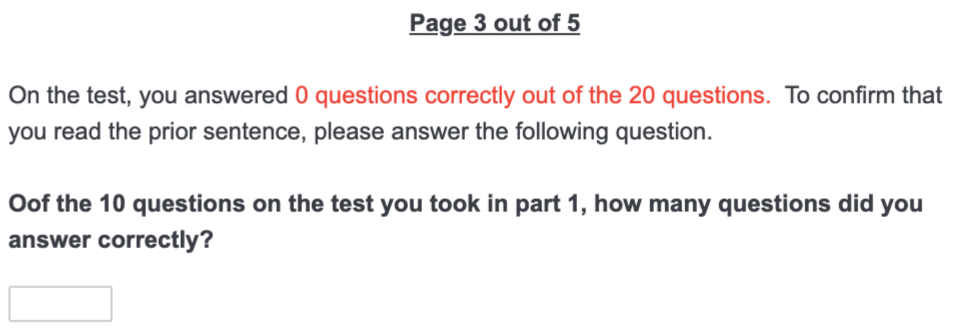
I performed well on the test.

If given an option, I would choose to take a class that involves topics like those covered on the test.

I would succeed in a class that involves topics like those covered on the test.

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



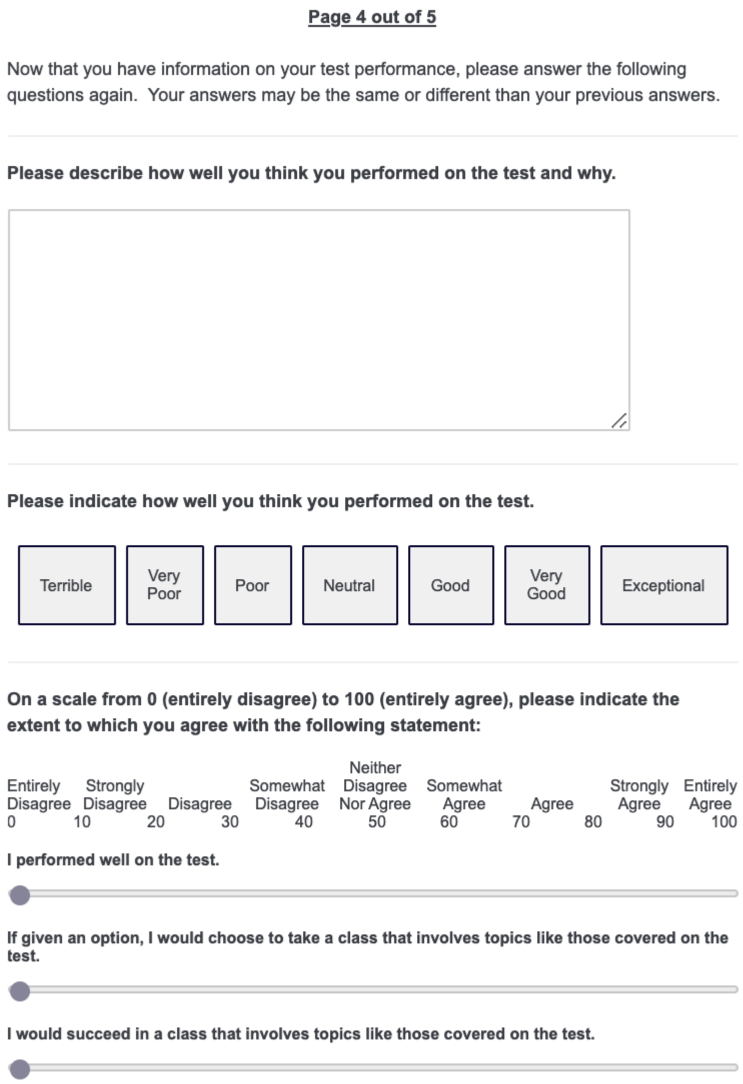
Page 3 out of 5

On the test, you answered 0 questions correctly out of the 20 questions. To confirm that you read the prior sentence, please answer the following question.

Oof the 10 questions on the test you took in part 1, how many questions did you answer correctly?

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



Page 4 out of 5

Now that you have information on your test performance, please answer the following questions again. Your answers may be the same or different than your previous answers.

Please describe how well you think you performed on the test and why.

Please indicate how well you think you performed on the test.

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

I performed well on the test.

If given an option, I would choose to take a class that involves topics like those covered on the test.

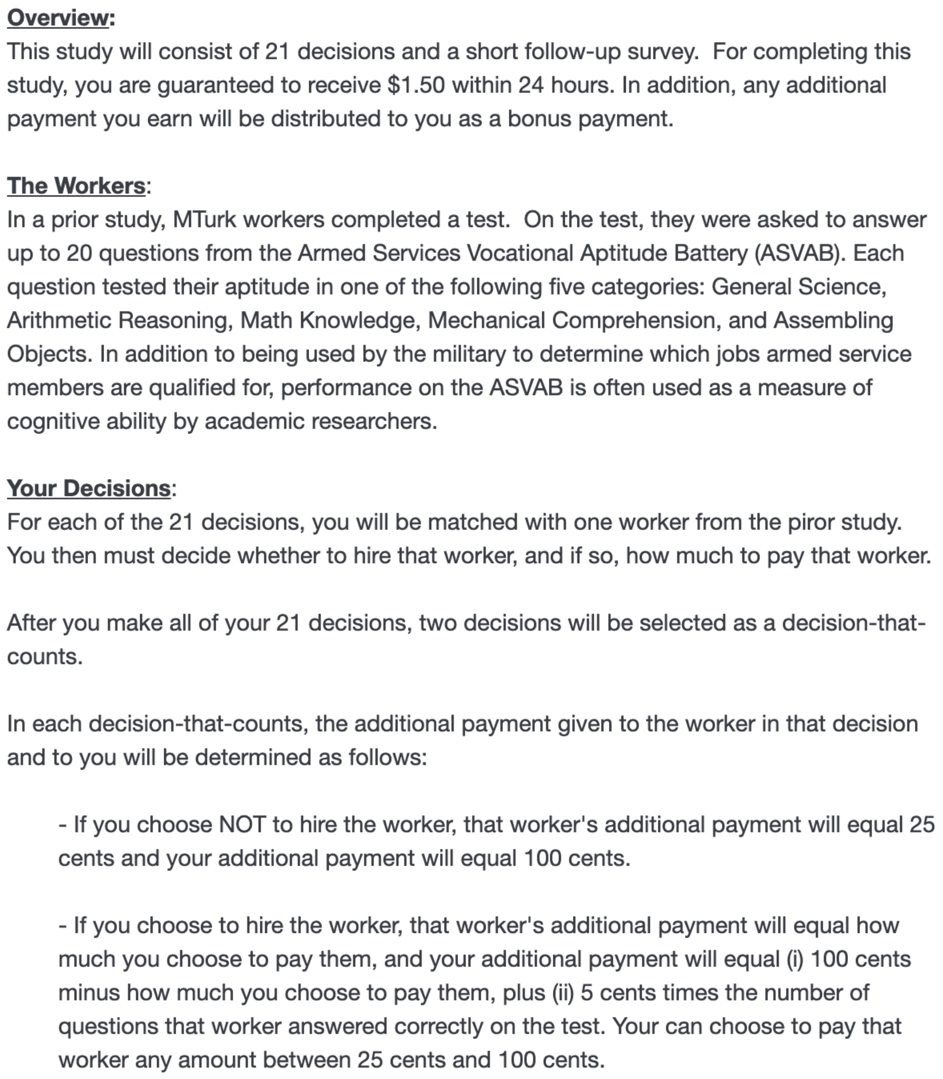
I would succeed in a class that involves topics like those covered on the test.

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.

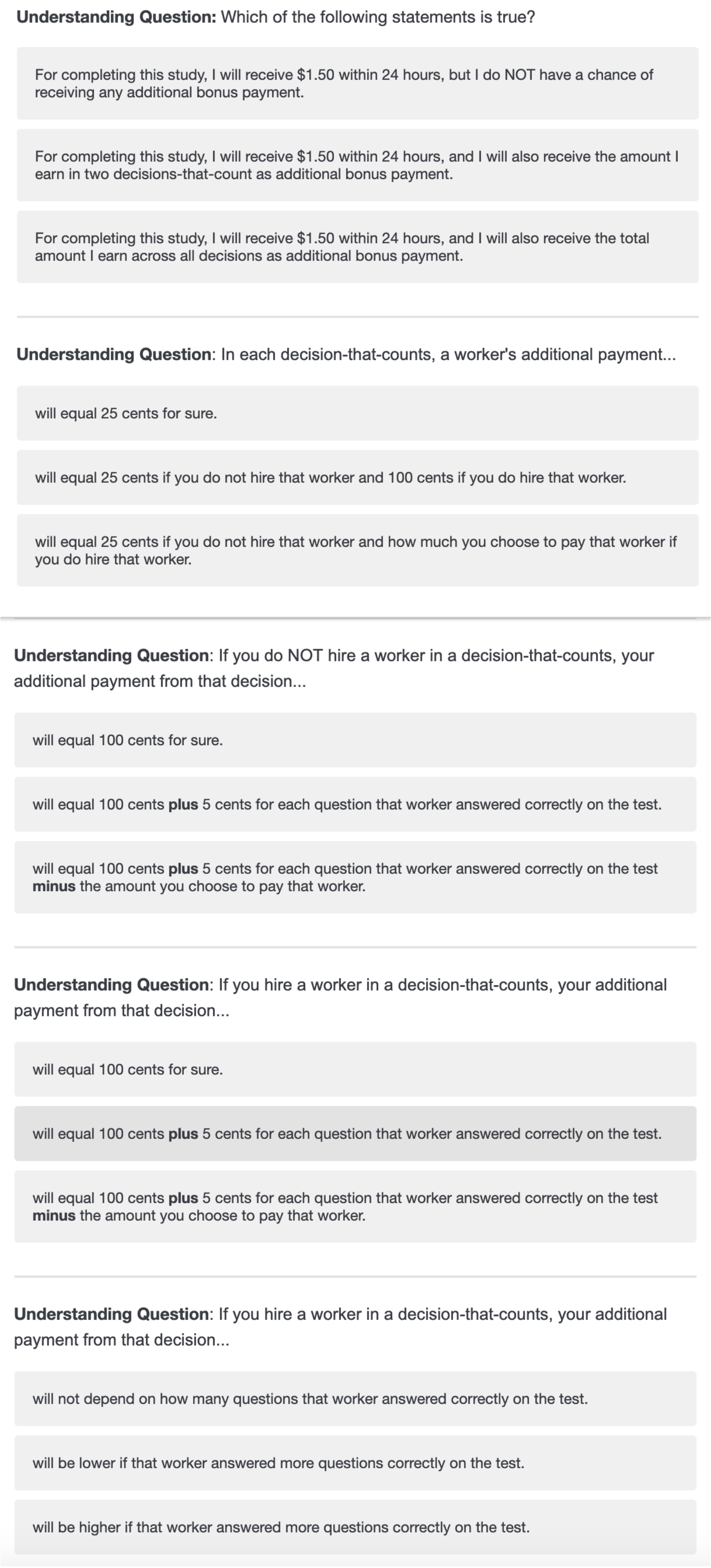
After you make all of your 21 decisions, two decisions will be selected as a decision-that-counts.

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



Understanding Question: Which of the following statements is true?

For completing this study, I will receive $1.50 within 24 hours, but I do NOT have a chance of receiving any additional bonus payment.

For completing this study, I will receive $1.50 within 24 hours, and I will also receive the amount I earn in two decisions-that-count as additional bonus payment.

For completing this study, I will receive $1.50 within 24 hours, and I will also receive the total amount I earn across all decisions as additional bonus payment.

Understanding Question: In each decision-that-counts, a worker's additional payment...

will equal 25 cents for sure.

will equal 25 cents if you do not hire that worker and 100 cents if you do hire that worker.

will equal 25 cents if you do not hire that worker and how much you choose to pay that worker if you do hire that worker.

Understanding Question: If you do NOT hire a worker in a decision-that-counts, your additional payment from that decision...

will equal 100 cents for sure.

will equal 100 cents **plus** 5 cents for each question that worker answered correctly on the test.

will equal 100 cents **plus** 5 cents for each question that worker answered correctly on the test **minus** the amount you choose to pay that worker.

Understanding Question: If you hire a worker in a decision-that-counts, your additional payment from that decision...

will equal 100 cents for sure.

will equal 100 cents **plus** 5 cents for each question that worker answered correctly on the test.

will equal 100 cents **plus** 5 cents for each question that worker answered correctly on the test **minus** the amount you choose to pay that worker.

Understanding Question: If you hire a worker in a decision-that-counts, your additional payment from that decision...

will not depend on how many questions that worker answered correctly on the test.

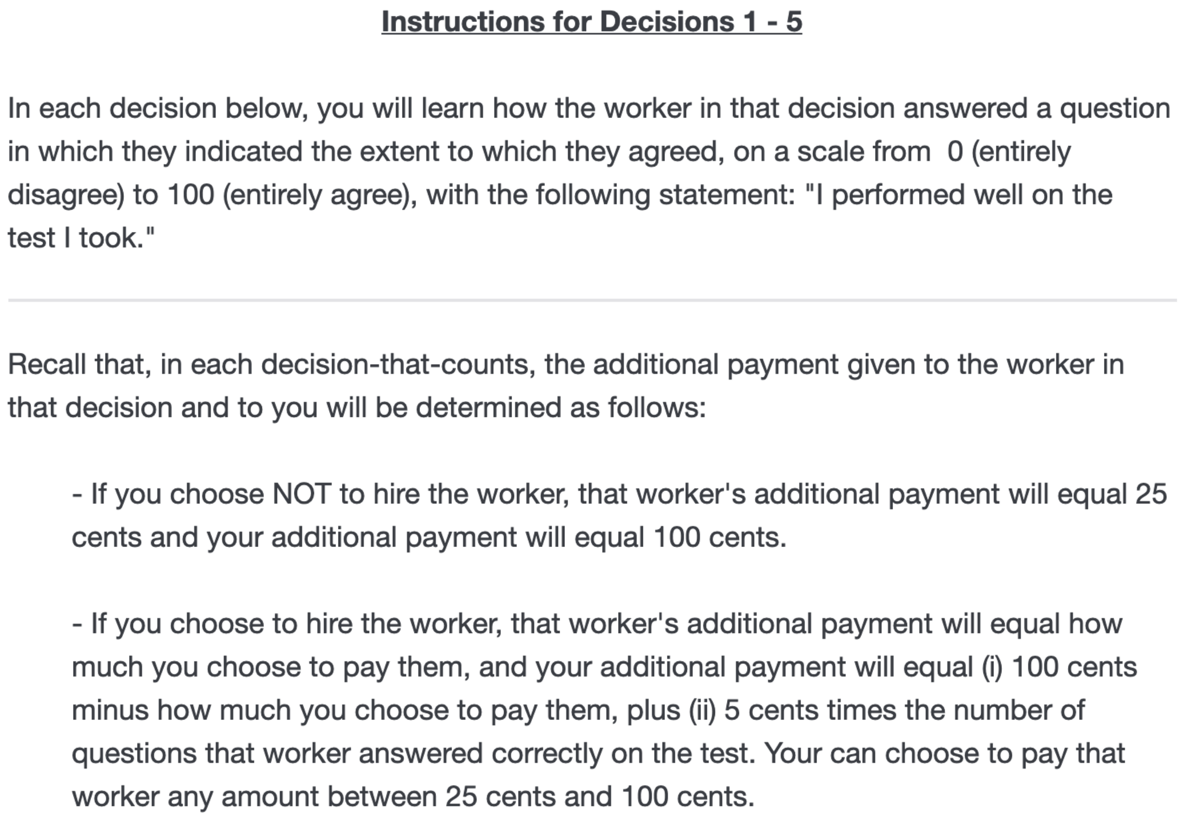
will be lower if that worker answered more questions correctly on the test.

will be higher if that worker answered more questions correctly on the test.

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



Instructions for Decisions 1 - 5

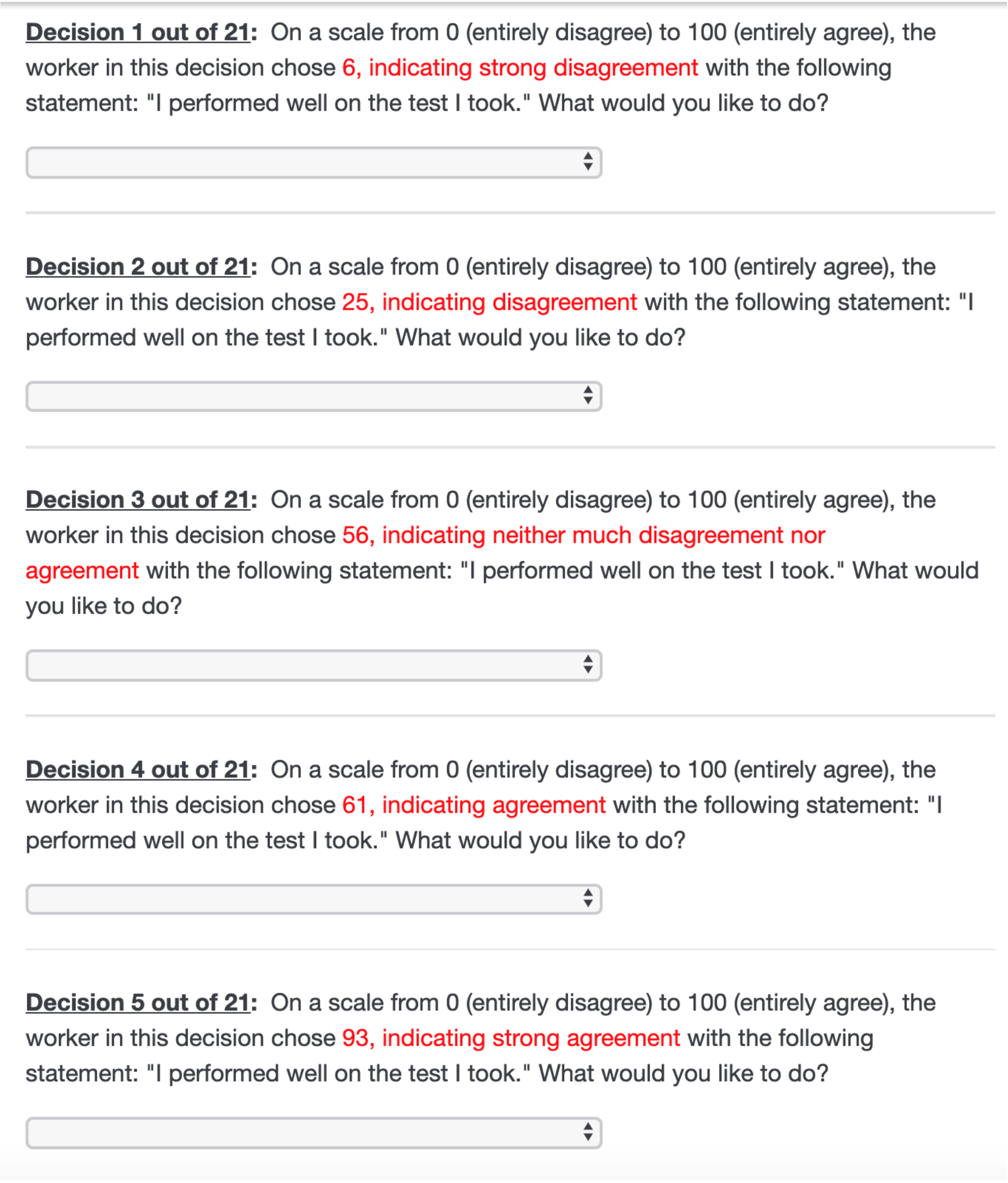
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I performed well on the test I took."

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.35: *Performance Evaluation* Decisions



Decision 1 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 6, indicating strong disagreement with the following statement: "I performed well on the test I took." What would you like to do?

Decision 2 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 25, indicating disagreement with the following statement: "I performed well on the test I took." What would you like to do?

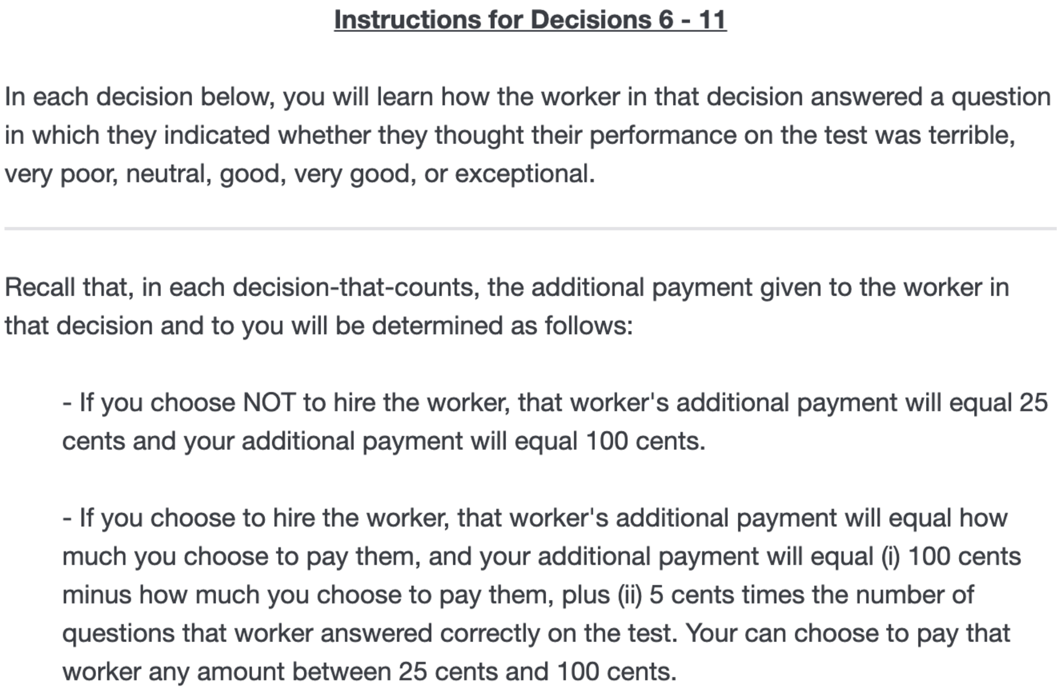
Decision 3 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 56, indicating neither much disagreement nor agreement with the following statement: "I performed well on the test I took." What would you like to do?

Decision 4 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 61, indicating agreement with the following statement: "I performed well on the test I took." What would you like to do?

Decision 5 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 93, indicating strong agreement with the following statement: "I performed well on the test I took." What would you like to do?

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

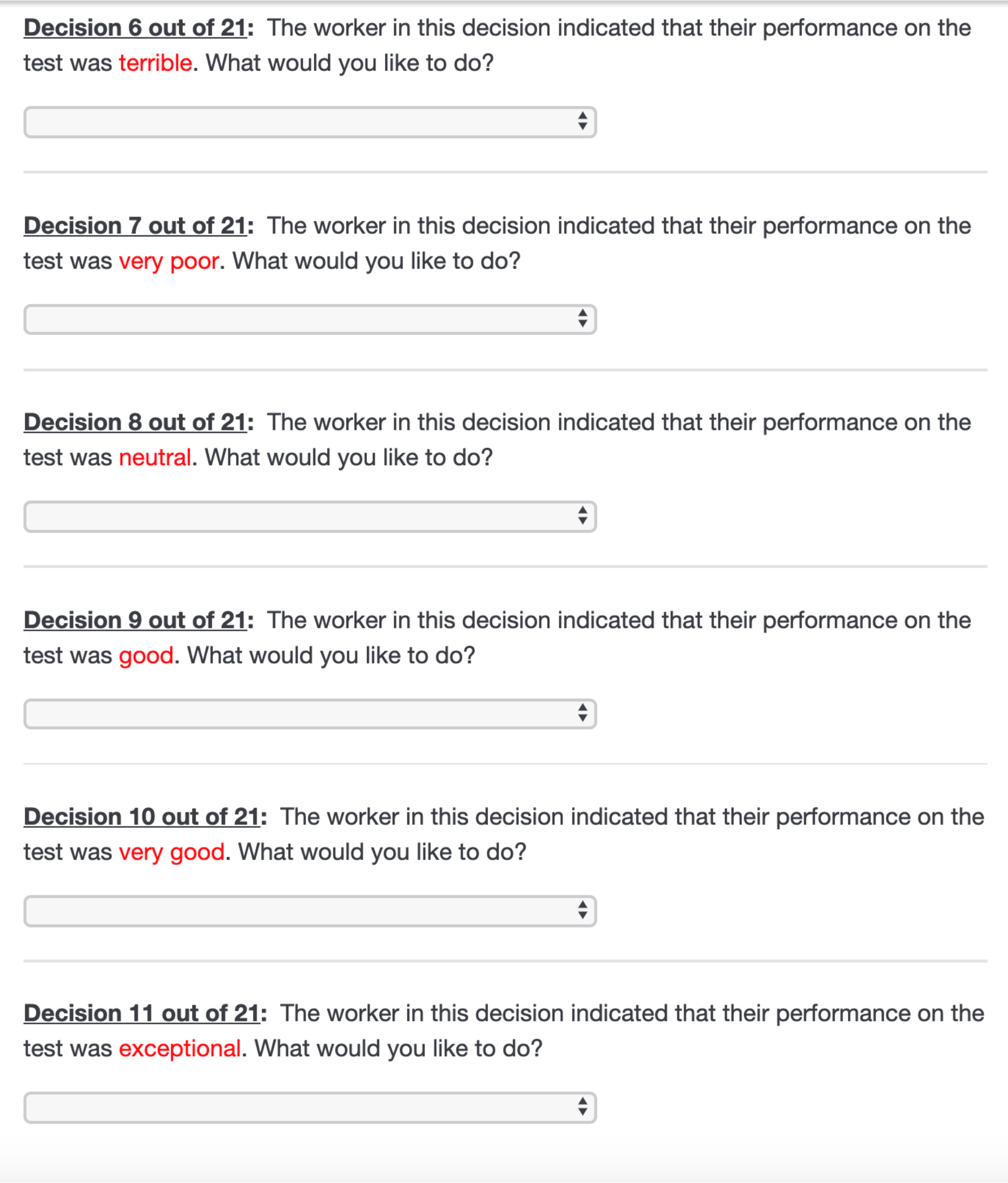
In each decision below, you will learn how the worker in that decision answered a question in which they indicated whether they thought their performance on the test was terrible, very poor, neutral, good, very good, or exceptional.

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



Decision 6 out of 21: The worker in this decision indicated that their performance on the test was terrible. What would you like to do?

Decision 7 out of 21: The worker in this decision indicated that their performance on the test was very poor. What would you like to do?

Decision 8 out of 21: The worker in this decision indicated that their performance on the test was neutral. What would you like to do?

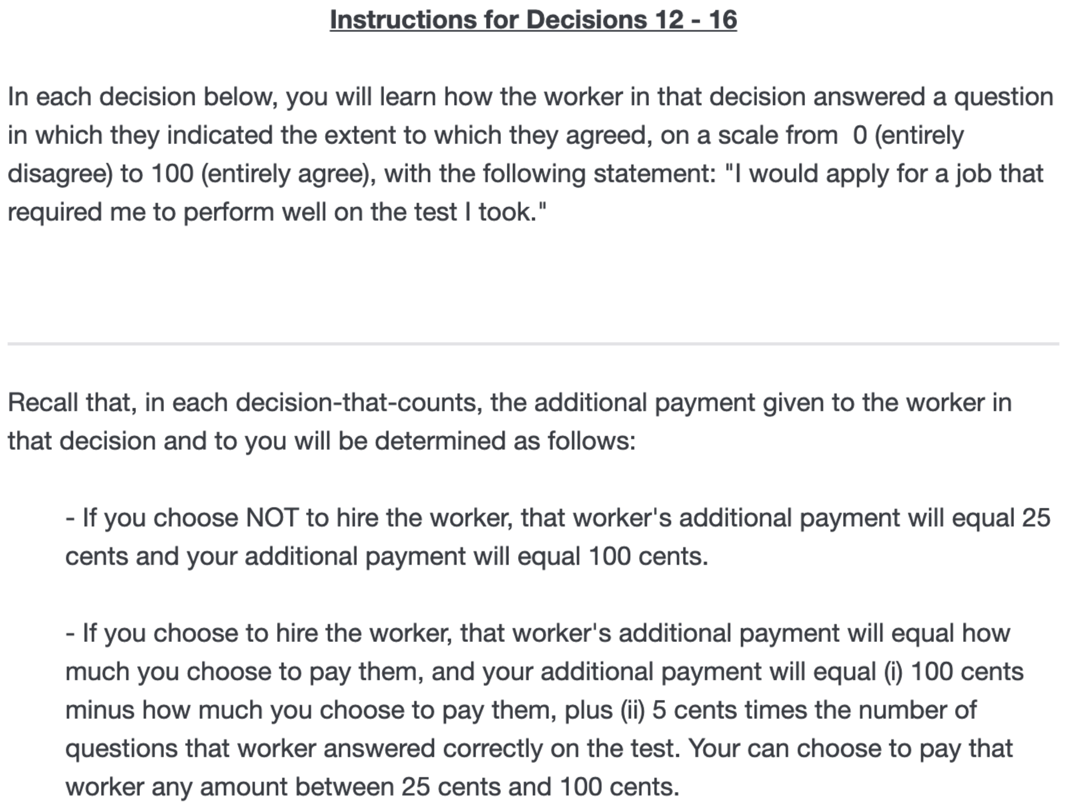
Decision 9 out of 21: The worker in this decision indicated that their performance on the test was good. What would you like to do?

Decision 10 out of 21: The worker in this decision indicated that their performance on the test was very good. What would you like to do?

Decision 11 out of 21: The worker in this decision indicated that their performance on the test was exceptional. What would you like to do?

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

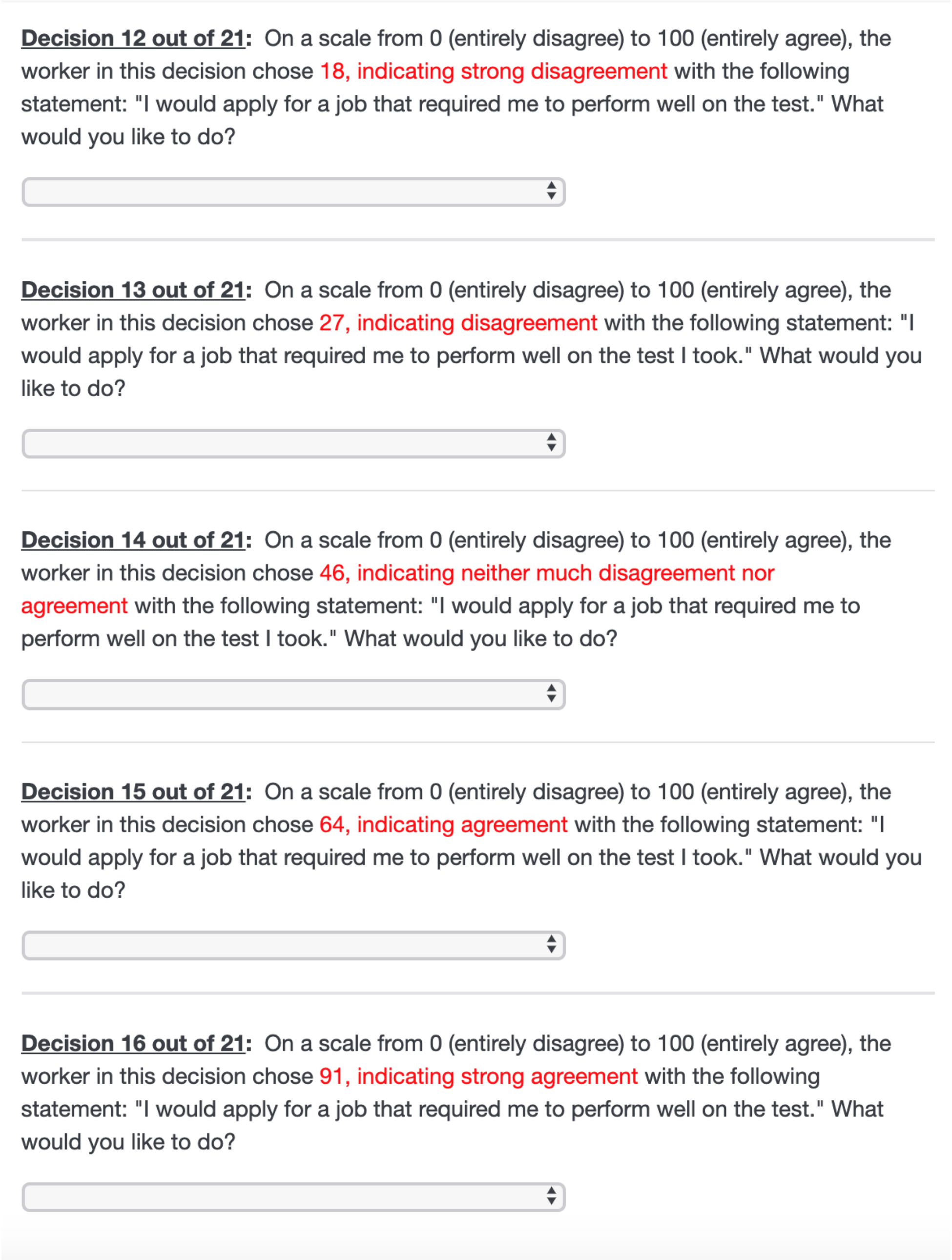
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I would apply for a job that required me to perform well on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



Decision 12 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 18, indicating strong disagreement with the following statement: "I would apply for a job that required me to perform well on the test." What would you like to do?

Decision 13 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 27, indicating disagreement with the following statement: "I would apply for a job that required me to perform well on the test I took." What would you like to do?

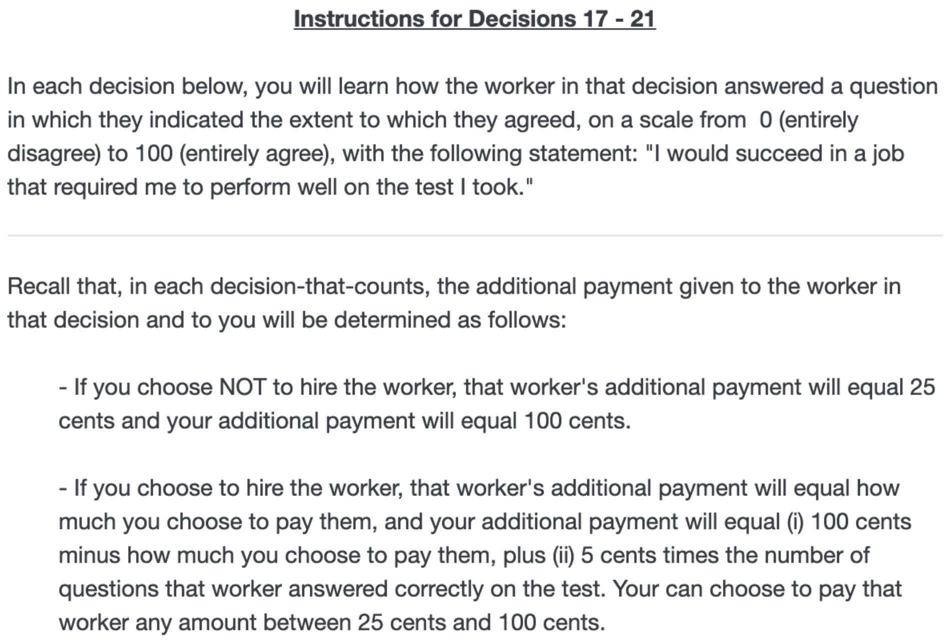
Decision 14 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 46, indicating neither much disagreement nor agreement with the following statement: "I would apply for a job that required me to perform well on the test I took." What would you like to do?

Decision 15 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 64, indicating agreement with the following statement: "I would apply for a job that required me to perform well on the test I took." What would you like to do?

Decision 16 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 91, indicating strong agreement with the following statement: "I would apply for a job that required me to perform well on the test." What would you like to do?

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

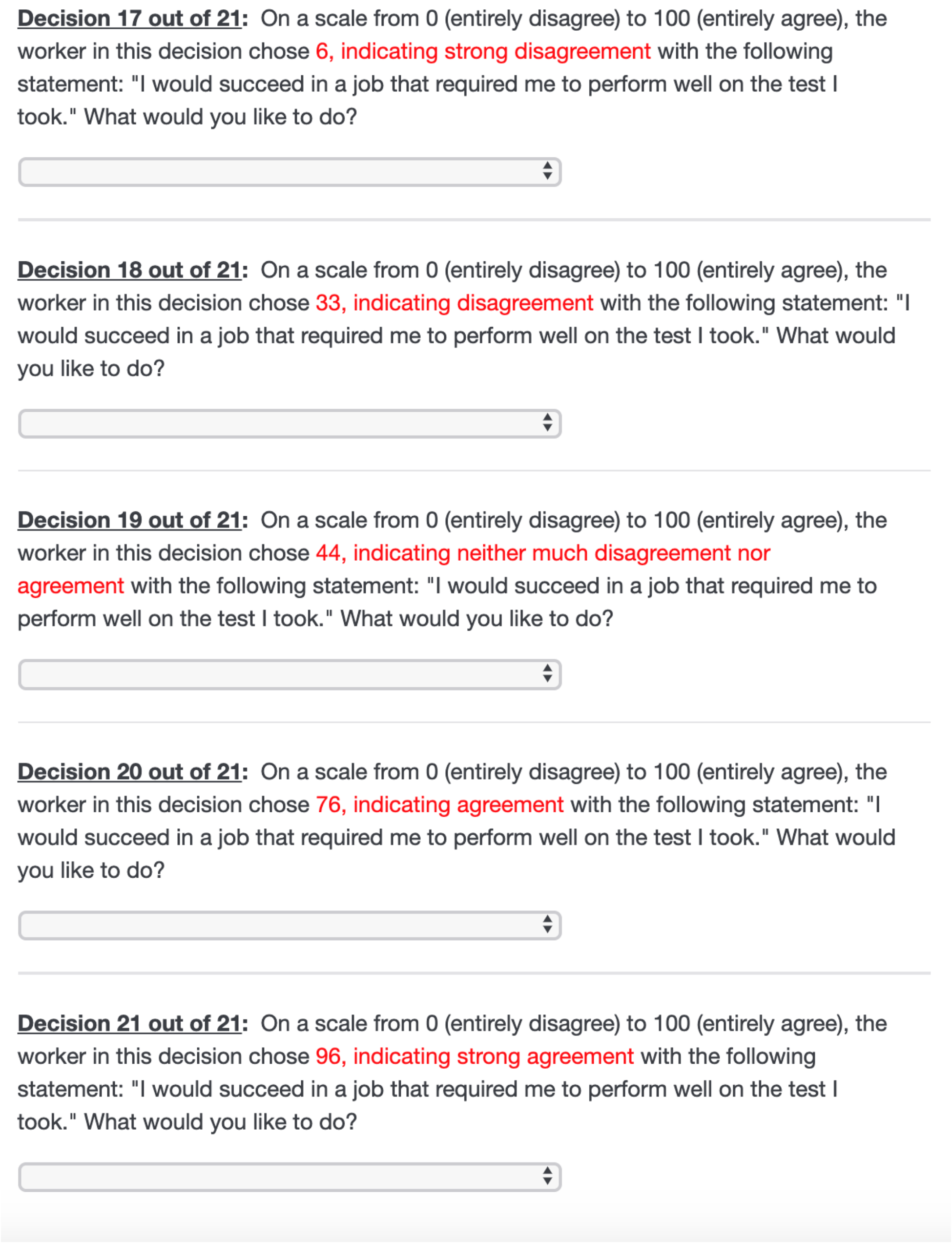
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I would succeed in a job that required me to perform well on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



Decision 17 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 6, indicating strong disagreement with the following statement: "I would succeed in a job that required me to perform well on the test I took." What would you like to do?

Decision 18 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 33, indicating disagreement with the following statement: "I would succeed in a job that required me to perform well on the test I took." What would you like to do?

Decision 19 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 44, indicating neither much disagreement nor agreement with the following statement: "I would succeed in a job that required me to perform well on the test I took." What would you like to do?

Decision 20 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 76, indicating agreement with the following statement: "I would succeed in a job that required me to perform well on the test I took." What would you like to do?

Decision 21 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 96, indicating strong agreement with the following statement: "I would succeed in a job that required me to perform well on the test I took." What would you like to do?