NBER WORKING PAPER SERIES

THE GENDER GAP IN SELF-PROMOTION

Christine L. Exley
Judd B. Kessler

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

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

The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The Gender Gap in Self-Promotion
Christine L. Exley and Judd B. Kessler
NBER Working Paper No. 26345
October 2019, Revised June 2020
JEL No. C91,D90,J16

ABSTRACT

In applications, interviews, performance reviews, and many other environments, individuals are explicitly asked or implicitly invited to evaluate their own performance and ability. In a series of experiments, involving over 4,000 participants, we find that women evaluate their performance less favorably than equally performing men. This gender gap in self-evaluations is notably persistent. It persists when we fully inform individuals about their absolute and relative performance (closing any gender gap in performance beliefs) and when we eliminate financial consequences of self-evaluations (removing incentives to distort self-evaluations). It is robust to providing information about the average self-evaluations of others and to introducing a chance that true performance will be revealed. However, there is no gap when men and women evaluate others rather than themselves, suggesting the gender gap is specifically driven by evaluating oneself. Given that self-evaluations of performance and ability can affect myriad economic outcomes, this gender gap may contribute to persistent gender gaps in educational and labor market environments.

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

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

1 Introduction

At various points in their educational and professional lives — in college and professional school applications, in job applications and interviews, in salary negotiations, in performance reviews, in informal conversations at work — individuals are asked to evaluate their performance and ability. How individuals subjectively evaluate their own performance and ability, and how they respond to explicit or implicit requests for self-evaluation, can directly impact their education and labor market outcomes.[1]

Consequently, one might be worried about the potential for a gender gap in self-evaluations. If women communicate self-evaluations that are less favorable than equally performing men, a gender gap in self-evaluations might contribute to observed gender gaps in education and labor market outcomes (Goldin, 2014; Blau and Kahn, 2017). However, there is scant research on self-evaluations and how they vary between equally performing men and women.

In this paper, we examine behavior in a controlled setting that allows us to compare the self-evaluations of equally performing men and women. Participants answer 20 questions from the Armed Services Vocational Aptitude Battery (ASVAB). Participants then complete a self-evaluation of their performance on the ASVAB test by answering several quantitative self-evaluation questions. Mirroring questions that are typical of self-evaluations in practice, participants are asked to indicate their agreement with statements such as "I performed well on the test" (on a scale from 0 to 100) and are asked to select which adjective describes their performance on the test from a Likert scale that ranges from "terrible" to "exceptional."

There are three key features of our setting. First, since we observe performance on the ASVAB test (i.e., the number of questions a participant answers correctly) and we elicit self-evaluations about that specific performance, we can compare the self-evaluations of equally performing men and women.[2] Second, since answers to our self-evaluation questions are quantitative (i.e., elicited on a Likert scale or on a scale from 0 to 100), we can examine the existence and magnitude of any gender difference in self-evaluations and how it changes across treatments. Third, the answers to our self-evaluation questions communicate perceptions for which there is no objective truth. There is no correct level of agreement with the statement "I performed well" and no correct answer when choosing how to describe one's performance from an ordered list of adjectives. This feature allows us to explore how people convey subjective information about their ability and performance, which is often how people communicate about themselves in practice. The lack of objective truth also helps

[1] Even outside of formal self-evaluations (e.g., that are part of job applications or promotion reviews), communicated self-evaluations can affect how an individual is viewed and can have important career implications. Consider an academic. Self-evaluations can feed into, among other things, how one writes graduate school applications, how one conveys his or her research ideas and technical skills, whether one gets the attention of desired advisors, how one is perceived in seminars, and how much credit one receives for joint work. Communicated self-evaluations of one's research can affect citations, prominence, as well as tenure and promotion decisions.

[2] Such a comparison is infeasible in work that considers more general attitudes of self-esteem and confidence — since these attitudes are not tied to a specific performance that allows individuals to be classified as equally performing — although this work is of clear importance; see Kling et al. (1999) for a survey.

to distinguish our work on our self-evaluations from prior work on how people answer questions about objective truths (e.g., beliefs about how many questions one answered correctly or about whether one is in the top half of performers).[3]

In our main study version, the *Self-Promotion* version, participants are aware that one of their answers to a self-evaluation question will be reported to a potential employer who will use that answer — and only that answer — to decide whether to hire them and how much to pay them. Data from employers confirm that self-promotion pays: more favorable self-evaluations increase the chance that participants are hired and the subsequent earnings they receive.[4]

We find a large 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 report evaluations that are 13 points lower than equally performing men. The average participant evaluates their performance as a 53 out of 100, so this 13-point gender gap represents nearly 25% of the mean. The gender gap is robust. It persists in all four self-evaluation questions we ask, it persists in all of the environments that we explore to investigate the underlying causes of the gender gap, and it persists in all of our attempts to close it.

In considering how we might close the gender gap in self-evaluations observed in the *Self-Promotion* version, we take advantage of the first design feature noted above: our self-evaluation questions are about specific performance on the ASVAB test. If our study had instead asked subjects to report *beliefs* about performance to a potential employer (e.g., beliefs about how many questions they answered correctly or whether they were in the top half of performers) rather than self-evaluations of that performance, and we observed a similar gender gap, there would be two main mechanisms to consider. A woman might report worse performance beliefs than an equally performing man because: (a) she believes she performed worse than the man (e.g., she believes she answered fewer questions correctly than the man or believes she is less likely to be in the top half of performers) and/or (b) she is relatively more averse to inflating her performance beliefs to a potential employer. When considering the gender gap in *self-evaluations*, these are the first two mechanisms that we consider. To test the role of the first mechanism, we design a treatment that perfectly informs participants of their absolute and relative performance, correcting their beliefs. To test the role of the second mechanism, we design a treatment that eliminates employers, removing their incentives to distort reports.

We examine the relevance of informing participants about their performance by asking participants to provide self-evaluations about *past* performance after we provide them with *perfect information* about their absolute and relative past performance (e.g., we tell them that they answered 15 out of 20 questions correctly on the ASVAB and thus scored better than than 80%, and worse than 12%, of prior participants). By telling equally performing men and women exactly how

[3] The lack of objective truth relates to work on verifiable versus unverifiable signals of support as in Kessler (2017).

[4] This study version and the title of the paper use the term "self-promotion" to emphasize that self-evaluations are communicated to employers who will make judgements based on them.

well they performed in absolute and relative terms, we mechanically close any gender gap in beliefs about absolute and relative performance.[5]

By comparing the gender gap in self-evaluations when individuals do not know their performance to the gender gap when individuals are informed, we can investigate whether the gender-gap in self-evaluations is related to the gender gap in beliefs about performance. Consistent with the prior literature on gender differences in beliefs about absolute and relative performance — often referred to as the literature on the gender gap in "confidence" (Lundeberg, Fox and Punćcohaŕ, 1994; Niederle and Vesterlund, 2007, 2011; Coffman, 2014; Niederle, 2016; Apicella, Demiral and Mollerstrom, 2017; Bordalo et al., 2019; Born, Ranehill and Sandberg, 2018; Isaksson, 2018; Coffman, Collis and Kulkarni, 2019*a*,*b*) — perfect information about absolute and relative performance (directionally) decreases the gender gap in self-evaluations, shrinking the gap by up to one-third. However, the remaining gender gap in self-evaluations is both substantial and significant. Informing participants of their absolute and relative past performance is not sufficient to eliminate the gender gap in self-evaluations about that same past performance.

We examine the relevance of eliminating incentives to distort reports by asking participants to complete self-evaluations in a version of our study that eliminates employers. The *Private* version of our study is nearly identical to the *Self-Promotion* version except that, in the *Private* version, there are no employers that participants have an incentive to impress, self-evaluations remain private, and payoffs do not depend on self-evaluations in any way.

By comparing the gender gap in the *Self-Promotion* version to the gender gap in the *Private* version, we investigate whether the gender gap in self-evaluations is due to a gender difference in the willingness to distort reports due to strategic incentives. In addition, since self-evaluations do not affect payoffs in the *Private* version, the comparison also allows us to investigate whether the gender gap is due to a gender difference in preferences about payoff outcomes or in beliefs about how self-evaluations map to payoff outcomes.

Consistent with strategic incentives causing participants to inflate self-evaluations — evidence of participants responding to the incentives in our *Self-Promotion* version — participants provide less favorable self-evaluations in the *Private* version than in the *Self-Promotion* version. However, the gender gap remains just as large in the *Private* version of our study, including when participants are informed about their absolute and relative performance. That is, men and women both provide higher self-evaluations when they have an incentive to do so, but the extent of this distortion is similar for both genders. That the gender gap in self-evaluation persists in the *Private* version of our study highlights that it is not driven by gender differences in willingness to distort self-evaluations in response to strategic incentives or by gender differences in preferences over payoffs or beliefs about how self-evaluations map to payoff outcomes.

[5] Two of our self-evaluation questions are specifically only about past performance, and the other two questions also relate to future, hypothetical performance. To assess the gender gap in self-evaluations when we close the gender gap in beliefs about absolute and relative performance, we focus on the first two self-evaluation questions.

We further consider how to close the gender gap in self-evaluations by investigating the robustness of our results to four additional study versions. First, in the *Self-Promotion (Risky)* version, we consider self-evaluations in an environment that is nearly identical to our *Self-Promotion* version except that participants are told that their actual performance (i.e., how many questions they answered correctly on the ASVAB) *could* be communicated to employers along with their self-evaluations. The possibility of true performance being communicated — which may make workers feel more constrained to provide appropriate self-evaluations since there is some chance of "being caught" if they inflate their self-evaluations too much — neither increases nor decreases the gender gap in self-evaluations. Second, in the *Private (Social Norms)* version, we consider self-evaluations in an environment that is nearly identical to our *Private* version except that participants are provided with the average self-evaluation of others who have the same performance as they do. This information — which may decrease any potential gender differences in beliefs about typical or potentially appropriate self-evaluations — does not attenuate the gender gap in self-evaluations. Third, in the *Private (Immediately Informed)* version, we consider self-evaluations in an environment that is nearly identical to our *Private* version except that the potential for consistency motives (e.g., women anchoring to lower self-evaluations before they learn their performance) is reduced. The gender gap in self-evaluations again persists.

Only our final version closes the gender gap in evaluations. Inspired by prior work that documents how women are better advocates for others than themselves in negotiations (Bowles, Babcock and McGinn, 2005), in the *Private (Other-Evaluation)* version, we consider evaluations in an environment that is nearly identical to our *Private (Immediately Informed)* version, except that participants are asked to evaluate the performances of other participants rather than themselves. In this version, the gender gap in evaluations — depending on the specific evaluation question — is either entirely or nearly eliminated. This result highlights that the gender gap in self-evaluations is about how men and women evaluate *themselves*. That is, it arises specifically from individuals evaluating their own performance and is not about men and women having different "standards" in general or different mappings from performance to performance evaluations. While men and women may agree about how favorably one should evaluate another subject answering 12 out of 20 questions correctly, a gender difference emerges when men and women self-evaluate their own performance of 12 out of 20.

To summarize, this paper explores self-evaluations and documents a robust gender gap in self-evaluations among equally performing men and women. When considering all our self-evaluation questions in all our versions where participants evaluate their own performance, we find a substantial and statistically significant gender gap in self-evaluations 56 out of 56 times. This 56/56 includes the many settings in which we both provide information, closing the gender gap in beliefs about absolute and relative performance, and eliminate employers, removing incentives to distort self-evaluations.

In light of the ample academic literature and policy initiatives devoted to closing gender gaps in economic outcomes, the persistence of the gender gap in self-evaluations is informative. It highlights the limitations of "change the women" approaches to closing gender gaps that depend on how individuals view or convey their performance.[6] Neither providing perfect information about absolute and relative performance nor providing information about how others answer self-evaluation questions eliminated the gender gap in self-evaluations. These findings suggest a potential limitation of strategies aiming to shrink gender gaps by providing information or changing beliefs. Consistent with work on culture (Gneezy, Leonard and List, 2009), the drivers of self-evaluations appear "deeply ingrained," resulting in persistent gender differences in evaluations of own performance despite little-to-no gender differences in evaluations of others' performance. Indeed, the similarity of the gender gap in the *Self-Promotion* and *Private* versions shows that the gender gap in self-promotion has little to do with the promotion aspect and instead is reflective of an underlying gender gap in self-evaluations. Other gender gaps observed in the literature — such as gaps in negotiation and group decision-making — may also relate to the gender gap in self-evaluations, since how one negotiates and whether or how one speaks up may convey self-evaluations. An alternative "change the system" approach may prove more promising. If the goal is to treat equally performing men and women equally, identifying that self-evaluations may have a built-in gender bias suggests that such self-evaluations should be deemphasized relative to more objective metrics in determining hiring and promotion decisions. We return to this discussion in the conclusion.

2 Design and Data

There are six primary versions of our study: the *Self-Promotion* version, the *Private* version, the *Self-Promotion (Risky)* version, the *Private (Social Norms)* version, the *Private (Informed Immediately)* version, and the *Private (Other-Evaluation)* version. Each version is detailed in one of the subsections 2.1–2.6. In all versions, participants first complete a 20-question ASVAB test that measures cognitive ability. Participants then complete self-evaluations of their performance on that test in the first five versions and complete evaluations of others' performances on that test in the *Private (Other-Evaluation)* version.

To examine the role of beliefs about absolute and relative performance, participants provide both "uninformed" and "informed" self-evaluations. Uninformed self-evaluations occur before participants are provided with information about their test performance. Informed self-evaluations occur after participants are informed of their absolute performance (i.e., number of questions they answered correctly on the test) and their relative performance (i.e., the percentile of their performance compared to other participants). In the *Private (Informed Immediately)* version, participants learn their absolute and relative performance immediately after taking the test and so only provide informed self-evaluations. In the *Private (Other-Evaluation)* version, participants are told about another subject's absolute and relative performance immediately after taking the test and only provide informed evaluations of that other subject's performance.

[6] For an example of how a "change the women" approach can backfire — in particular, by requiring women to lean-in and negotiate more often — see Exley, Niederle and Vesterlund (2020).

To allow for an examination of the role of incentives to distort reports and, more generally, the role of beliefs and preferences over payoff outcomes, the *Self-Promotion* version elicits self-evaluations in an environment in which employers make hiring decisions based on self-evaluations while the *Private* version elicits self-evaluations in a setting absent employers.

To examine the robustness of our results, we examine self-evaluations in three additional study versions. To potentially constrain participants to provide self-evaluations that are likely to be viewed as appropriate by employers, participants in the *Self-Promotion (Risky)* version learn that there is some chance that employers will learn both their self-evaluations and their absolute performance on the test. To reduce the ambiguity about the typical and potentially appropriate self-evaluations associated with a given performance, participants in the *Private (Social Norms)* version are informed of the average self-evaluation provided by previous participants who had the same test performance as they did. To mitigate consistency motives or anchoring effects that may arise from providing uninformed than informed self-evaluations, participants in the *Private (Immediately Informed)* version are immediately informed of their absolute and relative performance and then provide informed self-evaluations only (i.e., uninformed self-evaluations are not elicited).

To examine whether gender differences arise in evaluations of performance more broadly (e.g., due to a gender difference in "standards" or in mappings from performance levels to performance evaluations), the *Private (Other-Evaluation)* version asks participants to provide other-evaluations instead of self-evaluations.

A total of 3,293 participants on Amazon Mechanical Turk (MTurk) participated in one of these six versions of our study. Each participant was guaranteed a $2 completion fee for the 20-minute study. In addition, one part of each study was randomly selected to determine a possible bonus payment for each participant. After participants completed all parts of the study, they took a short follow-up survey that collected demographic information, including gender.[7],[8] Data collection occurred across four waves.[9]

Why did we collect data over four waves? We had four waves of data collection due to the persistence of the gender gap in self-evaluations across study versions and a desire to test the boundaries of this gap. In the first wave, we randomly assigned workers to either the *Self-Promotion* version, the *Private* version, or the *Self-Promotion (Risky)* version. Results from this wave allow us to iden-

[7] Gender was not mentioned prior to this question, so participants were not primed to think about their own gender when answering the self evaluation questions.

[8] To be eligible for any study version, participants must have previously completed at least 100 tasks on MTurk with a 95% or better approval rating from prior MTurk employers, and workers must be working from an United States IP address. Across all participants in all study versions that provide self-evaluations, 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.

[9] Data collection occurred in October 2018 for the first wave, November 2019 for the second wave, and April 2020 for the third and fourth waves. In the first three waves, we aimed to recruit 300 participants per study version. In the fourth wave, to be able to precisely identify a "null" effect in the *Private (Other-Evaluation)* version, we aimed to recruit 600 participants per study version. Realized sample size for each study version appear in Table 1.

tify the gender gap in self-evaluations and to test the two initial approaches to closing it proposed in our Introduction: perfectly informing participants of their absolute and relative performance prior to eliciting self-evaluations and removing incentives to distort self-evaluations. In subsequent waves, we ran new versions that were built off of the *Private* version, while always replicating one of our prior versions. Our focus on the *Private* version rather than the *Self-Promotion* version in these subsequent waves reflects our desire to limit the potential drivers of gender differences in self-evaluations (i.e., the gender gap could be driven by strategic incentives or preferences and beliefs over payoff outcomes in the *Self-Promotion* version but not in the *Private* version). In particular, participants are randomly assigned to either the *Private* version or the *Private (Social Norms)* version in the second wave, the *Private* version or the *Private (Immediately Informed)* version in the third wave, and the *Private (Immediately Informed)* version or the *Private (Other-Evaluation)* version in the fourth wave. The timing of these versions, and the number of participants run in each wave, is summarized in Table 1. As will be discussed in what follows, this data collection allows us to demonstrate the robustness of our results. We find a statistically significant gender gap in self-evaluations in 56 out of 56 of specifications, where a specification is defined by: self-evaluation question, whether participants are informed, study version, and wave. The results of each of these specifications are shown in Tables 7 and 8. We only eliminate the gender gap by having participants evaluate others rather than themselves. The gender gap is small or non-existent when evaluations are about others in the *Private (Other-Evaluation)* version.

Table 1: Study Versions by Wave

	Self-Promotion	Private	Self-Promotion (Risky)	Private (Social Norms)	Private (Immediately Informed)	Private (Other-Evaluation)
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)

Finally, an additional 300 participants were recruited to complete a version of our study as "employers," who are relevant for the *Self-Promotion* and *Self-Promotion (Risky)* versions of our study. Section 2.7 describes and presents results from this study version.[10]

[10] In addition, we use data from 100 participants from a prior study who completed the same ASVAB test described below, in order to provide information to study participants on their relative performance. We also analyze data from 399 MTurk workers who evaluated free-response comments generated by participants (as described below). Including these 499 participants and the 300 employers, this paper involves a total of 4,092 study participants.

2.1 The *Self-Promotion* Version

The *Self-Promotion* version of our study proceeds in four parts, described in sequence below, followed by a demographic survey. See Appendix B.1 for more details.

Part 1: Performance and Performance Beliefs

In the first part of the study, participants are asked to take a test comprising of 20 multiple choice questions from the Armed Services Vocational Aptitude Battery (ASVAB). They have up to 30 seconds to answer each question, and there are four questions each from the following five categories: General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. Participants are informed that "In addition to being used by the military to determine which jobs armed service members are qualified for, performance on the ASVAB is often used as a measure of cognitive ability by academic researchers." If the first part is randomly selected for payment, a participant's bonus payment is equal to 5 cents times the number of ASVB questions answered correctly.

As a measure of beliefs about their performance, after participants complete the 20 ASVAB questions, and before they continue to part 2, participants are asked: "Out of the 20 questions on the test you took in part 1, how many questions do you think you answered correctly?" This question is not incentivized, and participants can select any number from 0 to 20.[11]

Part 2: Uninformed Performance Evaluations

In the second part of the study, participants are asked five questions about their performance on the ASVAB. Participants are told that if the second part is randomly selected for payment, one of the responses to one of the questions will be shared with another MTurk 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 ASVAB test.[12]

To encourage participants to reflect on their performance, the first question is a free-response question that states: "Please describe how well you think you performed on the test that you took

[11] One could have imagined providing monetary incentives for answering this question correctly. However, in many cases in practice — particularly when individuals are asked to make self-evaluations — they are likely to form such beliefs about performance in the absence of such monetary incentives. In addition, we avoid concerns about the accuracy of belief elicitation because we do not use these beliefs as controls. As discussed below, we control for beliefs about absolute and relative performance by design, rather than statistically, which avoids any potential concerns about noise in belief elicitation.

[12] Note that employer earnings are based on the number of correct answers that the participant completed in part 1. This means that participants do not have to answer additional questions and the decision environment avoids any potential uncertainty that might arise about future performance.

in part 1 and why." The remaining four questions elicit quantitative self-evaluations that we analyze for the remainder of the paper.[13]

The first two self-evaluation questions focus solely on participants' *past performance* on the test. We first elicit a discrete self-evaluation, which we refer to as the *performance-bucket* evaluation. Participants are asked to indicate how well they think they performed by selecting from one of the following six answers: terrible, very poor, neutral, good, very good, and exceptional. We then elicit a more continuous self-evaluation, which we refer to as the *performance* evaluation. Participants are asked 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."

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 self-evaluations as the *willingness-to-apply* evaluation and the *success* evaluation, respectively.

Broadly, we use the first two self-evaluations as measures that allow us to cleanly consider underlying mechanisms, and the latter two self-evaluations as measures that speak to the robustness of our results. All of our results are robust to all four self-evaluation questions.

Part 3: Informed Performance Evaluations

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

However, to examine the role of beliefs about absolute and relative performance, before answering these self-evaluation questions, participants learn precise information about their *absolute* and *relative* performance on the ASVAB 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 amswered the same ASVAB questions as part of a prior study and told how many of those participants answered more questions correctly and how many answered fewer questions correctly (i.e., their relative performance). As an attention check, participants must correctly report how many of the 20 ASVAB questions they answered correctly before proceeding to answer the self-evaluation questions in part 3.

Part 4: Deservingness Question

In the fourth part of the study, participants are asked one question that measures deservingness for earnings from our experiment: "Out of a maximum amount of 100 cents, what amount of bonus

[13] The free-response question can also theoretically be interpreted as a self-evaluation. Analyzing this free-response data is fraught, however, as the text is hard to evaluate and can convey information such as gender and competence that makes measuring self-evaluation per se difficult. Nevertheless, we attempt to learn what we can from this data by having 399 MTurk participants evaluate the responses, and we summarize those findings in Appendix A.2.

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 as the part-that-counts, their bonus payment equals whatever amount they indicate from 0 to 100 cents. This question allows us to consider the potential gender difference in deservingness (i.e., how much participants believe they deserve to earn from the study) or in the desire to earn money from the experiment. Since this measure occurs *after* self-evaluations are provided, and may theoretically be influenced by self-evaluations, it is not an appropriate "control" variable in regressions with self-evaluations as the dependent variable. Rather, we consider it as an alternative dependent variable and discuss it in Section 3.3.

2.2 The *Private* Version

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

Given the lack of employers, the *Private* version eliminates the relevance of strategic incentives to provide more favorable self-evaluations in order to achieve higher study earnings. The lack of employers more generally eliminates the relevance of participants' beliefs and preferences over payoff for themselves and for the employers.

2.3 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 ASVAB test) along with their self-evaluation.[14] See Appendix B.3 for more details.

If participants expect that employers may learn their actual performance, the *Self-Promotion (Risky)* version may cause workers to feel constrained to provide self-evaluations that are more likely to be viewed as appropriate by their employers because of a desire to avoid "being caught" as having inflated their self-evaluations. More generally, the *Self-Promotion (Risky)* version helps us to relate to labor market settings where applicants or employees are aware that signals about true performance may be available to employers.

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 their informed self-evaluations. In particular, each of the four 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 B.4 for more details.

[14] This chance is ambiguous in the experimental instructions. In practice, there was a 1% chance that employers would be informed of this additional information, which resulted in them not being informed.

This additional information in the *Private (Social Norms)* version may mitigate gender differences in beliefs about what self-evaluations are typical or viewed as appropriate by others. More generally, the *Private (Social Norms)* version helps to relate our results to labor market settings in which workers observe peer self-evaluation behavior.

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 make informed self-evaluations. Thus, this study version does not elicit uninformed self-evaluations — skipping part 2 entirely — and only involves three parts. See Appendix B.5 for more details.

By only eliciting informed self-evaluations, the *Private (Immediately Informed)* version eliminates the potential role of consistency motives or anchoring effects that arise from the elicitation of uninformed self-evaluations before informed self-evaluations.

2.6 The *Private (Other-Evaluation)* Version

The *Private (Other-Evaluation)* version asks workers to provide evaluations about others rather than themselves. More specifically, the *Private (Other-Evaluation)* version proceeds exactly as the *Private (Immediately Informed)* version, except that participants are informed of the absolute and relative performance of another MTurk worker and asked to provide evaluations for that other MTurk worker. Unbeknownst to participants, this other MTurk worker is selected to have exactly the same performance on the test in the first part of the study as the participant. That is, a participant who answers X out of 20 questions correctly in part 1 is asked to provide informed evaluations for another participant who also answered X out of 20 questions correctly without being told that X out of 20 is also their score. See Appendix B.6 for more details.

The *Private (Other-Evaluation)* version helps identify whether there is a gender difference in standards or in evaluations of performance generally or whether a gender difference in evaluations is specific to own performance.

2.7 The *Employer* Version

We recruited 300 workers on MTurk to complete the *Employer* version of our study using the same criteria as in the main study versions (see footnote 8). Each employer received a guaranteed $1.50 completion fee for the 15-minute study. In addition, two of their decisions (out of 21 decisions in the study), are selected to determine a possible bonus payment for them and for associated "workers," participants in the *Self-Promotion* and *Self-Promotion (Risky)* study versions.

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

Employers make hiring and wage decisions via the strategy method. That is, their decisions involve each of the six possible answers selected from the performance-bucket evaluation question and five randomly selected answers (i.e., numbers from 0 to 100) from each of the three other evaluation questions.[15] Each employer's two decisions that are selected to determine bonus payments result in payments for themselves and for the two workers who provided the corresponding self-evaluations in those decisions. See Appendix B.7 for more details.

As expected, self-promotion pays. Employers respond to more positive self-evaluations by being more likely to hire workers and by paying them more. Table 2 shows how self-evaluations impact wages. In all specifications, the coefficient on *Evaluation* is positive and significant. Columns (1), (3), and (4) show that the wage given to workers increases by an average of 0.21 or 0.22 cents more for every point participants add to their evaluation on the 100-point scale in response to the performance evaluation question, the willingness-to-apply evaluation question, and the success evaluation question. Column (2) shows that the wage given to workers increases by 4.26 cents for each increase on the six-point Likert scale in the performance-bucket evaluation question. These results highlight that participants have an incentive to inflate their self-evaluations to increase their expected study earnings.

Table 2: *Employer* Version, Wage Regressions

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered by employer. Results are from OLS regressions of the wage received by the participant (25 cents if not hired and a chosen wage from 25–100 cents if hired). *Evaluation* is the self-evaluation provided by a participant in the evaluation question noted in that column. *Performance* indicates the extent of a participant's agreement (from 0–100) with the following statement: "I performed well on the test I took in part 1." *Performance-Bucket* indicates 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." *Willingness-to-Apply* indicates the extent of a 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." *Success* indicates the extent of a 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." Data are from the hiring decisions in the *Employer* version.

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

3 Results

In this section, we report on our experimental results. Section 3.1 confirms that our study environment captures many of the features of a "male-typed" setting. Despite women slightly outperforming men on the ASVAB test, we observe a large gender gap in beliefs about performance: women report that they answered fewer questions correctly on the test than equally performing men. Note that we chose to conduct our study in a male-typed setting, involving an analytical task, because the gender gaps in pay and in occupational and industry representation that motivate our study typically arise in male-typed settings.[16]

Section 3.2 details our main results from the *Self-Promotion* and *Private* versions. We document a substantial and significant gender gap in uninformed self-evaluations in the *Self-Promotion* version. We then show that eliminating the gender gap in beliefs about absolute and relative performance generates a slightly smaller, but still substantial and statistically significant, gender gap in informed self-evaluations. We further show that the gap is not driven by strategic incentives or beliefs or preferences over payoff outcomes. We find that eliminating employers in the *Private* version results in lower self-evaluations for both men and women, but the gender gap remains substantial and statistically significant when considering both uninformed and informed self-evaluations in the *Private* version.

Section 3.3 documents the robustness of our results. We find that the gender gap in self-evaluations persists in three additional study versions — the *Self-Promotion (Risky)* version, the *Private (Social Norms)* version, and the *Private (Immediately Informed)* version — as well as in replications of the *Private* version. We also show the impact of restricting our sample and allowing for heterogeneous treatment effects.

Section 3.4 discusses results from the *Private (Other-Evaluation)* version, which asks about other — rather than own — performance. This is the one study version in which the gender gap in evaluations is either eliminated or substantially reduced.

3.1 Our Study Environment

All of the participants in our study who provide self-evaluations or other-evaluations start by taking the ASVAB test and reporting their beliefs about how many questions they answered correctly. In this section, we report on how they perform and the beliefs that they report.[17]

As is common in the gender literature focusing on "male-typed" environments — the type of

[16] Bordalo et al. (2019) defines "male-typed" and "female-typed" settings. The paper considers: Arts and Literature, Business, Cars, Cooking, Disney Movies, Emotion Recognition, Kardashians, Mathematics, Rock and Roll, Sports and Games, Verbal Skills, and Videogames. Men believe they perform better than women in all settings except Disney Movies and Kardashians, the only two settings defined as "clearly female-typed" (see Bordalo et al. (2019), Figure 1). For more literature on gender stereotypes, see also Coffman (2014); Coffman, Collis and Kulkarni (2019*b*); Coffman, Flikkema and Shurchkov (2019). In societies with different gender norms, such as in matriarchal societies, one may expect gender gaps to reverse, as shown in Gneezy, Leonard and List (2009).

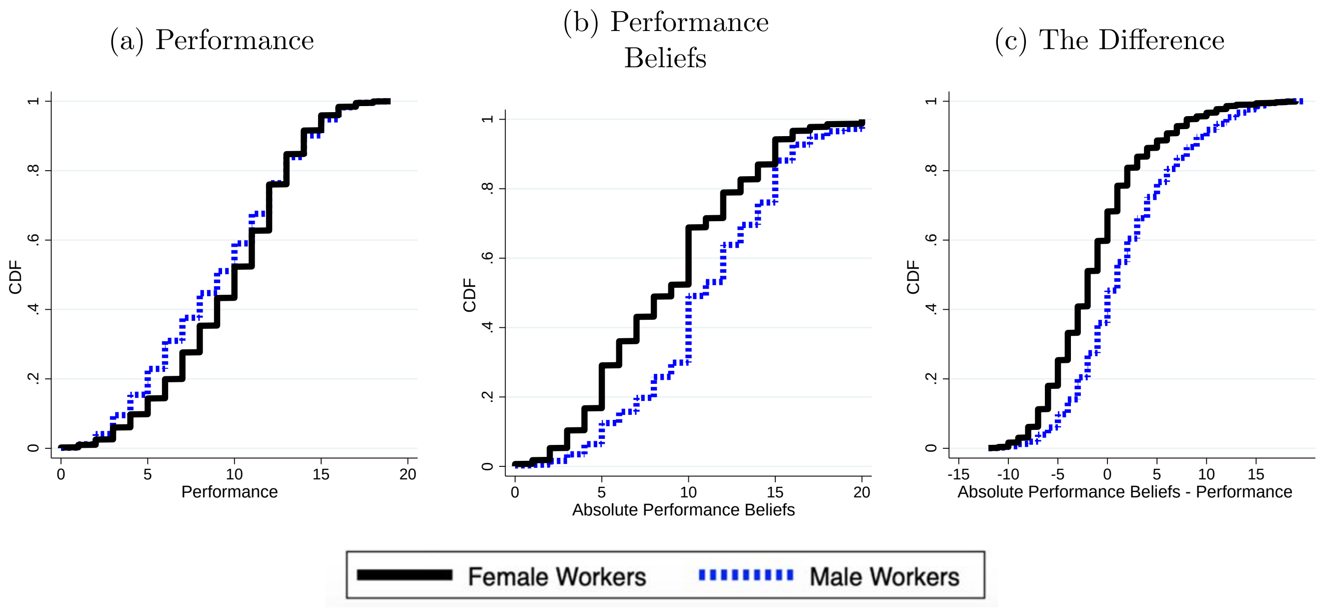
[17] The results reported in this section pool across all study versions, because participants answer the 20 ASVAB questions and report their beliefs about performance before receiving any information that is specific to their study version. However, the gender gap in beliefs about performance persists for each individual study version.

environments in which we observe pay and representation differences in the labor market — we find a large gender gap in beliefs about performance.

Panel A of Figure 1 shows CDFs of the number of ASVAB questions answered correctly by men and women. On average, women answer 9.79 questions correctly and men answer 9.13 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 on the test, Panels B and C of Figure 1 show that women believe they performed worse than men. Panel B shows raw beliefs about performance. On average, men believe they answered 11.03 questions correctly while women believe they answered only 8.81 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 actual performance and beliefs about 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 more women than men underestimating their actual performance and more men than women overestimating their actual performance.

Figure 1: Performance (Actual vs. Believed) Distributions


(a) Performance

(b) Performance Beliefs

(c) The Difference

These graphs show CDFs for the noted outcome. *Performance* is the number of questions a participant correctly answered out of the 20 ASVAB questions. *Performance Beliefs* is the number of questions a participant believes he or she correctly answered. *The Difference* equals *Performance Beliefs* − *Performance*. Data are from all study versions ($n = 3293$).

Table 3 presents the corresponding regression results. Column 1 shows that women outperform men on the ASVAB 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 score (Column 3), and when the outcome variable directly captures the difference between actual performance and beliefs about performance (Column 4). In the latter three columns, the coefficient on *Female* is negative, large, and statistically significant.

Table 3: Performance (Actual vs. Believed) Regressions

DV:	Performance	Performance Beliefs		The Difference
	(1)	(2)	(3)	(4)
Female	0.66***	-2.22***	-2.15***	-2.88***
	(0.14)	(0.15)	(0.15)	(0.18)
Constant	9.13***	11.03***		1.90***
	(0.09)	(0.09)		(0.12)
N	3293	3293	3293	3293
Performance FEs	No	No	Yes	No

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the noted dependent variable (DV). *Performance* equals the number of questions a participant correctly answered out of the 20 ASVAB questions. *Performance Beliefs* equals the number of questions a participant believes he or she correctly answered. *The Difference* equals *Performance Beliefs* – *Performance*. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance (0–20) on the ASVAB test. Data are from all study versions.

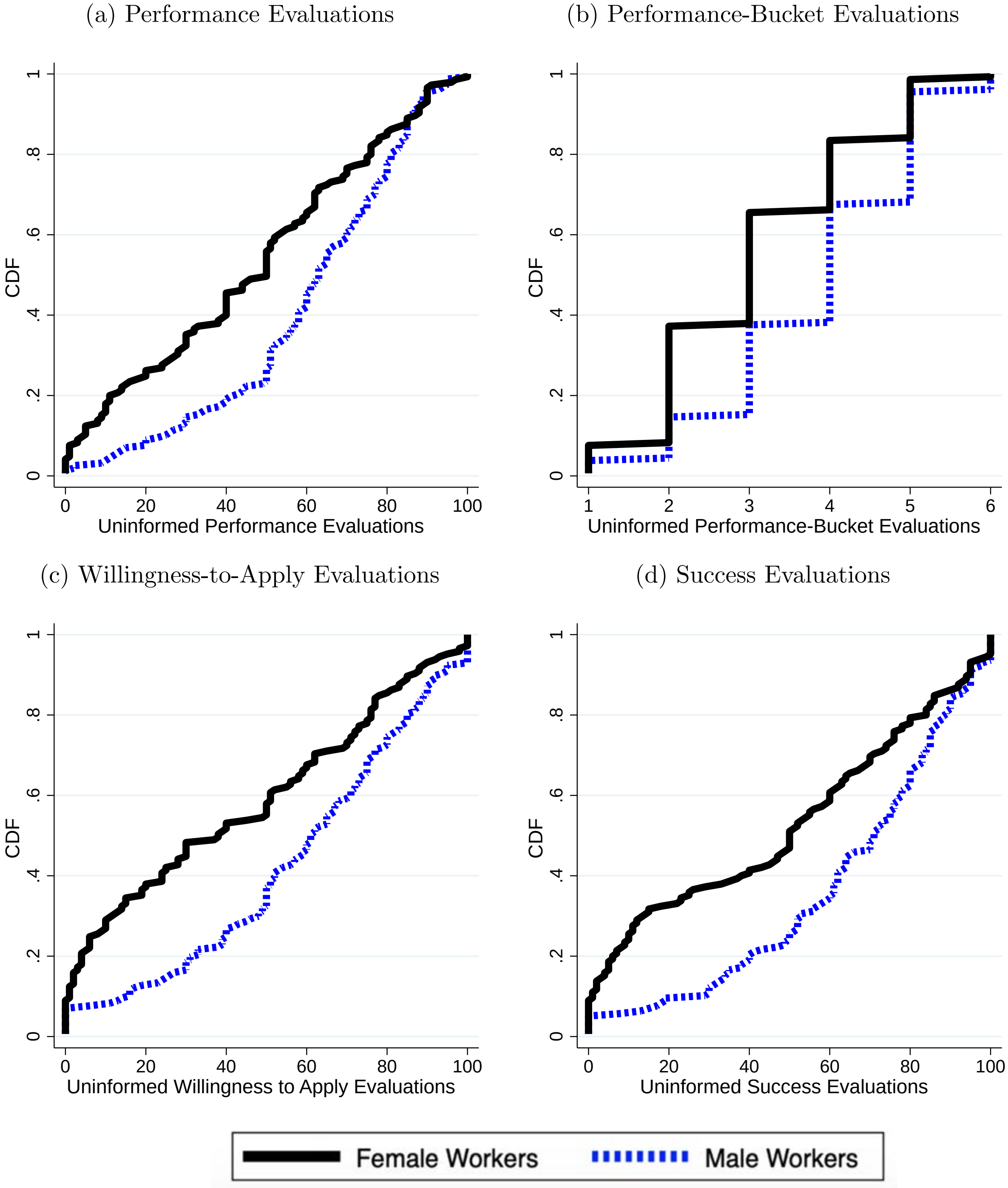
3.2 Main Results

3.2.1 Is there a gender gap in self-evaluations?

To assess whether there is a gender gap in self-evaluations, we turn first to the *Self-Promotion* version in which participants are told that one of their self-evaluations will be shared with a potential employer, and that this is all the employer will know when making a hiring and wage decision. Figure 2 shows raw responses to the four self-evaluation questions in part 2 of the *Self-Promotion* version. As described in Section 2, we refer to these as the uninformed self-evaluations because they occur before participants learn their absolute and relative performance.

All four panels show large gender gaps in uninformed self-evaluations. Women evaluate their performance less favorably than men. Panel A shows results from the 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). Women provide statistically significantly lower evaluations ($p < 0.01$ for the t-test and the Kolmogorov–Smirnov test). We obtain similar results in Panel B for the six-point Likert scale question: “Please indicate how well you think you performed on the test you took in part 1” ($p < 0.01$ for the t-test and the Kolmogorov–Smirnov test). Panels C and D show results from the self-evaluation questions that allow participants to hold preferences and beliefs about a related, hypothetical job. Participants respond to the statements “I would apply for a job that required me to perform well on the test I took in part 1” (Panel C) and “I would succeed in a job that required me to perform well on the test I took in part 1” (Panel D) on a scale from 0 (entirely disagree) to 100 (entirely agree). We again see statistically significant differences in self-evaluations ($p < 0.01$ for both t-tests and both Kolmogorov–Smirnov tests).

Figure 2: Uninformed Self-Evaluations in the *Self-Promotion* Version


(a) Performance Evaluations

(b) Performance-Bucket Evaluations

(c) Willingness-to-Apply Evaluations

(d) Success Evaluations

Table 4 confirms the statistical significance of these gender gaps when controlling for performance with fixed effects for each possible test score 0 to 20. The coefficient on *Female* remains negative, large, and statistically significant for all four self-evaluation questions. The performance fixed effects allow us to compare the self-evaluations of equally performing men and women. As detailed in Section 3.3, the regression results remain similar and statistically significant without performance fixed effects as well as when considering an Ordered Probit for the performance-bucket evaluations.

Table 4: Uninformed Self-Evaluations in the *Self-Promotion* Version

Evaluation:	Performance (1)	Performance- Bucket (2)	Willingness- to-Apply (3)	Success (4)
Female	-12.68***	-0.59***	-15.31***	-15.09***
	(2.96)	(0.13)	(3.46)	(3.46)
N	302	302	302	302
Performance FEs	Yes	Yes	Yes	Yes
Average	53.43	3.46	50.13	56.43
SD	27.36	1.27	31.88	31.97
Effect Size/SD	0.46	0.46	0.48	0.47
Effect Size as % of Average	24%	17%	31%	27%

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column and 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 (0–20) on the ASVAB test. Data are from uninformed self-evaluation questions in the *Self-Promotion* version.

3.2.2 Is the gap driven by beliefs about absolute and relative performance?

In Section 3.1, we document a gender gap in beliefs about absolute performance. To consider the role of beliefs about absolute performance, we could consider controlling for reported beliefs in our regressions. One limitation of this approach relates to measurement error. Another limitation is that beliefs about relative performance — not just absolute performance — could be relevant.

We overcome these limitations by focusing on a subset of self-evaluations for which we can correct beliefs about absolute and relative performance *by design*. First, we restrict our attention to the performance evaluation and performance-bucket evaluation questions since they only ask about *past* performance on the ASVAB test. The performance evaluation question only relates to this past performance because it asks participants to indicate their agreement on a scale from 0 to 100 with the following statement: “I performed well on the test I took in part 1.” The performance-bucket evaluation question only relates to this past performance because it asks participants to choose a Likert-scale response that indicates “how well you think you performed on the test in part 1.”

Second, we restrict our attention to the informed self-evaluations that are provided to these two questions in part 3 of our study. The informed self-evaluations are provided *after* participants are informed of their absolute performance (i.e., the number of questions they answered correctly on the ASVAB test) and their relative performance (i.e., the percent of participants who answered more questions correctly than them and the percent of participants who answered fewer questions correctly than them on the ASVAB test). Since we compare equally performing men and women who have both learned their (identical) score and their (identical) place in the performance distribution, beliefs about absolute and relative performance can no longer drive gender differences in informed self-evaluations. That is, after we have perfectly informed participants about their absolute and relative past performance — such that there can no longer be gender differences in beliefs about absolute and relative past performance — gender differences in self-evaluations of this same past performance cannot be driven by beliefs about absolute and relative past performance. Two main results follow.

The first result relates to prior literature that shows how beliefs about absolute and relative performance contribute to gender gaps in economic outcomes. We find that correcting beliefs about absolute and relative performance — and thus closing any gender gap in beliefs about absolute and relative performance — (directionally) decreases the gender gap in self-evaluations by up to one-third. The top panel of Table 5 presents regression results involving both the uninformed and informed self-evaluations in the *Self-Promotion* version. As indicated by the coefficient estimates on *Informed*Female* in Columns 1 and 2, correcting beliefs about absolute and relative performance insignificantly decreases the gender gap in self-evaluations in response to the performance question by 32% and in response to the performance-bucket question by 20%. When we are better powered — by pooling across all study versions involving self-evaluations in the bottom panel of Table 5 — the coefficient estimates on *Informed*Female* in Columns 1 and 2 become statistically significant and show that correcting beliefs about absolute and relative performance decreases the gender gap in self-evaluations in response to the performance question by 31% and in response to the performance-bucket question by 31% as well.

The second result is that the gender gap in self-evaluations remains large and statistically significant even when agents are informed. That is, correcting beliefs about absolute and relative performance does *not* eliminate the gender gap in self-evaluations. One way to convey this result is to note that the sums of the coefficient estimates on *Female* and *Informed*Female* are all statistically significantly negative in Table 5. More simply, the summary table shown later, Table 8, reports that the gender gap in *informed* self-evaluations remains substantial and statistically significant when considering only the *Self-Promotion* version (Columns 1 and 2, Panel 1) or when considering *any* of the other versions involving self-evaluations (Columns 1 and 2, Panels 2–8).

What role do beliefs about absolute and relative performance play in the self-evaluations provided in response to the willingness-to-apply and success questions?[18] There is some evidence that correcting beliefs about absolute and relative performance shrinks the gender gap for these self-evaluation questions as well. However, the coefficient estimates on *Female* and the sum of *Female* and *Informed*Female* in Columns 3 and 4 of Table 5 are all statistically significantly negative, and

[18] These questions ask participants about the extent of their agreement with statements reading "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." Consequently, for these questions, beliefs about *future* performance (in the referenced hypothetical job) could be relevant. Since informing participants about their absolute and relative *past performance* may not eliminate the role of beliefs about the relevant (future) absolute and relative performance, we focus on the other two self-evaluation questions, which are only about past performance, to be conservative in our examination of absolute and relative performance beliefs.

Table 5: The Role of Beliefs about Absolute and Relative Performance on Self-Evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
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
Performance FEs	Yes	Yes	Yes	Yes
All Versions with Uninformed and Informed Self-Evaluations				
Female	-13.89***	-0.67***	-16.59***	-15.38***
	(1.25)	(0.06)	(1.42)	(1.44)
Informed	-1.95***	-0.02	0.27	-0.91
	(0.60)	(0.03)	(0.59)	(0.57)
*Informed*Female*	4.34***	0.21***	2.14**	1.49*
	(0.94)	(0.05)	(0.90)	(0.88)
N	3600	3600	3600	3600
Performance FEs	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered by participant. Results are from OLS regressions of the responses provided to the evaluation question noted in each column and 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 (0–20) on the ASVAB test. Data in the top panel are from self-evaluations in the *Self-Promotion* version. Data in the bottom panel are from self-evaluations in all versions that contain uninformed and informed self-evaluation questions — that is, from all versions except for the *Private (Immediately Informed)* version and the *Private (Other-Evaluation)* version. Participants provide both uninformed and informed self-evaluations and so there are two observations for each participant.

Table 8 shows that the gender gap in *informed* self-evaluations remains substantial and statistically significant when considering only the *Self-Promotion* version (Columns 3 and 4, Panel 1) or when considering *any* of the other versions involving self-evaluations (Columns 3 and 4, Panels 2–8).

A few methodological notes are worth making. First, we provide perfect information about absolute and relative performance because doing so closes the gender gap in three types of confidence identified by Moore and Healy (2008). *Overestimation* suggests overconfidence in actual performance, *overplacement* suggests overconfidence in one's ranking relative to others, and *overprecision* suggests beliefs about an unknown state that are too precise. We counter *overestimation* by telling participants their absolute performance, we counter *overplacement* by telling participants their relative performance, and we counter *overprecision* by providing participants with perfect information on a known state of the world (i.e., their past performance) and asking them to provide self-evaluations about that known state. By addressing these types of confidence — and showing that the gender gap in self-evaluations shrinks but persists — we highlight that the gender gap in self-evaluations is driven in part — but only in part — by a gender gap in these types of confidence.[19]

Second, we view correcting beliefs about absolute and relative performance by design (for the performance and performance-bucket evaluations) as a methodological strength of our paper. It guards against inaccurate conclusions about the role of beliefs in driving our results. For example, imagine if we had instead attempted to control for these beliefs statistically, say by including a control variable for participants' reported beliefs about their absolute performance. As shown in the top panel of Table A.1, statistically controlling for beliefs about absolute past performance — either linearly or with fixed effects — in the *Self-Promotion* version would have suggested that these beliefs account for the *majority* of the gender gap in self-evaluations. But, as already shown in Table 5 by the coefficient estimates on *Informed*Female*, beliefs about absolute *and* relative past performance in the *Self-Promotion* version account for the *minority* of the gender gap in self-evaluations and fail to be statistically significant. Similar discrepancies in magnitudes also arise when we are better powered by pooling across all versions with uninformed and informed self-evaluations (see the bottom panel of Table A.1). The gender gap in self-evaluations is driven by much more than beliefs about absolute and relative performance, but statistically accounting for the role of beliefs about absolute performance would have suggested otherwise.

Third, correcting beliefs about absolute and relative performance by design is neither feasible nor desirable for most papers with interests that are different from ours. For instance, fully informing participants about their absolute performance (e.g., telling them they answered 12 out of 20 questions correctly) is not desirable if a paper seeks to understand how individuals form or update beliefs about that absolute performance. This is one reason why prior work on beliefs does not aim to eliminate beliefs by design. Since our paper is not interested in how individuals update their beliefs about absolute and relative performance, correcting these beliefs by design is helpful.

Fourth, while fully informing individuals of objective truths can eliminate gender differences in beliefs about those objective truths, there is no parallel strategy for eliminating gender differences in self-evaluations for which there is no objective truth. There are no objectively accurate responses to self-evaluation questions about the extent to which individuals agree they performed well or whether individuals believe their performance was "good" or "exceptional." While this lack of objective truth may pose a particular challenge to interventions targeted at closing the gender gap in self-evaluations, it reinforces our view that more work on self-evaluations is warranted.

3.2.3 Is the gap driven by incentives to distort reports or beliefs and preferences about payoff outcomes?

As noted in the Introduction, the gender gap in self-promotion could theoretically reflect gender differences in willingness to distort reports or gender differences in beliefs and preferences over

[19] For additional literature that considers the beliefs that underly these types of confidence, see Schotter and Trevino (2014). For work on biased beliefs more generally, see Ertac (2011); Mobius et al. (2011); Buser, Gerhards and Van der Weele (2018); Coutts (2018).

payoff outcomes. For examples: men could be more responsive to the strategic incentives to inflate self-evaluations to increase their chances of being hired, women could be more averse to inflating their self-evaluations due to a concern that they are misleading employers, or men and women may hold different beliefs about what level of self-evaluation is likely to get them hired and paid well.

To examine whether explanations like these drive the gender gap in self-evaluations that we observe, we compare the results from the *Self-Promotion* version to results from the *Private* version, which eliminates the role of employers and pays participants a fixed amount, thus removing incentives to distort reports and making irrelevant the role of participants' beliefs and preferences over payoff outcomes for themselves and employers. In particular, Table 6 presents results from the uninformed and informed self-evaluations in the *Self-Promotion* version and the *Private* version. Two main results follow.

First, self-evaluations are lower in the *Private* version than in the *Self-Promotion* version. The coefficient estimates on *Private* are significantly negative in 7 out of the 8 regressions and directionally negative in the remaining regression shown in Table 6. This finding is consistent with participants responding to strategic incentives to inflate their self-evaluations in the *Self-Promotion* version, where employers observe a self-evaluation before making a hiring and wage decision.

Second, the gender gap in self-evaluations persists even when the role of preferences and beliefs over payoff outcomes are eliminated. The coefficient estimates on *Private*Female* are small and insignificant and, more importantly, the sum of the coefficient estimates on *Female* and *Private*Female* are always significantly negative in Table 6. More simply, consider the results from the summary tables that are shown later: when considering the *Private* version only, Table 7 shows that the gender gap in uninformed self-evaluations is substantial and statistically significant (see Columns 1–4, Panel 2), and Table 8 shows that the gender gap in informed self-evaluations is also substantial and statistically significant (see Columns 1–4, Panel 2), Moreover, this finding is replicated many times when considering the subsequent *Private* versions we ran (see Columns 1–4, Panels 3–6 of Table 7 and Columns 1–4, Panels 3–8 of Table 8).

That the gender gap persists in the *Private* version highlights that the gender gap in self-evaluations does not rely on: gender differences in willingness to strategically answer self-evaluations to increase payment from employers, related to prior work on distorted and misreported beliefs about performance (Reuben, Sapienza and Zingales, 2014; Charness, Rustichini and Van de Ven, 2018; Soldà et al., 2019; Schwardmann and van der Weele, 2019); gender differences in risk aversion about how employers will respond to self-evaluations, related to work on gender differences in risk aversion over payoffs (Dwyer, Gilkeson and List, 2002; Eckel and Grossman, 2008; Croson and Gneezy, 2009); gender differences arising from a lack of control over payoffs given that employers determine wages, related to work on gender differences in locus of control (Cobb-Clark, 2015; Apicella, Demiral and Mollerstrom, 2020); or gender differences in other-regarding preferences towards employers, related to work on gender differences in preferred payoffs for others (Andreoni and Vesterlund, 2001; Croson and Gneezy, 2009; DellaVigna et al., 2013) and gender differences in deception (Dreber and

Johannesson, 2008; Childs, 2012; Erat and Gneezy, 2012; Houser, Vetter and Winter, 2012; Gylfason, Arnardottir and Kristinsson, 2013; Adams, Kuhn and Waddell, 2019).[20]

Table 6: The Role of Beliefs and Preferences over Payoffs

Evaluation:	Performance (1)	Performance-Bucket (2)	Willingness-to-Apply (3)	Success (4)
Self-Promotion and Private versions, Uninformed Self-Evaluations				
Female	-12.20***	-0.59***	-15.54***	-15.45***
	(2.94)	(0.13)	(3.43)	(3.41)
Private	-6.25**	-0.26**	-4.27	-6.93**
	(2.72)	(0.13)	(3.35)	(3.30)
*Private*Female*	-1.66	0.00	-2.30	-1.07
	(4.04)	(0.18)	(4.77)	(4.84)
N	606	606	606	606
Performance FEs	Yes	Yes	Yes	Yes
Self-Promotion and Private versions, Informed Self-Evaluations				
Female	-7.14**	-0.43***	-11.73***	-12.58***
	(2.86)	(0.13)	(3.37)	(3.29)
Private	-7.79***	-0.34**	-6.72**	-9.00***
	(2.85)	(0.14)	(3.34)	(3.24)
*Private*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 evaluation question noted in each column and 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 (0–20) on the ASVAB test. Data are from uninformed and informed self-evaluation questions in the *Self-Promotion* and *Private* versions.

3.3 Robustness of the gender gap in self-evaluations

In the previous section, we found that the gender gap in self-promotion is reflective of the gender gap in self-evaluations (i.e., it is not driven by the evaluations being shared with employers) and that the gender gap in self-evaluations persists even after we have closed any gender gap in beliefs about absolute and relative performance. We therefore want to further explore the robustness of the gender gap in self-evaluations. Does it persist in other environments? Can the provision of other information help close the gap? In the following subsections, we consider these questions and others by examining results from several additional study versions and from additional analyses of our data. Tables 7 and 8 show the results from our additional study versions along with the two study versions we have already discussed, reproduced in Panels 1 and 2 for reference.

[20] We repeat "related to" to emphasize that these gender gaps have not been studied in the context explored by our paper. Indeed, we note that "deception" (or lying or cheating) is not well-defined in our setting since there are no objectively "truthful" self-evaluations.

3.3.1 Is the gap robust to employers potentially learning true performance along with self-evaluations?

Participants in the *Self-Promotion (Risky)* version are aware that information about their actual performance could be communicated to employers along with their self-evaluation. This could make participants feel more constrained to provide appropriate self-evaluations since there is some chance of "being caught" if they inflate their self-evaluations too much. Nonetheless, Table 7 (Columns 1–4, Panel 3) shows that the gender gap in uninformed self-evaluations remains substantial and significant, and Table 8 (Columns 1–4, Panel 3) shows that the gender gap in informed self-evaluations remains substantial and significant.

3.3.2 Is the gap robust to providing information on the average self-evaluations of others?

As detailed in our Introduction, a defining characteristic of the self-evaluations we study involves the lack of objective truth. As such, a potential driver of our results could relate to men and women holding different beliefs about what self-evaluations are typical or socially appropriate. The *Private (Social Norms)* version decreases the scope for potential differences in beliefs about what self-evaluations are typical, since prior to providing informed self-evaluations, participants also learn the average self-evaluations provided by prior participants with the same performance as them. As shown in Table 8 (Columns 1–4, Panel 5), this information (along with the perfect information on their absolute and relative performance) proves ineffective at closing the gender gap, which remains large and statistically significant.

3.3.3 Is the gap robust to reduced consistency motives and anchoring effects?

When considering gender gaps in self-evaluation that may arise in the labor market, consistency could play a role. Initial self-evaluations — which could take place before participants get information about their performance — could affect subsequent self-evaluations. More generally, self-evaluations at one point in time may influence self-evaluations at a later point in time as individuals progress through their schooling and careers.

In the study versions discussed above, consistency motives or anchoring effects could influence our informed self-evaluations since we always elicit informed self-evaluations after participants provide uninformed self-evaluations. The *Private (Immediately Informed)* version decreases the scope for such consistency motives or anchoring effects by only asking participants to provide informed self-evaluations. As shown in Table 8 (Columns 1–4, Panel 7), this design change does not eliminate the gender gap in informed self-evaluations, a result that is also replicated when this study version is run for a second time (Columns 1–4, Panel 8).

3.3.4 Is the gap robust across study versions?

As evident from the many study versions discussed to this point, the gender gap in self-evaluations is quite robust. Including the replications that we ran, Table 7 (Columns 1–4, Panels 1–6) and Table 8 (Columns 1–4, Panels 1–8) show that separately considering each self-evaluation question, whether or not participants are informed, each study version, and each wave, generates 56 possible settings to look for a gender gap. We find a statistically significant gender gap in self-evaluations 56 out of 56 times.

Pooling across all of these versions involving self-evaluations, Table 7 (Columns 1–4, Panel 7) and Table 8 (Columns 1–4, Panel 9) show that the average gender gap (and the percent of the mean self-evaluation that it represents) is 13.71 (26%) in uninformed performance evaluations, 0.66 (19%) in uninformed performance-bucket evaluations, 16.63 (33%) in uninformed willingness-to-apply evaluations, 15.26 (28%) in uninformed success evaluations, 9.63 (18%) in informed performance evaluations, 0.47 (13%) in informed performance-bucket evaluations, 14.76 (28%) in informed willingness-to-apply evaluations, and 15.13 (27%) in informed success evaluations.

3.3.5 Is the gap robust to controlling for other demographic characteristics?

Table A.2 presents regressions that include the other demographic characteristics we observed in our follow-up survey (i.e., age, education, and political orientation) as controls. The gender gap in self-evaluations remains just as strong when we additionally control for these demographic characteristics, highlighting that the gender gap is not due to women and men in our sample differing in other observable ways.

3.3.6 Is the gap robust to excluding performance controls?

As detailed throughout our paper, we include performance fixed effects in our regressions so that we can compare self-evaluations among *equally performing* men and women. We view this as an important feature of our design. That said, Table A.3 (Columns 1–4, Panels 1–2) shows similar results when we do not include performance fixed effects.

3.3.7 Is the gap robust to excluding "inattentive" participants?

Table A.4 (Columns 1–4, Panels 1–2) shows similar results when we restrict our data to participants who correctly answered at least 6 out of 20 questions on the ASVAB test. Since each question involved selecting an answer from four options, we would expect participants responding randomly to correctly answer 5 out of 20 questions on average. That the gender gap in self-evaluations is similar when excluding participants who answered 5 or fewer questions correctly is consistent with our results not being driven by "inattentive" participants.

3.3.8 Is the gap robust to other distributional tests?

Table A.5 (Columns 1–4, Panels 1–3) show that the gender gap in self-evaluations — among the self-evaluation questions elicited on a 0 to 100 scale — remains statistically significant when considering quantile regressions estimated at the 25th percentile, the 50th percentile, and the 75th percentile. This provides evidence against the possibility of our results about the average self-evaluations provided by men and women being driven by "extreme" answers by men or women.[21]

[21] Note that quantile regressions are not presented for the performance-bucket question that is elicited on 6-point scale to avoid convergence issues given the discrete nature of this question and the inclusion of performance fixed effects.

3.3.9 Does the gender gap differ by other individual characteristics?

Tables A.6 investigates how the gender gap interacts with the other demographic characteristics we collected, and Table A.7 investigates how the gap interacts with ASVAB performance (replacing performance FEs with a linear control for performance). We see that the gender gap in self-evaluations is directionally (and sometimes significantly) larger for older individuals, not significantly influenced by education, significantly larger for individuals who feel more favorably about the Republican party, and significantly smaller for individuals with higher performances.[22] Nevertheless, median splits reveal that the gender gap in self-evaluations persists among those who are: older and younger, more educated and less educated, lean more Republican and lean less Republican, and higher performers and lower performers. Put differently, each of the 64 specifications that result from these 8 groups of individuals and 8 types of uninformed and informed self-evaluations questions yield a statistically significant gender gap in self-evaluations.

3.3.10 Is there a gender gap in a measure of deservingness?

In light of the gender gap in self-evaluations, one might wonder if women also believe they deserve to earn less money from the study than men. To consider this possibility, we consider results from the following question that is asked after all self-evaluations are elicited: "Out of a maximum amount of 100 cents, what amount of bonus payment, in cents, do you think you deserve for your performance on the test you took in part 1?" As shown in Appendix Figure A.1 and Table A.8, even when pooling across all study versions that involve self-evaluations, we find no statistically significant (nor economically meaningful) gender difference in response to this question, suggesting that women in our study do not simply feel less deserving or desire to earn less from the study.

[22] When considering heterogeneity by performance, recall from Figure 1 that very high ASVAB performances are rare: 75% of performances fall no more than 2.6 questions above average and 90% of performances fall no more than 4.6 questions above average.

Table 7: Uninformed Self-Evaluations in Each Study Version

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Self-Promotion Version, Wave 1				
Female	-12.68***	-0.59***	-15.31***	-15.09***
	(2.96)	(0.13)	(3.46)	(3.46)
N	302	302	302	302
Panel 2: Private Version Wave 1				
Female	-13.46***	-0.56***	-17.57***	-16.46***
	(2.93)	(0.13)	(3.51)	(3.61)
N	304	304	304	304
Panel 3: Self-Promotion (Risky) Version, Wave 1				
Female	-9.15***	-0.47***	-12.82***	-9.24***
	(2.93)	(0.13)	(3.29)	(3.32)
N	294	294	294	294
Panel 4: Private Version, Wave 2				
Female	-12.21***	-0.55***	-17.25***	-14.39***
	(3.18)	(0.15)	(3.54)	(3.53)
N	302	302	302	302
Panel 5: Private, Social Norms Version, Wave 2				
Female	-15.14***	-0.80***	-16.93***	-15.62***
	(3.28)	(0.16)	(3.71)	(3.71)
N	298	298	298	298
Panel 6: Private Version, Wave 3				
Female	-16.45***	-0.79***	-15.69***	-16.16***
	(3.18)	(0.15)	(3.92)	(3.87)
N	300	300	300	300
Panel 7: All Uninformed Self-Evaluations				
Female	-13.71***	-0.66***	-16.63***	-15.26***
	(1.25)	(0.06)	(1.42)	(1.44)
N	1800	1800	1800	1800
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 and 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 (0–20) on the ASVAB test. Data in each panel are from uninformed self-evaluations of the noted study version(s).

Table 8: Informed Self-Evaluations in Each Study Version

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: Self-Promotion Version, Wave 1				
Female	-7.01**	-0.40***	-10.73***	-11.73***
	(2.90)	(0.13)	(3.40)	(3.30)
N	302	302	302	302
Panel 2: Private Version, Wave 1				
Female	-8.01***	-0.33**	-13.25***	-13.15***
	(2.88)	(0.14)	(3.53)	(3.53)
N	304	304	304	304
Panel 3: Self-Promotion (Risky) Version, Wave 1				
Female	-7.24**	-0.36***	-9.11***	-8.07**
	(2.83)	(0.14)	(3.38)	(3.29)
N	294	294	294	294
Panel 4: Private Version, Wave 2				
Female	-7.58**	-0.42***	-14.15***	-14.37***
	(3.18)	(0.15)	(3.53)	(3.46)
N	302	302	302	302
Panel 5: Private, Social Norms Version, Wave 2				
Female	-11.93***	-0.62***	-16.39***	-15.77***
	(3.15)	(0.16)	(3.42)	(3.58)
N	298	298	298	298
Panel 6: Private Version, Wave 3				
Female	-12.70***	-0.52***	-16.55***	-15.87***
	(3.04)	(0.14)	(3.73)	(3.76)
N	300	300	300	300
Panel 7: Private, Immediately Informed Version, Wave 3				
Female	-7.61**	-0.47***	-11.42***	-12.48***
	(3.35)	(0.16)	(3.81)	(3.61)
N	299	299	299	299
Panel 8: Private, Immediately Informed Version, Wave 4				
Female	-8.54***	-0.42***	-16.63***	-18.66***
	(2.22)	(0.10)	(2.42)	(2.30)
N	597	597	597	597
Panel 9: All Informed Self-Evaluations				
Female	-9.63***	-0.47***	-14.76***	-15.13***
	(1.00)	(0.05)	(1.14)	(1.13)
N	2696	2696	2696	2696
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 and 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 (0–20) on the ASVAB test. Data in each panel are from informed self-evaluations of the noted study version(s).

3.4 Other-Evaluations

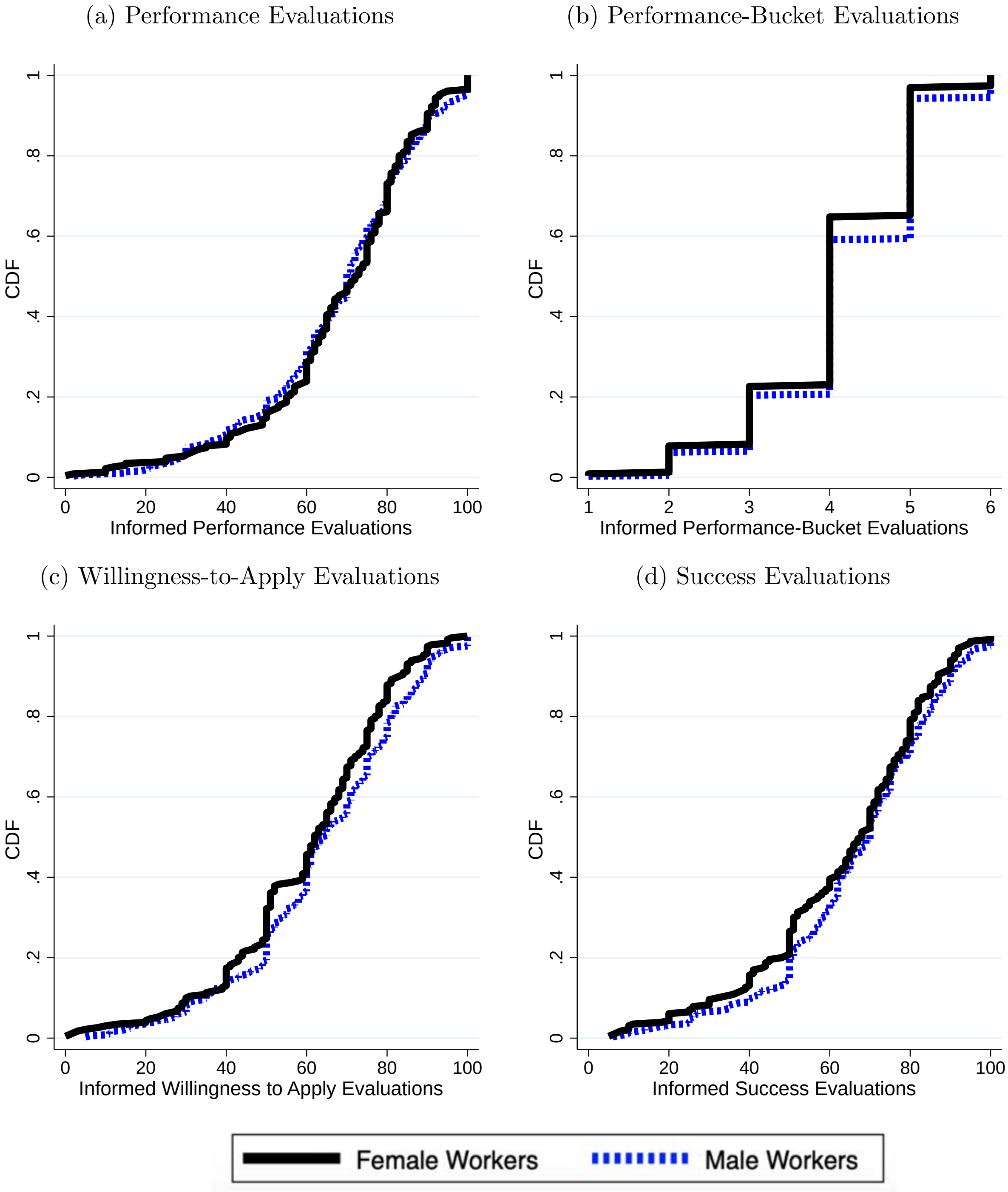
One potential explanation for a gender gap in self-evaluations is that men and women have different "standards" in general or a different mapping from performance to self-evaluations. For example, women may believe that scoring a 15 out of 20 (and being in the 80th percentile) is only "good" and worth a 70/100 on the 100-point scale while men believe such a score is "very good" and worth an 85/100.

On one hand, how we identify the gender gap in self-evaluations is synonymous with this explanation, and correspondingly, is not an explanation we seek to "rule out." That we document a gender gap in self-evaluations between equally performing men and women — even when they know how well they performed — directly implies that men and women have a different mappings from performance to self-evaluations.

On the other hand, we can investigate if a gender difference in mappings from performance to evaluations exists in general or is specific to evaluations of own performance. First, as shown in Table A.9, if we return to data from our *Employer* version and include controls for whether the employer is female, we find no gender difference in how male and female employers respond to the self-evaluations provided by workers, suggesting that men and women do not differ in how they view self-evaluations provided by others. Second, and more directly, we turn to our *Private (Other-Evaluation)* version, where we can examine if a gender gap exists in evaluations about the performance of others. In particular, we ask whether there is a gender gap in informed other-evaluations (uninformed other-evaluations would be difficult to interpret and were not collected).

Figure 3 and Table 9 show that there is little-to-no gender difference in other-evaluations. More specifically, the gender gap in other-evaluations is small and statistically insignificant when considering performance evaluations and performance-bucket evaluations. While the gender gap in other-evaluations is statistically significant when considering willingness-to-apply evaluations and marginally statistically significant when considering success evaluations, the magnitudes of these gaps are notably small. In particular, see Table 8 (Columns 3 and 4, Panel 8) to consider the gender gap in the corresponding self-evaluations from the *Private (Immediately Informed)* version that was run at the same time as the *Private (Other-Evaluation)* version. These gender gaps are on average equal to 16.63 and 18.77, implying that asking participants to evaluate others rather than themselves causes the gender gap in willingness-to-apply evaluations to decrease by 80% and the gender gap in success evaluations to decrease by 83%. If we include an interaction variable to capture the impact of evaluating someone else rather than oneself, we see that all of the decreases in the gender gap are statistically significant ($p < 0.01$).

Figure 3: Other-Evaluations



(a) Performance Evaluations

(b) Performance-Bucket Evaluations

(c) Willingness-to-Apply Evaluations

(d) Success Evaluations

Table 9: Other-Evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Female	0.29	-0.11	-3.54**	-3.17*
	(1.58)	(0.08)	(1.69)	(1.68)
N	597	597	597	597
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 and 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 (0–20) on the ASVAB test. Data are from informed other-evaluation questions in the *Private (Other-Evaluation)* version.

4 Conclusion

We have documented a gender gap in self-evaluations. When providing evaluations of their own performance, women systematically provide less favorable self-evaluations than equally performing men. The gap is only partially explained by the gender gap in beliefs about absolute and relative performance; we find that over two-thirds of the gap persists when participants are perfectly informed of their absolute and relative performance. The gap is not driven by differential willingness to distort self-evaluations or to gender differences in beliefs or preferences over payoffs; we find that it is present, and just as large, in study versions absent employers. In addition, we find that the gap is robust to: an environment where actual performance — along with self-evaluations — may be provided to employers, an environment where information on the average self-evaluations of others is provided, and an environment in which the potential for consistency motives or anchoring effects is reduced.

We focus our work on self-evaluations because we view it as an understudied behavior that could have important implications for labor market outcomes. To narrow in on self-evaluations themselves, to identify and compare equally performing men and women, and to exogenously manipulate potential underlying mechanisms (e.g., the role of beliefs about performance and the role of incentives to distort self-evaluations), we conducted this work in a series of experiments. An important avenue for future work, however, is to explore the extent to which self-evaluations impact labor market outcomes in field contexts. In many field settings, individuals are explicitly asked to complete self-evaluations: in applications to educational institutions, in job applications, and in performance reviews. Other contexts may invite implicit opportunities to provide self-evaluations. Individuals may communicate about their performance in written work products — for example, a recent article published in the British Medical Journal finds that women are less likely to use "positive" words in their titles and abstracts for papers on clinical research (Lerchenmueller, Sorenson and Jena, 2019).[23] Individuals may also communicate about their performance and ability in negotiations, in presentations and meetings, in group decision-making contexts, and in other conversations at work. Understanding how individuals construct and convey self-evaluations in all of these contexts is likely to be important and policy relevant.

Three additional avenues of future work appear promising to pursue. First, what explains the existence of the gender gap in self-evaluations? Our paper makes progress on this question by showing that the gender gap in evaluations is eliminated or substantially reduced when participants evaluate the performance of others rather than themselves. Our findings suggest that the gender gap in self-evaluations is deeply ingrained (i.e., it survives in every context we examine) and is specific to evaluations of own performance. How could the gender gap in self-evaluations be deeply ingrained? One possibility is that, consistent with the importance of culture (Gneezy, Leonard and List, 2009), the experiences of women and men in society lead women to internalize different benefits and costs to providing favorable self-evaluations (i.e., even if these benefits and costs are not relevant in our study).[24]

Second, given the prevalence of self-evaluations, how can impacts of the gender gap in self-evaluations on education and labor market outcomes be mitigated? Given our inability to close the gender gap with perfect information about absolute and relative performance and information about the average self-evaluations of others, the potential for information interventions to close the gap seems limited. That there is little to no gender differences in other-evaluations further suggests that men and women do not need more information to close the gap in mapping from performance to evaluations. Thus, the most effective approaches could involve "changing the system" rather than "changing the women," perhaps by decreasing the importance of self-evaluations in educational and work environments.[25]

Third, in light of the large literature on discrimination and gender-specific backlash (Riach and Rich, 2002; Bowles, Babcock and Lai, 2007; Rudman and Phelan, 2008; Blau and Kahn, 2017), how does making gender known influence self-evaluations and how employers view self-evaluations?[26]

[23] 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).

[24] For additional work on how gender gaps vary by culture, see Andersen et al. (2013) and Andersen et al. (2018). For work on the importance of gender norms and identity, see Akerlof and Kranton (2000), Bertrand (2011), and Bertrand, Kamenica and Pan (2015).

[25] While not on self-evaluations, prior literature does find evidence in support of some change-the-system approaches. For instance, He, Kang and Lacetera (2019) shows that the gender gap in willingness to enter competition is eliminated when individuals must "opt-out" of a competition rather than "opt-in" to a competition; Apicella, Demiral and Mollerstrom (2017) and Apicella, Demiral and Mollerstrom (2020) show that gender gaps in competitive entry are mitigated when women are asked to compete against themselves rather than others; Coffman, Collis and Kulkarni (2019*a*) shows that the gender gap in willingness to apply to an advanced job is eliminated when individuals are provided with clear guidance as to the conditions under which they should apply (e.g., if they have scored above some threshold on a skills-assessment test); and a reduction in ambiguity along with other contextual features has mitigated gender gaps in negotiations (Bowles and McGinn, 2008; Mazei et al., 2015; Leibbrandt and List, 2015).

[26] Bursztyn, Fujiwara and Pallais (2017) show how image concerns can cause women to downplay how they describe their career ambitions to others. For recent evidence on gender discrimination and how the ability of men and

Employers and workers in our experiment do not learn the gender — nor any other identifiable characteristics — of each other. We view future work on the interaction between self-evaluations and making gender known as important and note that the expected results could go in either direction. On one hand, if employers make their own mental corrections because they accurately expect that men will have higher self-evaluations conditional on performance, this may mitigate any gender gap in labor market outcomes that arises from a gender gap in self-promotion. However, Reuben, Sapienza and Zingales (2014) provides evidence against the empirical relevance of this possibility. That paper finds that men, more than women, tend to inflate their performance estimates (for which objective truths exist), but it finds that employers do not (fully) account for this. On the other hand, prior literature makes clear the potential for greater backlash for women relative to men, which could exacerbate gender gaps in self-evaluations — and their associated impact on education and labor market outcomes — when gender is known.

women are judged differently, see Reuben, Sapienza and Zingales (2014); Milkman, Akinola and Chugh (2015); Baert, De Pauw and Deschacht (2016); Bohnet and Bazerman (2016); Sarsons (2017*a*,*b*); Alston (2019); Bohren, Imas and Rosenberg (2019); Bohren et al. (2019); Coffman, Exley and Niederle (Forthcoming); Kessler, Low and Sullivan (2019); Sarsons et al. (Forthcoming)

References

Adams, Nathan R., Michael A. Kuhn, and Glen R. Waddell. 2019. "Confidence and Contrition: Is Cheating Internalized in Performance Assessments?" *Working Paper.*

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and identity." *The Quarterly Journal of Economics*, 115(3): 715–753.

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

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A List, and Sandra Maximiano. 2018. "On the cultural basis of gender differences in negotiation." *Experimental Economics*, 21(4): 757–778.

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.

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

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

Bertrand, Marianne, Emir Kamenica, and Jessica Pan. 2015. "Gender identity and relative income within households." *Quarterly Journal of Economics*, 130(2): 571–614.

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

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

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

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

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

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

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, and Kathleen L. McGinn. 2008. "Gender in job negotiations: a two-level game." *Negotiation Journal*, 24(4): 393–410.

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

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

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

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

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

Childs, Jason. 2012. "Gender differences in lying." *Economics Letters*, 114(2): 147–149.

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

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

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

Coffman, Katherine B., Manuela R. Collis, and Leena Kulkarni. 2019*a*. "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*b*. “Stereotypes and Belief Updating.” *Working Paper.*

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

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

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

Dreber, Anna, and Magnus Johannesson. 2008. “Gender differences in deception.” *Economics Letter*, 99(1).

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.

Erat, Sanjiv, and Uri Gneezy. 2012. “White Lies.” *Management Science*, 50(4): 723–733.

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

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

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

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

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

Gylfason, Haukur Freyr, Audur Arna Arnardottir, and Kari Kristinsson. 2013. “More on gender differences in lying.” *Economic Letters*, 119(1): 94–95.

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.*

Houser, Daniel, Stefan Vetter, and Joachim Winter. 2012. “Fairness and cheating.” *European Economic Review*, 56: 1645–1655.

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

Kessler, Judd B. 2017. “Announcements of support and public good provision.” *American Economic Review*, 107(12): 3760–87.

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

Kling, Kristen C, Janet Shibley Hyde, Carolin J Showers, and Brenda N Buswell. 1999. “Gender differences in self-esteem: a meta-analysis.” *Psychological bulletin*, 125(4).

Leibbrandt, Andreas, and John A. List. 2015. “Do women avoid salary negotiations? Evidence from a large-scale natural field experiment.” *Management Science*, 61(9): 2016–2024.

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.*

Mazei, Jens, Joachim Hüffmeier, Philipp Alexander Freund, Alice F. Stuhlmacher, Lena Bilke, and Guido Hertel. 2015. “A meta-analysis on gender differences in negotiation outcomes and their moderators.” *Psychological Bulletin*, 141(1): 85–104.

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

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

Moore, Don A, and Paul J Healy. 2008. “The trouble with overconfidence.” *Psychological review*, 115(2).

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

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

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

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

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

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

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

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

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

Schotter, Andrew, and Isabel Trevino. 2014. "Belief elicitation in the laboratory." *Annual Review of Economics*, 6(1): 103–128.

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

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

A Appendix

A.1 Additional Tables

Table A.1: Statistically controlling for the role of beliefs in uninformed self-evaluations

Evaluation	Performance		Performance- -Bucket		Willingness to Apply		Success	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Self-Promotion Version								
Female	-3.72	-3.81	-0.19*	-0.19	-7.14**	-6.71**	-6.70**	-6.48**
	(2.37)	(2.44)	(0.11)	(0.12)	(3.16)	(3.33)	(3.07)	(3.11)
Performance Beliefs	4.08***		0.18***		3.72***		3.82***	
	(0.29)		(0.01)		(0.38)		(0.38)	
N	302	302	302	302	302	302	302	302
All Versions with Uninformed Self-Evaluations								
Female	-5.39***	-4.73***	-0.28***	-0.25***	-8.65***	-8.01***	-7.11***	-6.30***
	(0.95)	(0.96)	(0.05)	(0.05)	(1.24)	(1.25)	(1.23)	(1.24)
Performance Beliefs	4.06***		0.19***		3.89***		3.98***	
	(0.13)		(0.01)		(0.15)		(0.15)	
N	1800	1800	1800	1800	1800	1800	1800	1800
Performance FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Belief FEs	No	Yes	No	Yes	No	Yes	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column and 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 (0–20) on the ASVAB test. Belief FEs are dummies for each possible reported absolute performance belief (0–20) on the ASVAB test. Data in the top panel are from uninformed self-evaluations in the *Self-Promotion* version. Data in the bottom panel are from uninformed self-evaluations in all versions that contain uninformed self-evaluation questions, that is from all versions except for the *Private (Immediately Informed)* version and the *Private (Other-Evaluation)* version.

Table A.2: Including additional controls, the gender gap in self-evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: All Uninformed Self-Evaluations				
Female	-12.44***	-0.60***	-15.17***	-13.82***
	(1.19)	(0.06)	(1.38)	(1.40)
Age	-0.30***	-0.01***	-0.31***	-0.24***
	(0.06)	(0.00)	(0.07)	(0.07)
Education (1-9)	4.34***	0.22***	4.67***	5.09***
	(0.43)	(0.02)	(0.50)	(0.51)
Republican Leaning (0-100)	0.14***	0.01***	0.12***	0.11***
	(0.02)	(0.00)	(0.02)	(0.03)
N	1800	1800	1800	1800
Panel 2: All Informed Self-Evaluations				
Female	-8.37***	-0.41***	-13.32***	-13.62***
	(0.96)	(0.05)	(1.10)	(1.10)
Age	-0.29***	-0.01***	-0.24***	-0.20***
	(0.04)	(0.00)	(0.05)	(0.05)
Education (1-9)	3.44***	0.16***	4.24***	4.50***
	(0.35)	(0.02)	(0.40)	(0.40)
Republican Leaning (0-100)	0.17***	0.01***	0.16***	0.13***
	(0.02)	(0.00)	(0.02)	(0.02)
N	2694	2694	2694	2694
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 and as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Age* equals each participant's age, demeaned by the average age. *Education (1-9)* is a number from 1 to 9 that corresponds with lower to higher levels of education, demeaned by the average level. *Republican Leaning (0-100)* is a number from 0 to 100 that indicates the extent to which a participant indicated feeling favorably about the Republican party, demeaned by the average number. Performance FEs are dummies for each possible performance (0–20) on the ASVAB test. Data are from all study versions involving self-evaluations except for the 2 participants who indicated "other" as their educational attainment, restricted to the set of self-evaluations noted in each panel.

Table A.3: Without performance fixed effects, the gender gap in self-evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: All Uninformed Self-Evaluations				
Female	-15.54***	-0.78***	-18.40***	-17.06***
	(1.25)	(0.06)	(1.42)	(1.44)
Constant	59.07***	3.80***	58.25***	62.03***
	(0.78)	(0.04)	(0.89)	(0.87)
N	1800	1800	1800	1800
Panel 2: All Informed Self-Evaluations				
Female	-10.71***	-0.55***	-16.16***	-16.21***
	(1.08)	(0.05)	(1.18)	(1.18)
Constant	58.01***	3.78***	59.12***	62.40***
	(0.65)	(0.03)	(0.70)	(0.68)
N	2696	2696	2696	2696
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 and 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 self-evaluations.

Table A.4: Excluding very low performers, the gender gap in self-evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: All Uninformed Self-Evaluations				
Female	-13.57***	-0.62***	-16.88***	-15.69***
	(1.31)	(0.06)	(1.54)	(1.55)
N	1490	1490	1490	1490
Panel 2: All Informed Self-Evaluations				
Female	-8.62***	-0.39***	-14.37***	-14.66***
	(1.04)	(0.05)	(1.22)	(1.22)
N	2175	2175	2175	2175
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 and 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 (0–20) on the ASVAB test. Data are from all study versions involving self-evaluations, restricted to the set of participants who answered at least 6 out of 20 questions correctly.

Table A.5: Quantile regressions, the gender gap in self-evaluations

Evaluation:	Performance	Willingness-to-Apply	Success
	(1)	(2)	(3)
Panel 1: All Uninformed Self-Evaluations (25th percentile)			
Female	-19.00***	-25.00***	-29.00***
	(2.45)	(2.64)	(3.17)
N	1800	1800	1800
Panel 2: All Informed Self-Evaluations (25th percentile)			
Female	-10.00***	-20.00***	-24.00***
	(1.68)	(2.17)	(2.34)
N	2696	2696	2696
Panel 3: All Uninformed Self-Evaluations (50th percentile)			
Female	-14.00***	-22.00***	-17.00***
	(2.15)	(2.44)	(2.31)
N	1800	1800	1800
Panel 4: All Informed Self-Evaluations (50th percentile)			
Female	-9.00***	-16.00***	-17.00***
	(1.28)	(1.85)	(1.88)
N	2696	2696	2696
Panel 5: All Uninformed Self-Evaluations (75th percentile)			
Female	-11.00***	-13.00***	-10.00***
	(1.38)	(1.88)	(1.79)
N	1800	1800	1800
Panel 6: All Informed Self-Evaluations (75th percentile)			
Female	-5.00***	-10.00***	-10.00***
	(1.14)	(1.41)	(1.45)
N	2696	2696	2696
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 evaluation question noted in each column and 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 (0–20) on the ASVAB test. Data are from all study versions involving self-evaluations.

Table A.6: By other demographics, the heterogeneity in the gender gap in self-evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: All Uninformed Self-Evaluations				
Female	-12.51***	-0.60***	-15.12***	-13.69***
	(1.19)	(0.06)	(1.39)	(1.40)
Age	-0.23***	-0.01***	-0.17*	-0.07
	(0.08)	(0.00)	(0.09)	(0.09)
Education (1-9)	4.45***	0.22***	4.36***	4.53***
	(0.55)	(0.03)	(0.67)	(0.66)
Republican Leaning (0-100)	0.19***	0.01***	0.19***	0.17***
	(0.03)	(0.00)	(0.03)	(0.03)
*Age *Female*	-0.12	-0.00	-0.28**	-0.33**
	(0.12)	(0.01)	(0.13)	(0.13)
*Education (1-9) *Female*	-0.23	-0.01	0.72	1.31
	(0.87)	(0.04)	(1.01)	(1.03)
Republican Leaning (0-100)	-0.11***	-0.01***	-0.14***	-0.12**
**Female*	(0.04)	(0.00)	(0.05)	(0.05)
N	1800	1800	1800	1800
Panel 2: All Informed Self-Evaluations				
Female	-8.34***	-0.41***	-13.24***	-13.52***
	(0.96)	(0.05)	(1.11)	(1.10)
Age	-0.24***	-0.01***	-0.15**	-0.10
	(0.06)	(0.00)	(0.07)	(0.07)
Education (1-9)	3.34***	0.17***	4.10***	4.32***
	(0.47)	(0.02)	(0.52)	(0.50)
Republican Leaning (0-100)	0.23***	0.01***	0.20***	0.17***
	(0.02)	(0.00)	(0.03)	(0.03)
*Age *Female*	-0.11	-0.00	-0.21**	-0.24**
	(0.09)	(0.00)	(0.10)	(0.10)
*Education (1-9) *Female*	0.21	-0.02	0.34	0.46
	(0.72)	(0.03)	(0.83)	(0.82)
Republican Leaning (0-100)	-0.14***	-0.01***	-0.11***	-0.08**
**Female*	(0.03)	(0.00)	(0.04)	(0.04)
N	2694	2694	2694	2694
Performance FEs	Yes	Yes	Yes	Yes

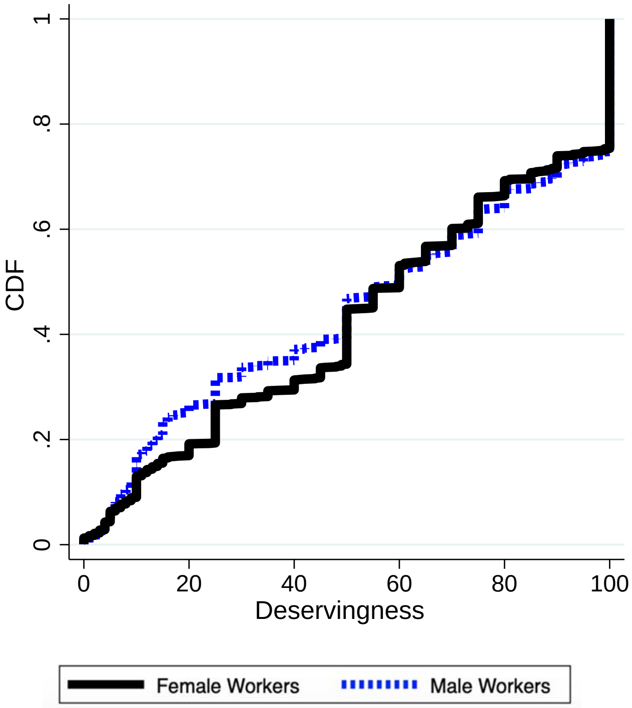
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column and as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Age* equals each participant's age, demeaned by the average age. *Education (1-9)* is a number from 1 to 9 that corresponds with lower to higher levels of education, demeaned by the average level. *Republican Leaning (0-100)* is a number from 0 to 100 that indicates the extent to which a participant indicated feeling favorably about the Republican party, demeaned by the average number. Performance FEs are dummies for each possible performance (0–20) on the ASVAB test. Data are from all study versions involving self-evaluations except for the 2 participants who indicated "other" as their educational attainment, restricted to the set of self-evaluations noted in each panel.

Table A.7: By performance, the heterogeneity in the gender gap in self-evaluations

Evaluation:	Performance	Performance-Bucket	Willingness-to-Apply	Success
	(1)	(2)	(3)	(4)
Panel 1: All Uninformed Self-Evaluations				
Female	-15.95***	-0.78***	-18.64***	-17.41***
	(1.30)	(0.06)	(1.45)	(1.47)
Demeaned Performance	-0.69***	-0.07***	-0.54***	-0.44**
	(0.19)	(0.01)	(0.21)	(0.20)
*Demeaned Performance*Female*	1.52***	0.09***	1.07***	1.12***
	(0.36)	(0.02)	(0.40)	(0.41)
Constant	58.96***	3.79***	58.17***	61.96***
	(0.78)	(0.04)	(0.89)	(0.87)
N	1800	1800	1800	1800
Panel 2: All Informed Self-Evaluations				
Female	-11.82***	-0.58***	-16.82***	-17.12***
	(1.08)	(0.05)	(1.19)	(1.18)
Demeaned Performance	0.42***	-0.02**	0.03	0.45***
	(0.15)	(0.01)	(0.16)	(0.16)
*Demeaned Performance*Female*	1.98***	0.10***	1.57***	1.45***
	(0.30)	(0.01)	(0.32)	(0.32)
Constant	58.13***	3.78***	59.13***	62.54***
	(0.64)	(0.03)	(0.70)	(0.68)
N	2696	2696	2696	2696
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 and as defined in the notes of Table 2. *Female* is an indicator for the participant being female. *Demeaned Performance* equals each participant's performance out of the 20 questions on the ASVAB, demeaned by the average performance. Data are from all study versions involving self-evaluations, restricted to the set of self-evaluations noted in each panel.

Figure A.1: Deservingness Measure Distributions



This graph shows the CDF of the deservingness measure 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." Data are from all study versions involving self-evaluations.

Table A.8: Deservingness Measure Regressions

	(1)	(2)
Female	1.96	-1.74
	(1.35)	(1.14)
Constant	56.56***	
	(0.90)	
Performance FEs	No	Yes
N	2696	2696

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the deservingness measure 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 (0–20) on the ASVAB test. Data are from all study versions involving self-evaluations.

Table A.9: *Employer Version*, Wage Regressions

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

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are clustered by employer. Results are from OLS regressions of the wage received by the participant (25 cents if not hired and a chosen wage from 25–100 cents if hired). *Female Employer* is an indicator for a female employer. *Evaluation* is the evaluation provided by each participant in the evaluation question noted in that column. *Performance* indicates 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." *Performance-Bucket* indicates 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." *Willingness-to-Apply* 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." *Success* 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." Data are from the hiring decisions in the *Employer* version.

A.2 The *Free-Response Evaluators* Versions

In February 2019, we recruited 400 workers on MTurk to complete the *Free-Response Evaluators* versions of our study using the same criteria as in the main study versions (see footnote 8). We collected data from 399 workers.[27] Each participant received a guaranteed $1.50 completion fee for the 15-minute study. In addition, one of their decisions, out of the 21 decisions in the study, was selected to determine a possible bonus payment for them and, if relevant, for an associated "worker."[28] After evaluators completed all decisions of the study, they took a short follow-up survey that collected demographic information.

The evaluators were randomly assigned either to make 21 hiring decisions (n=198) or to make 21 sets of predictions (n=201). Before making each decision or set of predictions, the evaluator was provided with the text entered by a 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. Evaluators were randomly assigned these 21 free responses from the set of eligible free responses written by the participants from the three versions of the study run in the first wave.[29]

Evaluators assigned to make hiring decisions were asked whether they would like to hire the participant who provided that free response and, if so, how much to pay them. The payoffs for the evaluator and associated participant are the same as described in the *Employer* version.[30] While similar to the *Employer* version, there are many more possible free responses than answers to the quantitative self-promotion questions, which means our analysis on hiring decisions is underpowered relative to the *Employer* version, since we only have at most a few evaluators reacting to each free response.

Evaluators assigned to make predictions were instead asked to predict whether the participant who wrote the free response was male or female and how many questions, out of 20, that participant answered correctly on the ASVAB. The payoffs for evaluators are determined as follows. One of the two predictions from one of the 21 sets was randomly selected. If the prediction was correct, the evaluator received a bonus payment of 50 cents.[31]

Relative to the *Employer* version, there are three important differences when considering the results

[27] One worker was excluded from participation for having previously participated in the study but was counted as being recruited.

[28] 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, in the *Free-Response Evaluators* versions, only select workers from the *Self-Promotion* and *Self-Promotion (Risky)* versions were matched with an evaluator and received corresponding payoffs, rather than everyone. Since we also wanted evaluators to provide data on the free responses from the *Private* version, evaluators 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).

[29] Not all of the free responses collected in the study were evaluated. First, the *Free-Response Evaluators* versions were run after the first wave but before the second wave, so free responses from the 2nd–4th waves did not yet exist. We consequently consider the 1800 free responses from the *Self-Promotion* version, the *Self-Promotion (Risky)* version, and the *Private* version run in the first wave. Second, a research assistant — blinded to sex and study version — deemed 130 of the 1800 potentially eligible free responses "ineligible" due to the answer not relating to the question asked or due to severe grammar and/or spelling issues that made an answer incomprehensible. Consequently, the evaluators 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 an evaluator.

[30] As explained in footnote 28, however, free responses from the *Private* version were never selected for payment.

[31] Unlike hiring decisions, the randomly selected prediction can come from a participant from any of our three study versions run in the first wave.

in the *Free-Response Evaluators* 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 the impact of a one unit increase on the 0–100 scales in the *Employer* version). Second, while evaluators are not informed of the gender of the associated worker, 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 predictions. Third, as noted above, given the large number of possible free responses, we are underpowered to consider the effect of specific free responses.

For these reasons, we favor the analysis of our quantitative self-evaluation questions presented in the main text to examine the gender gap in self-evaluation. Here, however, we investigate the hiring decisions and predictions from the *Free-Response Evaluators* versions to present several interesting (but inherently secondary) results. Given our power issues, we combine free responses from all three study versions (i.e., the *Self-Promotion*, *Private*, and *Self-Promotion (Risky)* versions).[32] In cases where multiple evaluators faced a decision about the same free response, we use the average decision provided by the evaluators (e.g., if a free response is predicted to be written by a female participant by one evaluator but a male participant by another evaluator, that participant is recorded as being predicted to be female with a 0.50 probability).

Table A.10 presents results from regressions testing whether the gender of the free response author affects the hiring decisions and predictions of evaluators. Columns (1) and (3) have no controls, and (2) and (4) have dummies for each level of performance. Panel 1 shows that evaluators predict that free responses provided by female participants come from lower-performing workers. This evidence is relatively consistent with our findings from the quantitative self-evaluation questions since women appear to provide less favorable subjective evaluations of their performance. Panel 2 shows that, although these evaluators are not informed of the gender of the participant associated with the free response, evaluators can infer gender — to some degree — when viewing the responses. Evaluators are significantly more likely to predict that a response was written by a female participant when it was indeed written by a female participant. Panel 3 shows that the relationship between the gender of the worker and evaluators' hiring decisions is inconclusive. Based on the free response answers, evaluators pay directionally, but not significantly, less to female worker. We note that there are several possible explanations for this last finding. For instance, a preference to hire workers believed to be higher performing (who are more likely to be male, per our first finding) may counteract a preference to hire workers believed to be female (who are more likely to be female, per our second finding). In other words, hiring decisions based off of the free responses may conflate performance beliefs and other preferences. As mentioned in footnote 13 in the main text of the paper, this difficulty with the free-response data contributes to our decision to focus our self-evaluation analysis on the quantitative self-evaluation questions we explore in the main text of the paper.

[32] The results are qualitatively similar when restricting to the data from each of these three versions, with one possible exception: the gender difference in the wage data is largely statistically insignificant but is sometimes directionally negative and sometimes directionally positive, depending on the study version.

Table A.10: Free Response Regressions

Sample:	Uninformed Free Responses		Informed Free Responses	
	(1)	(2)	(3)	(4)
Panel 1: DV = Predicted Performance				
Female	-0.82***	-0.67***	-0.51**	-0.35
	(0.23)	(0.22)	(0.24)	(0.23)
Constant	12.16***		12.36***	
	(0.17)		(0.18)	
N	749	749	773	773
Panel 2: DV = Predicted Probability Female				
Female	0.08***	0.08***	0.09***	0.09***
	(0.03)	(0.03)	(0.03)	(0.03)
Constant	0.37***		0.34***	
	(0.02)		(0.02)	
N	749	749	773	773
Panel 3: DV = Wage				
Female	-1.28	-1.44*	-0.96	-0.66
	(0.82)	(0.81)	(0.99)	(1.04)
Constant	33.58***		35.45***	
	(0.60)		(0.76)	
N	743	743	755	755
Performance FEs	No	Yes	No	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. SEs are robust. Results are from OLS regressions of the noted dependent variable (DV). *Predicted Performance* equals the number of questions that an evaluator predicts a participant correctly answered out of the 20 ASVAB questions. *Predicted Probability Female* equals the probability with which an evaluator predicted a participant to have been female. *Wage* equals the wage given to the participant by an evaluator. *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance (0–20) on the ASVAB test. Data in columns in (1) and (2) are from uninformed free responses elicited in part 2 and data in columns (3) through (4) are from informed free responses elicited in part 3 of all three study versions run in our first wave of data collection: the *Self-Promotion* version, the *Private* version, and the *Self-Promotion (Risky)* version.

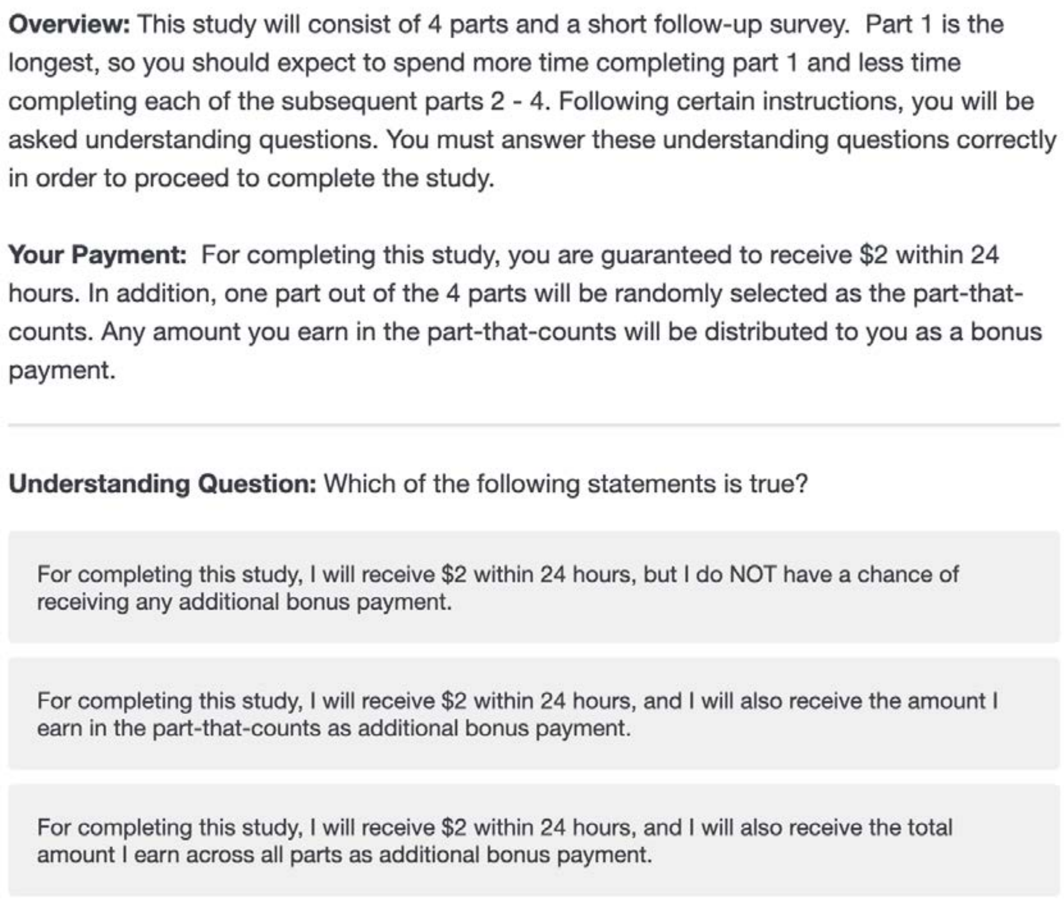
B Experimental instructions

B.1 Instructions for *Self-Promotion* 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 $2 study completion fee and of the opportunity to earn additional payment for themselves. Figure B.1 shows how this payment information is explained along with the understanding question that the participant must answer correctly to proceed.

Figure B.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?

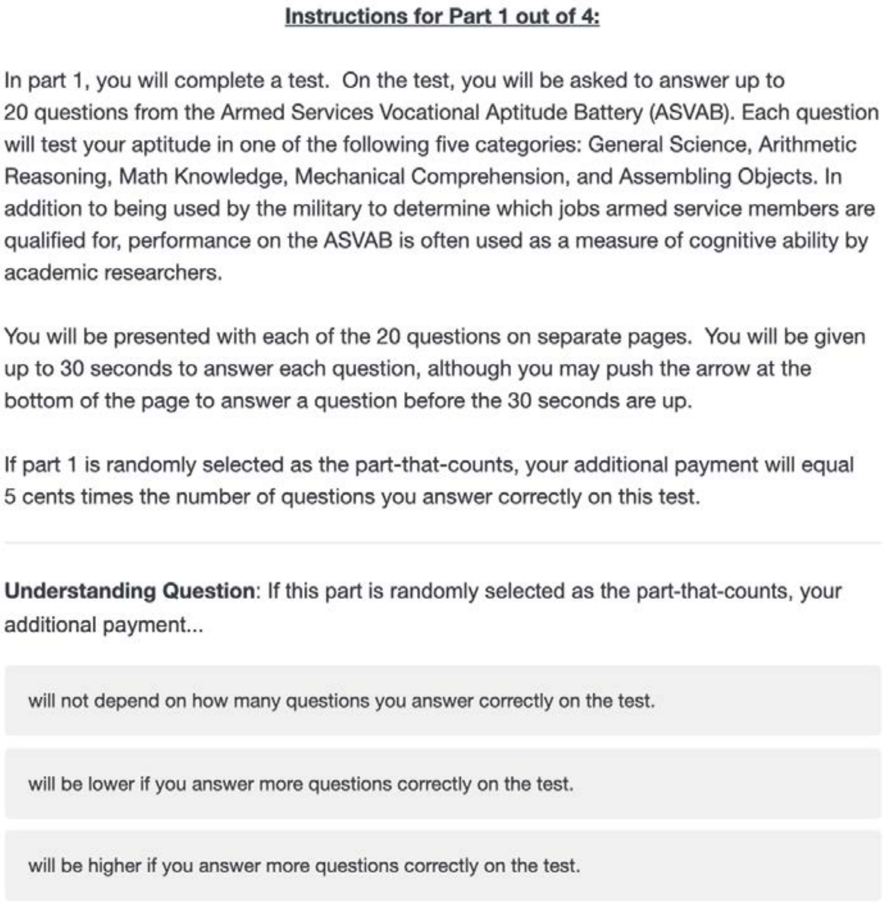
For completing this study, I will receive $2 within 24 hours, but I do NOT have a chance of receiving any additional bonus payment.

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 B.2 and an example of an ASVAB question is displayed in Figure B.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 B.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.

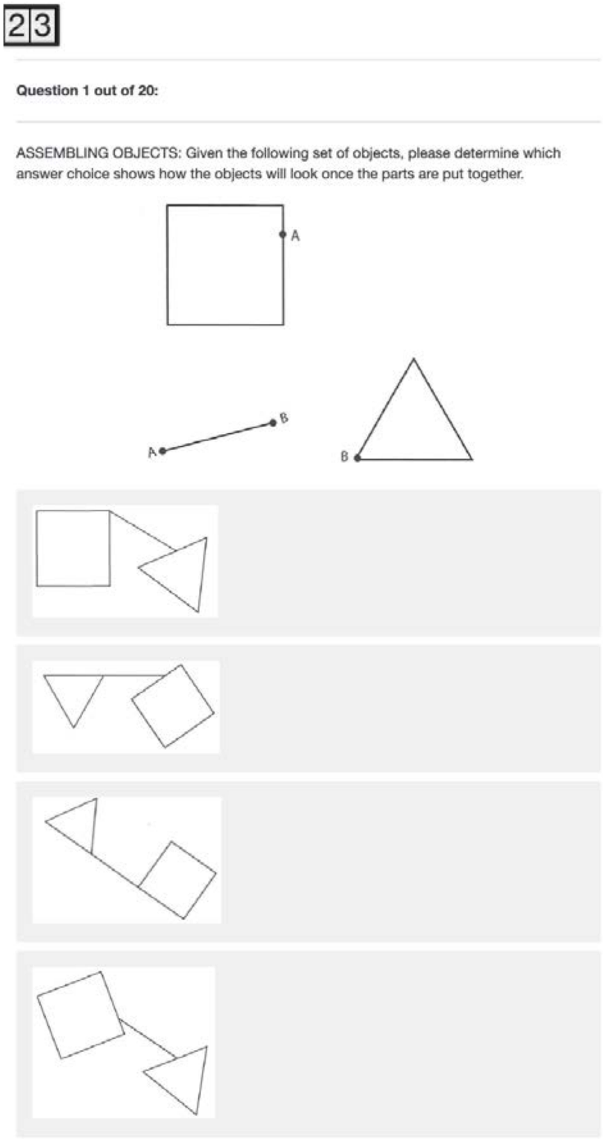
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

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 B.3: Part 1: Example ASVAB question



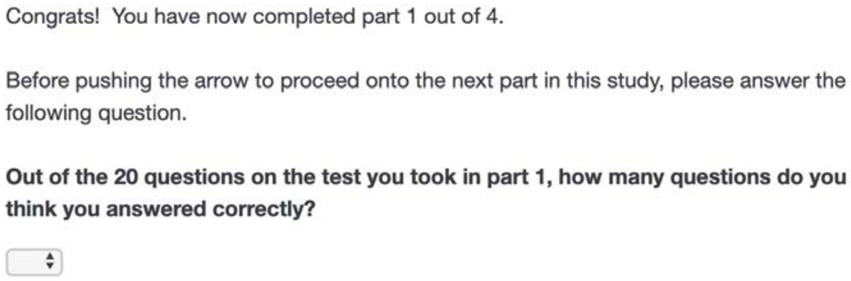
23

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 B.4.

Figure B.4: Absolute Performance Belief Question



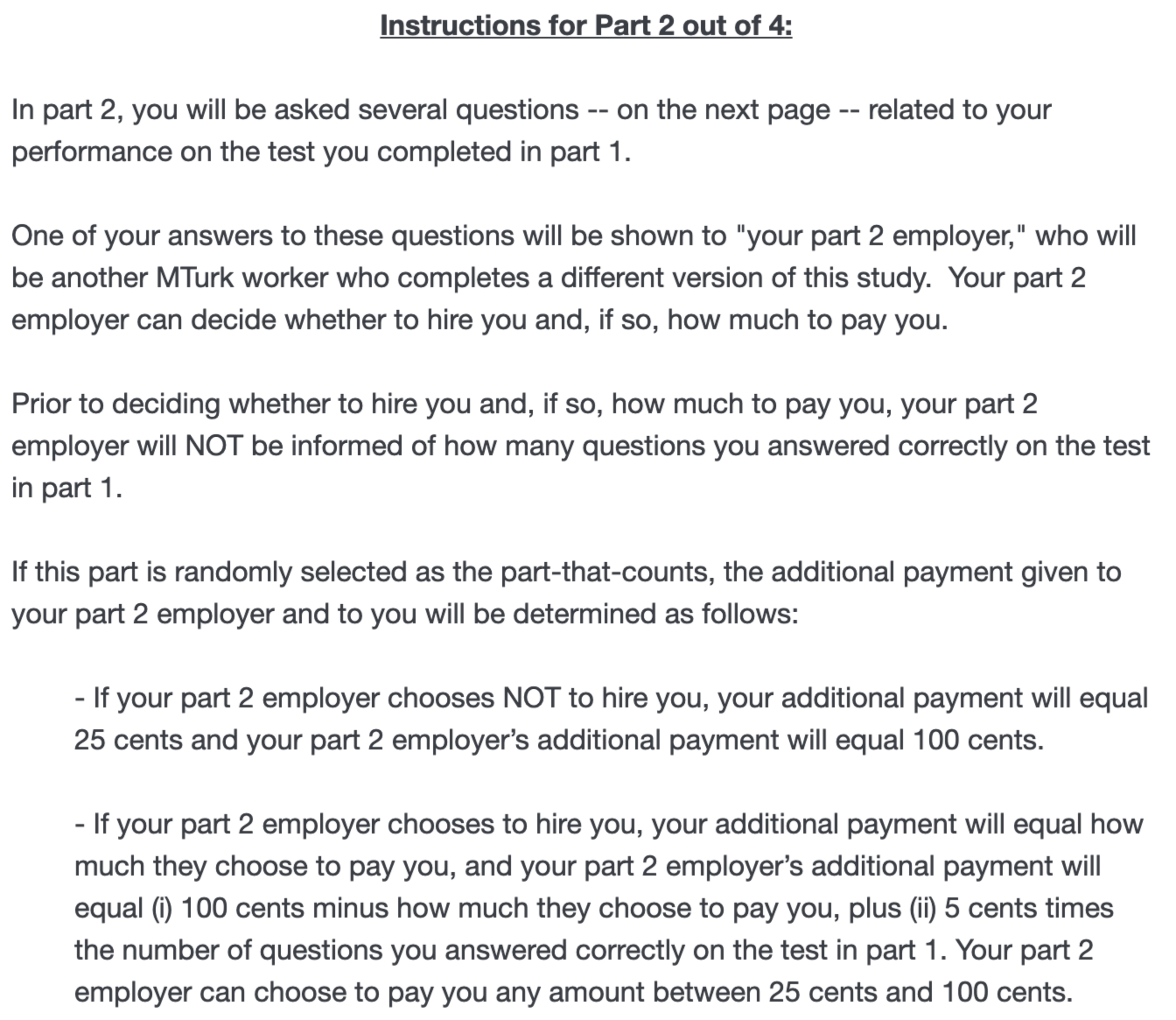
Congrats! You have now completed part 1 out of 4.

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 B.5), must correctly answer understanding questions about those instructions (see Figure B.6), and then are asked the self-evaluation questions (see Figure B.7).

Figure B.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.

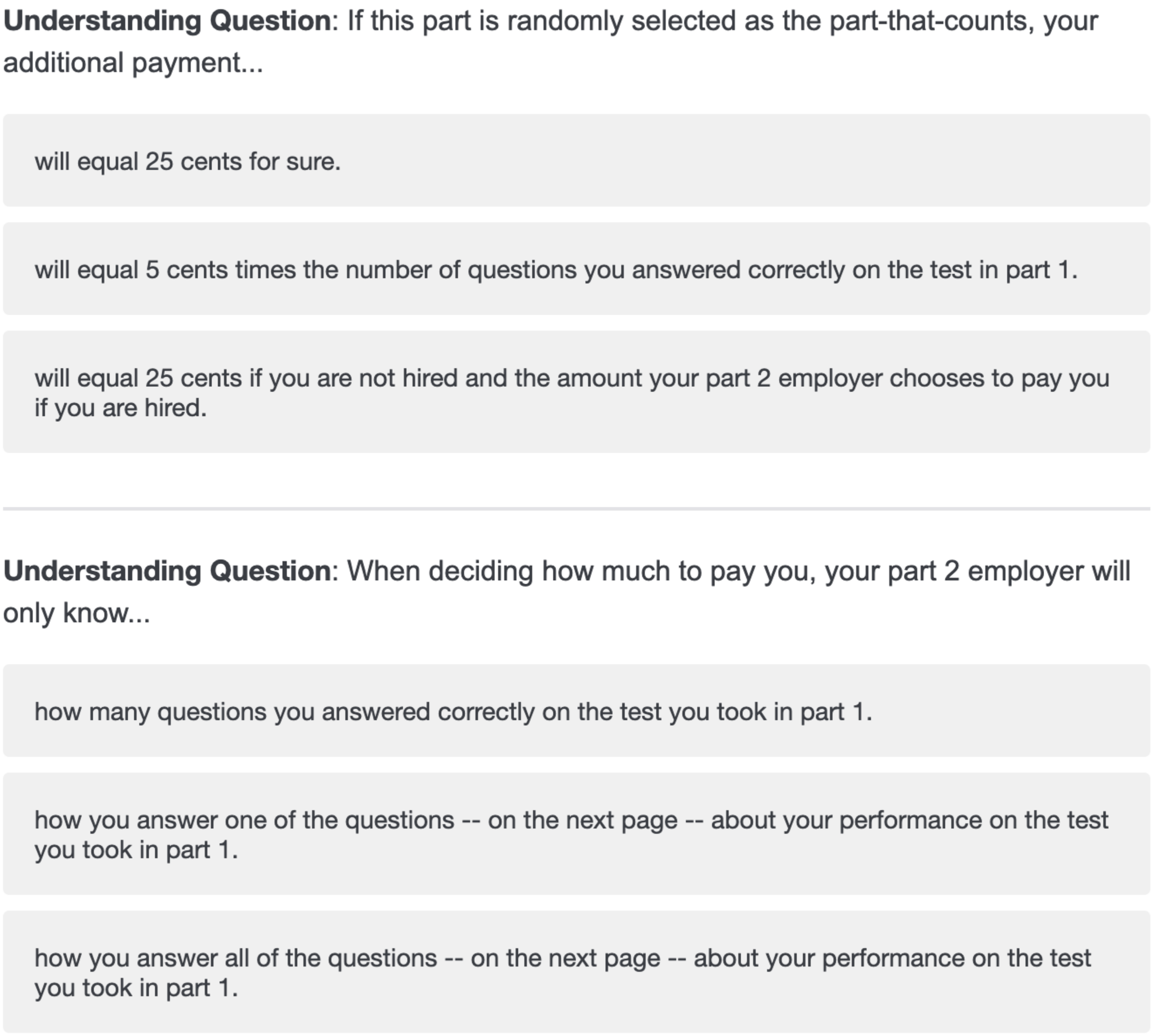
Prior to deciding whether to hire you and, if so, how much to pay you, your part 2 employer will NOT be informed of how many questions you answered correctly on the test in part 1.

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 B.6: Part 2 Understanding Questions



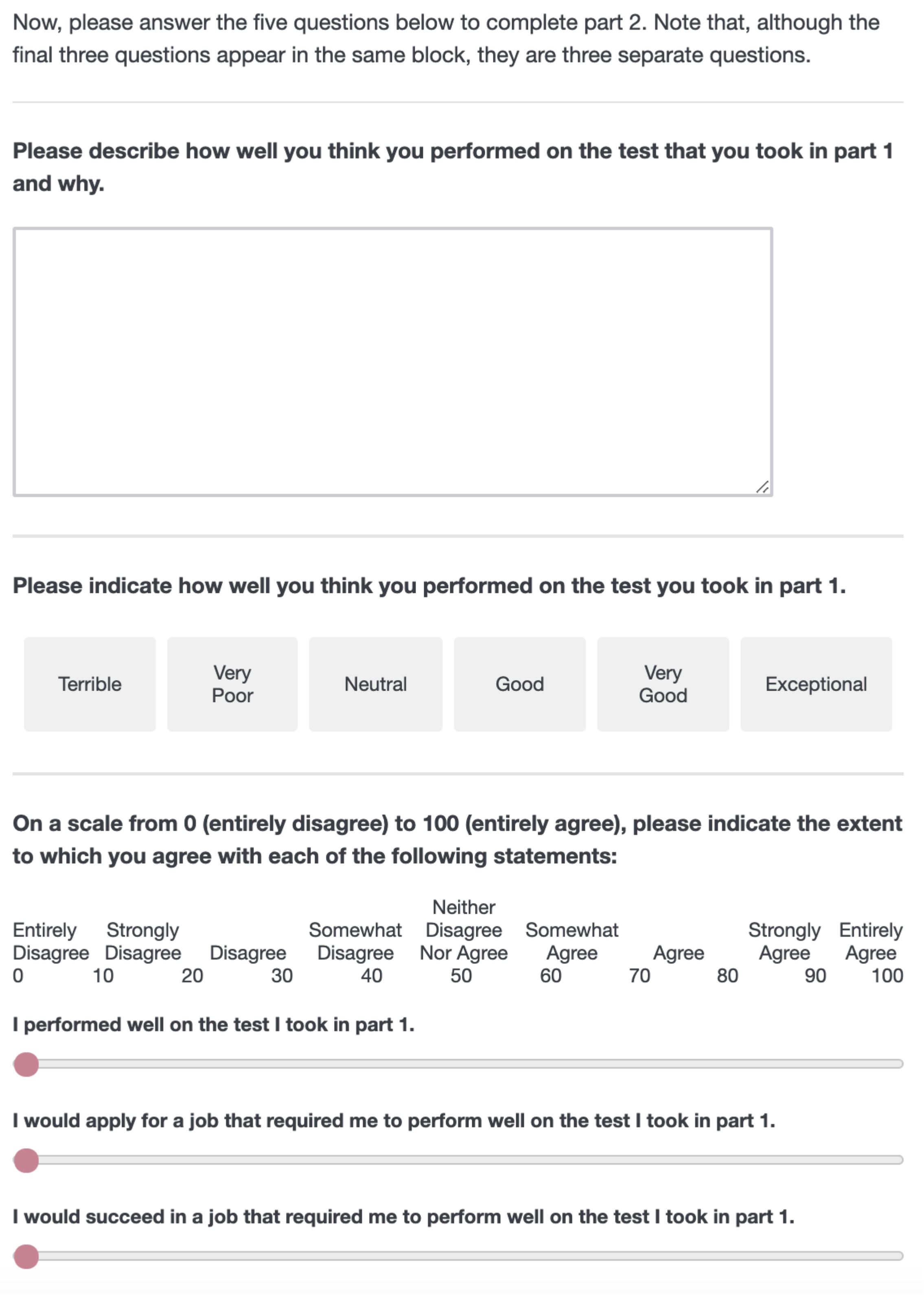
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

- 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 B.7: Part 2 Self-Evaluation Questions



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

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

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

Terrible	Very Poor	Neutral	Good	Very Good	Exceptional

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

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

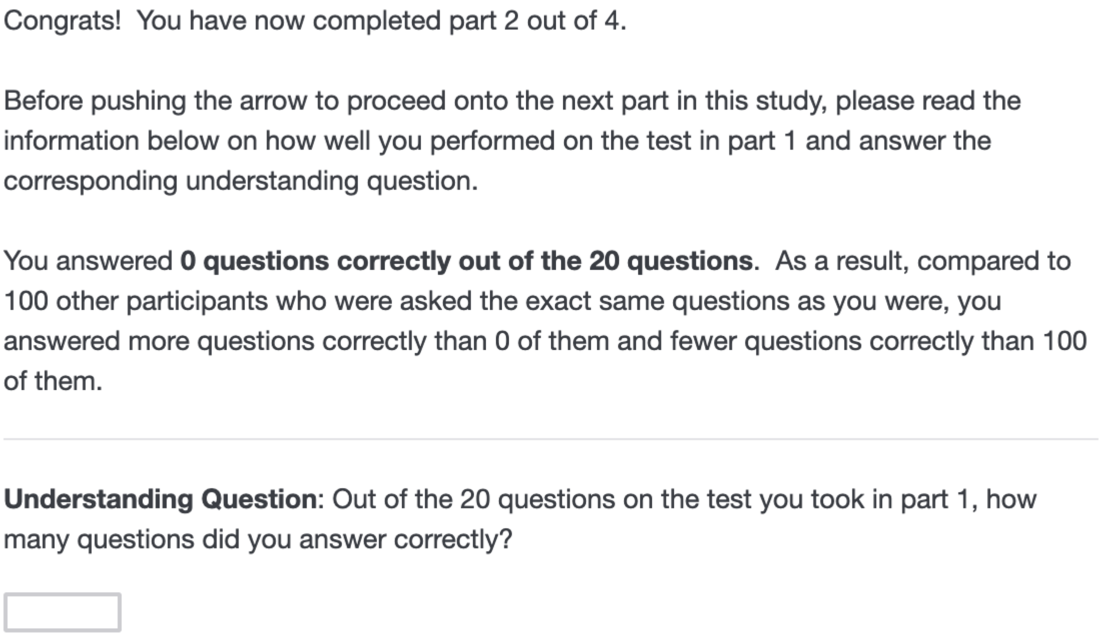
I performed well on the test I took in part 1.

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

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

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

Figure B.8: Absolute and Relative Performance Information



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

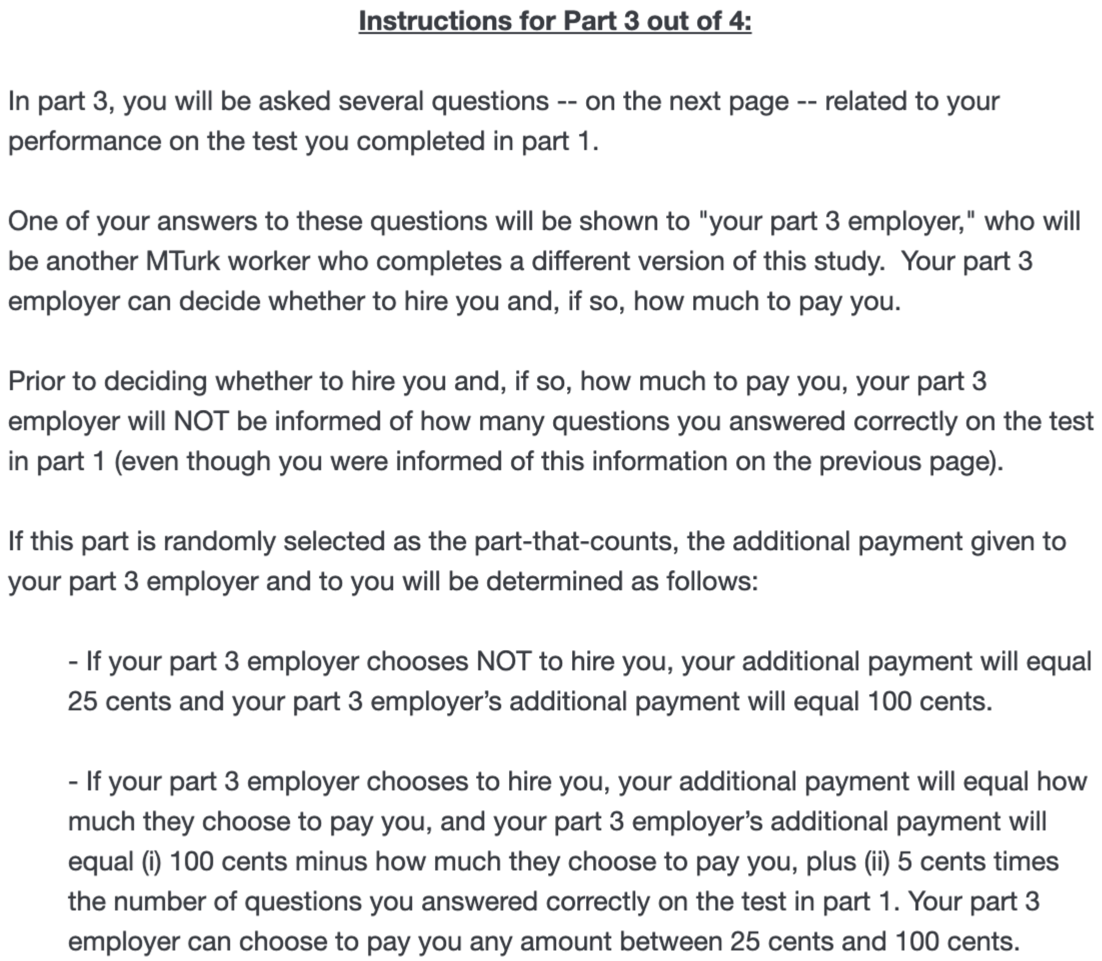
Before pushing the arrow to proceed onto the next part in this study, please read the information below on how well you performed on the test in part 1 and answer the corresponding understanding question.

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 B.9), understanding questions (see Figure B.10), and self-evaluation questions (see Figure B.11) as they were in part 2.

Figure B.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.

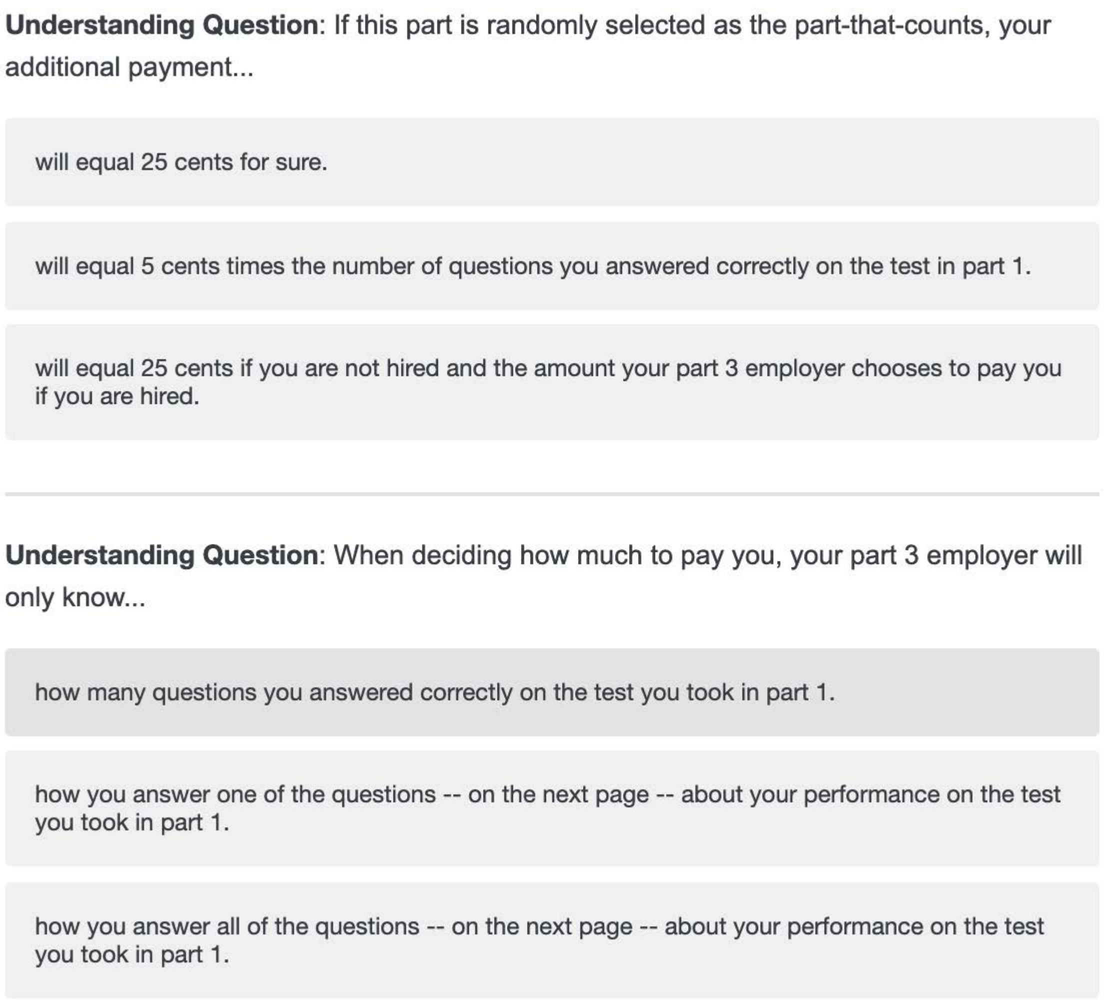
Prior to deciding whether to hire you and, if so, how much to pay you, your part 3 employer will NOT be informed of how many questions you answered correctly on the test in part 1 (even though you were informed of this information on the previous page).

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 B.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.

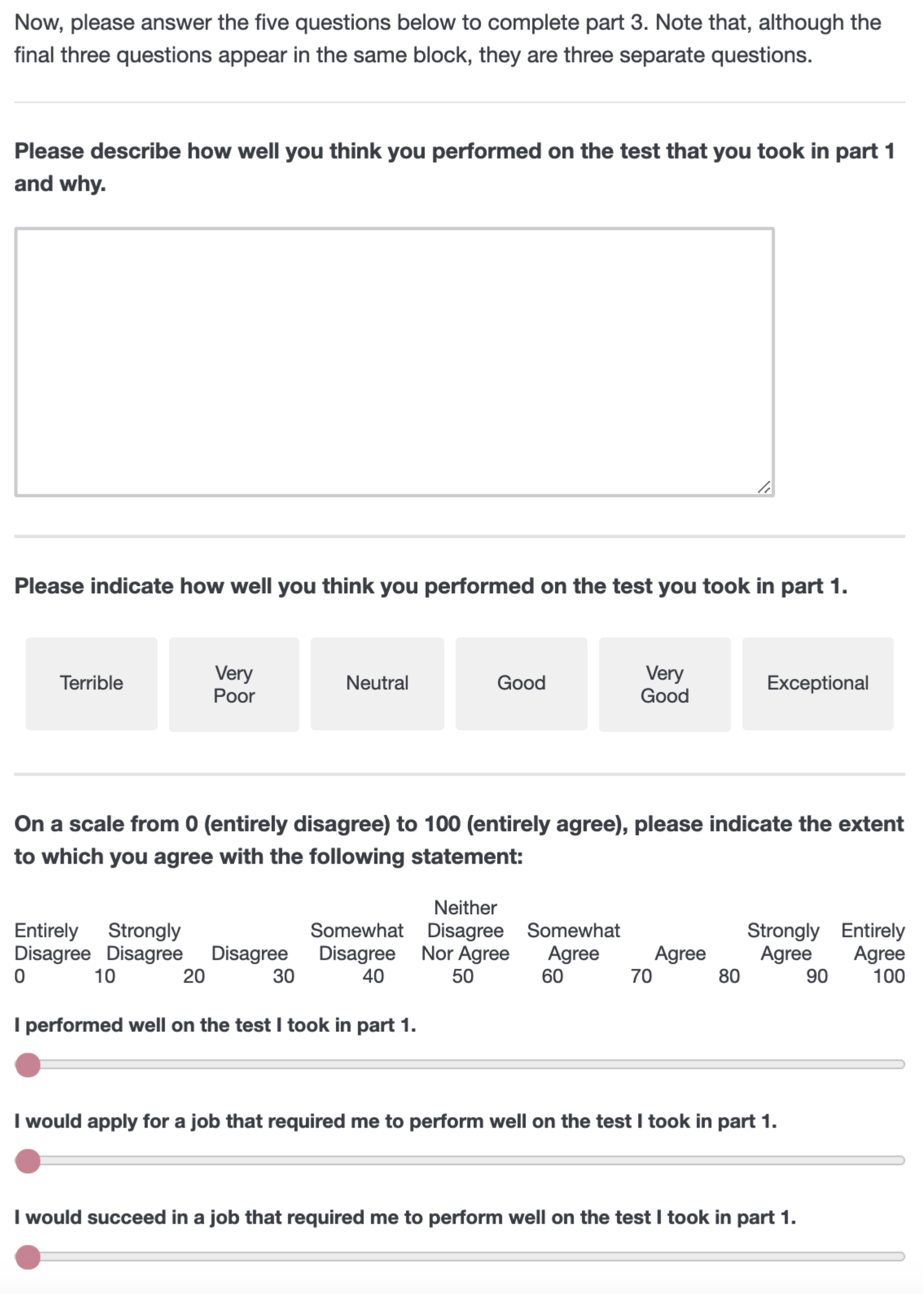
Understanding Question: When deciding how much to pay you, your part 3 employer will only know...

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 B.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:

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

0 10 20 30 40 50 60 70 80 90 100

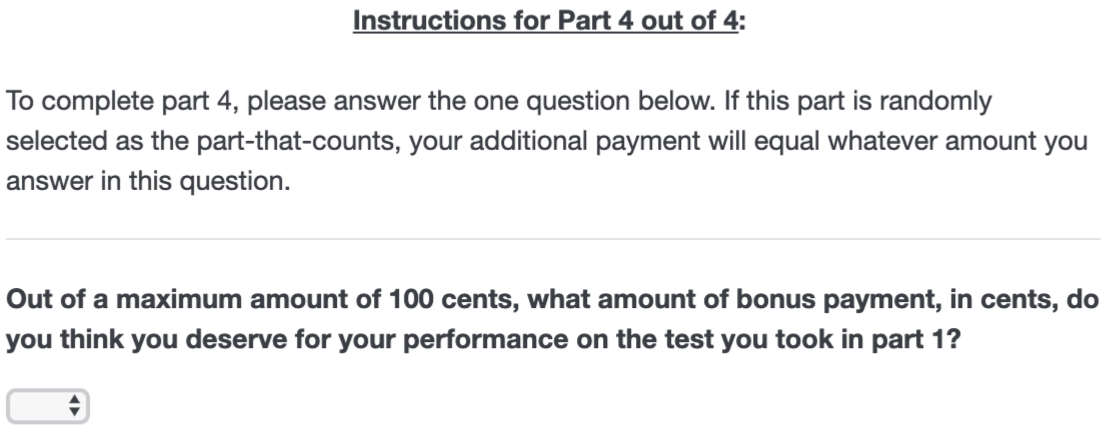
I performed well on the test I took in part 1.

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

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

Finally, participants receive instructions about and are asked to answer the deservingness question in Part 4 (see Figure B.12).

Figure B.12: Part 4 Instructions and Deservingness Question



Instructions for Part 4 out of 4:

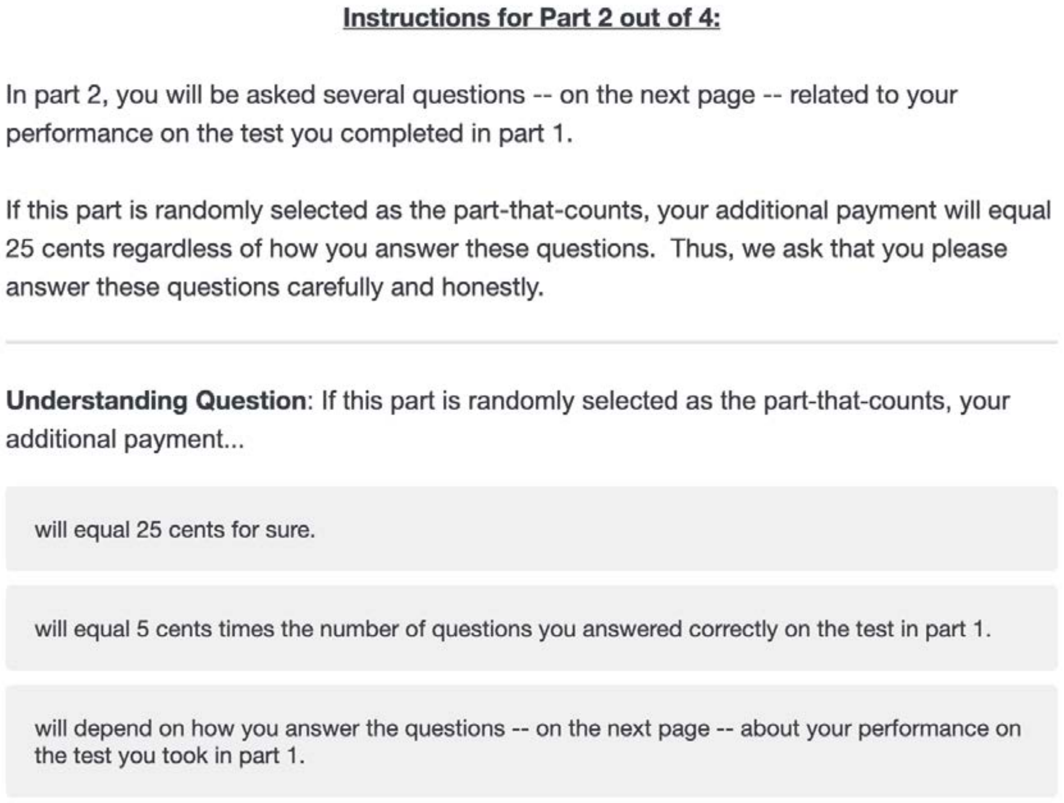
To complete part 4, please answer the one question below. If this part is randomly selected as the part-that-counts, your additional payment will equal whatever amount you answer in this question.

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?

B.2 Instructions for the *Private* version

The *Private* version run in the first wave 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 B.13 for these instructions and the corresponding understanding question. The *Private* versions run in the second and third waves 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 B.14 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 B.13: 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.

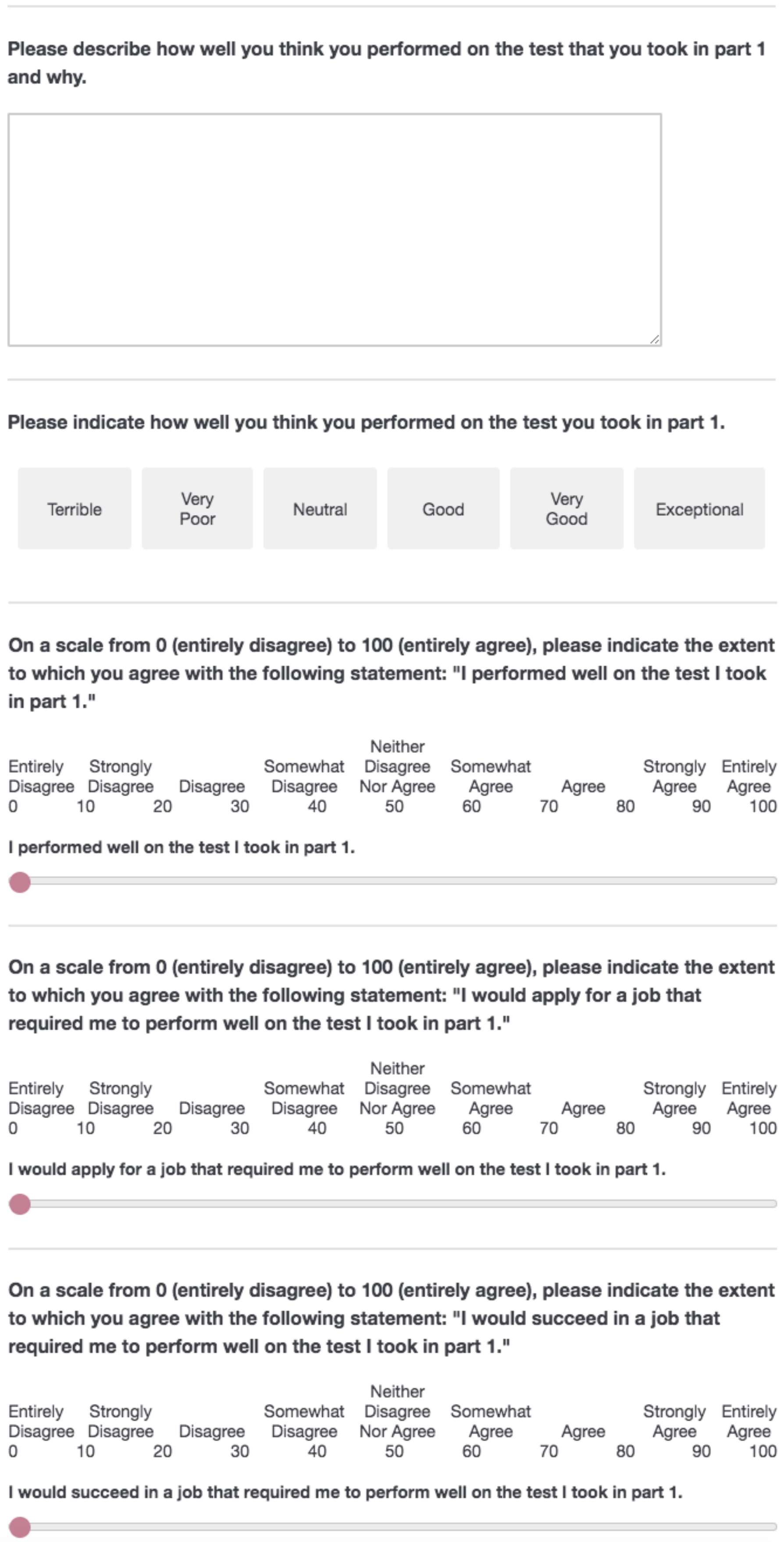
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

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 B.14: 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.

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 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."

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

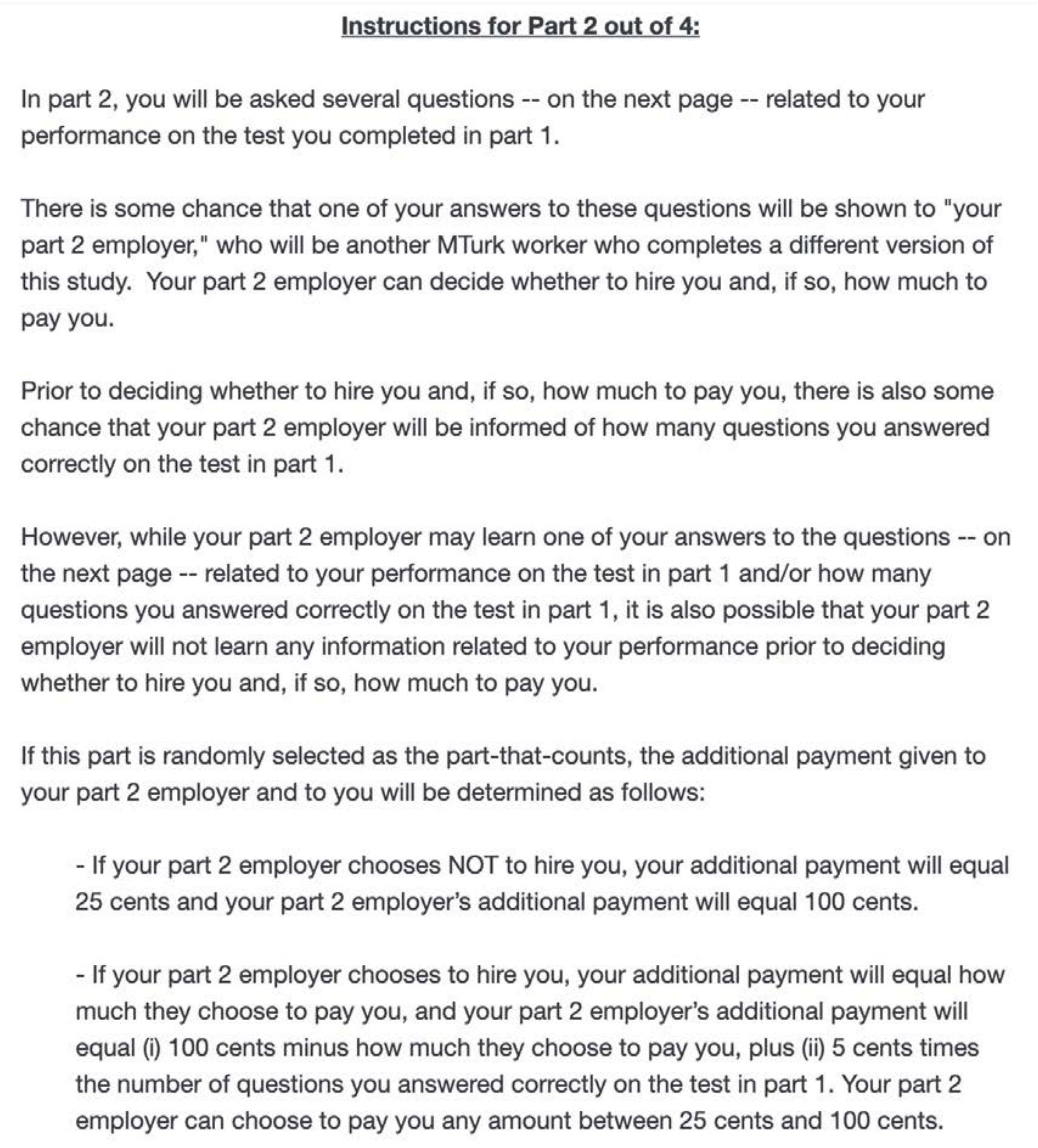
On a scale from 0 (entirely disagree) to 100 (entirely agree), please indicate the extent to which you agree with the following statement: "I would succeed in a job that required me to perform well on the test I took in part 1."

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

B.3 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 B.15 and B.16 for these instructions and the corresponding understanding questions, respectively.

Figure B.15: 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.

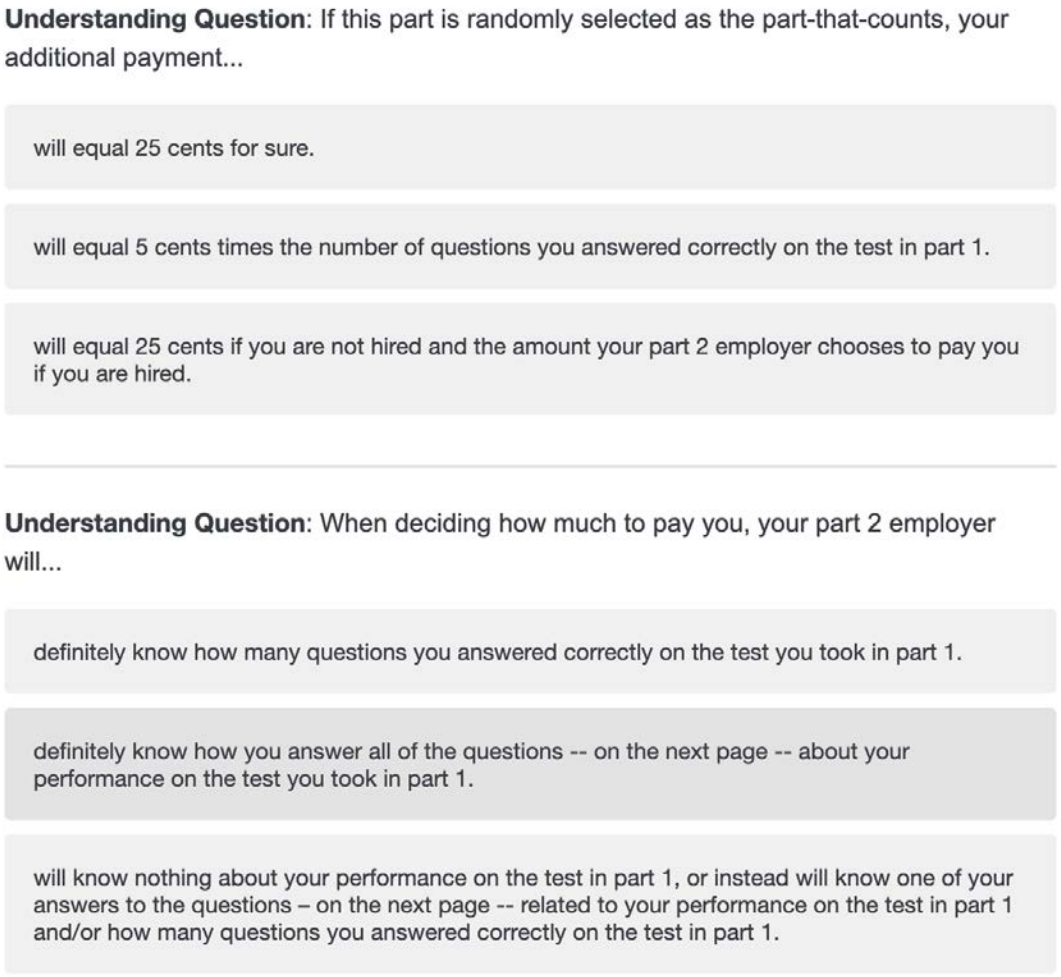
However, while your part 2 employer may learn one of your answers to the questions -- on the next page -- related to your performance on the test in part 1 and/or how many questions you answered correctly on the test in part 1, it is also possible that your part 2 employer will not learn any information 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 B.16: The *Self-Promotion (Risky)* version: Part 2 Questions about Performance



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.

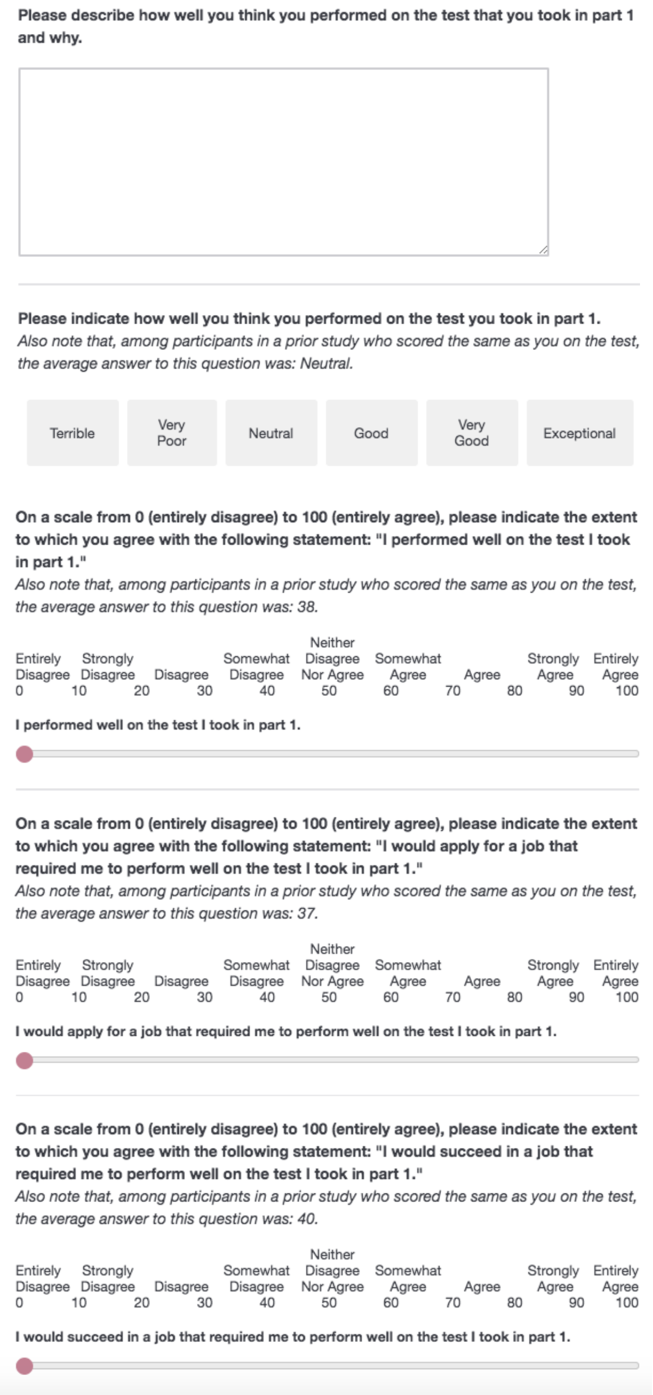
definitely know how you answer all of the questions -- on the next page -- about your performance on the test you took in part 1.

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.

B.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 B.17 for the corresponding screenshot of the part 3 questions.

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

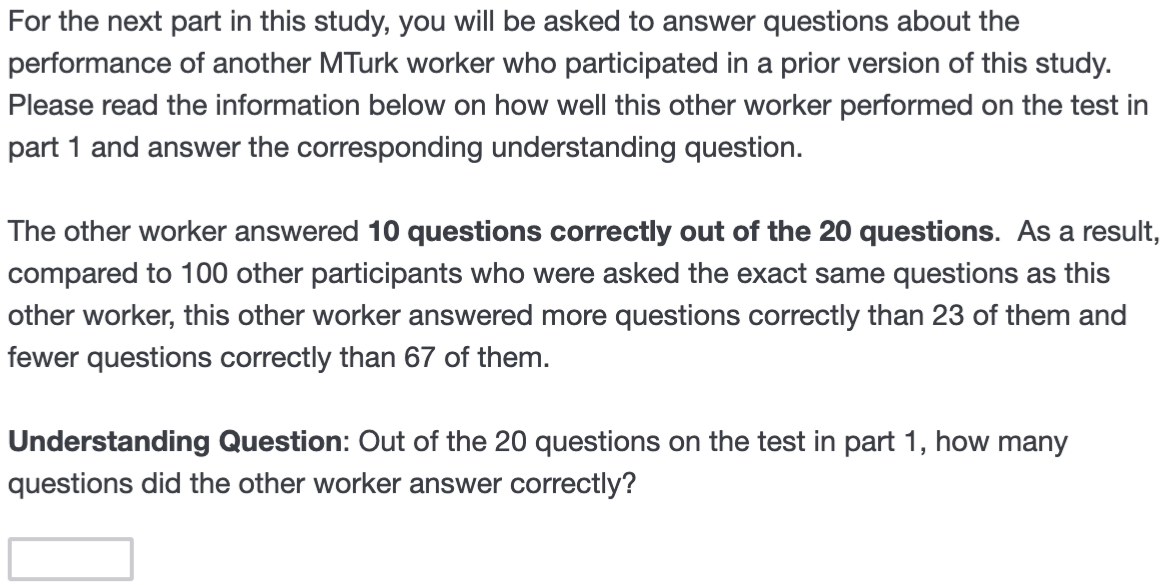
B.5 Instructions for the *Private (Immediately Informed)* version

The *Private (Immediately Informed)* version of the study proceeds in the same manner as the *Private* version of the study, except that uninformed self-evaluations are not elicited. 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, provide informed self-evaluations about that test in part 2, and answer the deservingness question in part 3.

B.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 worker (see Figure B.18) and then are asked to provide informed other-evaluations about this other MTurk worker rather than themselves (see Figures B.19 and B.20).

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

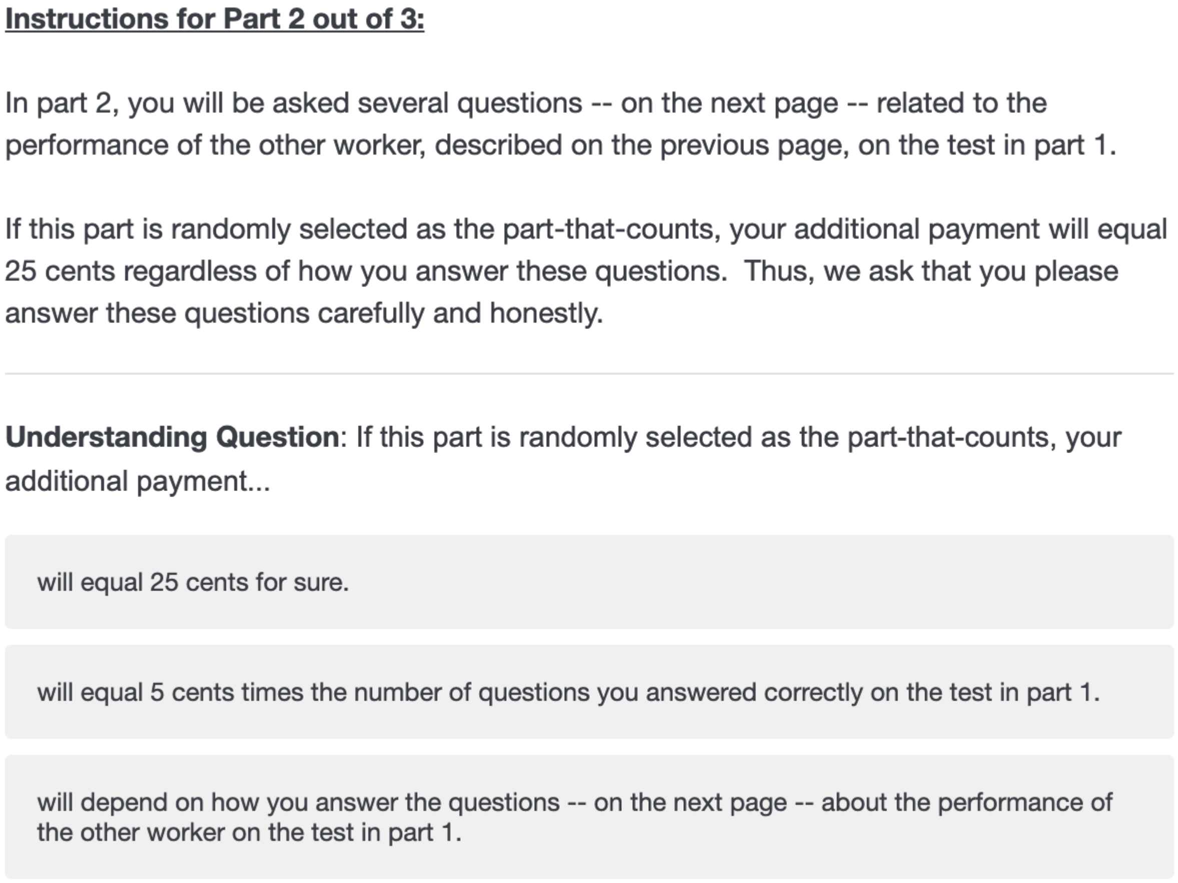


For the next part in this study, you will be asked to answer questions about the performance of another MTurk worker who participated in a prior version of this study. Please read the information below on how well this other worker performed on the test in part 1 and answer the corresponding understanding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

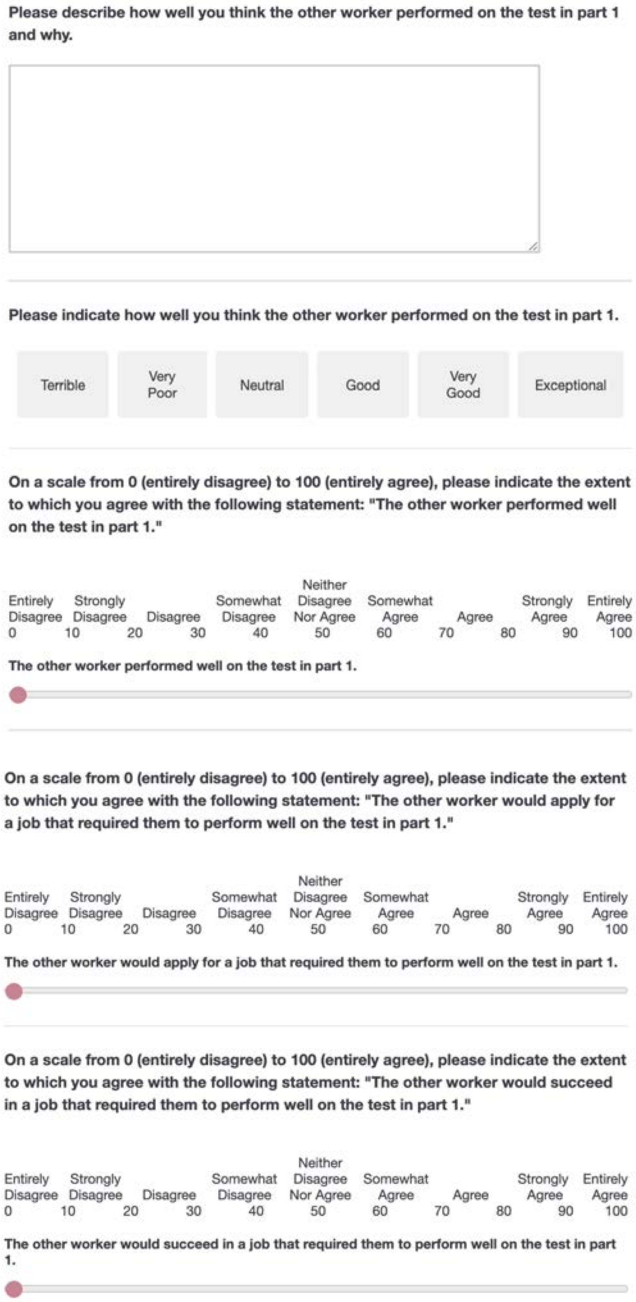
Understanding Question: If this part is randomly selected as the part-that-counts, your additional payment...

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

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."

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."

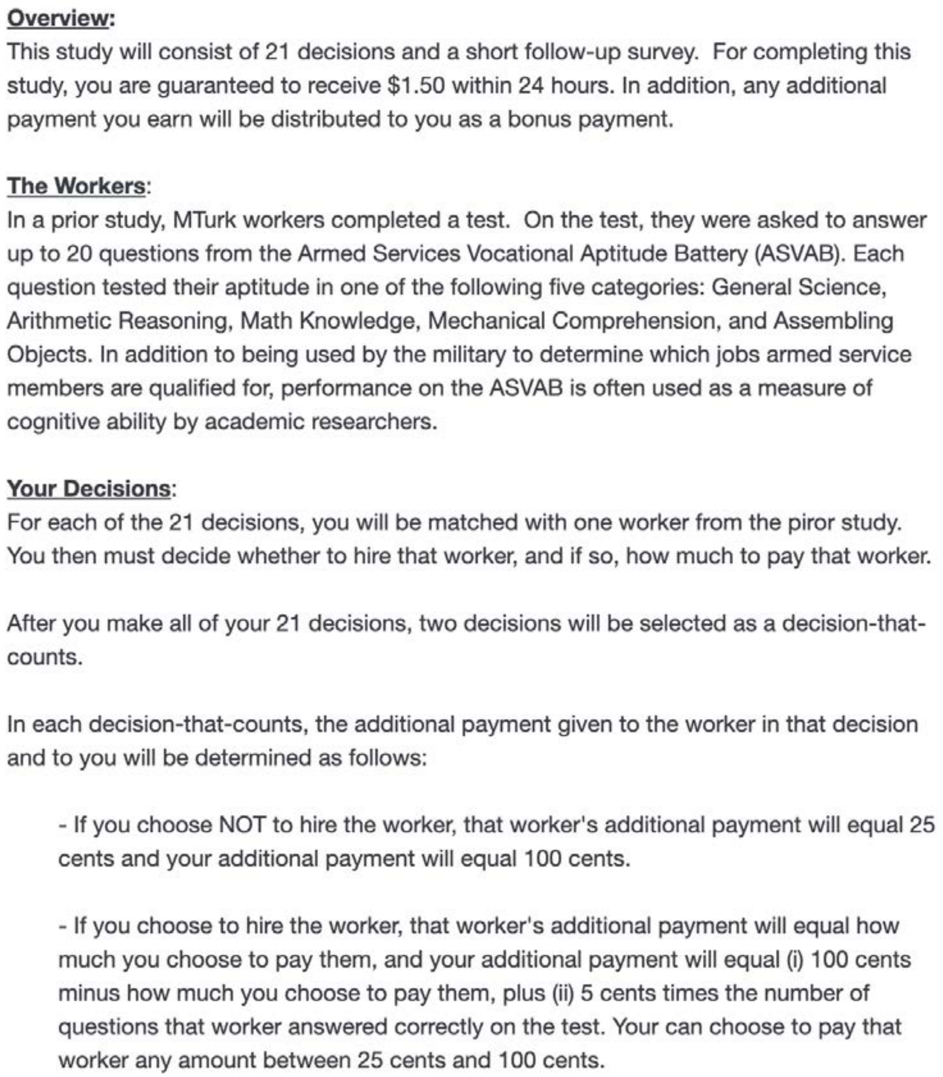
The other worker would succeed in a job that required them to perform well on the test in part 1.

B.7 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 B.21 shows how this payment information is explained. Figure B.22 shows the understanding questions that the participant must answer correctly to proceed.

Figure B.21: 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.

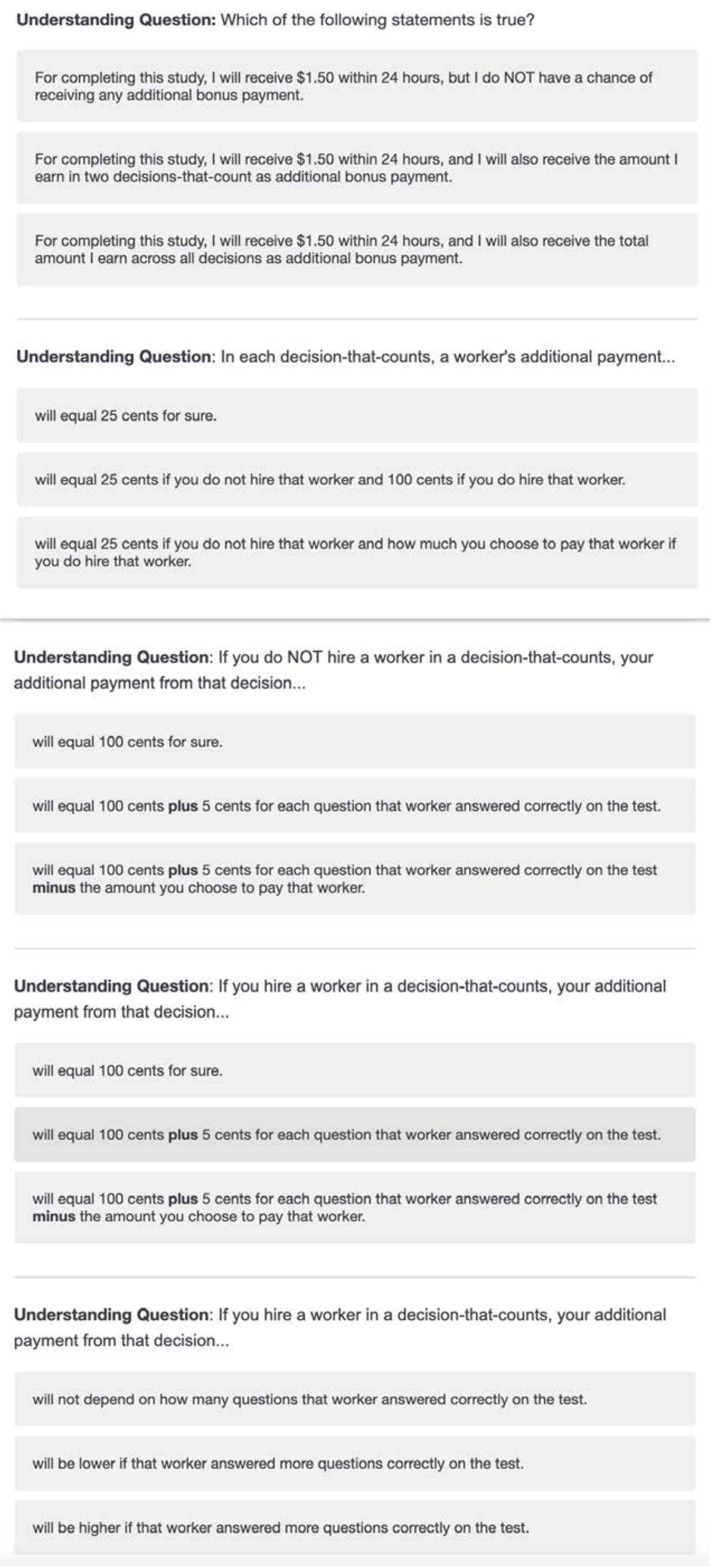
After you make all of your 21 decisions, two decisions will be selected as a decision-that-counts.

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 B.22: 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.

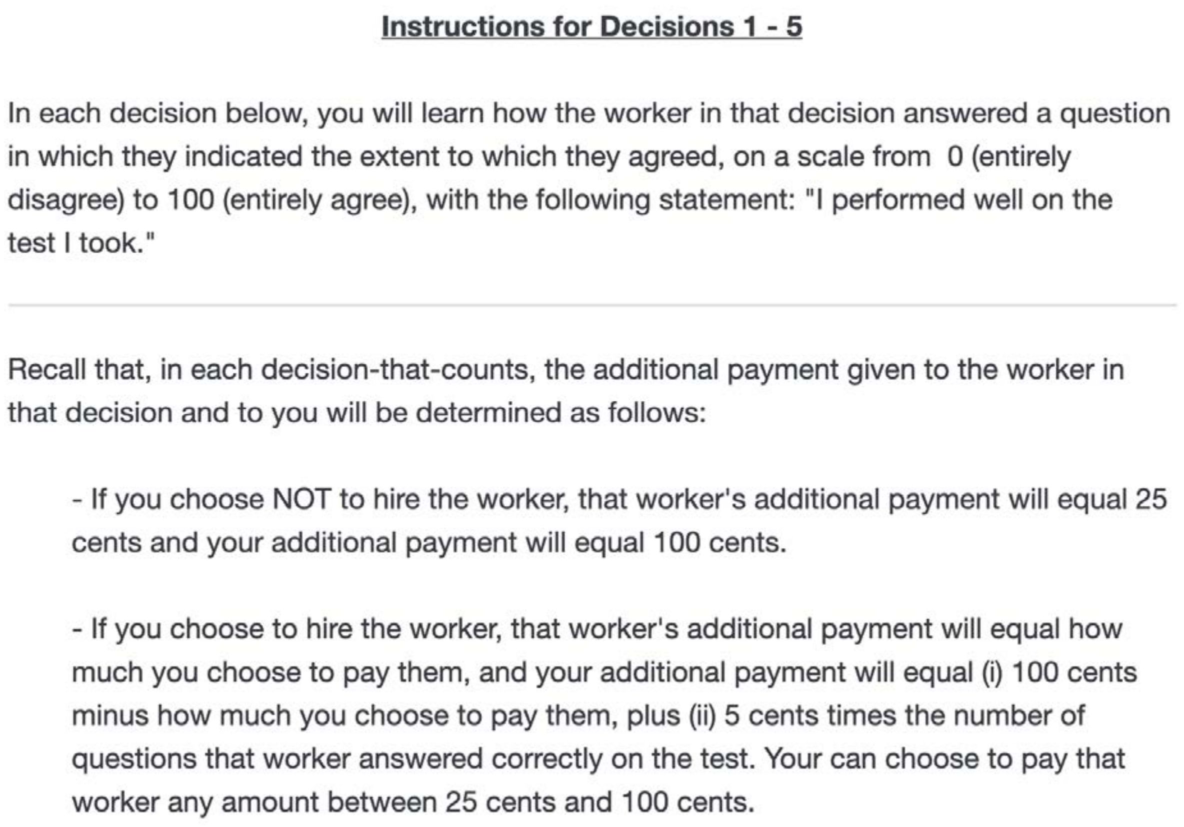
will be lower if that worker answered more questions correctly on the test.

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* evaluation question, the *willingness-to-apply* evaluation question and the *success* evaluation 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* evaluation 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 B.23 and B.24, respectively.

Figure B.23: Instructions for *Performance Evaluation* Decisions



Instructions for Decisions 1 - 5

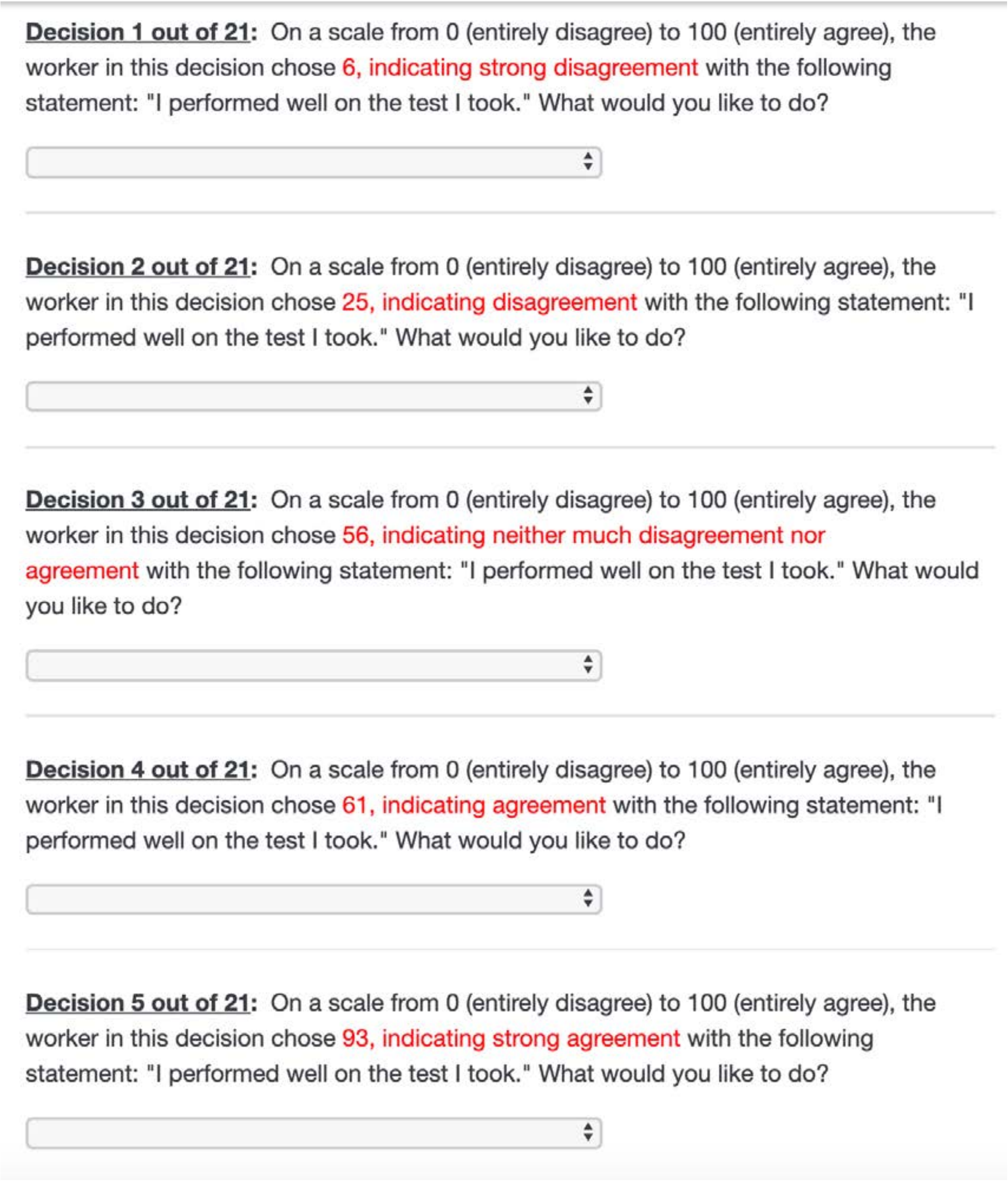
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I performed well on the test I took."

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 B.24: *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?

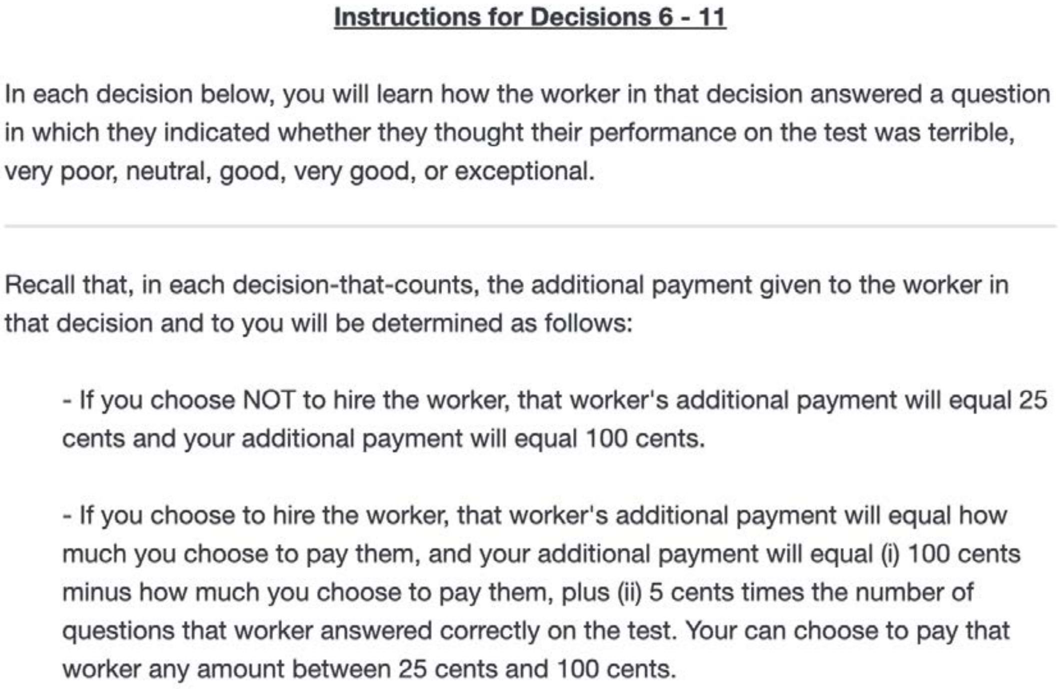
Decision 3 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 56, indicating neither much disagreement nor agreement with the following statement: "I performed well on the test I took." What would you like to do?

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 B.25 and B.26, respectively.

Figure B.25: Instructions for *Performance-Bucket Evaluation* Decisions



Instructions for Decisions 6 - 11

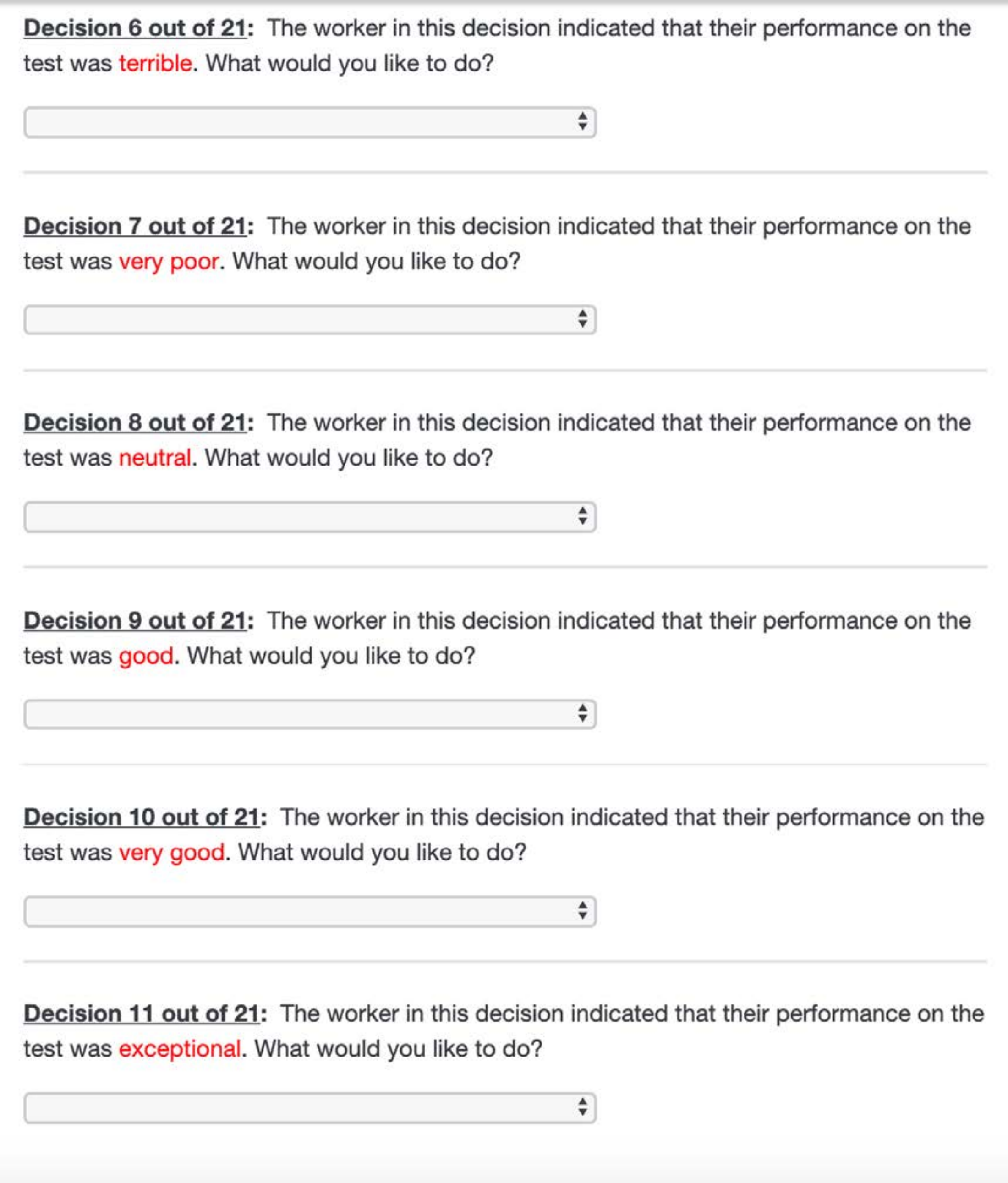
In each decision below, you will learn how the worker in that decision answered a question in which they indicated whether they thought their performance on the test was terrible, very poor, neutral, good, very good, or exceptional.

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 B.26: *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?

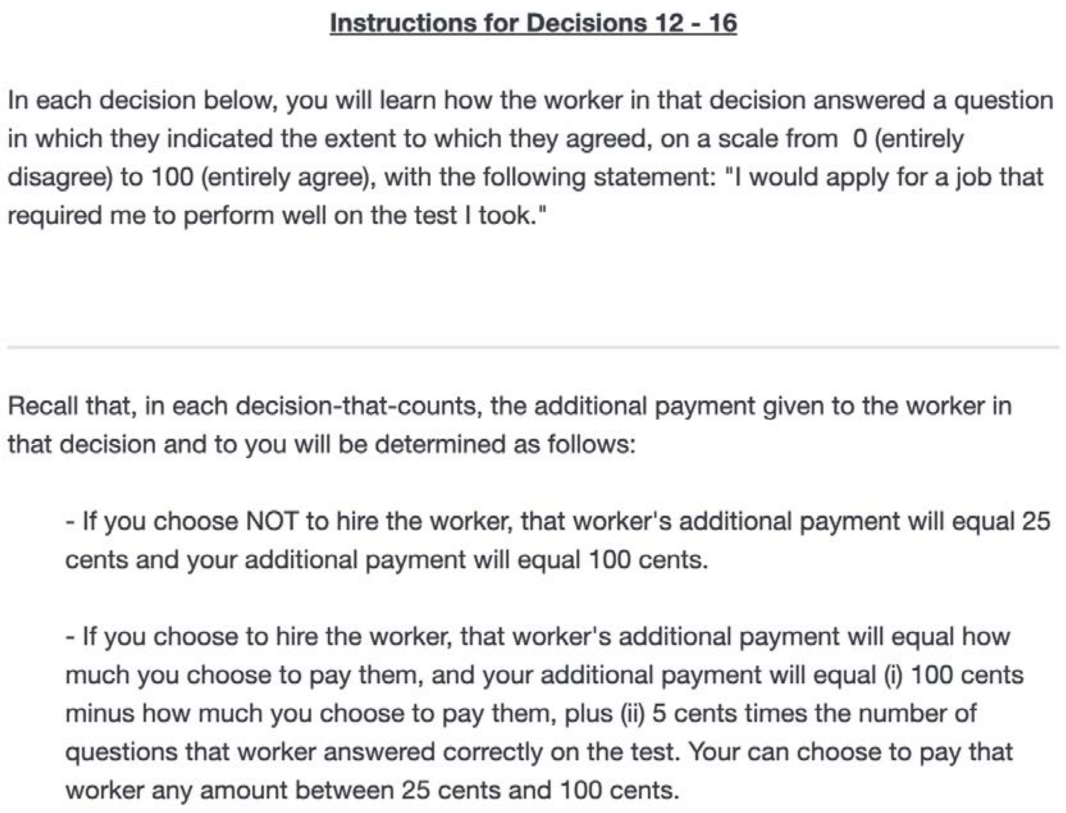
Decision 9 out of 21: The worker in this decision indicated that their performance on the test was good. What would you like to do?

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 B.27 and B.28, respectively.

Figure B.27: Instructions for *Willingness To Apply Evaluation* Decisions



Instructions for Decisions 12 - 16

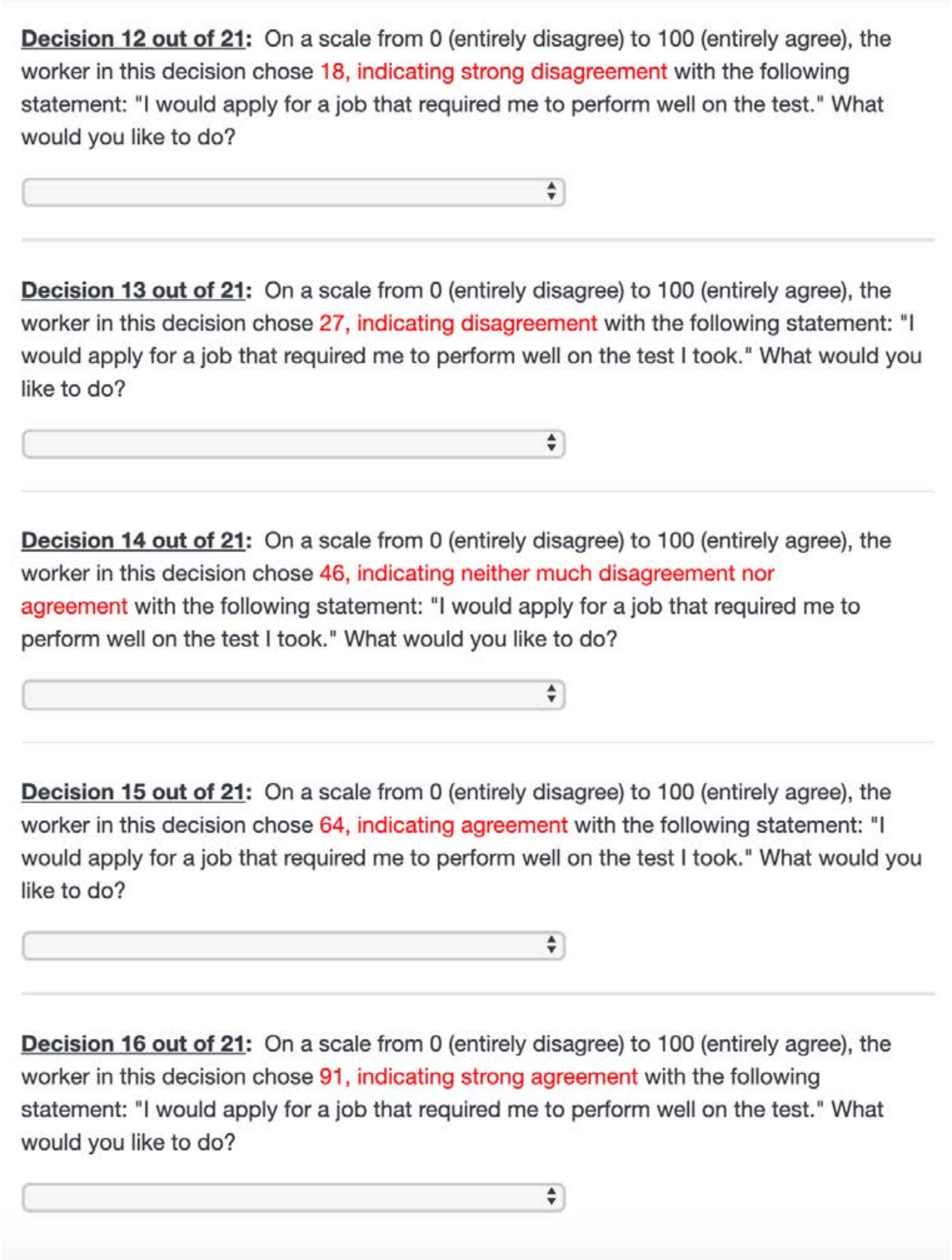
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I would apply for a job that required me to perform well on 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 B.28: *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?

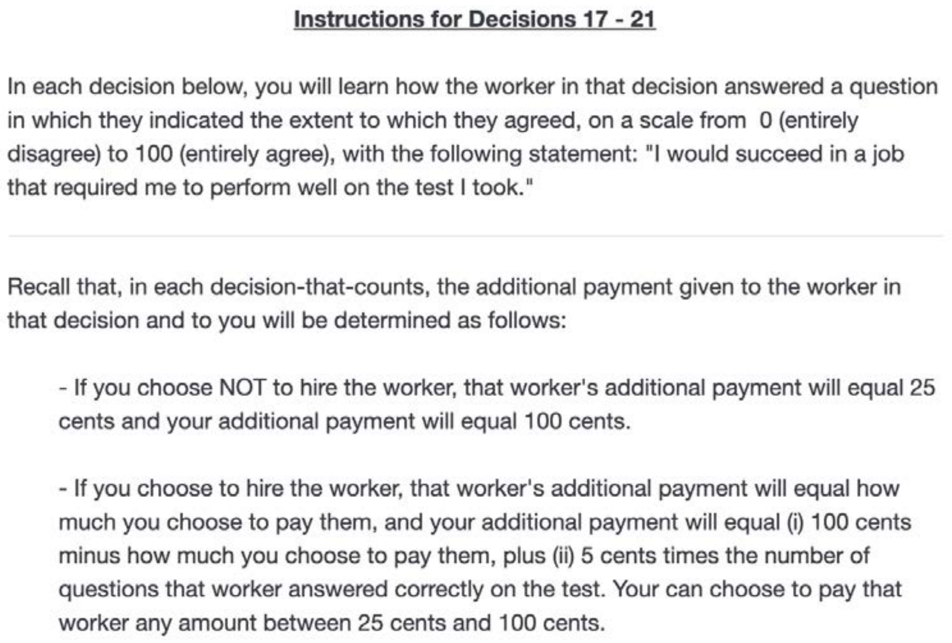
Decision 14 out of 21: On a scale from 0 (entirely disagree) to 100 (entirely agree), the worker in this decision chose 46, indicating neither much disagreement nor agreement with the following statement: "I would apply for a job that required me to perform well on the test I took." What would you like 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 B.29 and B.30, respectively.

Figure B.29: Instructions for *Success Evaluation* Decisions



Instructions for Decisions 17 - 21

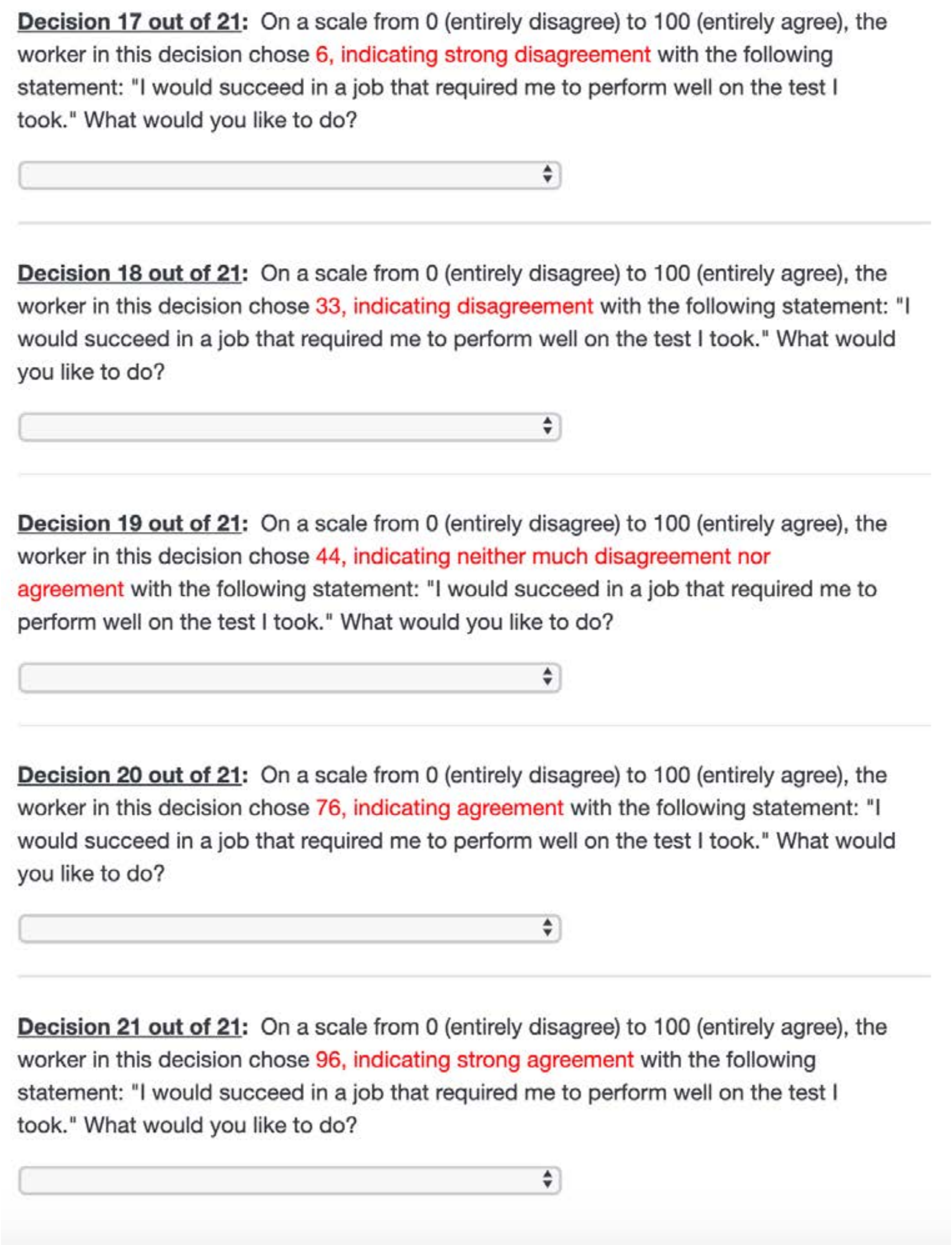
In each decision below, you will learn how the worker in that decision answered a question in which they indicated the extent to which they agreed, on a scale from 0 (entirely disagree) to 100 (entirely agree), with the following statement: "I would succeed in a job that required me to perform well on 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 B.30: *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?