THE GENDER GAP IN SELF-PROMOTION*

CHRISTINE L. EXLEY AND JUDD B. KESSLER

We run a series of experiments involving over 4,000 online participants and over 10,000 school-aged youth. When individuals are asked to subjectively describe their performance on a male-typed task relating to math and science, we find a large gender gap in self-evaluations. This gap arises when self-evaluations are provided to potential employers, and thus measure self-promotion, and when self-evaluations are not driven by incentives to promote. The gender gap in self-evaluations proves to be persistent and arises as early as the sixth grade. No gender gap arises if individuals are asked about their performance on a more female-typed task. *JEL Codes:* C91, D91.

I. INTRODUCTION

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

*We are grateful to the Character Lab Research Network for their partnership on this work. Funding for our research came from Harvard Business School, The Wharton School of the University of Pennsylvania, and the Wharton Behavioral Lab.

The Quarterly Journal of Economics (2022), 1345–1381. https://doi.org/10.1093/qje/qjac003. Advance Access publication on January 21, 2022.

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

When people convey evaluations of their performance, they frequently use subjective terms (e.g., asserting that they are "good" at math) rather than in more precise terms (e.g., asserting that they fall in the 90th percentile according to some observable metric). Thus, it is important to understand how individuals subjectively describe their performance and whether there is a gender gap in subjective descriptions. Indeed, prior work shows that women are less likely to report being "proficient" or "skilled" in programming languages on their résumés (Murciano-Goroff 2021), are less likely to use "positive" words in their titles and abstracts for papers on clinical research (Lerchenmueller, Sorenson, and Jena 2019), and are more likely to use narrow topic-specific—rather than broad—words in their research grant proposals (Kolev, Fuentes-Medel, and Murray 2019).[1]

Research on how people subjectively describe their performance faces three distinct challenges. First, subjective descriptions are often qualitative in nature and hence difficult to measure. Second, comparing the subjective descriptions of equally performing men and women requires observing subjective descriptions about a well-defined performance that can be precisely measured. Third, the ability to examine the underlying drivers of subjective descriptions is limited in settings where one cannot exogenously manipulate the environment.

The contributions of this study stem from our ability, through a carefully controlled experimental setting, to document a gender gap in subjective descriptions of performance—elicited using

1. For work on gender differences in communication and perceptions of that communication, see also Bohren, Imas, and Rosenberg (2018), Grossman et al. (2019), and Manian and Sheth (2021).

self-evaluation questions—among equally performing men and women and to narrow in on the drivers of this gap. Motivated by gender gaps in the labor market, we focus on self-evaluations about performance on a stereotypically male-typed task.

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

Motivated by the possibility that women describe their performance more negatively because they think they had a lower performance either in absolute or relative terms (Lundeberg, Fox, and Punćcohaŕ 1994; Niederle and Vesterlund 2007; Bordalo et al. 2019), we explore whether we also observe a gender gap in informed self-evaluations. Specifically, we investigate whether a gender gap persists when participants are provided with perfect information about their absolute and relative performance on the task. Results suggest that the gender gap in informed self-evaluations is somewhat smaller than the gender gap in (uninformed) self-evaluations, but we still find a substantial and statistically significant gender gap in informed self-evaluations.

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

reflect men, more so than women, strategically inflating their self-evaluations in response to incentives to self-promote.

To provide insight into this question, we investigate whether a gender gap in self-evaluations persists even absent any incentives to self-promote. In particular, we investigate whether a gender gap persists when self-evaluations are elicited privately and not shared with potential employers. Removing self-promotion incentives causes men to provide lower self-evaluations, but it also causes women to provide lower self-evaluations by a nearly identical amount. We thus observe statistically significant gender gaps in privately elicited self-evaluations that are just as large as the gender gaps when self-evaluations are provided to employers, implying that the gender gap in self-promotion reflects an underlying gender gap in self-evaluations even without any incentives to self-promote.

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

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

because of formative experiences—some of these experiences occur quite early in children's lives.

Adding to our understanding of the gender gaps in economic outcomes (Croson and Gneezy 2009; Bertrand 2011; Azmat and Petrongolo 2014; Niederle 2016), this article documents a gender gap in how individuals subjectively describe their performance on a male-typed task and investigates the drivers of this gap. Future work may investigate whether gender differences in subjective views about performance could relate to—and perhaps contribute to—gender differences in other outcomes. Akin to the role of confidence—as measured by absolute or relative performance—in helping explain the gender gaps in the the willingness to compete (Niederle and Vesterlund 2007; van Veldhuizen 2017), speak up (Coffman 2014), be a leader (Born, Ranehill, and Sandberg 2018), and claim credit (Isaksson, 2018), subjective assessments of one's own performance may cause women to not feel "good enough" to enter a competition or negotiation, apply for a job, or assert their expertise in stereotypically male domains. Such an explanation would correspond with prior work finding that female engineers ask for lower salaries, unless provided with information on the median salary requested (Roussille 2021), and that women are deterred from applying to jobs that subjectively describe the requisite management, analytical, computer, or technology skills (Coffman, Collis, and Kulkarni 2022; Abraham and Stein 2020).

II. Design, Data Collection, and Setting

We recruited 3,892 participants from online labor markets—Amazon's Mechanical Turk (MTurk) and Prolific—to participate in one of seven versions of our study across five waves of data collection, shown in the first five rows of Table I.[2] Each participant was guaranteed a completion fee plus a possible bonus payment from one randomly selected part of the study.[3] After participants

2. To be eligible for any study version, participants must have previously completed at least 100 tasks (on MTurk or Prolific) with a 95% or better approval rating and must be working from a U.S. IP address. The median age is 33 years old, the median educational attainment is a bachelor's degree, and the percentage of male participants is 59%. While participants were required to correctly answer understanding questions at various points to proceed in the study, no participants were excluded from our data analysis.

3. In all of our studies run on MTurk (i.e., data collected in waves 1–4), participants received a $2 completion fee for a 20-minute study. In our studies run

TABLE I
STUDY VERSIONS BY WAVE

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

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

completed all parts of the study, they took a short follow-up survey that collected demographic information, including gender. Gender was not mentioned prior to the follow-up survey, so participants were not primed to think about gender when answering self-evaluation questions.

Why did five waves of data collection occur? We collected data over five waves because of the persistence of the gender gap across study versions and because of our desire to test the boundaries of this gap. In the first wave, we randomly assigned workers to either the *Self-Promotion* version, the *Self-Promotion (Risky)* version, or the *Private* version. These study versions allowed us to test two potential drivers of gender differences in self-evaluations that we expected, given prior literature.

First, motivated by the vast literature on gender gaps in beliefs about performance and how such gaps contribute to gender gaps in behavior, we hypothesized that differences in beliefs

on Prolific (i.e., data collected in wave 5), participants received a $4 completion fee for a 25-minute study.

about performance could contribute to gender differences in self-evaluations. As further explained below, the three study versions in wave 1 allows us to examine the role of performance beliefs by comparing self-evaluations before and after participants are provided with perfect information about their absolute and relative performance on the task that their self-evaluations describe.

Second, motivated by the gender gaps in the labor market and prior literature on gender differences in reported beliefs about performance that arise in strategic contexts (Reuben, Sapienza, and Zingales 2014; Charness, Rustichini, and Van de Ven 2018), we hypothesized that incentives to inflate self-evaluations that would be shared with potential employers could contribute to gender differences in self-evaluations. The study versions in wave 1 allow us to test whether this is the case because the *Self-Promotion* and *Self-Promotion (Risky)* versions involve differing incentives to inflate self-evaluations, whereas the *Private* version removes all incentives to self-promote. Specifically, while self-evaluations may be communicated to potential employers in the *Self-Promotion* and *Self-Promotion (Risky)* versions, self-evaluations are privately elicited and not shared with potential employers in the *Private* version.

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

As will be discussed in what follows, our data collection and continual replication of our earlier findings—across time and across labor market platforms—highlights the robustness of our results. In addition, as noted in the final row of Table I, an additional 10,637 youth participated in a modified *Private* version of our study designed to explore the origins of the gender gap in self-evaluations. The design of this version, and the associated results, are discussed in Section V.

In addition to the data described already, 298 participants completed a version of our study as "employers," who are relevant for the *Self-Promotion* and *Self-Promotion (Risky)* versions of our

study.[4] As discussed in Section IV.C, results from the employers demonstrate that self-promotion pays. Participants who report higher self-evaluations in the *Self-Promotion* and *Self-Promotion (Risky)* versions of our study are paid more by employers.

II.A. The Self-Promotion Version

The *Self-Promotion* version of our study proceeds as follows: each participant completes a math and science test, provides their beliefs about their absolute performance on that test, provides responses to self-evaluation questions about their test performance, is informed of their absolute and relative test performance, provides informed responses to self-evaluation questions about their test performance, and then answers questions that elicit control and demographic information, including gender. More specifically, the *Self-Promotion* version has four parts, described in order below. See Online Appendix D.1 for screenshots and additional details.

1. *Part 1: Performance and Performance Beliefs.* In part 1 of the study, participants are asked to take a test comprised of 20 multiple choice questions. They have up to 30 seconds to answer each question. Given the gender gaps that motivate our study and the fact that gender gaps are often more prevalent in stereotypically male-typed tasks, we selected questions that related to math and science. Specifically, we selected four questions each from the following five categories on the Armed Services Vocational Aptitude Battery (ASVAB): General Science, Arithmetic Reasoning, Math Knowledge, Mechanical Comprehension, and Assembling Objects. By selecting questions from the ASVAB, we are also able to follow prior literature that uses performance on the ASVAB as a measure of cognitive ability (Frey and Detterman 2004) and to convey to participants why performance on questions like the ones they are answering are often informative.

4. In addition to the participants described in the main text, we use performance data from 200 participants to create reference groups to provide participants with information on relative performance (100 participants who completed the math and science test and 100 participants who completed the verbal test). We also analyze data from 600 MTurk workers who evaluated free-response comments generated by participants as described below and discussed in Online Appendix B. Including these 800 participants and the 298 employers described in the main text, this article involves a total of 4,990 study participants from online labor markets.

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

As a measure of beliefs about their absolute performance, after participants complete the 20 ASVAB questions, they are asked: "Out of the 20 questions on the test you took in part 1, how many questions do you think you answered correctly?" Participants can select any number from 0 to 20, and their answers are not incentivized. Their answers are not incentivized because we control for beliefs about absolute (and relative) performance by design, as described later, mitigating concerns about noise in this measure.

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

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

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

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

To encourage participants to reflect on their performance, the first question in part 2 is a free-response question that states: "Please describe how well you think you performed on the test that you took in part 1 and why." The remaining four are the quantitative self-evaluation questions that we analyze for the remainder of this article.[6]

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

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

The answers to these four self-evaluation questions allow us to quantify—on a 1–6 scale for the performance bucket question and on a 0–100 scale for the three other questions—how

6. One could also imagine analyzing responses to the free-response question. Analyzing responses to this question is fraught, however, as the text is hard to evaluate and can convey additional information that makes measuring the "positivity" of the response difficult. Nevertheless, we attempt to learn what we can from this data by having a total of 600 MTurk participants evaluate the free responses from wave 1. We summarize those findings in Online Appendix B.

participants subjectively describe their performance to a potential employer.

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

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

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

II.B. The Self-Promotion (Risky) Version

To explore the robustness of the gender gap in self-promotion, we ran the *Self-Promotion (Risky)* version. The *Self-Promotion (Risky)* version proceeds exactly as the *Self-Promotion* version except that participants are told that there is some chance that their

employers will learn their actual performance (i.e., be informed of how many questions they answered correctly on the test) along with one of their answers to a self-evaluation question.[7] See Online Appendix D.2 for screenshots and additional details.

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

II.C. The Private Version

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

Given the lack of employers, the *Private* version eliminates the relevance of strategic incentives to provide more favorable responses to self-evaluation questions to achieve higher financial returns. Put differently, it eliminates the incentives to self-promote that were present in the *Self-Promotion* version. In addition, in the *Private* version, gender differences in response to self-evaluation questions cannot be driven by potential gender differences in risk aversion, gender differences arising from lack of control over payoffs, or gender differences in preferences toward employers (e.g., caring about employers' earnings).

II.D. The Private (Social Norms) Version

The *Private (Social Norms)* version proceeds exactly as the *Private* version except that participants are provided with additional information when providing responses in part 3 (i.e., after they receive performance information). In particular, each of the four self-evaluation questions now includes a message that reads:

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

"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 Online Appendix D.4 for screenshots and additional details.

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

II.E. The Private (Immediately Informed) Version

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

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

II.F. The Private (Other-Evaluation) Version

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

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

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

II.G. The Private (Verbal) Version

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

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

II.H. Our Study Environment

In this section, we present data on performance on the math and science test and on the beliefs that participants report about their absolute performance (i.e., how many questions they think they answered correctly on the test). Because our results are very similar across study versions, and because participants always take the test and report beliefs about their absolute performance before encountering any version-specific variation, we pool across all study versions from waves 1–5 in which participants take the math and science test (i.e., all versions except the *Private (Verbal)* version). We find results consistent with our setting being "male-typed" in that women think they answered significantly fewer questions correctly than equally performing men.

Figure I, Panel A shows CDFs of the number of test questions answered correctly by male participants and by female participants. On average, women answer 9.94 questions correctly and men answer 9.34 questions correctly. The mean difference is statistically significant ($p < .01$) and the distributions are statistically significantly different (a Kolmogorov-Smirnov test yields $p < .01$).

Despite women performing better than men, Figure I, Panels B and C show that women believe they performed worse on the

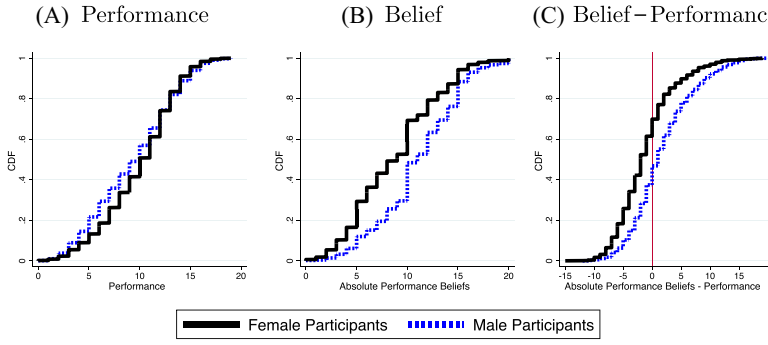

FIGURE I

Performance and Absolute Performance Belief Distributions

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

test than men. Panel B shows raw beliefs about performance. On average, men believe they answered 11.05 questions correctly and women believe they answered only 8.77 questions correctly. The mean difference is statistically significant ($p < .01$), and the distributions are statistically significantly different (a Kolmogorov-Smirnov test yields $p < .01$). Panel C shows the difference between beliefs about performance and actual performance. Again, the mean difference is statistically significant ($p < .01$), and the distributions are statistically significantly different (a Kolmogorov-Smirnov test yields $p < .01$). Looking at where the CDFs cross zero, we see that the gender gap in beliefs about performance is driven by the majority of women underestimating their performance and the majority of men overestimating their performance.

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

and actual performance (column (4)). In the latter three columns, the coefficient on *Female* is negative, large, and statistically significant.

These results highlight that women believe they answered fewer questions correctly than equally performing men. We consider the role of such beliefs in the gender gap in self-evaluations that we observe. As noted in Section II.A, rather than using these reported beliefs as statistical controls, we instead control for beliefs by design.

III. Results

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

The first two subsections document persistent gender gaps in self-evaluations when participants are asked about their performance on the math and science test. Focusing on results from the *Self-Promotion* and *Self-Promotion (Risky)* versions, Section III.A documents a large gender gap in self-evaluations that are provided to potential employers, which we refer to as the gender gap in self-promotion. Focusing on results from the *Private*, *Private (Social Norms)*, and *Private (Immediately Informed)* study versions, Section III.B documents a large gender gap in self-evaluations even without any incentives to self-promote.

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

III.A. The Gender Gap in Self-Promotion on a Math and Science Task

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

TABLE II

RESULTS FROM EVALUATIONS (BEFORE PERFORMANCE INFORMATION IS PROVIDED)

Question:	Performance (1)	Performance bucket (2)	Willingness to apply (3)	Success (4)
Panel A: Self-Promotion version, wave 1 ($N = 302$)				
Female	−12.68***	−0.59***	−15.31***	−15.09***
	(2.96)	(0.13)	(3.46)	(3.46)
Panel B: Self-Promotion (Risky) version, wave 1 ($N = 294$)				
Female	−9.15***	−0.47***	−12.82***	−9.24***
	(2.93)	(0.13)	(3.29)	(3.32)
Panel C: Private version, wave 1 ($N = 304$)				
Female	−13.46***	−0.56***	−17.57***	−16.46***
	(2.93)	(0.13)	(3.51)	(3.61)
Panel D: Private version, wave 2 ($N = 302$)				
Female	−12.21***	−0.55***	−17.25***	−14.39***
	(3.18)	(0.15)	(3.54)	(3.53)
Panel E: Private (Social Norms) version, wave 2 ($N = 298$)				
Female	−15.14***	−0.80***	−16.93***	−15.62***
	(3.28)	(0.16)	(3.71)	(3.71)
Panel F: Private version, wave 3 ($N = 300$)				
Female	−16.45***	−0.79***	−15.69***	−16.16***
	(3.18)	(0.15)	(3.92)	(3.87)
Panel G: Private (Immediately Informed) version, wave 3: no evaluations				
Panel H: Private (Immediately Informed) version, wave 4: no evaluations				
Panel I: Private (Other-Evaluation) version, wave 4: no evaluations				
Panel J: Private version, wave 5 ($N = 294$)				
Female	−13.05***	−0.59***	−18.77***	−19.18***
	(2.61)	(0.11)	(3.30)	(3.17)
Panel K: Private (Verbal) version, wave 5 ($N = 305$)				
Female	1.15	−0.12	1.99	−0.36
	(2.40)	(0.11)	(3.19)	(3.02)
Panel L: All evaluations of own math and science performance ($N = 2{,}094$)				
Female	−13.83***	−0.67***	−17.28***	−16.12***
	(1.13)	(0.05)	(1.31)	(1.32)
Performance fixed effects	Yes	Yes	Yes	Yes

Notes. * $p < .10$, ** $p < .05$, *** $p < .01$. Standard errors are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each column before participants are informed of their absolute and relative performance. Responses to the performance question indicate the extent of each participant's agreement (0–100) with the following statement: "I performed well on the test I took in part 1." Responses to the performance bucket question indicate which Likert-scale response (coded from 1 for the lowest to 6 for the highest) a participant selects when asked to "indicate how well you think you performed on the test in part 1?" Responses to the willingness to apply question indicate the extent of each participant's agreement (0–100) with the following statement: "I would apply for a job that required me to perform well on the test I took in part 1." Responses to the success question indicate the extent of each participant's agreement (0–100) with the following statement: "I would succeed in a job that required me to perform well on the test I took in part 1." *Female* is an indicator for the participant being female. Performance fixed effects are dummies for each possible performance out of the 20 questions on the test. Data in each panel are from the noted study version(s).

TABLE III

RESULTS FROM INFORMED EVALUATIONS (AFTER PERFORMANCE INFORMATION IS PROVIDED)

Question:	Performance	Performance bucket	Willingness to apply	Success
Panel A: Self-Promotion version, wave 1 ($N = 302$)				
Female	−7.01**	−0.40***	−10.73***	−11.73***
	(2.90)	(0.13)	(3.40)	(3.30)
Panel B: Self-Promotion (Risky) version, wave 1 ($N = 294$)				
Female	−7.24**	−0.36***	−9.11***	−8.07**
	(2.83)	(0.14)	(3.38)	(3.29)
Panel C: Private version, wave 1 ($N = 304$)				
Female	−8.01***	−0.33**	−13.25***	−13.15***
	(2.88)	(0.14)	(3.53)	(3.53)
Panel D: Private version, wave 2 ($N = 302$)				
Female	−7.58**	−0.42***	−14.15***	−14.37***
	(3.18)	(0.15)	(3.53)	(3.46)
Panel E: Private (Social Norms) version, wave 2 ($N = 298$)				
Female	−11.93***	−0.62***	−16.39***	−15.77***
	(3.15)	(0.16)	(3.42)	(3.58)
Panel F: Private version, wave 3 ($N = 300$)				
Female	−12.70***	−0.52***	−16.55***	−15.87***
	(3.04)	(0.14)	(3.73)	(3.76)
Panel G: Private (Immediately Informed) version, wave 3 ($N = 299$)				
Female	−7.61**	−0.47***	−11.42***	−12.48***
	(3.35)	(0.16)	(3.81)	(3.61)
Panel H: Private (Immediately Informed) version, wave 4 ($N = 597$)				
Female	−8.54***	−0.42***	−16.63***	−18.66***
	(2.22)	(0.10)	(2.42)	(2.30)
Panel I: Private (Other-Evaluation) version, wave 4 ($N = 597$)				
Female	0.29	−0.11	−3.54**	−3.17*
	(1.58)	(0.08)	(1.69)	(1.68)
Panel J: Private version, wave 5 ($N = 294$)				
Female	−7.74***	−0.24**	−12.91***	−14.24***
	(2.26)	(0.10)	(3.09)	(3.01)
Panel K: Private (Verbal) version, wave 5 ($N = 305$)				
Female	−0.93	−0.05	−1.34	−1.36
	(1.94)	(0.09)	(2.76)	(2.61)
Panel L: All evaluations of own math and science performance ($N = 2{,}990$)				
Female	−9.83***	−0.47***	−15.12***	−15.59***
	(0.94)	(0.04)	(1.08)	(1.07)
Performance fixed effects	Yes	Yes	Yes	Yes

Notes. * $p < .10$, ** $p < .05$, *** $p < .01$. Standard errors are robust. Results are from OLS regressions of the responses provided to the question noted in each column, as defined in the notes of Table II, after participants are informed of their absolute and relative performance (or the other participant's absolute and relative performance in Panel I). *Female* is an indicator for the participant being female. Performance FEs are dummies for each possible performance out of the 20 questions on the test. Data in each panel are from the noted study version(s).

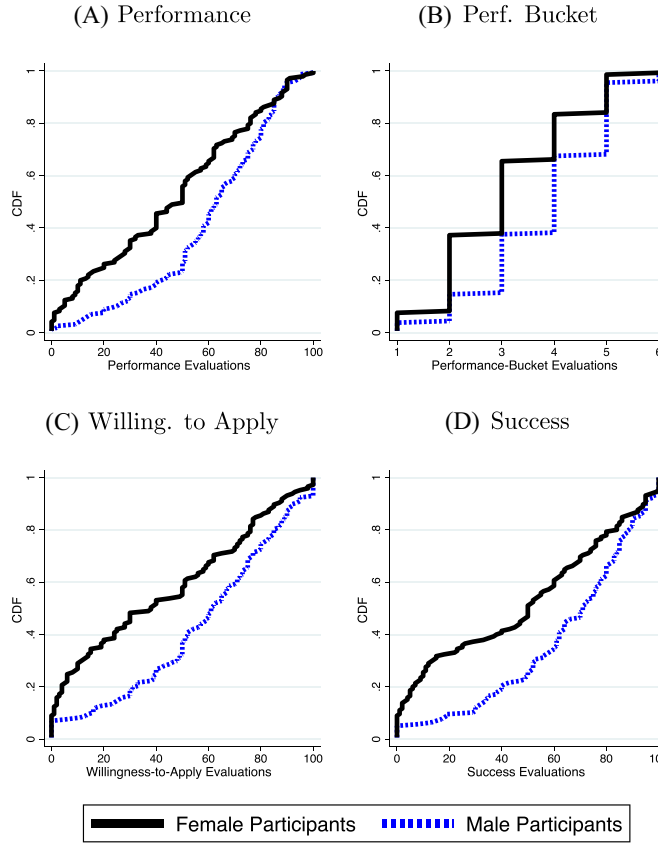

FIGURE II

In the *Self-Promotion* Version, CDFs Showing the Gender Gap in Self-Promotion

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

is, a gender gap in how individuals promote or describe their performance to others.

Figure II shows raw responses to the four self-evaluation questions from part 2 of the *Self-Promotion* version. These responses are provided before participants learn their absolute and relative performance on the test. Women provide significantly lower responses to each question ($p < .01$ for each corresponding t-test and for each Kolmogorov-Smirnov test).

Table II, Panel A confirms the statistical significance of these gender gaps in self-evaluations when controlling for performance with fixed effects for each possible test score (0 to 20) to allow us to compare equally performing men and women. The coefficient on *Female* is negative, large, and statistically significant for all four questions. Column (1) presents results from the performance question that asks participants to respond to the statement "I performed well on the test I took in part 1" on a scale from 0 (entirely disagree) to 100 (entirely agree). The average responses provided by women are 12.68 points lower than those provided by men, which represents a 24% decrease relative to the mean. Column (2) presents results from the performance bucket question that asks participants to "Please indicate how well you think you performed on the test you took in part 1" on a six-point Likert scale. The average responses provided by women are 0.59 points lower, which represents a 17% decrease relative to the mean. Columns (3) and (4) present results from the more "context-rich" questions that may relate to participants' preferences and beliefs about a related, hypothetical job. Column (3) presents results from the willingness-to-apply question that asks participants to respond to the statement "I would apply for a job that required me to perform well on the test I took in part 1" on a scale from 0 (entirely disagree) to 100 (entirely agree). The average responses provided by women are 15.31 points lower, which represents a 31% decrease relative to the mean. Column (4) presents results from the success question that asks participants to respond to the statement "I would succeed in a job that required me to perform well on the test I took in part 1" on a scale from 0 (entirely disagree) to 100 (entirely agree). The average responses provided by women are 15.08 points lower, which represents a 27% decrease relative to the mean. Thus, across all four questions, there is a substantial and statistically significant gender gap in self-evaluations among equally performing men and women.

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

To explore whether this gender gap reflects women thinking they had a lower performance (in absolute or relative terms) than equally performing men—particularly in light of the gender gap in beliefs about absolute performance as detailed in Section II.H—we turn to results from the self-evaluation questions in part 3 of the *Self-Promotion* version. Because these questions are asked after participants are informed of their absolute and relative performance on the test—and thus after we close any gender gap in beliefs about absolute and relative performance "by design"—we refer to the responses to these questions as informed self-evaluations.

Table III, Panel A presents results from responses after participants have learned their absolute and relative performance. These results reveal substantial and statistically significant gender gaps in informed self-evaluations. When considering the questions asked on the 0–100 scale, the gender gap in informed self-evaluations is 7.01 points for the performance question, 10.73 for the willingness-to-apply question, and 11.73 for the success question. When considering the question asked on the 1–6 scale, the gender gap in informed self-evaluations is 0.40.

The gender gap in informed self-evaluations makes clear that the gap is not just a result of women thinking they had a lower performance—in terms of absolute or relative performance—than men. The gender gap also arises when participants are perfectly informed of their absolute and relative performance on the task (i.e., closing any gender gap in beliefs about absolute and relative performance on the task). Put differently, we document a gender gap in self-evaluations that cannot be attributed to gender differences in "confidence," if confidence is modeled as a person's beliefs about their absolute and relative performance, an implicit definition often adopted in prior literature. That said, one may naturally wish to consider confidence more broadly, particularly in the case of the willingness-to-apply and success questions, and hence still consider our results as potentially relating to a gender gap in confidence. Indeed, one could even consider self-evaluations to directly measure a subjective form of confidence.

While the gender gap persists when participants are informed of their absolute and relative performance, the gender gap in informed self-evaluations appears smaller than the gender gap in (uninformed) self-evaluations that are elicited before participants are informed of their absolute and relative performance. As shown

in Online Appendix Table A.2, this is a result of men and women responding somewhat differently to information on their performance. While men inconsistently respond to this information (see the coefficient estimates on *Informed*), women directionally increase their self-evaluations in response to this information (the sum of the coefficient estimates on *Informed* and *Informed***Female* is directionally positive for all four questions). In addition, women directionally increase their self-evaluations in response to this information more so than men (the coefficient estimates on *Informed***Female* are always directionally positive). Although none of these effects are statistically significant in the *Self-Promotion* version on its own, as also shown in Online Appendix Table A.2, similar and statistically significant patterns follow when we pool across all study versions in which we elicit self-evaluations before and after performance information is provided.

An important and interesting question for future work relates to the persistence of the gender gap in self-evaluations across different promotion incentives—beyond those we explored in our *Self-Promotion* version. We take a first step in this direction by presenting results from the *Self-Promotion (Risky)* version. Table II, Panel C and Table III, Panel C show that the gender gap in self-evaluations and the gender gap in informed self-evaluations remain substantial and significant under the slightly different promotion incentives in the *Self-Promotion (Risky)* version.

III.B. The Gender Gap in Self-Evaluations on a Math and Science Task

The gender gaps in self-evaluations that are provided to potential employers in the *Self-Promotion* and *Self-Promotion (Risky)* versions—that is, the gender gaps in self-promotion—could arise due to the incentives to promote one's performance to potential employers or could instead reflect an underlying gender gap in self-evaluations that is present even without any promotion incentives.

To examine the relevance of promotion incentives—and assess whether there is an underlying gender gap in self-evaluations even absent any incentives to self-promote—we turn to the *Private* version. In the *Private* version, self-evaluations no longer serve as a measure of self-promotion because they are not shared with potential employers. Participants receive a fixed payment

regardless of their self-evaluations, eliminating any incentives to self-promote. More broadly, this version allows us to measure any underlying gender gap in self-evaluations that cannot be driven by gender differences relating to strategic incentives (Reuben, Sapienza, and Zingales 2014; Charness, Rustichini, and Van de Ven 2018), risk aversion over payoffs (Dwyer, Gilkeson, and List 2002; Eckel and Grossman 2008), lack of control over payoffs (Cobb-Clark 2015; Apicella, Demiral, and Mollerstrom 2020), or preferences over others' payoffs (Andreoni and Vesterlund 2001; DellaVigna et al. 2013).

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

Results from the *Private* version show that this underlying gender gap is large. When participants are not informed about

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

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

To further investigate the robustness and drivers of the gender gap in self-evaluations, we consider results from the additional study versions that do not involve any self-promotion incentives. In our second wave of data collection, we replicated the gender gap in the *Private* version—when participants are not informed about their performance (see Table II, Panel D) and when participants are informed about their absolute and relative performance (see Table III, Panel D). We also show that the gender gap arises in the *Private (Social Norms)* version, when participants are not informed about their performance (see Table II, Panel E, which is essentially another replication of the *Private* version, since subjects have not yet received additional information) and when participants are informed about their absolute and relative performance as well as the average answers to self-evaluation questions provided by others who had the same performance as them (see Table III, Panel E). The gender gap in informed self-evaluations is just as large in the *Private (Social Norms)* version as in the *Private* version. Thus, the gender gap in self-evaluation persists even when information on what may be typical or socially appropriate is provided.

In our third wave of data collection, we again replicate the gender gap in the *Private* version—when participants are not informed about their performance (see Table II, Panel F) and when participants are informed about their absolute and relative performance (see Table III, Panel F). We also show that the gender gap arises in the *Private (Immediately Informed)* version when participants are immediately informed about their absolute and relative performance and then asked self-evaluation questions (see Table III, Panel G). Even when participants are not asked self-evaluation questions before being informed of their performance—and thus when we remove any related consistency or anchoring effects—we still observe a gender gap after participants are informed about their absolute and relative performance.

III.C. No Gender Gap in Other-Evaluations on a Math and Science Task

Given the robust gender gaps in self-evaluations on a math and science task, one may wonder whether similar gender differences emerge when participants are asked to evaluate the performance of others on the same task or whether, like in prior findings related to negotiation and competition (Bowles, Babcock, and McGinn 2005; Cassar, Wordofa, and Zhang 2016), this change in focus mitigates gender differences. To investigate this, in our fourth wave of data collection we replicate the gender gap in the *Private (Immediately Informed)* version when participants are informed about their absolute and relative performance (see Table III, Panel H). However, we find small, often statistically insignificant gender gaps in the *Private (Other-Evaluation)* version when participants are informed about another participant's absolute and relative performance and then asked the four evaluation questions about that other participant's performance (see Table III, Panel I).

III.D. No Gender Gap in Self-Evaluations on a Verbal Task

Given the gender gaps in pay and in occupational and industry representation that motivate our study, the main task that participants face is a stereotypically male-typed task relating to math and science skills. Given prior work on gender stereotypes and how the type of task can influence gender differences in beliefs (Bordalo et al. 2019; Coffman, Collis, and Kulkarni 2021),

competitions (Günther et al. 2010; Shurchkov 2012; Dreber, von Essen, and Ranehill 2014), group decision making (Coffman 2014; Coffman, Flikkema, and Shurchkov 2021), and test-taking (Saygin and Atwater 2021), one may expect that the gender gap we observe in the male-typed task might be mitigated or even reversed when we consider a more stereotypically female-typed task.

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

These findings suggest that the gender gap in self-evaluations may be more prevalent in male-typed tasks than in female-typed tasks and highlights the value of future work exploring whether gender gaps in self-evaluations arise across a wider range of tasks. Together with the evidence in Section III.C, these findings also make clear that the gender gap in self-evaluations arising in response to the math and science task is not driven by women subjectively evaluating performance differently than men in general (e.g., having different "standards" in general), because it does not persist when participants are asked about their own performance relating to verbal skills or when they are asked about someone else's performance on the math and science task.

IV. Discussion

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

IV.A. The Robustness of the Gender Gap

We examine the gender gap in self-evaluations—on a math and science task—across a range of settings. Separately consid-

ering each self-evaluation question, whether participants are informed, each study version, and each wave, we have 64 possible settings to look for a gender gap. Table II (Panels A–F and J) and Table III (Panels A–H and J) report these 64 tests. We find a statistically significant gender gap 64 out of 64 times. Not surprisingly, when we pool across all self-evaluations relating to the math and science task, the gender gaps in self-evaluations persist, regardless of whether participants are uninformed about their performance (see Table II, Panel L) or informed about their performance (see Table III, Panel L).

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

IV.B. Heterogeneity Analyses

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

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

Second, Online Appendix Table A.12 shows statistically significantly more favorable self-evaluations among younger participants, more educated participants, and more Republican-leaning participants. Online Appendix Table A.12 also shows that the gender gaps are larger among more Republican-leaning participants. To garner additional insights about what drives differences in self-evaluations across groups—and shed light on the potential role of culture—future work might investigate the relationship between self-evaluations and these demographics, as well as other factors such as socioeconomic status, race,

where someone lives, and where someone grew up. We hope that future work also gathers data from countries beyond the United States.

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

IV.C. Consequences of the Gender Gap

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

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

9. Each employer received a guaranteed $1.50 completion fee for the 15-minute study and was recruited using the same criteria as detailed in note 2.

implemented to determine the bonus payments for the employer and for 2 corresponding workers. See Online Appendix D.9 for screenshots and additional details.

As shown in Online Appendix Table A.17, employers are more willing to hire workers who provide more positive self-evaluations. Columns (1), (3), and (4) show that this willingness increases by 1 percentage point for every point on the 0–100 scale in response to the performance question, the willingness-to-apply question, and the success question. Column (2) shows that this willingness increases by an average of 18 percentage points for each increase on the six-point Likert scale in the performance bucket question. Online Appendix Table A.18 shows similar results when we instead consider employers' wage decisions. We do not observe any significant differences by the gender of the employer.

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

2. *The Predictor Results.* If employers anticipate the gender gap in self-evaluations, one might hypothesize smaller economic consequences from the gap because employers can account for women providing less favorable subjective evaluations than men. To assess whether the gender gap in self-evaluations is anticipated, we added eight incentivized questions to the end of the study versions we ran in wave 5 of data collection (see screenshots in Online Appendix D.7). Each question asked participants to predict the average performance of male and female workers in the *Self-Promotion* version after learning the average self-evaluation responses provided by those male and female workers.

As shown in Online Appendix Table A.19, participants do not correctly predict that male and female workers have a similar

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

average performance (equal to about 10) in the *Self-Promotion* version. Rather, when considering predictions based on answers to each self-evaluation question, both male and female participants predict that the average performance of men is significantly higher than the average performance of women. Thus, we find no evidence of predictors correcting the gap when making assessments about workers. Future work—in the laboratory and in the field—should investigate whether this applies more broadly in other settings, such as when experience helps employers get better at identifying the gender gap and perhaps correcting for it. In light of the large literature on discrimination and gender-specific backlash (Riach and Rich 2002; Bowles, Babcock, and Lai 2007; Rudman and Phelan 2008), future work should also investigate the effect of making gender known on self-evaluations.

V. The Gender Gap in Self-Evaluations among Youth

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

To gain insight into the age at which gender gaps in self-evaluations emerge, we ran an additional experiment involving 10,637 middle school and high school students. These students were recruited through the Character Lab Research Network, a network of schools and researchers that partner to run studies that help "advance scientific insights that help kids thrive." Our sample is balanced by gender (48% of students are male) and skewed toward middle school students, giving us particular power at relatively younger ages.

These students completed a *Private* version of our study with four main modifications to accommodate this population and the recruitment process. First, the test for youth only involved the 10 easiest questions from our math and science test. Second, in the willingness-to-apply question, we asked them about their

TABLE IV

THE GENDER GAP IN EVALUATIONS AMONG YOUTH

	Among Students in Grade						
	6	7	8	9	10	11	12
Performance question							
Female	−10.52***	−11.81***	−11.05***	−11.80***	−12.14***	−11.40***	−10.44***
	(1.26)	(1.04)	(0.79)	(1.45)	(1.41)	(1.49)	(1.74)
Performance bucket question							
Female	−0.47***	−0.56***	−0.51***	−0.59***	−0.52***	−0.53***	−0.45***
	(0.07)	(0.05)	(0.04)	(0.07)	(0.07)	(0.08)	(0.09)
Willingness question							
Female	−6.82***	−6.48***	−3.68***	−3.86**	−6.92***	−0.29	−5.77**
	(1.60)	(1.31)	(1.00)	(1.82)	(1.88)	(1.98)	(2.38)
Success question							
Female	−9.42***	−9.85***	−7.19***	−7.41***	−8.40***	−4.58***	−7.29***
	(1.52)	(1.24)	(0.93)	(1.73)	(1.76)	(1.69)	(2.16)
Informed performance question							
Female	−4.00***	−7.10***	−6.98***	−6.51***	−9.55***	−6.75***	−6.24***
	(1.45)	(1.19)	(0.91)	(1.66)	(1.73)	(1.74)	(2.10)
Informed performance bucket question							
Female	−0.15**	−0.33***	−0.27***	−0.27***	−0.33***	−0.26***	−0.22**
	(0.07)	(0.06)	(0.05)	(0.09)	(0.09)	(0.09)	(0.11)
Informed willingness question							
Female	−4.54***	−4.02***	−2.35**	−3.43*	−6.65***	0.00	−5.62**
	(1.74)	(1.38)	(1.03)	(1.87)	(1.87)	(1.98)	(2.39)
Informed success question							
Female	−5.02***	−7.42***	−4.94***	−4.61**	−7.12***	−5.10***	−8.20***
	(1.68)	(1.36)	(1.01)	(1.83)	(1.93)	(1.88)	(2.32)
N	1,521	2,208	3,367	1,031	989	871	650
Perf. fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. * $p < .10$, ** $p < .05$, *** $p < .01$. Standard errors are robust. Results are from OLS regressions of the responses provided to the evaluation question noted in each row among students in the grade indicated by the column (additional details on the question wording can be found in Online Appendix D.8). *Female* is an indicator for the participant being female in the administrative data provided by Character Lab Research Network. Performance fixed effects are dummies for each possible performance out of the 10 questions on the test.

willingness to take a class that involved topics like those covered on the test. Third, in the success question, we asked them about their likelihood of success in a hypothetical class that involved topics like those covered on the test. Fourth, when we provided information on performance, we only provided absolute performance information (we did not have prior performance data on youth to provide relative information). See Online Appendix D.8 for screenshots and additional details.

As seen in Table IV, the gender gap persists across all questions and across all grades. There is some evidence that the

gender gap in willingness to take a class is smaller for older students, perhaps because what classes they have left to take in school is already determined. The clear takeaway, however, is that the gender gap in self-evaluations is robust to this very different setting and that it appears as early as sixth grade, among the youngest students that we study.

Following much of the heterogeneity analysis presented in Section IV.B, Online Appendix Tables A.20–A.25 present parallel results exploring heterogeneity based on performance, beliefs about absolute performance, other demographics, and GPA. Online Appendix Table A.20 reveals that, unlike our prior results, the gender gaps for youth are larger among higher performers. Online Appendix Tables A.21 and A.22 reveal that—while beliefs are positively and significantly correlated with self-evaluations—evidence on how they correlate with the size of the gap is mixed. Online Appendix Table A.23 shows that relative to the 34% of students who are non-Hispanic whites, students from racial minority groups provide less positive responses to the self-evaluation questions about performance and more positive responses about their willingness to take a class. But the gender gap does not appear to systematically differ by race. Online Appendix Table A.24 reveals that the 39% of students who qualify for a free or reduced price lunch (FRPL) provide somewhat less favorable self-evaluations but that FRPL status does not correlate with the gender gap. Finally, Online Appendix Table A.25 shows that GPA is positively and significantly correlated with answers to the self-evaluation questions, and the gap is—if anything—larger among those with a higher GPA.

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

VI. Conclusion

This article documents a large gender gap in self-evaluations on a male-typed task relating to math and science: women subjectively describe their performance less favorably than equally performing men. We first show a substantial and robust gender

gap in self-evaluations that will be shared with potential employers, which we take as evidence of a gender gap in self-promotion. We then show that this gap is not specific to settings with promotion incentives. When self-evaluations are elicited privately, the gender gap remains just as large. Finally, by focusing on settings in which self-evaluations are elicited privately, we show that the gender gap in self-evaluations is robust to a variety of environments and arises early (as evident from our results with over 10,000 middle school and high school students). A notable exception to the robustness of these results is that we do not observe a gender gap in self-evaluations when we ask participants about their performance on a test assessing verbal ability.

We end the article by highlighting the many exciting and important avenues for future work. A first pathway relates to further exploring settings beyond those relating to math and science. That we do not observe a gender gap in self-evaluations when participants are asked about their performance on a verbal test suggests that a gender gap in self-evaluations is less likely in female-typed domains. But because this study only privately elicits self-evaluations in this female-typed domain, future work is needed to assess the effect of communicating self-evaluations to employers in female-typed domains. For example, focusing on a female-dominated profession, Biasi and Sarsons (2022) find that female public school teachers are less willing to negotiate than male public school teachers.

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

A third avenue relates to investigating the effect of the information structure on self-evaluations and how employers respond to them. We find that women are less likely to be hired than equally performing men when a potential employer only learns their answer to one self-evaluation question, in the *Self-Promotion*

version of our study, and when information on performance might be shared with employers, in the *Self-Promotion (Risky)* version. Future work may investigate the effect on self-evaluations when employers have additional information on performance—or signals of performance—or when gender is known (see discussion in Section IV.C).

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

A fifth avenue relates to examining policy interventions to mitigate any consequences of the gender gap in self-evaluations. Akin to the findings in Kessel, Mollerstrom, and van Veldhuizen (2021), future work may investigate the effectiveness of informing individuals of the gender gap in self-evaluations and the associated financial consequences when self-evaluations are communicated to employers.[11] In addition, given the potential difficulty of altering how men and women subjectively view their performance—particularly in the short run if such perceptions are deeply ingrained—promising approaches may require "changing the system" rather than "changing the women."[12] Future work should investigate the effect of relying less on subjective self-evaluations for hiring and promotion.

HARVARD UNIVERSITY, UNITED STATES
UNIVERSITY OF PENNSYLVANIA, UNITED STATES

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

12. For work on potential downsides to a "changing the women" approach, Exley, Niederle, and Vesterlund (2020) show that forcing women to take actions they would not choose themselves backfires in the context of choosing when to negotiate. For excellent recent work on change-the-system approaches, see Apicella, Demiral, and Mollerstrom (2017), He, Kang, and Lacetera (2019), and Carlana, La Ferrara, and Pinotti (2022).

Supplementary Material

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online.

Data Availability

Data and code replicating the tables and figures in this article can be found in Exley and Kessler (2022) in the Harvard Dataverse, https://doi.org/10.7910/DVN/YSWKHY.

References

Abraham, Lisa, and Alison Stein, "Words Matter: Experimental Evidence from Job Applications," RAND Corporation Working Paper, 2020.

Andersen, Steffen, Seda Ertac, Uri Gneezy, John A. List, and Sandra Maximiano, "Gender, Competitiveness, and Socialization at a Young Age: Evidence from a Matrilineal and a Patriarchal Society," *Review of Economics and Statistics*, 95 (2013), 1438–1443.

Andreoni, James, and Lise Vesterlund, "Which Is the Fair Sex? Gender Differences in Altruism," *Quarterly Journal of Economics*, 116 (2001), 293–312.

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

———, "Compete with Others? No, Thanks. With Myself? Yes, Please!," *Economics Letters*, 187 (2020), 108878.

Azmat, Ghazala, and Barbara Petrongolo, "Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?," *Labour Economics*, 30 (2014), 32–40.

Bertrand, Marianne, "New Perspectives on Gender," *Handbook of Labor Economics*, 4 (2011), 1543–1590.

Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz, "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors," *American Economic Journal: Applied Economic*, 2 (2010), 228–255.

Biasi, Barbara, and Heather Sarsons, "Flexible Pay, Bargaining, and the Gender Gap," *Quarterly Journal of Economics*, 137 (2022), 215–266.

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

Bohren, Aislinn, Alex Imas, and Michael Rosenberg, "The Language of Discrimination: Using Experimental versus Observational Data," *AEA Papers and Proceedings*, 108 (2018), 169–174.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Beliefs about Gender," *American Economic Review*, 109 (2019), 739–773.

Born, Andreas, Eva Ranehill, and Anna Sandberg, "A Man's World? The Impact of a Male Dominated Environment on Female Leadership," University of Gothenburg Working Paper in Economics no. 744, 2018.

Bowles, Hannah Riley, Linda Babcock, and Lei Lai, "Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt to Ask," *Organizational Behavior and Human Decision Processes*, 103 (2007), 84–103.

Bowles, Hannah Riley, Linda Babcock, and Kathleen L. McGinn, "Constraints and Triggers: Situational Mechanics of Gender in Negotiation," *Journal of Personality and Social Psychology*, 89 (2005), 951–965.

Cárdenas, Juan-Camilo, Anna Dreber, Emma Von Essen, and Eva Ranehill, "Gender Differences in Competitiveness and Risk Taking: Comparing Children in Colombia and Sweden," *Journal of Economic Behavior & Organization*, 83 (2012), 11–22.
Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti, "Goals and Gaps: Educational Careers of Immigrant Children," *Econometrica*, 90 (2022), 1–29.
Cassar, Alessandra, Feven Wordofa, and Y. Jane Zhang, "Competing for the Benefit of Offspring Eliminates the Gender Gap in Competitiveness," *Proceedings of the National Academy of Sciences*, 113 (2016), 5201–5205.
Charness, Gary, Aldo Rustichini, and Jeroen Van de Ven, "Self-Confidence and Strategic Behavior," *Experimental Economics*, 21 (2018), 72–98.
Cobb-Clark, Deborah A., "Locus of Control and the Labor Market," *IZA Journal of Labor Economics*, 4 (2015).
Coffman, Katherine B. Baldiga, "Evidence on Self-Stereotyping and the Contribution of Ideas," *Quarterly Journal of Economics*, 129 (2014), 1625–1660.
Coffman, Katherine B., Manuela Collis, and Leena Kulkarni, "Stereotypes and Belief Updating," Harvard Business School Working Paper, 2021.
———, "Whether to Apply?," Harvard Business School Working Paper, 2022.
Coffman, Katherine B., Clio Bryant Flikkema, and Olga Shurchkov, "Gender Stereotypes in Deliberation and Team Decisions," *Games and Economic Behavior*, 129 (2021), 329–349.
Croson, Rachel, and Uri Gneezy, "Gender Differences in Preferences," *Journal of Economic Literature*, 47 (2009), 448–474.
DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao, "The Importance of Being Marginal: Gender Differences in Generosity," *American Economic Review: Papers & Proceedings*, 103 (2013), 586–590.
Dreber, Anna, Emma Von Essen, and Eva Ranehill, "Outrunning the Gender Gap—Boys and Girls Compete Equally," *Experimental Economics*, 14 (2011).
———, "Gender and Competition in Adolescence: Task Matters," *Experimental Economics*, 17 (2014), 154–172.
Dwyer, Peggy D., James H. Gilkeson, and John A. List, "Gender Differences in Revealed Risk Taking: Evidence from Mutual Fund Investors," *Economics Letters*, 76 (2002), 151–158.
Eckel, Catherine C., and Philip J. Grossman, "Men, Women and Risk Aversion: Experimental Evidence," in *Handbook of Experimental Economics Results,* C. Plott and V. Smith, eds. (New York: Elsevier, 2008), 1061–1073.
Exley, Christine L., and Judd B. Kessler, "Replication Data for: 'The Gender Gap in Self-Promotion'," (2022), Harvard Dataverse, https://doi.org/10.7910/DVN/YSWKHY.
Exley, Christine L., Muriel Niederle, and Lise Vesterlund, "Knowing When to Ask: The Cost of Leaning In," *Journal of Political Economy*, 128 (2020), 816–854.
Frey, Meredith C., and Douglas K. Detterman, "Scholastic Assessment or g? The Relationship between the Scholastic Assessment Test and General Cognitive Ability," *Psychological Science*, 15 (2004), 373–378.
Gneezy, Uri, and Aldo Rustichini, "Gender and Competition at a Young Age," *American Economic Review*, 94 (2004), 377–381.
Goldin, Claudia, "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, 104 (2014), 1091–1119.
Grossman, Philip J., Catherine Eckel, Mana Komai, and Wei Zhan, "It Pays to Be a Man: Rewards for Leaders in a Coordination Game," *Journal of Economic Behavior & Organization*, 161 (2019), 197–215.
Günther, Christina, Neslihan Arslan Ekinci, Christiane Schwieren, and Martin Strobel, "Women Can't Jump? An Experiment on Competitive Attitudes and Stereotype Threat," *Journal of Economic Behavior & Organization*, 75 (2010), 395–401.
He, Joyce, Sonia Kang, and Nicola Lacetera, "Leaning In or Not Leaning Out? Opt-Out Choice Framing Attenuates Gender Differences in the Decision to Compete," NBER Working Paper no. 26484, 2019.

Hernandez-Arenaz, Iñigo, and Nagore Iriberri, "A Review of Gender Differences in Negotiation," in *Oxford Research Encyclopedia of Economics and Finance* (Oxford: Oxford University Press, 2019), https://doi.org/10.1093/acrefore/9780190625979.013.464.
Isaksson, Siri, "It Takes Two: Gender Differences in Group Work," Norwegian School of Economics Working Paper, 2018.
Kessel, Dany, Johanna Mollerstrom, and Roel van Veldhuizen, "Can Simple Advice Eliminate the Gender Gap in Willingness to Compete?," *European Economic Review* 138 (2021), https://doi.org/10.1016/j.euroecorev.2021.103777.
Kolev, Julian, Yuly Fuentes-Medel, and Fiona Murray, "Is Blinded Review Enough? How Gendered Outcomes Arise Even under Anonymous Evaluation," NBER Working Paper no. 25759, 2019.
Lerchenmueller, Marc J., Olav Sorenson, and Anupam B. Jena, "Gender Differences in How Scientists Present the Importance of their Research: Observational Study," *British Medical Journal*, 367 (2019), I6573.
Lundeberg, Mary A., Paul W. Fox, and Judith Punćcohaŕ, "Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments," *Journal of Educational Psychology*, 86 (1994), 114–121.
Manian, Shanthi, and Ketki Sheth, "Follow My Lead: Assertive Cheap Talk and the Gender Gap," *Management Science*, 67 (2021), 6880–6896.
Michelmore, Katherine, and Sharon Sassler, "Explaining the Gender Wage Gap in STEM: Does Field Sex Composition Matter?," *Russell Sage Foundation Journal of the Social Sciences*, 2 (2016), 194–215.
Murciano-Goroff, Raviv, "Missing Women in Tech: The Labor Market for Highly Skilled Software Engineers," *Management Science* (2021), https://doi.org/10.1287/mnsc.2021.4077.
Niederle, Muriel, "Gender," in *Handbook of Experimental Economics*, vol. 2, John Kagel and Alvin E. Roth, eds. (Princeton, NJ: Princeton University Press, 2016), 481–553.
Niederle, Muriel, and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete Too Much?," *Quarterly Journal of Economics*, 122 (2007), 1067–1101.
———, "Gender and Competition," *Annual Review of Economics*, 3 (2011), 601–630.
Reuben, Ernesto, Paola Sapienza, and Luigi Zingales, "How Stereotypes Impair Women's Careers in Science," *Proceedings of the National Academy of Sciences*, 111 (2014), 4403–4408.
Riach, P. A., and J. Rich, "Field Experiments of Discrimination in the Market Place," *Economic Journal*, 112 (2002), F480–F518.
Roussille, Nina, "The Central Role of the Ask Gap in Gender Pay Inequality," UC Berkeley Working Paper, 2021.
Rudman, Laurie A., and Julie E. Phelan, "Backlash Effects for Disconfirming Gender Stereotypes in Organizations," *Research in Organizational Behavior*, 28 (2008), 61–79.
Saygin, Perihan O., and Ann Atwater, "Gender Differences in Leaving Questions Blank on High-Stakes Standardized Tests," *Economics of Education Review*, 84 (2021), 102162.
Shurchkov, Olga, "Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints," *Journal of the European Economic Association*, 10 (2012), 1189–1213.
Sutter, Matthias, and Daniela Glätzle-Rützler, "Gender Differences in the Willingness to Compete Emerge Early in Life and Persist," *Management Science*, 61 (2015), 2339–2354.
van Veldhuizen, Roel, "Gender Differences in Tournament Choices: Risk Preferences, Overconfidence or Competitiveness?," Econstor Discussion Paper No. 14, 2017.