Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Whether to Apply

Katherine B. Coffman, Manuela R. Collis, Leena Kulkarni

To cite this article:
Katherine B. Coffman, Manuela R. Collis, Leena Kulkarni (2023) Whether to Apply. Management Science

Published online in Articles in Advance 15 Sep 2023

. https://doi.org/10.1287/mnsc.2023.4907

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–21
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Whether to Apply

Katherine B. Coffman,[a,*] Manuela R. Collis,[b] Leena Kulkarni[c]

[a] Harvard Business School, Boston, Massachusetts 02163; [b] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada; [c] Array Technology Group LLC, New York, New York 10022

*Corresponding author

Contact: kcoffman@hbs.edu, https://orcid.org/0000-0003-4293-2220 (KBC); manuela.collis@rotman.utoronto.ca, https://orcid.org/0000-0003-3410-4126 (MRC); lkulkarni@hsph.harvard.edu, https://orcid.org/0000-0002-6593-0052 (LK)

Received: March 15, 2022
Revised: September 20, 2022; December 21, 2022
Accepted: February 4, 2023
Published Online in Articles in Advance: September 15, 2023

https://doi.org/10.1287/mnsc.2023.4907

Abstract. Labor market outcomes depend, in part, upon an individual's willingness to put him- or herself forward for different opportunities. We use a series of experiments to explore gender differences in willingness to apply for higher-return, more challenging work. We find that, in male-typed domains, qualified women are significantly less likely to apply than similarly well-qualified men. We provide evidence both in a controlled setting and in the field that reducing ambiguity surrounding required qualifications increases the rate at which qualified women apply. The effects are mixed for men. Our results point to a way to increase the pool of qualified women applicants.

History: Accepted by Yan Chen, behavioral economics and decision analysis.
Funding: This work was funded by the National Science Foundation [Grant 1713752] and Harvard Business School.
Supplemental Material: The e-companion and data are available at https://doi.org/10.1287/mnsc.2023.4907.

Keywords: gender • economics • behavior • behavioral decision making • organizations

"Why are you not a full professor – given your eminence?"

[Silence]

"I never applied."
Donna Strickland, Nobel Laureate in Physics, 2018

1. Introduction

An important body of work documents the impact of gender bias and discrimination on women's careers (see Riach and Rich 2002 for an overview). Women are less likely to be interviewed for high-status jobs (Fernandez and Mors 2008) and promotions (Ginther and Kahn 2009, Ibarra et al. 2010, Zahidi and Ibarra 2010). Evidence from the laboratory reinforces these findings, with many studies showing that employers in simulated labor markets are more likely to hire men than women for male-typed jobs (Reuben et al. 2014, Bohnet et al. 2016, Coffman et al. 2021). Once a female worker is hired, she is subject to bias in both formal job evaluation processes (Heilman 2001) and more informal mentoring (Ibarra et al. 2010).

Of course, when considering the sources of gender gaps in labor market outcomes, discrimination and bias are only one side of the coin. Decisions made by employees themselves also have the potential to have large impacts on gender gaps in outcomes. Candidates decide what types of education and training to pursue, which jobs to apply to, and whether to put themselves up for promotion. Gender differences at these crucial decision nodes could contribute to gender gaps in outcomes. Indeed, social scientists have documented that occupational segregation plays an important role in explaining gender gaps in wages (Altonji and Blank 1999). Choosing an industry, however, is just one of many important choices an employee makes.

In this paper, we study the decisions of candidates about whether to apply for different opportunities. These decisions are likely to be key not only at the hiring stage but also as careers advance, presenting opportunities for promotion. We ask whether women are as likely as men to see themselves as qualified for challenging, higher-paying positions and whether they apply for promotions or opportunities at similar rates, conditional on their degree of qualification.

Past work provides potential reasons why qualified women may be less likely to apply. Careful laboratory evidence suggests that, conditional on having the same ability, women have more pessimistic beliefs about their own ability in male-typed domains compared with men both in objective terms (Niederle and Vesterlund 2007, Coffman 2014, Bordalo et al. 2019) and subjective terms (Exley and Kessler 2022). In the field, Murciano-Goroff (2021) found that, conditional on having the same level of skill, female software engineers are less likely to self-report that skill on their resume compared with men.

This suggests that even if men and women both have the same skills and share the same view as to what is required for a given position, women may be less likely to believe they possess that qualification. Even conditional on holding the same beliefs, differences in competitive preferences could also drive differences in behavior (Niederle and Vesterlund 2007), as could differences in preferences for more challenging work (Niederle and Yestrumskas 2008).

Clever field experiments have explored some factors that impact job seekers' application decisions. Consistent with the factors mentioned above, Flory et al. (2015) show that an opening that is framed as more male-typed, more competitive, or with more pay uncertainty deters female candidates more than male candidates. Similarly, in a field experiment with a high-skilled population, Samek (2019) found that competitive compensation schemes deter women more so than men. Gee (2018) reported that revealing the number of other applicants increases applications from women more than men, showing that social information can play a role. Role models can also influence application decisions: Del Carpio and Guadalupe (2022) found that women are more likely to opt into a tech skills training program when presented with an example of a female success story. In a field experiment focused on attracting men to female-typed fields, Delfino (2022) documents that the perceived gender composition of the workforce has no impact on application decisions, but information on worker effectiveness draws in talented men. Recent field experiments have also studied how explicit appeals to diversity impact application decisions. Kuhn et al. (2020) found that a request for women to apply increases the number of women who apply; Flory et al. (2021) revealed that signaling interest in workplace diversity increases the rate at which racial minorities apply (and are hired), with weaker impacts on women. Clearly, there are many factors that influence job seeker behavior, including with respect to gender.

Our focus will be primarily on the role that believed qualifications play in shaping behavior. A simultaneous project by Abraham and Stein (2022) has explored how language used in job postings impacts the application behavior of men and women. In a large, randomized control trial, they varied how demanding the qualifications in a job posting are. In particular, they removed optional qualifications (i.e., "PhD preferred"), removed adjectives describing the qualifications (i.e., remove "excellent" in front of "coding skills"), and lowered qualifications (i.e., replace "SQL fluency" with "experience with SQL"). This significantly increases the overall application rate while leaving the fraction of female applicants unchanged.

Our paper builds on this important body of work by attempting to understand better the decision of whether to apply, to identify the factors that may contribute to gender gaps, and to propose and test potential policy solutions. We do this in contexts where we have detailed information on the pool of potential applicants, allowing us to observe who is selecting in and out. We choose to focus on male-typed domains, speaking to field contexts where women are often under-represented. Our experiments zoom in on the role of uncertainty about qualifications, because this is an aspect that not only seems likely to be relevant for understanding gender differences but also may be addressable by potential employers.

Our main hypothesis is that women will be less likely to apply than similarly qualified men. We expect this to be driven, at least in part, by ambiguity surrounding whether an individual is qualified for a given opening. The idea of ambiguity as a driver of gender gaps has been proposed in other contexts, including in negotiation. Bowles et al. (2005) showed that reducing situational ambiguity about what is reasonable or appropriate reduces gender differences in negotiation outcomes. Here, we explore whether ambiguity about where "the bar" is—in terms of required qualifications—affects beliefs about one's own qualification level and decisions to put oneself forward for different opportunities.

Consider an individual deciding whether to apply for an opening; the individual may ask him- or herself, among other things, "Am I qualified for this position?" The answer to this question likely depends not only on the candidate's self-assessment of his or her own aptitude (what are my skills, strengths, and talents) but also on his or her assessment of what the bar is (that is, what level of skills, strengths, and talents is the employer looking for?). These assessments are often made under considerable uncertainty.

In these environments, there may be gender differences in the likelihood of seeing oneself as above the bar. Alternatively, or in addition, women may perceive larger (reputational, psychological, or backlash-driven) costs to applying if below the bar. In the face of uncertainty, women may not be as likely to view applying as worth the risk. Each of these factors could produce a gender gap in application decisions, even conditional on holding the same qualifications. We explore how changing the degree of ambiguity around what the bar is impacts the application decisions of men and women.

Our first experiment is a field experiment on the online labor market platform UpWork. Serving as a potential employer, we create job opportunities to which participants can apply. In our baseline condition, we find that qualified women are significantly less likely to apply to our more demanding and more lucrative job opportunity than equally qualified men. In two treatment conditions, we provide more clarity on what "the bar" is. We find that qualified women are more likely to apply when the bar is clearer, only directionally in one treatment condition and significantly so in the other. Qualified men do not adjust their behavior across our

treatment conditions. As a result, the gender gap in application rates among qualified candidates is reduced when the desired qualifications for the opportunity are less ambiguous. This creates a larger, more gender-diverse pool of qualified applicants. At the same time, our treatments reduce the number of unqualified applicants.

We follow up this field experiment with a well-powered, preregistered replication study on Prolific. In this more controlled setting, we elicit more detailed data from our participants, providing not only a robustness test of our findings from the field but also shedding additional light on the mechanisms at work. Again, we find that qualified women are significantly less likely than qualified men to apply in our baseline condition. Also, our treatment conditions significantly increase the rate at which qualified women apply, increasing the number of qualified female applicants in the candidate pool. In these ways, the findings for women in our more stylized experiment are consistent with our findings in the field. However, in this setting, we find that qualified men also apply more in our treatment conditions compared with the baseline. As a result, the treatments do not reduce the gender gap in application rates.

Using our more detailed data, we show that women view themselves as significantly less well-qualified than men, conditional on having the same objectively measured qualifications. These perceptions are correlated with application behavior. Our treatments increase the extent to which qualified women (and men) perceive themselves as well-qualified. This is likely the case because the bar is clearer; indeed, participants in the experiment report that required qualifications are significantly more objective, specific, and clear in our treatment conditions compared with the baseline condition.

We gather evidence from two other controlled studies that speak to believed qualifications and application decisions. In one study, we show that when evaluating real job advertisements, women assess themselves as marginally less well-qualified for the opening than men do on average. This gap is smaller for advertisements with more clearly stated required qualifications. In the other study, we use a simulated labor market to explore beliefs and behavior. There, our analysis shows that women view themselves as significantly less likely to receive a promotion conditional on applying compared with similarly well-qualified men; however, we find no significant differences in application decisions in this context. In the main text, we focus on our UpWork field study and the preregistered Prolific replication of the UpWork experiment, which was designed to overcome the shortcomings of these other studies; full details of these other studies can be found in Online Appendix C.

Together, our results suggest that talented women may be less likely to put themselves forward for opportunities compared with equally qualified men. Across both our field experiment and our controlled replication, we find that reducing ambiguity about the bar can increase the rate at which qualified women apply. This may be a low-cost way for employers to grow the set of qualified, female applicants.

2. Growing the Pool of Qualified Applicants in the Field

From August to November 2017, we ran a field experiment on an online employment platform called UpWork. UpWork (previously Elance-oDesk) is the largest global freelancing website (UpWork n.d.a). UpWork facilitates matchmaking between freelance workers and potential employers. To implement our field experiment, we act as employers on UpWork, posting job advertisements, inviting a pool of workers to view and apply to our ads, and then tracking application rates. We make job offers to the most qualified workers that apply to each ad and provide them the opportunity to complete the job for the advertised pay. Freelancers are unaware of their participation in an experiment at the time that they make the decision of whether to apply to the job opening.

We start with some institutional context about the setting. Freelancers who register with UpWork can advertise their services by creating a profile. This profile is publicly available and can be searched for and viewed on the UpWork website. A profile can include the following information: the freelancer's first name and last initial, photo, state of residence, hourly rate, self-reported education, self-reported skills, self-reported work experience, number of jobs completed on UpWork, hours worked on UpWork, reviews from previous UpWork employers, and availability status.

UpWork offers hundreds of standardized "Skills Tests" with topics ranging from Adobe to XML ("Skills Tests" n.d.). UpWorkers are encouraged to take as many tests as they would like and have the option to retake a test after 180 days. These tests serve as free, verified evaluations of skills. Freelancers choose whether to take any given test(s) and whether to display the results on their profile.

We take advantage of these skills tests in the design and implementation of our experiment. In particular, we construct our desired qualifications around test scores on either the Management Skills Test (Wave 1 of the experiment) or the Analytical Skills Test (Wave 2 of the experiment). We choose these tests both because they have a relatively large number of freelancers who have taken them and because they are stereotypically more male-typed domains. We focus on male-typed domains to better proxy the male-typed environments that have struggled with female under-representation and lack of advancement (i.e., business, STEM).

We start by identifying all available UpWorkers that are residents of the United States and have displayed on

their profile either the results of the Management Skills Test (Wave 1 of the experiment) or the Analytical Skills Test (Wave 2 of the experiment). This gives us a pool of workers that have completed a test of interest. We then compile the profile information for each of the UpWorkers in this pool, creating a data set with a wealth of information about each worker. We attempt to capture all commonly available profile features, including posted hourly rate, state, hours worked on UpWork, jobs completed on UpWork, indicator of whether they are currently available, measure of current availability (more than 30 hours/week, less than 30 hours/week, as needed), education level (indicators for profile listed a college degree, an MBA, or another graduate degree), a set of indicators for skills in different job categories, and the total number of tests that they have chosen to display on their profile.[1] On top of that, we enter into the data set an indicator of freelancer gender.[2]

The data set also contains the freelancer's score on the test of interest (either Management Skills or Analytical Skills) on a normalized 1–5 scale. This is a score computed by UpWork. Only workers who choose to display their scores appear in our data set. We also record the number of minutes it took the freelancer to complete the test. Recall that freelancers are able to retake these tests if 180 days have passed since the last time they took the test. We do not observe the number of times a freelancer took the test.

Having created this data set, we reach out to each worker in our data set via UpWork, inviting them to apply to our positions. Every invitation contains the following information. Freelancers are informed of two jobs. One job is an "intermediate job," whereas the other job represents the more challenging but also better compensated "expert job." Both jobs require writing essay-style answers to two questions and are advertised to take one hour; we inform freelancers that the questions will be "intermediate level" or "expert level," respectively. We offer pay of $70 for the intermediate job and $155 for the expert job.

All participants are presented with both options and are invited to apply to either of the two jobs (workers can choose either job to apply to but are told that they can apply to no more than one). All freelancers receive generic information on desired characteristics of a successful applicant for the expert job that reads as follows: "We are looking for candidates with [management expertise/experience in analytical thinking], as demonstrated through education, past work experience, and test scores. Successful applicants will also have strong writing and communication skills."

Each worker is randomly assigned to one of three treatments. In our control treatment, workers are provided with no additional information on the desired qualifications. In our positive treatment, freelancers are provided with a descriptive statement about the desired qualifications. The job description states that "we expect that most successful applicants to the expert-level job will have a [Management/Analytical] Skills test score above [3.75/4.05]". In our normative treatment, freelancers are provided with a prescriptive statement about whether to apply for the expert job. The job description states that "we invite applicants with a [Management/Analytical] Skills test score above [3.75/4.05] to apply for the expert-level job."[3] Interested freelancers then apply through the UpWork website, specifying to which job, expert or intermediate, they are applying. We then make hiring decisions using a predetermined algorithm that has assigned weights to our listed desired qualifications.[4] We do not provide information about the exact hiring process to workers.

A few features of our design are worth noting. First, our job ads are set to private on the UpWork platform. Only freelancers in our pre-created data set are able to view and apply to a job. They are able to do so for exactly one of the three treatments, as randomly assigned by the researchers. This ensures that no contamination across treatment occurs. Furthermore, by directly reaching out to workers rather than simply posting the jobs, we hope to boost response rates.

Second, we employ this design with two jobs because we worried that by directly contacting workers and inviting them to apply, we might already be de-biasing workers; our invitation alone might suggest to workers that indeed they are qualified for our opening. To remedy this, we use two jobs, an intermediate job and an expert job, and use the decision to apply to the expert job as our outcome of interest. In this way, even if we are signaling to workers that they are likely a good fit for one of our positions because of our invitation, it is still the case that they face a less obvious decision about whether to apply to the expert or intermediate job.

Third, our treatments use language that is common to application and admissions contexts. Our positive treatment reflects language used in some college and graduate school admissions. Although top schools rarely issue strict cutoffs or qualifications, they sometimes provide information on what typical scores look like for successful applicants. For instance, MIT undergraduate admissions provides the middle 50% score range of admitted students for SAT and ACT scores (MIT Admissions 2021; https://perma.cc/6LA2-HPMH). Our normative treatment borrows language common to job advertisements, where employers "invite" candidates with particular qualifications to apply (see, for instance, this posting from Deloitte; Indeed.com 2020, https://perma.cc/Z4ZA-Q9L3).

Finally, we make a discretionary decision about what the right test score qualification was for our experiment. We use scores within each test sample that are challenging to achieve (just under 25% of candidates in our pool have a test score at or above the stated qualification) but

still allow for a somewhat reasonable sample size of participants who are "qualified" according to our test score qualification.

2.1. Hypotheses

A basic conceptual model helps to illustrate our hypotheses. Consider an individual's decision of whether to apply for a position, such as a new job or a promotion. Denote the payoff to applying and receiving the position (i.e., salary or recognition) as A_1, the payoff to applying and not receiving the position as A_0, the cost of applying as c, and the perceived probability of receiving the position conditional on applying as P. Abstracting away from specific hiring rules, we think about P as a weakly increasing function of the likelihood of being qualified. Define the payoff to not applying, the outside option, as A.[5] Then, a risk-neutral individual will apply for the promotion if and only if

$$P(A_1) + (1 - P)(A_0) - c \geq A$$

This simple expression illustrates several factors that could drive a gender difference in willingness to apply. First, there could be gender differences in real or perceived payoffs or costs. For instance, it could be the case that men expect a new opportunity to be more lucrative, women face higher costs of applying, or women have better outside options. Second, women could be more risk averse than men, leading to a gender difference in willingness to apply (Croson and Gneezy 2009, Niederle 2016). Although these channels merit further study, they are not the main focus of our studies.

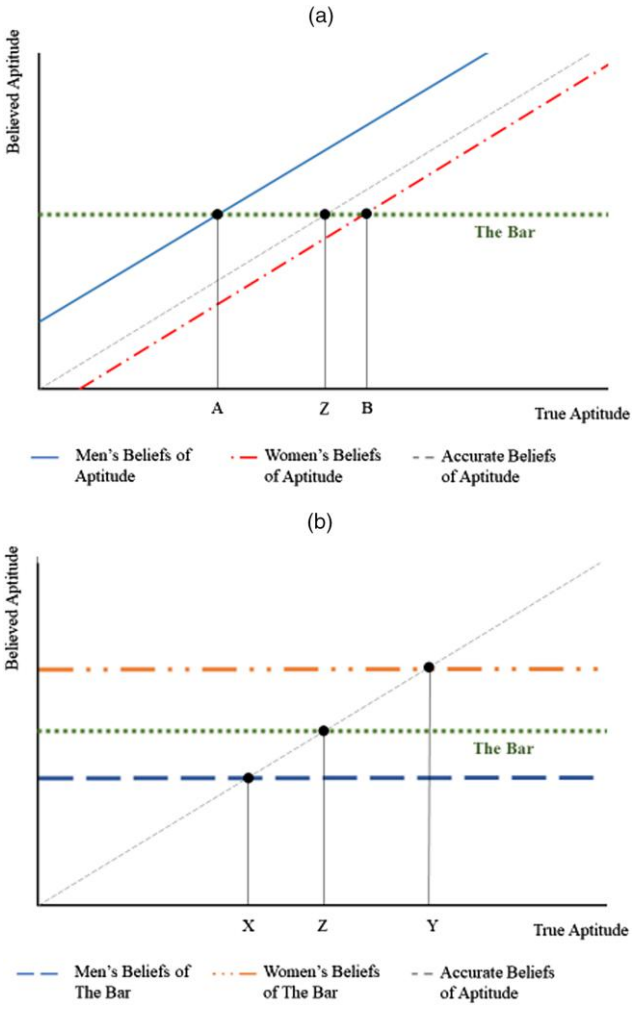
Instead, our focus is the believed likelihood of being qualified. Gender differences in the likelihood of being qualified may emerge based upon differences in self-perception, differences in beliefs of the "the bar," or any combination of the two. To make this concrete, we illustrate two simple cases in Figure 1. For the purposes of the figure, assume that $P = 1$ if the candidate believes his or her ability is above the bar, and $P = 0$ otherwise, and that payoffs and costs are such that all candidates will choose to apply if and only if $p > 0$. The x-axis plots true aptitude for the position, whereas the y-axis plots self-perceived aptitude.

Figure 1(a) illustrates a gender gap in self-perceived aptitude, such that conditional on true aptitude, men are overconfident, and women are underconfident. With this pattern of beliefs, men with true aptitudes between points **A** and **Z** will apply despite not being qualified; whereas their self-perceived talents are above the bar, their true talents are not. Women with true aptitudes between **Z** and **B** will choose not to apply despite being qualified. For all aptitudes between **A** and **B**, we observe a gender gap in application decisions; men apply, whereas women do not.

Differences in beliefs about how high the bar is may also produce a gender gap. We illustrate this case in Figure 1(b). Although the true bar is the same for men and women—the horizontal dotted green line—men and women may hold different beliefs about how high the bar is likely to be, particularly in the face of uncertainty. Imagine a simple case where women perceive the bar to be higher than it truly is (illustrated as the dash-dot-dot orange line), whereas men perceive it to be lower (the dashed blue line). Then, even if individuals have perfect information about their own aptitudes (graphed as the 45-degree line), we may observe a gender gap in applications. Men with aptitudes between **X** and **Z** will apply despite not being qualified, whereas women with aptitudes between **Z** and **Y** will not apply despite being qualified. For all aptitudes between **X** and **Y**, men apply, whereas women do not, again generating a gender gap in applications conditional on measured qualifications.

Beyond these two specific examples, we hypothesize that differences in beliefs about the likelihood of being

Figure 1. (Color online) How Beliefs Can Drive Gender Gaps in Application Rates



Note. (a) Gender differences in beliefs of own aptitude; (b) Gender differences in beliefs of where the bar is.

qualified may generate gender gaps in application decisions. Because ambiguity provides the necessary wiggle room for differences in perceptions, we expect these gender gaps to be larger in more ambiguous environments.

Turning to our setting, our prediction is that indeed qualified women will be less optimistic about being qualified for the expert job in our control condition, leading to a gender gap in application rates in this ambiguous environment. By reducing ambiguity, we expect our treatments to better align men's and women's beliefs about the bar, bringing their beliefs closer to the true bar. Thus, we expect that unqualified men will now be less likely to apply to the expert job relative to the control, whereas qualified women will be more likely to apply relative to the control. As a result, we would expect the treatments to narrow the gender gap in application rates. Of course, the extent to which we observe this effect will depend on initial beliefs about the bar.

In our view, both treatments increase the objectivity, specificity, and clarity of desired qualifications for the expert job relative to the control. Whereas the positive treatment simply describes the qualification, the normative treatment takes things farther; it explicitly encourages candidates with the qualification to apply. A candidate (with a score above the threshold) worried about whether applying is the socially appropriate or right thing to do may be additionally reassured by the normative treatment, even relative to the positive treatment. In this way, the normative treatment may be a more aggressive intervention relative to the positive treatment. If it is, we would expect that the normative treatment has a larger impact on candidate sorting even relative to the positive treatment.

2.2. Results

Table B.1 in the Online Appendix provides descriptive statistics of the freelancers in our sample.[6] Men and women vary in many dimensions in our sample, although these differences appear relatively well-balanced across treatment (Online Appendix Table B.2). Women have more experience on UpWork and are more likely to advertise writing skills, administrative support skills, and customer service skills. Men, on the other hand, post greater hourly rates (in line with work by Dubey et al. 2017 and Foong et al. 2018) and are more likely to advertise skills in Web development, IT, data science, engineering, design, and accounting. This could reflect true differences in skills, although we caution that Murciano-Goroff (2021) found that women are less likely to advertise skills on resumes in the tech domain, even given the same level of experience and skill.

Men outperform women on average in both tests: management skills (male mean 3.55, female mean 3.42, p value from two-tailed t-test < 0.01) and analytical skills (male mean 3.73, female mean 3.57, $p < 0.01$). Also, a greater fraction of men than women are qualified for our expert job according to their test score (i.e., have a test score greater than or equal to the stated test score threshold in our treatments, 29% vs. 19%, $p < 0.01$). See Online Appendix Figure B.1 for the full distribution of test scores by gender and treatment. Our regression analysis will consider gender differences conditional on observed score.

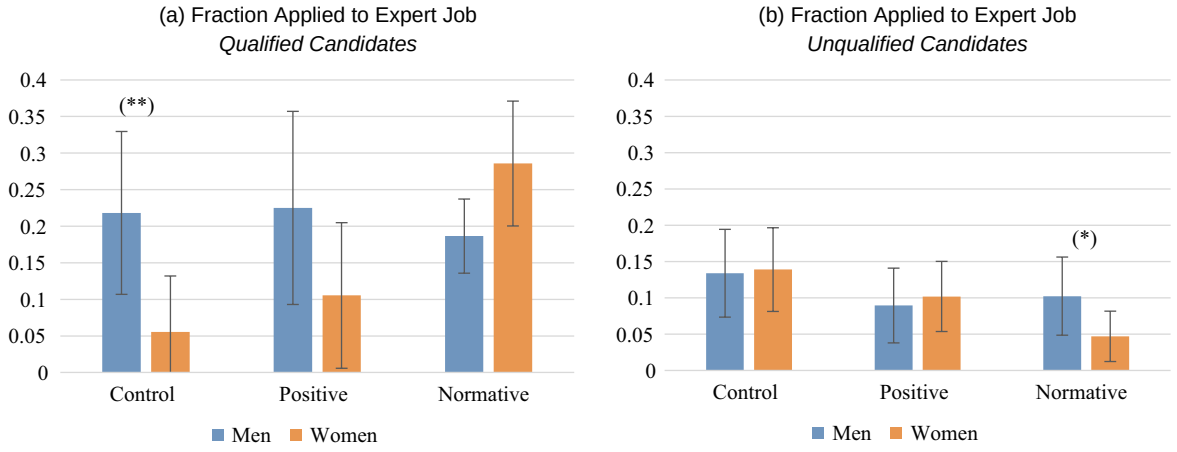
Overall, 20% of men and 18% of women in our sample applied to one of our job postings.[7] This aggregate rate is relatively constant across the three treatments, with 20% of men and 19% of women applying in the control, 22% of men and 19% of women applying in the positive treatment, and 20% of men and 17% of women applying in the normative treatment. Of the 209 participants who applied to our job postings, most (130) applied to the expert job.

Our main question of interest is how application rates to the expert job vary. This is the job for which we introduce variation in the clarity of "the bar." We expect that reduced ambiguity in desired qualifications should (weakly) increase the likelihood of qualified candidates applying to the expert job. Being better informed about where the bar is, if you are above it, should increase your willingness to apply. On the other hand, reduced ambiguity should (weakly) decrease the likelihood of unqualified candidates applying.

We first consider the rates of application to the expert job among qualified candidates—those candidates who have a test score at least as high as the test score threshold. Figure 2(a) demonstrates the results by gender and treatment. In our control treatment, we observe a gender gap in application rates, with just 6% of qualified women applying for the expert job compared with 22% of qualified men. The fraction of all qualified men who applied to the expert job is quite steady across treatment, ranging from 19% to 23%. The reduced ambiguity of our treatments does not draw in additional qualified men; qualified men applied at similar rates regardless of how much information they had about the bar. The application decisions of qualified women, however, appear more responsive to information. We observe the highest rate of application in the normative treatment: 29%. Figure 2(b) shows that application rates to the expert job among unqualified candidates are low, with no clear systematic differences by treatment or gender.

In Table 1, we predict the decision to apply to the expert job from treatment, using the control treatment as our reference category. We control for all profile information included in our summary statistics table.[8] We analyze the full sample in Columns I–III. Consistent with the raw data, when we do not condition on qualification level, we see that overall, our treatments have no significant impact on application rates for men or women (Column I, Column III). Of course, this may mask any competing patterns across unqualified and qualified candidates. In fact, in Column II, we show that

Figure 2. (Color online) Proportion of Freelancers Who Applied for Expert Job by Qualification Level



relative to the control, both treatments decrease application rates among unqualified candidates, whereas they increase them among qualified candidates. These effects are insignificant in the positive treatment and significant in the normative treatment (we cannot reject that these effects are the same across the two treatments).

In Columns IV and V, we analyze the decisions of unqualified candidates (those with test scores less than the threshold). Contrary to our expectation, we do not observe that unqualified women are less likely to apply than unqualified men in our control treatment. Overall, we find that both treatments decrease application rates

Table 1. Application Rates to Expert Job in the Field

	OLS predicting decision to apply for expert job						
	All participants			All unqualified		All qualified	
	I	II	III	IV	V	VI	VII
Positive treatment	−0.026	−0.043	−0.039	−0.046*	−0.057	0.044	0.0067
	(0.024)	(0.027)	(0.035)	(0.026)	(0.038)	(0.061)	(0.081)
Normative treatment	−0.030	−0.067**	−0.033	−0.070***	−0.044	0.098	−0.00076
	(0.024)	(0.028)	(0.034)	(0.026)	(0.038)	(0.060)	(0.074)
Female	−0.029	−0.026	−0.039	−0.0062	0.0023	−0.075	−0.20**
	(0.021)	(0.021)	(0.035)	(0.023)	(0.038)	(0.055)	(0.086)
Qualified		−0.057					
		(0.046)					
Positive × qualified		0.066					
		(0.057)					
Normative × qualified		0.15***					
		(0.056)					
Female × positive			0.024		0.020		0.10
			(0.049)		(0.052)		(0.12)
Female × normative			0.0050		−0.047		0.28**
			(0.048)		(0.052)		(0.12)
Controls	Y	Y	Y	Y	Y	Y	Y
Observations	1,083	1,083	1,083	827	827	256	256
Adjusted R^2	0.035	0.039	0.034	0.037	0.037	0.012	0.026
p value tests of equality:							
Positive = normative	0.87	0.39	0.86	0.35	0.73	0.39	0.93
Positive × above = normative × above		0.15					
Female positive = female normative			0.70		0.20		0.16

Notes. Qualified candidates are those with a test score greater than or equal to the advertised threshold. Controls are posted hourly rate, hours worked, jobs worked, total tests posted, normalized test score, time taken to complete the test, college degree dummy, MBA dummy, other graduate degree dummy, dummies for each category of availability (>30 hours/week, <30 hours/week, as needed), dummies for each self-reported skill, and a dummy for being in the second wave of the experiment.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$.

to the expert job (Column IV), in line with our prediction. Splitting by gender, we estimate that unqualified men's decisions are not significantly impacted by our treatments. For women, we estimate that, relative to the control treatment, the normative treatment deters applications from unqualified women by 9pp ($p < 0.01$; sum of coefficients on *normative* and *female* × *normative*, Column V). However, in an interacted model, we cannot reject that the deterrence effect is of a similar size for men and women (insignificant coefficient on *female* × *normative*, Column V).[9]

In Columns VI and VII, we focus on qualified candidates. Overall, we estimate that our two treatments directionally increase the rate at which qualified candidates apply for the expert job. Our treatments have no impact on qualified men's decisions. Women's decisions, however, do vary by treatment. Qualified women are 10pp more likely to apply in our positive treatment relative to the control ($p = 0.24$) and 28pp more likely to apply in our normative treatment relative to the control ($p < 0.01$). In our control condition, qualified women are 20pp less likely to apply than qualified men. Both treatments directionally reduce this gap by drawing in more qualified women.

Overall, our evidence suggests that clearer qualifications seem to improve candidate sorting into the expert job, primarily among women. From a firm's perspective, the impact of more clearly stated qualifications on the potential pool seems positive: a larger, more gender diverse pool of qualified applicants and fewer unqualified applicants.

The normative treatment has a directionally larger impact than the positive treatment, although we note that the differences are not statistically significant. With that caveat in mind, we offer some speculative thoughts on how the distinct elements of the normative treatment might produce a stronger effect. It may be that the explicit ask—inviting someone with that score to apply—more successfully overcomes hesitations about what the social norms, employer expectations, or right course of action is. Recall that Kuhn et al. (2020) also found that an explicit ask—inviting women specifically to apply—increased female applications in their setting. And Bowles et al. (2005) documented how "strong" situations "where everyone has the same understanding of how they are supposed to respond" produce smaller gender gaps in negotiation outcomes than "weak," more ambiguous situations. Our normative treatment may more successfully produce the type of "strong" situation that minimizes gaps in the willingness to apply context. Of course, more work is needed to investigate the precise channels at play in each of these treatments and whether they indeed produce differential effects.

The results of our field experiment suggest that reduced ambiguity about where the bar is draws in qualified female candidates, narrowing the gender gap in applications among qualified candidates. But although the natural setting of the field has benefits in terms of external validity, it does come with shortcomings. First, we observe only the application decision from our participants. These limited data make it challenging to explore the underlying mechanism behind our results. Second, our intention is to focus on supply-side decisions about when to apply, absent anticipation of employer bias. In the field setting, we cannot rule out that candidates may have anticipated the potential for discrimination, impacting their application decisions. In a more controlled setting, we can more convincingly shut down this potential channel.

And, finally, one key shortcoming of the field study is the limited statistical power. Although our overall sample is not small, we have a smaller number of qualified candidates, smaller still once we cut by gender. If we compute the minimum detectable effect, using our standard errors from Table 1 with a power of 80% and rejection rate of 5%, we are powered only to detect treatment effect of 0.21–0.23 for qualified men and 0.34 for the interaction of female and our treatments. This suggests a need for caution in interpreting these results. By running a well-powered, preregistered replication, we can explore whether our findings from the field are robust in a new but related setting.

3. Replication in a Controlled Environment

We conduct a follow-up experiment to probe the robustness of our results and better unpack the mechanisms underlying behavior in these types of settings. Our design closely mimics the UpWork experiment, with the addition of post-application decision measures from participants. We preregistered this study, and we collected the data in December 2021 on Prolific (AEARCTR-0008223).[10]

3.1. Design

In our experiment, participants first build a "resume" by completing a brief screening task and then decide whether to apply to an expert-level or intermediate-level short-term job. Once again, we serve as the employer, making hiring decisions for the job. Hired participants are invited back to the Prolific platform to complete the job as a second, separate study.

At the outset of the experiment, participants are told that the study consists of building a resume and then deciding whether to apply for a job. To parallel our field experiment, we have participants build a resume consisting of multiple components, including a test score, education information, and details of their work quality and history. This resume is simple and standardized to ensure straightforward comparability across participants.

To establish test scores, we ask participants to complete a skills test during the first part of the experiment. The test consists of 10 multiple-choice questions. We draw the questions from four categories of the Armed Service Vocational Aptitude Battery (ASVAB): General Science, Arithmetic Reasoning, Math Knowledge, and Mechanical Comprehension. The advantage of this test is that it is a reputable cognitive skills test and consists of mostly questions that are mostly hard to "Google" quickly. Furthermore, they cover stereotypically male-typed domains, matching our field setting and the labor market settings we aim to speak to. Participants are given 20 seconds per question and earn \$0.15 per correctly answered question in additional payments. All questions appear on a separate page, and the order is randomized.

For education and work experience, we take advantage of Prolific's built-in screening feature. Prolific users complete a sociodemographic survey prior to signing up to complete studies. We use information from this survey to complete participants' resumes. In particular, we construct a resume for each participant that includes their skills test score from the first part of their experiment, their highest-achieved education level as reported on their Prolific profile, and indicators that they have completed at least 100 Prolific studies with an approval rate above 95%.[11] This ensures that there is no study-specific reporting bias.

Note that, by design, only two factors vary across resumes: test scores and education. We restrict the pool of eligible participants to individuals who completed 100 Prolific studies and obtained an approval rate of 95%[12]. Thus, although the resume states whether participants completed 100 Prolific studies and obtained an approval rate of 95% or higher, by design every participant fulfills these criteria.[13] This generates a high degree of similarity across resumes for many of our applicants, minimizing the risk of large gender differences in resume characteristics and helping to maximize statistical power.

After the skills test, we provided participants with additional information about the two short-term jobs available to them. We used near-identical language to our UpWork field experiment. One job was an "intermediate-level job," whereas the other job represented the more challenging but also better compensated "expert-level job." Both jobs required writing essay-style answers to one question and were advertised to take 15 minutes; just as on UpWork, we were explicit that the jobs required answers to "intermediate-level" and "expert-level" questions, respectively. We offered pay of \$5 for the intermediate job and \$10 for the expert job. We also provided information on how hiring decisions would be made. In particular, we told participants that we would first screen the pool of applicants to each job and determine the set of qualified applicants. Then, we would randomly hire one percent of the set of qualified applicants.[14]

We outlined both job opportunities, and then the participants decided. Participants could apply to one of the jobs; they also had the option to explicitly apply to neither.[15] Participants received generic information on desired characteristics of a successful applicant for the expert job that read as follows: "We are looking for candidates with expertise in analytical thinking, as demonstrated through education, past work experience, and test scores."

Prior to making their application decision, participants viewed their resume. They saw their test score, indicators that they had completed at least 100 Prolific studies and had an approval rate of 95% or greater and were told that we would also consider their education as listed on their Prolific profile. We were explicit that these pieces of information were the only factors that would determine hiring decisions.

We randomly assigned each worker to one of three treatments. In our control treatment, participants saw no additional information on the desired qualifications. In our positive treatment, participants received a descriptive statement about the desired qualifications: "We expect that most successful candidates to the expert-level job will have an ASVAB skills test score above 5.5." In our normative treatment, participants received a prescriptive statement about whether to apply for the expert job: "We invite candidates with an ASVAB skills test score above 5.5 to apply for the expert-level job." We chose a cutoff of 5.5 based upon pilot data that suggested that this was likely to generate a sample where approximately half of participants were above this test score threshold, helping to maximize power. We chose a non-integer cutoff to make it straightforward to classify every participant as either above or below the threshold.

The controlled setting allowed us to ask several follow-up questions that spoke to mechanisms. In particular, we elicited participants' beliefs of how well-qualified they were, of how high the bar was for the expert job, and of how objective, specific, and clear the required qualifications for the expert job were.

The first set of post-application decision questions asked the participant about their perceived probability that they would be considered qualified for the expert job. This question did not depend upon whether they chose to apply to the job. We simply asked them to consider the information on the resume they built and provide their estimate of the likelihood that, based upon this information, they would be considered qualified for the expert job.

Next, we asked them what they expected would be the lowest skills test score among hired candidates for the expert job. This question assessed their beliefs of what the bar was. We also asked them how confident they felt about that guess, where they used a 1–5 scale to indicate not at all sure to completely sure.

Finally, participants assessed how objective, specific, and clear the required qualifications for the expert job

opportunity were. They indicated their answers on a 1–6 scale. Once they answered that question, we provided them with the language used in the other two treatment conditions and asked them the same question: "In your opinion, how objective, specific, and clear are the [below] required qualifications for the expert-level job opportunity?" That is, a participant who was assigned to the normative treatment version first provided their assessment of their job ad. Then, on the following page, they saw the language used for the control version and the language used for the positive version, and they indicated how objective, specific, and clear they found each of these two other ads. This provided us with both across-subject and within-subject evaluations of how the clarity of the bar varied across our three treatment conditions. Once participants completed the main portion of the survey, they answered a brief sociodemographic questionnaire.[16]

We advertised the study as 10 minutes with a completion fee of $1.85 and the opportunity to earn up to $1.50 in additional pay. The actual median hourly pay (additional pay included) was $17.75 as measured by the time spent on our Qualtrics survey. We restricted the pool of participants to individuals who resided in the United States, were age 18 or older, were fluent in English, and completed 100 or more Prolific studies with an approval rating of at least 95%. We incorporated understanding questions and screens for attentiveness. Full instructions are provided in Online Appendix A.

3.2. Results

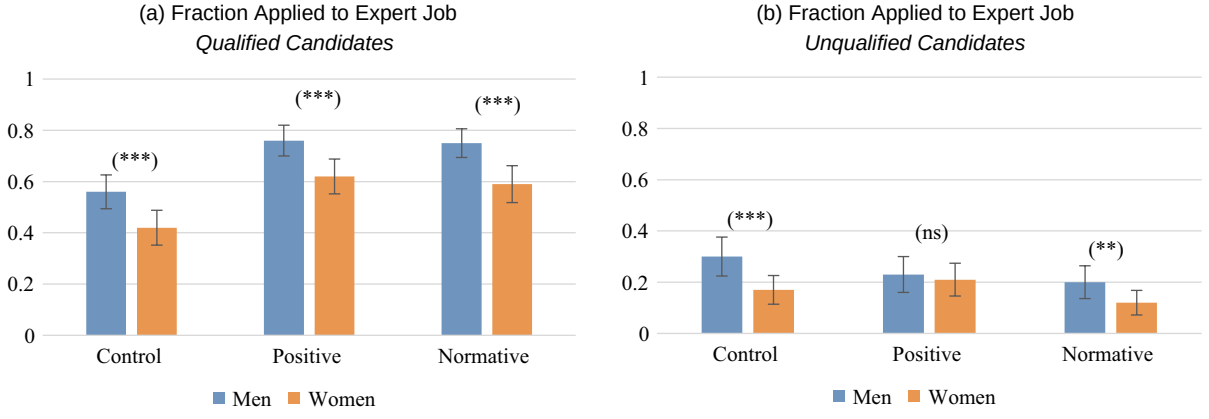
We collected 2,400 observations. After exclusions, our final sample size was 2,243, with no fewer than 142 participants in any given cell (treatment × gender × qualified).[17] In line with our field study and our preregistration, we define qualified as having an ASVAB score strictly above the threshold used in our treatments, 5.5. Online Appendix Table B.5 presents summary statistics for our participants, both overall and conditional on being "qualified."

Women in our study, both overall and conditional on being qualified, are younger and have less educational attainment on average. They have also completed fewer jobs on Prolific. There is a modest but statistically significant difference in average ASVAB scores by gender, with men answering 5.96 questions correctly on average and women answering 5.64 questions correctly on average ($p < 0.01$); this gap is 7.21 to 6.99 among the qualified participants ($p < 0.01$). These gender differences appear balanced across treatments; see Online Appendix Table B.6 and Figure B.2.[18] In our regression analysis, we include control variables to capture these differences.

Figure 3(a) presents the raw data on application rates to the expert job among our qualified candidates (candidates with an ASVAB score above 5.5). In our control treatment, we observe a significant gender difference in application rates to the expert job. Whereas 56% of qualified men applied, only 42% of qualified women applied. Recall that we observe a similarly large gender gap in application rates in the control treatment of our UpWork experiment.

Both the positive and the normative treatments significantly increased the proportion of qualified women who applied to the expert job, bumping up the proportion to 62% ($p < 0.01$ vs. control) and 59% ($p < 0.01$ vs. control), respectively. Similar to what we found on UpWork, clearer, more objective information about qualifications drew in more qualified women. On UpWork, this increase in the rate at which qualified women applied narrowed the gender gap in our treatments, because the behavior of men was essentially unchanged. But, in this sample, we observe that men also responded to our treatments. The positive and normative treatments increased the proportion of qualified men who applied to 76% ($p < 0.01$ vs. control) and 75% ($p < 0.01$ vs. control), respectively. As a result, there was no

Figure 3. (Color online) Proportion Who Applied for Expert Job in Prolific Study by Qualification Level



Note. *$p < 0.10$, **$p < 0.05$, and ***$p < 0.01$ from test of proportions comparing application rate of men and women within treatment.

change in the gender gap across the treatments. Instead, there were simply more qualified applicants (both men and women) for the expert job.

Figure 3(b) presents the evidence for unqualified candidates. Application rates for the expert job were generally low among unqualified candidates. In our control treatment, we observe that unqualified men were significantly more likely to apply than unqualified women: 30% versus 17%. Our treatments somewhat decreased the rate at which unqualified men applied. The treatments had little impact on the application decisions of unqualified women.

Table 2 predicts the decision to apply from treatment, controlling for participant characteristics, paralleling Table 1. Column I reveals that, overall, women were 10 percentage points less likely to apply to the expert job ($p < 0.01$) and that our treatments, through their large impacts on qualified candidates, increased the rate at which individuals applied to the expert job. Column II confirms that the treatments were pulling in qualified candidates significantly more than they were pulling in unqualified applicants. Overall, there was no significant interaction of either treatment with gender (Column III). Zooming in on unqualified candidates, we see that the normative treatment significantly reduced the rate at which unqualified applicants applied, whereas the positive treatment had no impact (Column IV). Column V suggests that the deterrence effects of the treatments on unqualified candidates were not significantly different by gender.

We estimate that our treatments had a large impact on qualified candidates. Column VI estimates that qualified candidates were 20 percentage points more likely to apply in the positive and normative treatments than they were in the control ($p < 0.01$). However, unlike what we observed in UpWork, these effects are observed for both qualified men and qualified women (Column VII). As a result, neither treatment significantly reduced what is a meaningful gender gap in application rates among qualified candidates in the control of 12pp ($p < 0.01$).[19]

Following their application decision, we asked participants to estimate the likelihood that they would be considered qualified for the expert job just based on the resume we consider. Table 3 estimates this believed likelihood conditional on gender, treatment, and observables. Unqualified women believe they are 7pp less likely to be qualified than similarly unqualified men (Column I, $p < 0.01$). Column II indicates that this gap is not significantly changed by our treatments. We see a similar gender gap among qualified candidates. Qualified women believe they are 6pp less likely to be qualified than similarly qualified men (Column III, $p < 0.01$). Although our

Table 2. Applications to Expert Job on Prolific

	OLS predicting decision to apply for expert job						
	All participants			All unqualified		All qualified	
	I	II	III	IV	V	VI	VII
Positive treatment	0.098***	−0.023	0.079**	−0.015	−0.068	0.20***	0.19***
	(0.023)	(0.035)	(0.033)	(0.031)	(0.047)	(0.032)	(0.044)
Normative treatment	0.075***	−0.081**	0.081**	−0.077**	−0.10**	0.20***	0.19***
	(0.023)	(0.034)	(0.032)	(0.031)	(0.045)	(0.031)	(0.042)
Female	−0.10***	−0.10***	−0.11***	−0.064**	−0.11**	−0.12***	−0.12***
	(0.019)	(0.019)	(0.032)	(0.026)	(0.044)	(0.027)	(0.044)
Qualified indicator		0.043					
		(0.042)					
Positive × qualified		0.22***					
		(0.046)					
Normative × qualified		0.28***					
		(0.045)					
Female × positive			0.038		0.097		0.011
			(0.046)		(0.063)		(0.064)
Female × normative			−0.012		0.045		0.0030
			(0.045)		(0.062)		(0.063)
Controls	Y	Y	Y	Y	Y	Y	Y
Observations	2,243	2,243	2,243	967	967	1,276	1,276
Adjusted R^2	0.211	0.238	0.210	0.035	0.035	0.112	0.111
p value: Positive = normative	0.32	0.09	0.95	0.05	0.46	0.96	0.96
p value: Positive × above = normative × above		0.23					
p value: Female positive = female normative			0.28		0.41		0.91

Notes. Qualified candidates are those with a test score greater than or equal to the advertised threshold (5.5). Controls are ASVAB test score, Prolific approval rate, total studies completed on Prolific, indicator for college as highest education attainment, indicator for graduate degree as highest educational attainment, indicator for community or technical school education, indicator for Black or African American, indicator for Latino or Latina, indicator for Asian, indicator for attending high school in the United States, and year of birth.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$.

Table 3. Perceptions of How Well-Qualified Candidates Feel

	OLS predicting believed likelihood of being qualified for expert job			
	All unqualified		All qualified	
	I	II	III	IV
Positive treatment	−2.35	−3.95	10.2***	10.3***
	(1.97)	(2.91)	(1.80)	(2.51)
Normative treatment	−2.81	−4.36	11.4***	9.73***
	(1.92)	(2.84)	(1.77)	(2.38)
Female	−6.49***	−8.37***	−6.16***	−7.33***
	(1.61)	(2.73)	(1.53)	(2.52)
Female × positive		2.93		−0.22
		(3.94)		(3.61)
Female × normative		2.82		3.70
		(3.86)		(3.56)
Controls	Y	Y	Y	Y
Observations	967	967	1,276	1,276
Adjusted R^2	0.068	0.067	0.133	0.132

Notes. Qualified candidates are those with a test score greater than or equal to the advertised threshold (5.5). Controls are ASVAB test score, Prolific approval rate, total studies completed on Prolific, indicator for college as highest education attainment, indicator for graduate degree as highest educational attainment, indicator for community or technical school education, indicator for Black or African American, indicator for Latino or Latina, indicator for Asian, indicator for attending high school in the United States, and year of birth.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$.

treatments are effective at increasing the perceived likelihood of being qualified among qualified candidates, they do not reduce the gender gap (Column IV). This is consistent with the patterns we saw in terms of application decisions, suggesting that beliefs about the likelihood of being qualified are relevant for decisions.

To explore this formally, we rerun Specifications IV–VII in Table 2, adding the believed likelihood of being qualified as a regressor in Table 4.[20] Doing so doubles adjusted R^2, pointing to the importance of beliefs. Interestingly, once we control for these beliefs, the overall gender gap in applications is reduced.[21] This highlights that one reason why we observe lower application rates for women, across both the treatments and control, is that women feel less qualified than men. However, it is worth noting that even conditional on

Table 4. Do Beliefs Explain Applications for Expert Job on Prolific?

	OLS predicting decision to apply to expert job			
	All unqualified		All qualified	
	I	II	III	IV
Positive treatment	−0.0028	−0.047	0.13***	0.12***
	(0.030)	(0.044)	(0.030)	(0.041)
Normative treatment	−0.062**	−0.080*	0.12***	0.13***
	(0.029)	(0.043)	(0.029)	(0.039)
Female	−0.029	−0.066	−0.079***	−0.076*
	(0.025)	(0.041)	(0.025)	(0.041)
Female × positive		0.082		0.012
		(0.059)		(0.059)
Female × normative		0.030		−0.022
		(0.058)		(0.058)
Believed Likelihood of being qualified (0–100)	0.0053***	0.0053***	0.0067***	0.0067***
	(0.00049)	(0.00049)	(0.00046)	(0.00046)
Controls	Y	Y	Y	Y
Observations	967	967	1,276	1,276
Adjusted R^2	0.141	0.141	0.241	0.240

Notes. Qualified candidates are those with a test score greater than or equal to the advertised threshold (5.5). Controls are ASVAB test score, Prolific approval rate, total studies completed on Prolific, indicator for college as highest education attainment, indicator for graduate degree as highest educational attainment, indicator for community or technical school education, indicator for Black or African American, indicator for Latino or Latina, indicator for Asian, indicator for attending high school in the United States, and year of birth.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$.

their beliefs of how well-qualified they are, qualified women continue to apply at lower rates than qualified men across both the treatments and the control. We return to this point in Table 6.

Table 4 also presents evidence that our treatment effects for qualified applicants operate at least partly through beliefs. In Table 2, we estimated that our treatments increased applications to the expert job by roughly 20pp. Once we control for beliefs, the estimated treatment effects are reduced by approximately one-third. What explains the remaining impact of the treatments? It could relate to noise in the measurement of beliefs, leading to attenuation. Or our treatments may impact behavior in ways beyond the belief of being qualified, perhaps through reducing uncertainty about what is appropriate or expected.

We also asked participants their beliefs about what "the bar" is for the expert job in terms of minimum required ASVAB score. The full histograms of beliefs by gender and treatment are presented in Online Appendix Figure B3.[22] We do not see gender differences, suggesting that women do not perceive a higher bar than men do on average in this setting despite applying at lower rates. In the control treatment, men estimated the bar to be a score of 7.2 on average (SD 1.67), and women estimated 7.3 (SD 1.66) on average. The modal answer in the control treatment is 8. Our treatments significantly lowered beliefs of the bar among both men and women, to approximately 6.6 (SD 1.45) for men and 6.5 (SD 1.47) for women in the positive treatment and 6.6 (SD 1.35) for men and 6.5 (SD 1.53) for women in the normative treatment, reflecting our communicated information.[23] Note that our treatments not only reduced the average believed score required but also reduced variance; the standard deviation in the control is 1.66, whereas the standard deviation in the positive and normative treatments is 1.46 and 1.43, respectively.

We hypothesized that our treatments would increase how sure individuals felt about their beliefs of what the bar is. In our control treatment, individuals indicate a 2.7 on the $1-5$ scale, where 1 indicates completely unsure and 5 indicates completely sure (2.76 for men, 2.67 for women, n.s.). Uncertainty is similar in the positive treatment (2.82 for men, 2.68 for women, not a significant increase for either gender). In the normative treatment, individuals are significantly more confident about their guess of the bar than in the control, although the differences are modest (average of 2.98 for men, a 0.22 increase, $p < 0.01$; average of 2.83 for women, a 0.16 increase, $p < 0.05$). Pooling across the three conditions, we observe that women express significantly less confidence in their beliefs of where the bar is (gap of 0.13 points, $p < 0.01$), with similar-sized gaps across both the control and treatment conditions.

Finally, we asked participants directly about the objectivity, clarity, and specificity about required qualifications in each treatment. Recall that each participant first saw this Likert-scale question about their own assigned treatment (before seeing any alternative treatment language). Then, after answering that question, they were shown the language used in each of the other two treatments and asked to make this judgment for each of the other two treatments. As a result, we can analyze perceptions about the amount of ambiguity in each treatment fully across subjects, using only their answers to the question about their own treatment or also using within-subject data.

In Table 5, we regress a participant's answer to the question of how objective, specific, and clear the required qualifications for the expert job were on a dummy for which treatment they were evaluating. They assessed this on a $1-6$ scale, where 6 was extremely objective, specific, and clear. The first column uses only the across-subject data, whereas the second column uses all three observations per participant, clustering standard errors at the individual level. In both specifications, we see a clear ordering of the treatments in terms of the amount of ambiguity. The control treatment (with a mean ranking of 3.63 from participants assigned to the control) is perceived as least objective, clear, and specific, as expected. The positive and normative treatments are each seen as significantly more objective, clear, and specific, with the normative perceived as significantly more of an improvement than the positive treatment.

Although the positive and normative treatments had similar impacts on application behavior in the Prolific setting, it is interesting to relate these data back to our UpWork setting. On UpWork, our normative treatment

Table 5. Perceptions of Clarity of Desired Qualifications

	OLS predicting how objective, clear, and specific were desired qualifications ($1-6$)	
	Across-subject only	All
Positive language	0.41***	1.20***
	(0.066)	(0.030)
Normative language	0.55***	1.33***
	(0.065)	(0.031)
Controls	Y	Y
Observations (clusters)	2,243	6,729 (2,243)
Adjusted R^2	0.040	0.228
Test positive = normative	$p = 0.031$	$p < 0.001$

Notes. Qualified candidates are those with a test score greater than or equal to the advertised threshold (5.5). Controls are gender, an indicator for being qualified, ASVAB test score, Prolific approval rate, total studies completed on Prolific, indicator for college as highest education attainment, indicator for graduate degree as highest educational attainment, indicator for community or technical school education, indicator for Black or African American, indicator for Latino or Latina, indicator for Asian, indicator for attending high school in the United States, and year of birth. We also control for treatment assignment in Column II.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$; ****$p < 0.001$.

Table 6. Beliefs and Behavior

	Application rates to expert job			
	Overall	Control	Positive	Normative
Strictly above believed bar				
Men	82.3% (N = 328)	74.7% (N = 71)	84.5% (N = 116)	84.4% (N = 141)
Women	71.1% (N = 273)	56.5% (N = 62)	78.0% (N = 118)	72.0% (N = 93)
p value M vs. W	p = 0.001	p = 0.03	p = 0.20	p = 0.02
Score equal to believed bar				
Men	71.8% (N = 227)	77.0% (N = 87)	63.9% (N = 72)	73.5% (N = 68)
Women	63.4% (N = 194)	65.4% (N = 78)	66.1% (N = 56)	58.3% (N = 60)
p value M vs. W	p = 0.07	p = 0.10	p = 0.80	p = 0.07
Strictly below believed bar				
Men (N = 573)	24.8% (N = 573)	24.2% (N = 223)	25.8% (N = 155)	24.6% (N = 195)
Women (N = 648)	14.2% (N = 648)	13.5% (N = 252)	15.6% (N = 187)	13.9% (N = 209)
p value M vs. W	$p < 0.001$	p = 0.003	p = 0.02	p = 0.006

Notes. p values are from tests of proportions comparing men and women within treatment. "Above believed bar" refers to individuals whose score is greater than their stated belief of the minimum required score. "Below believed bar" refers to individuals whose score is strictly less than their stated belief of the minimum required score.

had a statistically significant impact on the application rates of qualified women, whereas the positive treatment did not (although the two effects were statistically indistinguishable). The Prolific data reveal that the normative treatment is "stronger" in terms of reduced ambiguity, at least as perceived by these participants, suggesting a potential explanation for the UpWork pattern. However, it is interesting to note that this does not seem to translate into a differential impact of the treatments within Prolific. In the next section, we dig deeper into factors beyond uncertainty about the bar that may contribute to these patterns.

3.3. Additional Analysis

In this section, we build on our preregistered analysis with additional exploration of how application decisions relate to perceptions of where the bar is. We hypothesized that gender differences in perceptions of the bar could generate differences in willingness to apply. In our Prolific study, we observe that even when men and women have similar perceptions of what the bar is, men apply at significantly higher rates than similarly qualified women. This is true across our treatments. Understanding where these persistent differences emerge can help provide insights into the drivers of this gender gap.

Participants in our study provide their belief of what minimum ASVAB score would be required to be considered qualified for the expert job, and they know their own score. Thus, we can ask whether participants whose scores exceed their believed bar indeed chose to apply and whether there are gender differences. Similarly, we can probe the behavior of participants whose score falls at or below their believed bar. Recall that participants knew their score at the time of their application decision. This analysis asks how participants acted on believing that their score was above or below their perceived bar.

Table 6 provides these results. We see evidence of gender differences for individuals who believed they are above the bar, at the bar, and below the bar. First, consider individuals whose score is above what they believe the bar to be. Overall, we find that 82% of those men applied compared with just 71% of women ($p <$ 0.01). This gender gap is largest in our control treatment, where 75% of men who believed they had a score above the bar applied but just 57% of women did (p = 0.03). In our view, this is a striking result that merits investigation in future work; given that they believe they have the required qualification, why do nearly half of these women choose not to apply for the expert job?

We see that our treatments both increase the number of individuals who believe they have a score above the bar while also increasing the rate at which those individuals apply. This is true for both men and women. In our treatment conditions, approximately 85% of men who believe they have a score above the bar apply. For women, our treatments increase the application rate of women who believe they are qualified from 56% in the control to 78% in the positive treatment and 72% in the normative treatment, directionally shrinking the gender gap in application rates among this subpopulation.

Together, these results suggest that our treatments are effective even among the population of individuals who already believed they were above the bar. This may be because our treatments increase believed certainty about being above the bar. Indeed, if we compare self-reported certainty about where the bar is among the subsample of individuals with scores above their believed bar, certainty is significantly higher in the treatments than in the control (2.48 in control vs. 2.94 in normative, $p < 0.01$, and vs. 2.70 in positive, $p < 0.05$).

This increased certainty may propel more individuals to apply by reducing the perceived risk in applying. More work is needed to understand these important patterns.

We can also explore the behavior of individuals whose scores are equal to their believed bar. Here, we again see a gender difference; 72% of these men and 63% of these women applied to the expert job overall ($p = 0.07$). This gap does not seem to interact strongly with our treatments. Finally, we can turn to individuals whose believed scores are strictly below the bar. Application rates to the expert job are low among this subpopulation, but with persistent gender differences. Across each of our treatments, approximately 25% of men whose score did not meet what they believed the required qualification to be still chose to apply, whereas approximately 14% of women made this same decision ($p < 0.001$ overall).

This analysis provides further insight into the gender gaps we observe. In our setting, it seems to be the case that men and women have similar beliefs about what the bar is and yet make different decisions conditional on those beliefs. Our treatments increase application rates among men and women by shifting their beliefs of the bar, but the gender differences remain. This seems to be in part because even once beliefs about the bar are corrected, men and women continue to make different decisions about whether to apply conditional on believing they are above (or below) this bar.

In principle, these patterns could reflect differences in the cost of applying or in outside options. Our design aimed to limit the impact of these factors by paying above-market wages, offering an implied hourly wage above the standard hourly rate on Prolific ($40 vs. $12). And the most relevant outside option for participants, both men and women, is the intermediate job—the job they chose to apply for instead of the expert job. Finally, for both men and women, the (monetary and time) cost of applying for the expert job is limited to forgoing an application to the intermediate job, suggesting that large differences in application costs are unlikely. For these reasons, it seems unlikely that differences in outside options or application costs are the drivers of these results.

Risk preferences may be an important factor. If men are more risk-loving than women, they could apply at higher rates even when holding the same beliefs and facing the same outside options and costs (Croson and Gneezy 2009, Niederle 2016). Alternatively, given the male stereotype of the domain, it may be that qualified women are more likely to discount their scores relative to men, believing they are less reflective of their talent than luck, reducing application rates. Similarly, unqualified men may be more likely than women to explain away their negative score to luck rather than talent, increasing their application rates relative to women. The score attribution channels would be consistent with past work showing these types of gender biases in responses to feedback (Buser and Yuan 2019, Shastry et al. 2020, Coffman et al. 2023).

Another explanation relates to the cost of rejection. Whereas the monetary cost of not getting the expert job is the same for men and women (the foregone chance to apply for and receive the intermediate job), there may be other nonmonetary costs to consider. For instance, women may anticipate larger backlash associated with rejection, driven by a fear of violating gender norms that associate assertiveness and ambition more with men (Sczesny et al. 2018). It could also be that the cost of rejection is larger for women. This could be connected to past findings that women are deterred more after negative feedback in competitive settings (Buser and Yuan 2019). Understanding how these other factors contribute to gender differences in application behavior is an important topic for future work.

4. Discussion

In this section, we focus on synthesizing the results of our two studies. Although there are many similarities in the results across the two contexts, there are also a few important differences. Most notably, our interventions increase the rate at which qualified men apply on Prolific, but not in our field setting. This is directly tied to the impact of our interventions on the gender gap. In this section, we focus on understanding why this difference might emerge. This question has important implications for understanding whether ambiguity-reducing interventions in other contexts are likely to reduce a gender gap in application rates.

We ground this conversation in the conceptual model we introduced in Section 2. We argued that gender differences in the believed likelihood of being qualified may emerge based upon differences in perceptions of where "the bar" is relative to one's own ability. We hypothesized that under sufficient ambiguity, men and women may hold different beliefs about the bar, with women more pessimistic about their own chances of being qualified. By reducing ambiguity, we argued that our treatments might better align men's and women's perceptions of the bar, drawing in qualified women and reducing the gender gap. This argument depends critically on the idea that men and women begin with different perceptions of the bar.

Recall that on Prolific, men and women in the control treatment have indistinguishable beliefs about where the bar is; both overestimate the bar, on average. Given these starting conditions, our interventions, aimed at correcting beliefs, seem unlikely to have a differential impact on men and women; instead, the interventions should just increase the rate at which qualified candidates apply. Indeed, this is what we find.

One plausible hypothesis is that a larger degree of initial ambiguity in the field might give rise to a gender

gap in initial perceptions of the bar in a way that the less ambiguous, more controlled context does not. Although we do not have a precise quantitative measure to compare, we think it is likely that the field features more ambiguity than Prolific. Consider the two control treatments. In both settings, candidates have limited information about desired qualifications for the expert job. But on Prolific, candidates know that we consider only the three resume items: test score, education, and a Prolific approval rate of 95% across at least 100 studies. Of these dimensions, only two vary across participants: test score and education. On UpWork, there is substantially more ambiguity about what might be relevant given their profile; their resumes are more detailed, complex, and varied across individuals.

In this way, the field may more readily provide the necessary wiggle room for an initial gap in beliefs about the bar. It could be the case that on UpWork, there is an initial gender gap in how high individuals believe the bar to be, more like the illustration in Figure 1. Although we can only speculate, this would be consistent with a piece of evidence from one of our follow-up studies that also uses real job ads, reported in Online Appendix C, where we find that women assess themselves and others as less likely to be qualified for the opening.

One interesting possibility is that, in the face of uncertainty, women might believe the bar to be higher because of experienced or anticipated discrimination.[24] If women have past experience seeing less qualified men hired instead of them or simply believe this may be a possibility, they may rationally expect a higher bar, particularly in more subjective evaluation settings. We argue that these factors are much greater in the field. On Prolific, we provide an objective, simplified resume for each candidate, less open to interpretation, and we explicitly shut down the possibility of discrimination.[25] This may help to better align men's and women's beliefs about the bar, even in our control condition, by reducing anticipated discrimination.

Similarly, it could be the case that men and women differentially attend to required qualifications. That is, women may be more concerned with what the bar is and whether they are likely to be above it, paying more attention to this aspect of the job posting. This could potentially drive women to respond more to our treatments than men do, because they are more attuned to information on required qualifications. Of course, to help explain our results, this attentiveness gap would have to be larger on UpWork than on Prolific. This may indeed be the case; Prolific participants are likely accustomed to reading carefully and following instructions in research study contexts, and they must pass an attention check in our study. UpWork candidates operate in a more naturalistic field setting, leaving open the possibility of differential attention.

A related consideration is whether the risks of applying and being rejected are perceived differently across the two platforms. It could be that in the regulated academic research study environment on Prolific, the perceived reputational or backlash costs of being rejected are minimal compared with the more ambiguous field context of UpWork. It is also possible that the psychological costs of being rejected are larger in the field, where the failure is more likely to be tied to one's "real" job and resume. For this to explain our results, these concerns would have to (i) be felt more acutely by women than men on average and (ii) loom larger in the field. Although more research is needed, we think it is interesting to consider the possibility that our treatments, through clearer, more objective information about the bar, potentially reduce the role that anticipated discrimination, fear of backlash, or distaste for rejection plays in application decisions.

4.1. Ceiling Effects

Above, we speculated that that the more ambiguous field environment may give rise to a (larger) gap in initial beliefs about the bar, driving the differences across the two studies. Here, we consider a different hypothesis related to initial conditions. We ask whether the initial conditions on UpWork are such that we could reasonably expect treatment effects for men. In our control condition, we observe that, conditional on applying to either the intermediate or expert job, 92% of qualified men apply to the expert job, whereas only 50% of qualified women do. This leaves open the question of how much "room" there is to draw in additional qualified men to the expert opening; it may be that a large majority of men who are interested, available, and aware of our ad may already apply to the expert job in the control treatment, limiting our scope for intervention among qualified men.[26,27]

On Prolific, however, we know that there are a large number of both men and women who are qualified but choose to apply to the intermediate job rather than the expert job in the control, in part because they overestimate the bar. It may be that on UpWork, either because of ex ante optimistic beliefs they hold about the bar and/or their propensity to apply to the expert job even in the control condition, it is harder to draw in additional qualified men.

4.2. Other Explanations

Before concluding, we touch briefly on other explanations that we think are less likely to explain the differences we observe across the two platforms. First, could the fact that our field study consists of both an analytical skills wave and a management skills wave whereas we use just analytical skills on Prolific be a critical difference? No; even if we use only the analytical skills data

from the field, we are left with the same pattern of results.

Second, could it be sociodemographic differences across the two samples? Although we do not have the same set of observables to compare across the two populations, we do observe that the participants on UpWork have higher educational attainment on average. When we restrict the Prolific sample to include only those who report an undergraduate and/or advanced degree, our main results are largely unchanged, suggesting that it is not the education gap that drives the pattern. Although this does not rule out that some other demographic difference plays a role, we do not have a convincing hypothesis for what might vary across the samples that interacts with both our treatments and gender.

Third, could risk preferences help to reconcile the results? The Prolific job is advertised as 15 minutes, implying that an hourly wage is $20 for the intermediate job and $40 for the expert job. On Upwork, the implied hourly wages are $70 and $155, respectively. These differences in stakes across the platforms were intended to set wages that were high but not completely abnormal on the respective platforms. Of course, there are also likely differences in beliefs of receiving the position across the platforms. Thus, comparing concretely the implied "gambles" faced by participants across the settings is challenging. With that said, we think it is unlikely that differences across the two settings are driven primarily by risk aversion. It would need to be the case that men (but not women) hold risk preferences such that increasing their likelihood of being qualified does not change their application decision on UpWork, either because they are risk-loving enough to already enter, or because they are risk-averse enough to never enter. But their risk preferences also must be such that on Prolific, an increased likelihood of being qualified indeed increases both of their likelihoods of applying.[28] Although this seems unlikely, direct data on risk preferences would be necessary to formally rule out this possibility.

Finally, although we attempted to maintain parallelism across the two studies in our language and design choices, we cannot rule out that arguably minor differences in our instructions and implementation generate the differential behavior of men across the two settings. Perhaps the most pertinent language difference across the two settings was that the field study included an additional statement, in all treatments, that strong writing and communication skills were expected.[29] It seems unlikely, but not impossible, that this statement or other minor differences in the designs led men to fail to react to our treatments.

4.3. Takeaways

In some contexts, simply increasing the application rate of talented female candidates (independent of the response of men) may work to diversify the hired pool.[30] Nonetheless, better understanding the impact of these policies not just on women but on the gender gap is essential. We observe that our treatments are more effective at closing the gender gap in the field, where initial ambiguity is high, anticipated discrimination is a possibility, and the costs of rejection may loom large. Of course, these features characterize many field contexts of interest. In our view, this makes it critical for future work to better understand exactly how these factors contribute to gender gaps in application rates and precisely how they interact with ambiguity-reducing interventions.

5. Conclusion

A large literature explores the factors that contribute to gender gaps in labor market outcomes. Within this rich literature, however, supply-side decisions focused on when individuals choose to put themselves forward for different opportunities are understudied. This paper takes a step toward exploring this important question, asking whether there are gender differences in application decisions.

Across complementary contexts, we explore the extent to which men and women choose to apply for a given opportunity. In our baseline conditions, we see evidence that, on average, talented women apply at significantly lower rates than talented men despite having observably similar qualifications. In our controlled experiment, we identify further evidence that women view themselves as significantly less likely to be considered qualified for the position as compared with equally well-qualified men. This supply-side difference in willingness to put oneself forward for an opportunity has potentially important implications for labor market advancement.

We show that exogenously reducing ambiguity about the required qualifications increases the rate at which qualified candidates apply for the position. In our controlled experiment, we see this effect for both qualified men and qualified women. As a result, our treatments attract a larger pool of qualified applicants. In the field, the treatment effects are concentrated among women, working to directionally reduce a gender gap in qualified applicants. We hypothesize that larger initial ambiguity in the field may contribute to the across-study differences, but future research should more rigorously investigate this issue. To better understand the mechanism behind our results and to inform policy, we need clear evidence on the conditions that lead ambiguity reduction to not only increase the rate at which qualified women apply but also to close the gender gap. In any case, our results suggest that there may be soft-touch employer interventions that can improve the representation of women in the applicant pool in male-typed domains, helping to draw in qualified female

candidates. This seems like a promising and low-cost path to explore.

Although improving clarity around a job's qualification requirements may be an effective and feasible intervention for increasing the rate at which qualified women apply, one avenue for future work is understanding the persistence of the gender gap in contexts like our controlled study. In particular, it seems worth delving deeper into why many women who do seem to believe they are above the bar choose not to apply. In future work, it would also be useful to consider behavior in more female-typed domains to understand whether the patterns we observe generalize. It could be that it is not women in general who are less likely to apply but rather that individuals are less likely to apply in more gender-incongruent areas or areas less consistent with their identity.[31]

Of course, many hiring decisions are substantially more complicated than those studied in our experiments and may involve evaluating candidates across a range of dimensions, some qualitative and some quantitative. Our policy suggestion is most obvious to translate for quantitative dimensions: better specifying desired years of experience, minimum GRE score, number of projects completed in the past, etc. Assuming that the employer has a bar in mind for these dimensions (that is, the employer only wants to hire people above that bar), it seems that our type of intervention could be helpful. Candidates below that bar should be less likely to apply, and the employer may draw in qualified people who, for example, didn't realize that "extensive experience" meant only X years. Assuming that performance on these quantitative dimensions is not systematically negatively correlated with performance on other dimensions, better sorting on at least one dimension should weakly improve the pool of applicants. Whereas our experiments analyzed quantitative cases where it was straightforward to specify a bar, our hypothesis is that more general forms of ambiguity reduction around desired qualifications could produce similar effects.

Although extrapolating from one context to others always presents challenges, we think our results may offer useful lessons for many settings of interest. An important next step could be studying these questions in a more traditional, salaried employment setting. Even though candidates in these contexts are likely to have more experience with job application processes, one could also imagine this being a case where learning is difficult. If qualified candidates choose not to apply, they miss out not only on the job but also on the opportunity for feedback about whether they were above the bar.

Acknowledgments

The authors thank the department editor Yan Chen, the associate editor, and three anonymous referees for helpful comments and guidance.

Endnotes

[1] UpWork assigns each job to one of the following categories: Web/Mobile/Software Development, IT & Networking, Data Science & Analytics, Engineering & Architecture, Design & Creative, Writing, Translation, Legal, Administrative Support, Customer Service, Sales & Marketing, and Accounting & Consulting. Note that each of these categories has up to 83,000 subcategories. We captured self-reported capabilities at the job category level.

[2] Gender determinations were done as follows. First, we made three predictions of the gender using different methods. (i) One member of the research team predicted gender based on name and photo prior to treatment assignment. Next, we predicted gender with use of the (ii) 1990 Census (IPUMs) and (iii) 1940–1970 Social Security Administration (SSA) name files. For (ii) and (iii), a name is assigned a gender if 90 percent of individuals with the freelancer's stated name are classified as either male or female by the data source. In most cases, the three sources were in agreement. When all three methods (researcher, IPUMs, SSA) were not in agreement or when IPUMs and/or SSA produced an unclassified result, we had another researcher code the gender (blind to treatment and results, $n = 231$). In those cases, we go with the gender given by the majority of the predictors (of the two researchers, the IPUMs, and the SSA), with a minimum of two predictors having to be in agreement. Otherwise, we drop the observation ($n = 9$).

[3] Like Abraham and Stein (2022), our interventions target qualifications. Their removal of optional qualifications and descriptors lowers perceptions of "the bar" while having a less obvious impact on how much ambiguity there is about the bar. Less information may lead to more ambiguity, but it is hard to know. Conversely, our treatments seek to reduce ambiguity about the bar, moving beliefs closer to the truth. Because it is unclear how to assess the impact of their intervention on ambiguity, it is challenging to directly compare the results.

[4] We computed a "hiring score" ranging from 0 to 100 for each worker that was a function of the desired qualifications communicated to them within the job advertisement, assigning a weighted score based upon their experience (100 points if they completed any job on UpWork, 0 points if they had no UpWork experience; weight: 10%), education (as indicated by degrees held, 0 points for no stated education, 60 points for completed college education, 80 points for a Masters degree, 90 points for an MBA degree, 100 points for an MBA and another graduate degree; weight: 20%), and test score on the test of interest (their skills test score converted into a 100-point scale; weight: 70%). We made job offers to the two workers with the best hiring scores for each posting (two intermediate offers and two expert offers within each treatment, for each wave, for a total of 24 offers). Freelancers who received job offers were simultaneously told of the experiment and offered the opportunity to withdraw their data. We had no freelancers request removal; 20 of the 24 workers we made offers to accepted the job and completed it for pay. Note that only workers who applied for the expert job were eligible for the expert job; we selected the best two hiring scores within the set of workers who applied to each particular posting.

[5] In many cases, we may have that $A_0 = A$, consistent with a rejected applicant simply returning to their outside option. But we allow for the case where applying for but not receiving the position changes the payoff relative to the outside option; for instance, a candidate may have to give up their current position in order to apply for a new one. It could also be the case that A_0 incorporates psychological or reputational costs of applying for and not receiving the position.

[6] By construction, all workers in our sample have completed and displayed either the Management Skills or Analytical Skills test; what does this mean for selection into our sample? UpWork actively encourages their freelancers to complete skills test (UpWork n.d.b).

Freelancers can earn points for every addition they make to their profile. Such additions can be in the form of a profile photo, employment history, or skills tests. Freelancers who have earned enough points receive a badge ("Rising Star" or "Top Rated"). From conversations among freelancers, there seems to be some consensus that tests are valuable mostly to freelancers who are newer to the platform (UpWork Community Forum 2019); freelancers take the tests to help establish a reputation before they have completed jobs or earned ratings on the site. To the extent that we are selecting on some characteristic, this selection is the same across treatment condition. We also control for the total number of tests taken by the freelancer, capturing an intensive-margin measure of this characteristic. Unfortunately, in 2019, UpWork retired skills tests; thus, at the time of drafting the paper, we were unable to conduct a systematic comparison of workers with and without skills tests displayed.

[7] Of the 209 participants who applied for our position, 14 did not apply for strictly one job. For all 14 participants who either (i) failed to specify which of the two jobs they wished to apply for or (ii) explicitly applied to both jobs, our research team contacted them via the UpWork platform after their initial application and asked them to clarify which job they were choosing to apply for. Nine of those 14 individuals then specified one application decision (intermediate or expert). Four participants remained unspecified in their choice, and one participant remained an applicant for both jobs. We code these five workers as having applied for the intermediate job and as having applied for the expert job. Table B.4 in the Online Appendix consists of a robustness check of the results where we drop these 14 observations. The results remain directionally unchanged.

[8] The main findings are robust to using logistic regression (Online Appendix Table B.3) and to excluding the indicator variables for self-reported skills (results upon request).

[9] We note that only 14% of unqualified applicants apply for the expert job in our control group, potentially limiting our ability to identify deterrence effects.

[10] Note that prior to running this preregistered replication, but after running the UpWork study, we ran two other follow-up studies that explored beliefs of how well-qualified men and women perceive themselves to be and decisions to put themselves forward for promising opportunities. These studies are detailed in full in Online Appendix C. In the first of these follow-ups, we used real job advertisements to collect men's and women's perceptions of how well-qualified they feel for different positions and how these perceptions vary with different features of the ad. We found that women feel less well-qualified for positions on average, and this gap is smaller for advertisements with more clearly stated qualifications. In the second follow-up, we constructed a simulated labor market with an MTurk sample. We observed application decisions to a "promotion" opportunity and beliefs of the likelihood of being promoted conditional on applying. We found a gender gap in beliefs of probability of promotion that was reduced when more clearly stated qualifications were used. However, we found no significant gender differences in application decisions. The UpWork replication study reported below was designed to address the shortcomings of these other studies. We preregistered the decision to report this replication in the main text while moving the other studies to the Online Appendix.

[11] The education question on the Prolific screening survey reads, "*Which of these is the highest level of education you have completed?*" Approval rating measures what percentage of completed studies have been accepted by the researcher and is an indicator for how good the quality of the participant's work has been as judged by previous researchers or study issuers.

[12] This inclusion criterion is a common practice in online experiments, intended to improve expected data quality by screening out individuals without a proven track record of successful past participation in studies.

[13] In theory, participants know that all other participants also fulfill these criteria because the eligibility criteria are common knowledge. However, this is not made salient to them during the study.

[14] That is, we determine the set of qualified applicants who apply for the expert job, then select 1% of those applicants to hire for the expert job. Similarly, we determine the set of qualified applicants who apply for the intermediate job, then select 1% of those applicants to hire for the intermediate job. This procedure minimizes budget expenditure by limiting hires, without creating the impression that only the very top tier of candidates has a chance of being hired.

[15] Participants who indicated that they did not want to apply for either of the two jobs were later asked about the reason for their decision. Of our sample, 200 applied for neither job (103 men and 97 women). Of those 200, 38 indicated that they didn't have the time or interest to complete any of the jobs (23 men and 15 women), 20 indicated that the jobs didn't pay sufficiently well (16 men and 4 women), 130 indicated that they didn't think they were qualified enough (55 men and 75 women), and 12 indicated that the reason for not applying was not listed (9 men and 3 women). In line with our preregistration, our final sample excludes participants who applied to neither, except for the 130 participants who answered, "*I do not think I am qualified for either job.*"

[16] Note that, as part of our study, we asked participants about their gender identity; we use this response as gender in our data set. We intentionally did not use the participant's sex as indicated on Prolific. After a TikTok video went viral in June 2021, female participation on Prolific increased dramatically, leading to a gender imbalance in the participant pool. It has been speculated that researchers' attempts to recruit balanced samples by restricting female signups may have led some women to misreport their sex on their Prolific profile for eligibility purposes.

[17] We exclude 68 observations from individuals who did not self-report their gender as either man or woman, 15 observations from those who failed our attention check, and 70 observations from those who chose to apply for neither job for reasons unrelated to believed qualifications. These exclusions follow our preregistration plan. In addition, beyond these preregistered exclusions, we excluded four participants for whom we were unable to verify and match their reported Prolific ID.

[18] The balance table also reveals that fewer participants were randomized into the positive treatment. This was an unintended imbalance. It appears to have resulted from Qualtrics randomization, which was not constrained to evenly assign treatments.

[19] In the Online Appendix, we present the robustness checks of Table 2. We reproduce our results using logistic regressions instead of linear probability models (see Online Appendix Table B.7). We also re-run the specifications of Table 2 on a restricted sample of participants with scores "close" to the threshold, scores of either 5 or 6 (see Online Appendix Table B.8). As one might expect, we find directionally larger treatment effects among this subsample, but the change does not change the estimated impacts on gender gaps.

[20] Note that the analysis in Table 3 was not preregistered but was added in response to a helpful referee suggestion.

[21] It becomes indistinguishable from zero among unqualified applicants (Column I). It is reduced by roughly one-third for qualified applicants (Column III) but remains statistically significant at eight percentage points.

[22] One participant stated a required bar of "75." We treat this guess as missing in our analysis of beliefs of the bar.

[23] The remaining heterogeneity in beliefs in our treatment conditions could be driven by a variety of factors. For instance, some participants may not have fully understood the hiring rule, the advertisement, or specific beliefs question we asked. More subtly, participants could have had different beliefs of what the realized

distribution of scores would be among applicants, a function of the believed distribution of performances on the test and application decisions among participants.

[24] Alston (2022) provided evidence from a simulated labor market that, indeed, individuals anticipate gender discrimination in hiring in male-typed domains.

[25] On Prolific, we informed the participant that the resume items are "the only information we will consider in our hiring decisions" and that "we will consider all applicants that apply to the [intermediate-level job/expert-level job] and select the individuals who are qualified for the [intermediate-level/expert-level job]. We will then choose 1 percent of these qualified applicants to complete the job." The full language can be found on page 20 of the instructions.

[26] Our treatments can operate in two ways: drawing in more overall applicants or shifting the proportion of applicants from the intermediate to the expert job. We see some evidence of both channels for women. Our treatments increased the rate at which qualified women applied for either opening, intermediate or expert (from 11% in the control to 21% in the positive treatment and 36% in the normative treatment). In the control, only half of qualified women who applied chose the expert job; this was the same in the positive treatment (50%) but rose to 80% in the normative treatment. The rate at which qualified men applied to either opening varied less clearly with treatment (proportion of qualified men who applied for either opening: 24% in control, 33% in positive, 24% in normative). And if anything, our treatments seemed to decrease the proportion of qualified men who chose the expert job conditional on applying: 92% in control, 69% in positive, 79% in normative).

[27] It may also be the case that qualified men on UpWork have more lucrative outside options than qualified women. Although we aimed to limit this concern by offering highly competitive pay for our expert-level job (in excess of the profile-reported hourly rate for 98% of our sample), we cannot rule out that differences in outside options contribute to these patterns. Note that differences in outside options are even less likely to play a role on Prolific, given that individuals had the explicit option to opt out of applying for either job if they felt it did not pay sufficiently well.

[28] Holt and Laury (2002) found that risk aversion increases as stakes increase. Interestingly, they also found that whereas women make more risk-averse choices than men in lower-stakes conditions, these differences are reduced under higher stakes. Applying their results to exactly our setting is challenging, but it seems suggestive that, if anything, risk preferences between men and women might be more similar under the UpWork stakes.

[29] Because we have no credible way to assess writing and communication skills from the limited resume constructed during the Prolific study, we eliminated this sentence from the design.

[30] This would be the case if, for instance, a firm felt as though they simply did not receive enough applications from talented women to diversify their hiring. A firm that would have hired N individuals could hire $N + Y$, with Y representing the number of new women that are hired from the additional qualified women that apply, or that firm could continue to hire N individuals but replace a man with an equally well-qualified women from the set of new applicants. In both of these cases, where a firm is specifically looking to target qualified women, increasing the rate of female applicants may be enough to diversify hires. If, on the other hand, the firm is expected to hire proportionally from the pool of qualified applicants, holding fixed the number of hires, then increasing the rate at which talented women apply without closing the gender gap will fail to diversify the set of hired candidates. Similarly, if firms are more likely to hire qualified men than qualified women, a policy that increased the number of qualified men that applied, even if it also increased the number of qualified women that applied, could even reduce the gender diversity of the hired pool. Understanding more about hiring patterns, given a particular applicant pool, is critical for understanding the ultimate impact of these types of policies on employment and advancement.

[31] In a field experiment in rural India, Oh (2022) found that, indeed, identity—in the form of caste—shapes labor supply decisions.

References

Abraham L, Stein A (2022) Words matter: Experimental evidence from job applications. Working paper, RAND Corporation, Santa Monica, CA.

Alston M (2022) The (perceived) cost of being female: An experimental investigation of strategic responses to discrimination. Working paper, Department of Economics, Florida State University, Tallahassee, FL.

Altonji JG, Blank RM (1999) Race and gender in the labor market. Ashenfelter OC, Card D, eds. *Handbook of Labor Economics*, Chapter 48, vol. 3 (Elsevier, Amsterdam), 3143–3259.

Bohnet I, van Geen A, Bazerman M (2016) When performance trumps gender bias: Joint vs. separate evaluation. *Management Sci.* 62(5):1225–1234.

Bordalo P, Coffman KB, Gennaioli N, Shleifer A (2019) Beliefs about gender. *Amer. Econom. Rev.* 109(3):739–773.

Bowles HR, Babcock L, McGinn K (2005) Constraints and triggers: Situational mechanics of gender in negotiation. *J. Personality Soc. Psychol.* 89(6):951–965.

Buser T, Yuan H (2019) Do women give up competing more easily? Evidence from the laboratory and the Dutch math olympiad. *Am. Econ. J. Appl. Econom.* 11(3):225–252.

Coffman KB (2014) Evidence on self-stereotyping and the contribution of ideas. *Quart. J. Econom.* 129(4):1625–1660.

Coffman KB, Collis MR, Kulkarni L (2023) Stereotypes and belief updating. *J. Eur. Econom. Assoc.*, Forthcoming.

Coffman KB, Exley CL, Niederle M (2021) The role of beliefs in driving gender discrimination. *Management Sci.* 67(6):3551–3569.

Croson R, Gneezy U (2009) Gender differences in preferences. *J. Econom. Lit.* 47(2):448–474.

Del Carpio L, Guadalupe M (2022) More women in Tech? Evidence from a field experiment addressing social identity. *Management Sci.* 68(5):3196–3218.

Delfino A (2022) Breaking gender barriers: Experimental evidence on men in pink-collar jobs. Working paper, Bocconi University, Milan.

Dubey A, Abhinav K, Hamilton M, Kass A (2017) Analyzing gender pay gap in freelancing marketplace. In *Proceedings of the 2017 ACM SIGMIS Conference on Computers and People Research*. SIGMIS-CPR '17 (ACM, New York), 13–19.

Exley C, Kessler J (2022) The gender gap in self-promotion. *Quart. J. Econom.* 137(3):1345–1381.

Fernandez RM, Mors ML (2008) Competing for jobs: Labor queues and gender sorting in the hiring process. *Soc. Sci. Res.* 37(4): 1061–1080.

Flory JA, Leibbrandt A, List JA (2015) Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *Rev. Econom. Stud.* 82(1):122–155.

Flory JA, Leibbrandt A, Rott C, Stoddard O (2021) Increasing workplace diversity: Evidence from a recruiting experiment at a Fortune 500 company. *J. Hum. Resour.* 56(1):73–92.

Foong E, Vincent N, Hecht B, Gerber EM (2018) Women (still) ask for less: Gender differences in hourly rate in an online labor marketplace. *Proc. ACM Hum.-Comput. Interact.* 2(CSCW):1–21.

Gee LK (2018) The more you know: Information effects on job application rates in a large field experiment. *Management Sci.* 65(5): 2077–2094.

Ginther DK, Kahn S (2009) Does science promote women? Evidence from academia 1973–2001. Freeman RB, Goroff DL, eds. *Science and Engineering Careers in the United States: An Analysis of*

Markets and Employment. A National Bureau of Economic Research Conference Report (University of Chicago Press, Chicago).

Heilman ME (2001) Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *J. Soc. Issues.* 57(4):657–674.

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92(5):1644–1655.

Ibarra H, Carter NM, Silva C (2010) Why men still get more promotions than women. *Harvard Bus. Rev.*, September 1, 2010. https://hbr.org/2010/09/why-men-still-get-more-promotions-than-women.

Kuhn P, Shen K, Zhang S (2020) Gender-targeted job ads in the recruitment process: Facts from a Chinese job board. *J. Dev. Econom.* 147(November):102531.

Murciano-Goroff R (2021) Missing women in tech: The labor market for highly skilled software engineers. *Management Sci.* 68(5): 3262–3281.

Niederle M (2016) Gender. John K, Roth AE, eds. *The Handbook of Experimental Economics 2* (Princeton University Press, Princeton, NJ), 481–562.

Niederle M, Vesterlund L (2007) Do women shy away from competition? Do men compete too much? *Quart. J. Econom.* 122(3):1067–1101.

Niederle M, Yestrumskas AH (2008) Gender differences in seeking challenges: The role of institutions. Working paper, Stanford University, Stanford, CA.

Oh S (2022) Does identity affect labor supply? CESifo Working Paper No. 9487, Paris School of Economics, Paris.

Reuben E, Sapienza P, Zingales L (2014) How stereotypes impair women's careers in science. *Proc. Natl. Acad. Sci. USA.* 111(12): 4403–4408.

Riach PA, Rich J (2002) Field experiments of discrimination in the market place. *Econom. J. (Lond.).* 112(483):F480–F518.

Samek A (2019) A university-wide field experiment on gender differences in job entry decisions. *Management Sci.* 65(7):3272–3281.

Sczesny S, Nater C, Eagly AH (2018) Agency and communion: Their implications for gender stereotypes and gender identities. Abele A, Wojciszeke B, eds. *Agency and Communion in Social Psychology* (Routledge, London), 103–116.

Shastry GK, Shurchkov O, Xia LL (2020) Luck or skill: How women and men react to noisy feedback. *J. Behav. Exp. Econom.* 88:101592.

Skills Tests. n.d. UpWork Help Center. Accessed February 4, 2019, http://support.UpWork.com/hc/en-us/articles/211063198-Skills-Tests; Permalink: https://perma.cc/B2UA-KBV8.

UpWork n.d.a UpWork. Accessed December 5, 2018, https://www.UpWork.com.

Upwork n.d.b Upwork Skill Certification. https://support.upwork.com/hc/en-us/articles/360052581373-Upwork-Skill-Certifications; Permalink: https://perma.cc/92CF-HPBD.

UpWork Community Forum (2019) Removing skill tests on July 16th. June 21, 2019. https://community.UpWork.com/t5/Announcements/Removing-Skill-Tests-on-July-16th/m-p/609790#M31055; Permalink: https://perma.cc/K4S2-YJL6.

Zahidi S, Ibarra H (2010) The corporate gender gap report 2010. World Economic Forum, Geneva.