WHETHER TO APPLY

Katherine B. Coffman*
Manuela R. Collis
Leena Kulkarni

March 2022

Abstract: Labor market outcomes depend, in part, upon an individual's willingness to put 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 more mixed for men. Our results suggest a path for increasing the pool of qualified women applicants.

---

*Corresponding author: kcoffman@hbs.edu 445 Baker Library, Harvard Business School, Boston, MA. MC: manuela.collis@rotman.utoronto.ca . LK: lkulkarni@hsph.harvard.edu .

Acknowledgements: This work was funded by the National Science Foundation and Harvard Business School.

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

[Silence]

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

I. 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, Carter, and Silva 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 (Bohnet, van Geen, and Bazerman 2016; Reuben, Sapienza, and Zingales 2014; Coffman, Exley, and Niederle 2021). Once a female worker is hired, she is subject to bias in both formal job evaluation processes (Heilman 2001) and in more informal mentoring (Ibarra, Carter, and Silva 2010). Firms are devoting significant attention to reducing these biases, with the hope of achieving greater gender diversity throughout their ranks.

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 be a factor. 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.

Outside of controlled experiments, many of these decisions about whether to apply may be made in the face of (anticipated) bias, making it hard to isolate the role for willingness to apply from bias. We take advantage of controlled frameworks to separate between these stories, allowing us to focus on the role of candidate perceptions and decisions, absent employer biases. We ask whether women are as likely as men to see themselves as qualified for challenging, higher-paying positions, and whether they apply at similar rates conditional on their degree of qualification.

Past literature on gender differences 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 to men, both in objective terms (Niederle and Vesterlund 2007, Coffman 2014; Bordalo et al. 2019) and subjective terms (Exley and Kessler 2020). In the field, Murciano-Goroff (2020) finds that conditional on having the same level of skill, female software engineers are less likely to self-report that skill on their resume compared to men. This suggests that even if men and women both have the same skills and share the same view as to what is required to qualify for a given position, women may be less likely to believe they possess that qualification (holding all else equal). 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 male and female job-seekers' probability of application. Consistent with the factors mentioned above, Flory, Leibbrandt, and List (2015) find 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) finds that competitive compensation schemes deter women more so than men. Gee (2018) finds that showing job-seekers the number of other applicants increases applications from women more than men, showing that providing social information can play a role in the decision to complete an application. Female role models can also influence application decisions: Del Carpio and Guadalupe (2020) find that women are more likely to opt into a tech skills training program when presented with an example of a female success story. Finally, Kuhn, Shen, and Zhang (2020) use data on a large job ad platform and find that an explicit request for women to apply increases the number of women who apply.

Closest to our work is a simultaneous project by Abraham and Stein (2020) exploring how the language used in job postings impacts the application behavior of men and women. In a large, randomized control trial, they vary the language around how demanding and intense the required qualifications for a given position are. In particular, their treatment "softens qualifications," removing optional qualifications from the posting and using less demanding language for remaining qualifications. They find that when qualifications are softened, more individuals apply, and it reduces the skills gap between male and female applicants. That is, in comparison to the control treatment, it is no longer case that the female applicants that apply are significantly more skilled than the male applicants.

Our paper builds on this body of important 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. Our main hypothesis is that uncertainty surrounding whether or not an individual is qualified for

an opening may produce gender gaps in willingness to apply. The idea of ambiguity as a driver of gender gaps has been proposed in other contexts, including in negotiation. Bowles, Babcock, and McGinn (2005) show that reducing situational ambiguity, for instance, 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 own qualification level and decisions to put oneself forward for different opportunities.

Consider an individual deciding whether or not to apply for an opening; she may ask 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 her own aptitude (what are my skills, strengths, and talents), but also on 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. 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 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, pre-registered replication study on Prolific, in an attempt to better understand the mechanisms at work behind our results. In this more controlled setting, we also find that qualified women are significantly less likely than qualified men to apply in our baseline condition. And, 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. As a result, the treatments do not reduce the gender gap in application rates.

The follow-up experiment also provides additional insights into the mechanisms behind our results. We find 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 to 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, we find that women view themselves as significantly less likely to receive a promotion conditional on applying compared to 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 pre-registered 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 Appendix C.

Together, our results suggest that talented women may be less likely to put themselves forward for opportunities compared to equally qualified men. Across both our field experiment and our controlled replication, we find that reducing the amount of 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.

II. 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.). Upwork facilitates match-making 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 or not 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.

In addition, freelancers have the opportunity to take standardized tests of their skills and aptitudes in different domains. Upwork offers hundreds of so called "Skills Tests" for free with the topics of those tests 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. For each test, Upwork provides information on the number of freelancers who have already taken this test, and their corresponding scores. These tests essentially serve as verified evaluations of capabilities, and freelancers have the option of displaying the results of these tests 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 experiment) or the Analytical Skills Test (Wave 2 of 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 choose to study male-typed domains to better proxy the male-typed environments that have been historically characterized by female under-representation and lack of advancement (i.e. business, STEM).

We started 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 experiment) or the Analytical Skills Test (Wave 2 of experiment). This gives us a pool of workers that have completed a test of interest. We compiled the profile information for each of the Upworkers in this pool, creating a dataset 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/wk, less than 30, 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 they have chosen to display on their profile.[1] On top of that, we enter into the dataset an indicator of freelancer gender.[2]

The dataset 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, but workers have discretion over whether to display their score. Only workers who choose to display their score appear in our dataset. We also record the number of minutes it took the freelancer to complete the test, made available by Upwork alongside the worker score. Freelancers are able to re-take 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 dataset, we reach out to each worker in our dataset 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 level job" while the other job represents the more challenging but also better compensated "expert level job". Both jobs require writing essay-style answers to two questions and are advertised to take one hour. We offer pay of $70 for the intermediate job and $150 for the expert job.

All participants are presented with both options and are invited to apply to either of the two jobs (the worker 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

[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 sub-categories. 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).

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 level job. The job description states that "we invite applicants with a [Management / Analytical] Skills test score of [3.75 / 4.05] to apply for the expert-level job." Freelancers who were interested in our positions were able to contact us to apply through the Upwork website. We then made hiring decisions using a pre-determined algorithm that assigned weights to our listed desired qualifications.[3]

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. While 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).

Both treatments increase the objectivity, specificity, and clarity of desired qualifications for the expert level job relative to the control. While the positive treatment simply describes the qualification, the normative treatment takes things a step farther: it explicitly encourages candidates with the qualification to apply. A candidate (with a test 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. In this way, the normative treatment may be a more aggressive intervention relative to the positive treatment.

A few features of our design are worth noting. First, we reach out to workers rather than simply post the jobs in order to boost response rates, increasing the extent to which our ads are visible to workers and ensuring unique and random assignment to treatment. Each freelancer in our dataset is able to view and apply to a job for exactly one of the three treatments.

[3] 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 have 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 receive job offers are 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 to the expert-level job were eligible for the expert-level job; we selected the best two hiring scores within the set of workers who applied to each particular posting.

Second, we chose 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 level job and an expert level job, and use the decision to apply to the expert level 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 level or intermediate level job.

Finally, we had to 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 our participants 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.

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 2020). 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 mostly valuable to freelancers who are newer to the platform (Upwork Community 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 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.

Results

Table B1 in the Appendix provides descriptive statistics of the freelancers in our sample. Men and women vary in many dimensions in our sample. 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, though we should caution that Murciano-Goroff (2020) finds 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 qualification tests – the Management Skills test (male mean 3.55, female mean 3.42, p-value from two-tailed t-test < 0.01) and the Analytical Skills test (male mean 3.73, female mean 3.57, p-value from two-tailed t-test < 0.01). And, a greater fraction of men than women are qualified for our expert level 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).

Overall, 20% of men and 18% of women in our sample apply to one of our job postings.[4] This aggregate rate is relatively constant across the three treatments, with 20% of men and 19% women applying in the control, 21% of men and 18% of women applying in the Positive treatment, and 20% of women and 17% of women applying in the Normative treatment. Of the 209 participants who apply to our job postings, most (130) apply to the expert-level 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 1, Panel 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-level job, compared to 22% of qualified men. The fraction of all qualified men who apply to the expert level job is quite steady across treatment, ranging from 19 – 23%. The reduced ambiguity of our treatments does not draw in additional qualified men; qualified men apply at similar rates regardless of how much information they have 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 1, Panel B shows that application rates to the expert job among unqualified candidates are low, with no clear systematic differences by treatment or gender.

[4] Of the 209 participants who apply to our position, 14 did not apply to strictly one job. For all 14 participants who either (i) failed to specify which of the two jobs they wished to apply to, 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 to. Nine of those 14 individuals specified one application decision (intermediate or expert level). Four participants remained unspecified in their choice and 1 participant remained an applicant for both jobs. We code these five workers as having applied to the intermediate job *and* as having applied to the expert job. Table B2 in the Appendix consists of a robustness check of the results where we drop these 14 observations. The results remain directionally unchanged.

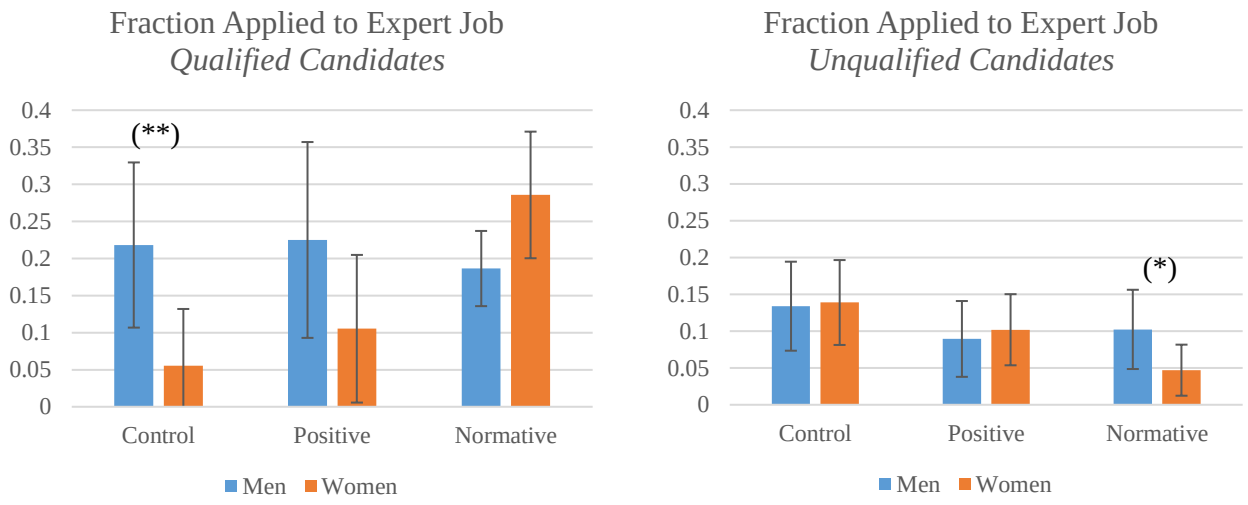

Figure 1. Proportion of Freelancers who Apply to Expert Job, By Qualification Level

(*) indicates p<0.10, (**) indicates p<0.05, (***) indicates p<0.01 from test of proportions comparing men and women within treatment

In Table 1, we predict the decision to apply to the expert level job from treatment assignment, using the control treatment as our reference category. We control for all profile information included in our summary statistics table.[5] We start by analyzing 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 relative to the control treatment, both the positive and normative treatments decrease application rates among unqualified candidates, while increasing 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 - V, we analyze the decisions of unqualified candidates (those with test scores less than the threshold). Overall, we find that both qualification treatments decrease application rates to the expert-level job (Column IV). 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). However, in an interacted model, we cannot reject that the deterrence effect is of a similar size for men and women (Column V).

[5] Results are unchanged if we exclude the indicator variables for self-reported skills.

Table 1. Application Rates to Expert Job in the Field

	OLS Predicting Decision to Apply to Expert Job						
	All Participants			All Unqualified		All Qualified	
	I	II	III	IV	V	VI	VII
Positive	-0.026	-0.043	-0.039	-0.046*	-0.057	0.044	0.0067
Treatment	(0.024)	(0.027)	(0.035)	(0.026)	(0.038)	(0.061)	(0.081)
Normative	-0.030	-0.067**	-0.033	-0.070***	-0.044	0.098	-0.00076
Treatment	(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 x		0.066					
Qualified		(0.057)					
Normative x		0.15***					
Qualified		(0.056)					
Female x			0.024		0.020		0.10
Positive			(0.049)		(0.052)		(0.12)
Female x Normative			0.0050		-0.047		0.28**
			(0.048)		(0.052)		(0.12)
Controls	Y	Y	Y	Y	Y	Y	Y
Observations	1083	1083	1083	827	827	256	256
Adj. R-squared	0.035	0.039	0.034	0.037	0.037	0.012	0.026

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 hrs/wk, < 30 hrs/wk, as needed), dummies for each self-reported skill, and a dummy for being in the second wave of experiment. * indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

In Columns VI - VII, we focus on qualified candidates. Overall, we estimate that our two qualifications treatments directionally increase the rate at which qualified candidates apply to our expert-level job.

Consistent with Figure 1, we estimate that 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.

In this setting, the normative treatment has a directionally larger impact than the positive treatment, though we should note that the differences are not statistically significant. 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) document 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.

The results of our field experiment suggest that reduced ambiguity about where the bar is can draw in more qualified female candidates, narrowing the gender gap in applications among qualified candidates. We now turn to a more controlled setting, where we can elicit more detailed data from our participants, in an attempt to shed further light on this result.

III. Replication in a Controlled Environment

We conducted 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 pre-registered this study, and we collected the data in December 2021 on Prolific (AEARCTR-0008223).[6]

[6] Note that prior to running this pre-registered 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 Appendix C. In the first of these follow-ups, we use 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 find that women feel less well-qualified for positions on average, and this gap is smaller for advertisements with more clearly stated

Design

We conduct a simple experiment during which 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 to 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 to which the answers are hard to "Google" quickly. Further, they cover stereotypically male-typed domains, matching our field setting. 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 percent.[7] This ensures that there is no study-specific reporting bias.

---

qualifications. In the second follow-up, we construct a simulated labor market with an MTurk sample. We observe application decisions to a "promotion" opportunity and beliefs of the likelihood of being promoted conditional on applying. We find a gender gap in beliefs of probability of promotion that is reduced when more clearly stated qualifications are used. However, we find no significant gender differences in application decisions. The UpWork replication study reported below was designed to address shortcomings of these other studies. We pre-registered the decision to report this replication in the main text while moving the other studies to the Appendix.

[7] 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.

Note, that by design, only two factors vary across resumes: skills 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%. Thus, while 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.[8] 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 provide participants with additional information about the two short-term jobs available to them. We use near-identical language to our Upwork field experiment. One job is an "intermediate level job" while the other job represents the more challenging but also better compensated "expert level job". Both jobs require writing essay-style answers to one question and are advertised to take 15 minutes. We offer pay of $5 for the intermediate job and $10 for the expert job. We also provide information on how hiring decisions will be made. In particular, we tell participants that we will first screen the pool of applicants to each job and determine the set of qualified applicants. Then, we will randomly hire one percent of the set of qualified applicants.[9]

We outline both job opportunities, then participants decide. Participants can apply to one of the jobs or have the option to explicitly apply to neither.[10] Participants receive generic information on desired characteristics of a successful applicant for the expert job that reads 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 view their resume. They see their test score, indicators that they have completed at least 100 Prolific studies and have an approval rate of 95 percent or greater, and are told that we will also consider their education as listed on their Prolific profile. We are explicit that these pieces of information are the only factors that will determine hiring decisions.

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

[9] That is, we determine the set of qualified applicants who apply to the expert-level job, then select 1% of those applicants to hire for the expert-level job. Similarly, we determine the set of qualified applicants who apply to the intermediate-level job, then select 1% of those applicants to hire for the intermediate-level 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.

[10] Participants who indicate that they don't want to apply to either of the two jobs are later asked about the reason for their decision. Of our sample, 200 apply to neither job. Of those 200, 38 indicated that they don't have the time or interest to complete any of the jobs, 20 indicated that the jobs don't pay sufficiently well, 130 indicate that they don't think they are qualified enough, and 12 indicate that the reason for not applying is not listed. In line with our pre-registration, our final sample excludes participants who apply to neither, except for the 130 participants who answered "*I do not think I am qualified for either job.*"

We randomly assign each worker to one of three treatments. In our control treatment, participants see no additional information on the desired qualifications. In our positive treatment, participants receive 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 receive a prescriptive statement about whether to apply for the expert level 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 from the screening test questions that indicated 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 allows us to ask several follow-up questions that speak to mechanisms. In particular, we elicit participants' beliefs of how well-qualified they are, of how high the bar is for the expert-level job, and of how objective, specific, and clear the required qualifications for the expert-level job were.

The first set of post-application decision questions ask the participant about their perceived probability that they would be considered qualified for the expert-level job. This question does not depend upon whether or not they chose to apply to the job. We simply ask 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-level job.

Next, we ask them what they expect will be the lowest skills test score among hired candidates for the expert job. This question assesses their beliefs of what the bar is. We also ask them how confident they feel about that guess, where they use a 1 – 5 scale to indicate not at all sure to completely sure.

Finally, participants assess how objective, specific, and clear the required qualifications for the expert-level job opportunity were. They indicate their answer on a 1 – 6 scale. Once they answer that question, we provide them with the language used in the other two treatment conditions and ask 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 is assigned to the *Normative* treatment version first provides their assessment of their job ad. Then, on the following page, they see the language used for the *Control* version and the language used for the *Positive* version, and they indicate how objective, specific, and clear they find each of these two other ads. This provides us with both across-subject and within-subject evaluations of how the clarity of the bar varies across our three treatment conditions. Once participants complete the main portion of the survey, they answer a brief sociodemographic questionnaire.

We conducted the study on Prolific.co and advertised it as a 10-minute study with a completion fee of USD 1.85 and the opportunity to earn up to USD 1.50 in additional pay. The actual median hourly pay (additional pay included) is USD 17.75 as measured by the time spent on our Qualtrics survey. We restrict the pool of participants to individuals who reside in the United States, are age 18 or older, are fluent in English, completed 100 or more Prolific studies, and have an approval rating on Prolific of 95%. We incorporated understanding questions, an attention check, and an open question.[11] Full instructions are provided in Appendix A.

Results

We collected 2,400 observations. After exclusions, our final sample size is 2,243; the distribution of participants across cells can be found in Table 2.[12,1314] Appendix Table B3 presents summary statistics for our participants, both overall and conditional on being "qualified." In line with our field study (and our pre-registration), we define qualified as having an ASVAB score strictly above the threshold used in our treatments, 5.5.

Table 2. Distribution of Participants across cells

	Qualified Men	Qualified Women	Unqualified Men	Unqualified Women	Totals
Control	234	208	147	184	773
Positive	201	199	142	162	704
Normative	248	186	156	176	766
Totals	683	593	445	522	2,243

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

[11] This is done to screen out bots.

[12] Note, that as part of our study, we ask participants about their gender identity; we use this response as gender in our dataset. We intentionally did not use the participant's sex as indicated on Prolific.co since doubts about that screening question's validity have been raised. After a TikTok video went viral in June 2021, female participation on Prolific increased dramatically, leading to a gender imbalance in the participant pool. Researchers' attempts to recruit balanced samples by restricting female sign-ups may have led some women to intentionally mis-report their sex on their Prolific profile to increase their eligibility for studies.

[13] 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 to neither job for reasons unrelated to believed qualifications. These exclusions follow our pre-registration plan. In addition, beyond these pre-registered exclusions, we exclude 4 participants for whom we were unable to verify and match their reported Prolific ID.

[14] This was an unintended imbalance. It appears to have resulted from Qualtrics randomization, which was not constrained to evenly assign treatments.

on average and women answering 5.64 questions correctly on average ($p<0.001$); this gap is 7.21 to 6.99 among the qualified participants ($p<0.001$). In our regression analysis, we include control variables to capture these differences.

Figure 2, Panel A presents the raw data on application rates to the expert-level 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-level job. While 56% of qualified men apply, only 42% of qualified women apply. Recall that we observed a similarly large gender gap in application rates in the Control treatment of our UpWork experiment.

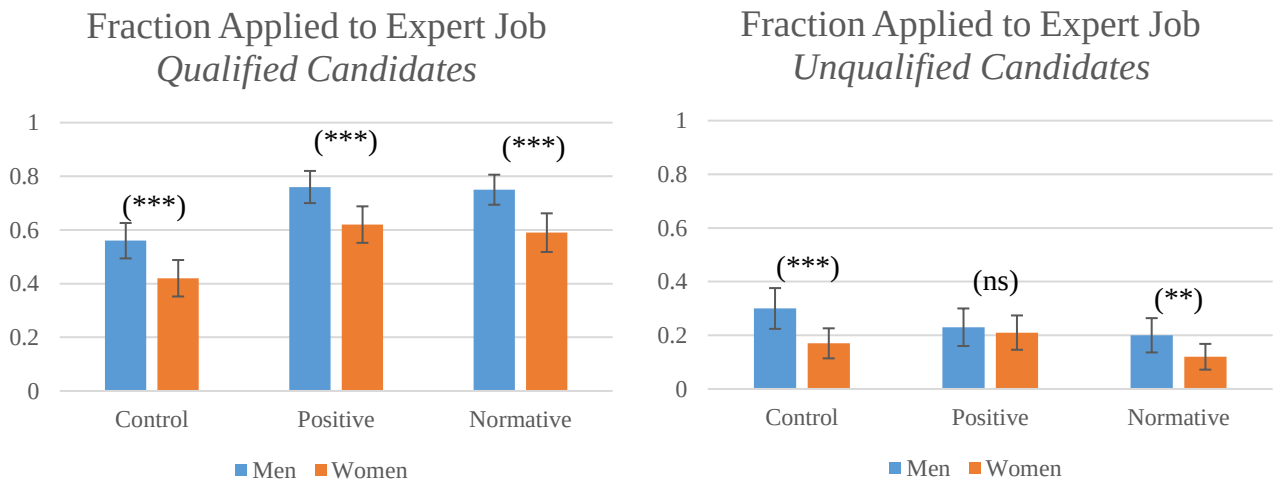

Figure 2. Proportion who Apply to Expert Job in Prolific Study, By Qualification Level
(*) indicates $p<0.10$, (**) indicates $p<0.05$, (***) indicates $p<0.01$ from test of proportions comparing men and women within treatment

Both the Positive and the Normative treatments significantly increase the proportion of qualified women who apply to the expert-level job, bumping up the proportion to 62% ($p<0.001$ versus control) and 59% ($p<0.001$ versus control), respectively. Similar to what we found on UpWork, clearer, more objective information about qualifications indeed draws in more qualified women. On UpWork, this increase in the rate at which qualified women applied narrowed the gender gap in our treatments, as the behavior of men was essentially unchanged. But, in this sample, we observe that men *also* respond to our treatments. The Positive and Normative treatments increase the proportion of qualified men who apply, to 76% ($p<0.001$ versus control) and 75% ($p<0.001$ versus control), respectively. Because men and women are responding similarly to the treatments, there is no change in the gender gap across the treatments. Instead, there are simply more qualified applicants (both men and women) to the expert-level job.

Figure 2, Panel B presents the evidence for unqualified candidates. Application rates to the expert-level job are generally low among unqualified candidates. In our Control treatment, we observe that unqualified men are significantly more likely to apply than unqualified women: 30% versus 17%. Our treatments somewhat decrease the rate at which unqualified men apply. The treatments have little impact on the application decisions of unqualified women.

Table 3 predicts the decision to apply from treatment, controlling for participant characteristics, paralleling Table 1. Column I reveals that, overall, women are 10 percentage points less likely to apply to the expert-level job ($p<0.01$), and that our treatments, through their large impacts on qualified candidates, increase the rate at which individuals apply to the expert-level job. Column II confirms that the treatments are pulling in qualified candidates significantly more than they are pulling in unqualified applicants. Overall, there is no significant interaction of either treatment with gender (Column III). Zooming in on unqualified candidates, we see that the Normative treatment significantly reduces the rate at which unqualified applicants apply, while the Positive treatment has no impact (Column IV). Column V suggests that the deterrence effects of the treatments on unqualified candidates are not significantly different by gender.

We estimate that our treatments have a large impact on qualified candidates. Column VI estimates that qualified candidates are 20 percentage points more likely to apply in the Positive and Normative treatments than they are 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 reduces what is a meaningful gender gap in application rates among qualified candidates in the Control of 12 percentage points ($p<0.01$).

One key reason for conducting the follow-up experiment was to gather more information on the potential mechanisms behind the gender gap in application rates and the impact of our treatments. Following their application decision, we ask participants to estimate the likelihood that they would be considered qualified for the expert-level job, conditional on the resume we consider. Table 4 estimates this believed likelihood conditional on gender, treatment, and observables. Columns I and II consider unqualified candidates. Unqualified women believe they are 7 percentage points 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 6 percentage points less likely to be qualified than similarly qualified men (Column III, $p<0.01$). While our 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. If we correlate the decision to apply with believed likelihood of being qualified, we estimate a correlation of 0.53.

Table 3. Applications to Expert-Level Job on Prolific

	OLS Predicting Decision to Apply to Expert Job						
	All Participants			All Unqualified		All Qualified	
	I	II	III	IV	V	VI	VII
Positive	0.098***	-0.023	0.079**	-0.015	-0.068	0.20***	0.19***
Treatment	(0.023)	(0.035)	(0.033)	(0.031)	(0.047)	(0.032)	(0.044)
Normative	0.075***	-0.081**	0.081**	-0.077**	-0.10**	0.20***	0.19***
Treatment	(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		0.043					
Indicator		(0.042)					
Positive x		0.22***					
Qualified		(0.046)					
Normative x		0.28***					
Qualified		(0.045)					
Female x			0.038		0.097		0.011
Positive			(0.046)		(0.063)		(0.064)
Female x Normative			-0.012		0.045		0.0030
			(0.045)		(0.062)		(0.063)
Controls	Y	Y	Y	Y	Y	Y	Y
Observations	2243	2243	2243	967	967	1276	1276
Adj. R-squared	0.211	0.238	0.210	0.035	0.035	0.112	0.111

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. * indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

We also ask participants their beliefs about what "the bar" is for the expert-level job in terms of minimum required ASVAB score. The full histograms of beliefs by gender and treatment are presented in Figure 3 below.[15] We do not see gender differences, suggesting 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 estimate the bar to be a score of 7.2 on average, and women estimate 7.3 on average. The modal answer in the Control treatment is 8. Our treatments significantly lower beliefs of the bar among both men and women, to

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

approximately 6.6/6.5 for men/women in the Positive treatment and 6.5 for men and women in the Normative treatment, with modes of 6 in both treatments, reflecting our communicated information. Beliefs are also less variant in our treatment conditions: the standard deviation in the Control is 1.66, while the standard deviation in the Positive and Normative treatments is 1.46 and 1.43, respectively.

Table 4. 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	-2.35	-3.95	10.2***	10.3***
Treatment	(1.97)	(2.91)	(1.80)	(2.51)
Normative	-2.81	-4.36	11.4***	9.73***
Treatment	(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 x		2.93		-0.22
Positive		(3.94)		(3.61)
Female x Normative		2.82		3.70
		(3.86)		(3.56)
Controls	Y	Y	Y	Y
Observations	967	967	1276	1276
Adj. R-squared	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. indicates $p<0.10$, ** indicates $p<0.05$, ***indicates $p<0.01$, ****indicates $p<0.001$.

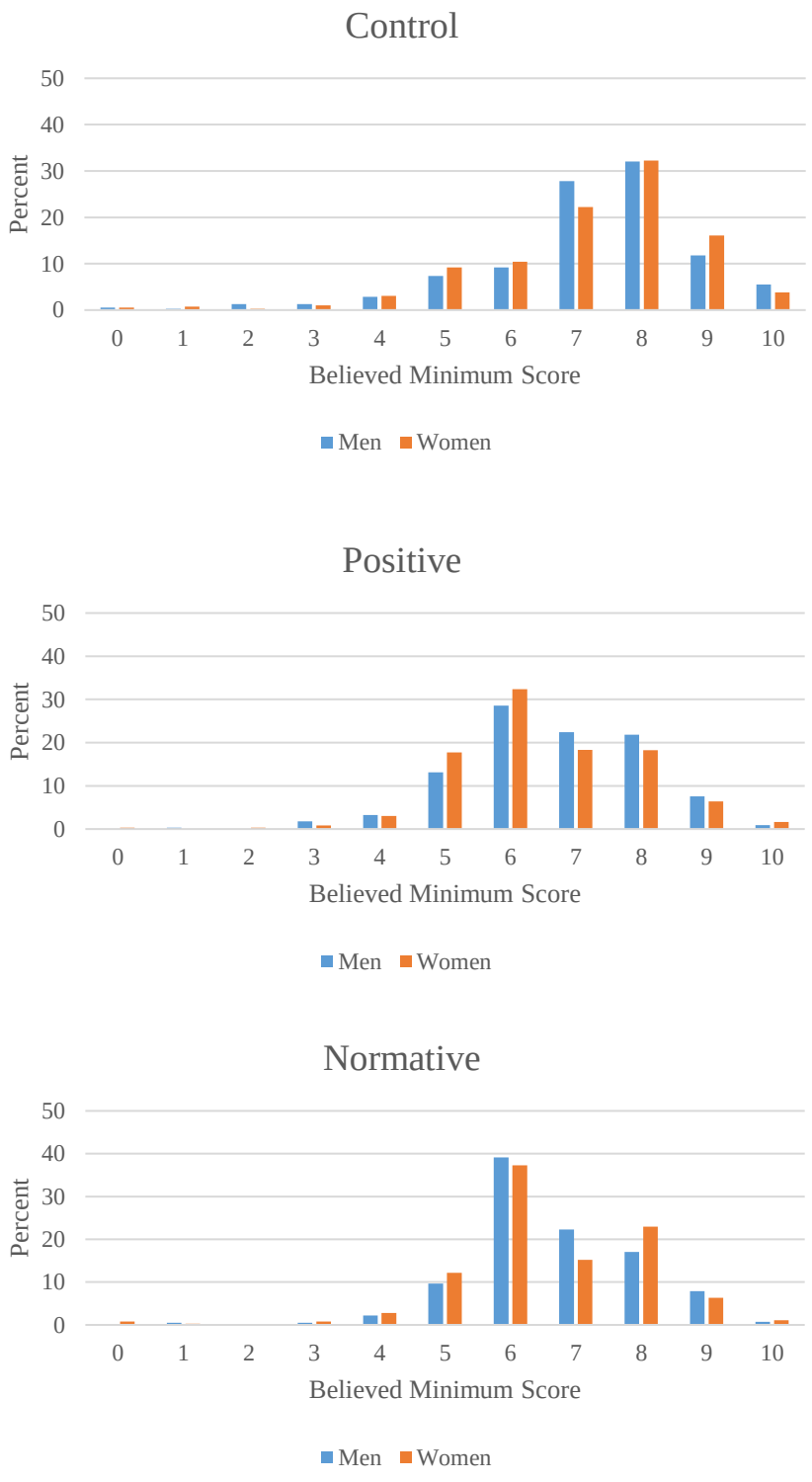

Figure 3. Beliefs of "The Bar"

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, though the differences are modest (average of 2.98 for men, a 0.23 increase, $p<0.01$; average of 2.83 for women, a 0.16 increase, $p<0.05$). Overall this data suggests that many more qualified candidates realize they are qualified after receiving information about the bar in our treatment conditions, with no strong gender differences in these effects.

Finally, we ask participants directly about the objectivity, clarity, and specificity about required qualifications in each treatment. Recall that each participant first sees this Likert-scale question about their own assigned treatment (before seeing any alternative treatment language). Then, after answering that question, they are 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-subject, 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 language they were evaluating. They assess this on a 1-6 scale where 6 is extremely objective, specific, and clear. We include our standard controls. The first column uses only the across-subject data, while 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 (which is given a ranking of 3.63 by participants assigned to the Control treatment) 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 more of an improvement than the Positive treatment.

The evidence is consistent with our proposed mechanism of our treatments reducing the amount of ambiguity around the bar. And, while the Positive and Normative treatment have quite similar impacts on application behavior in the Prolific setting, it is interesting to relate this data back to our UpWork setting. On UpWork, our Normative treatment had a statistically significant impact on the application rates of qualified women, while the Positive treatment did not (though the two effects were statistically indistinguishable). The Prolific data reveals the Normative treatment is "stronger" in terms of reduced ambiguity, at least as perceived by these participants, suggesting an explanation for the UpWork result.

Table 5. Perceptions of Clarity of Desired Qualifications

	OLS Predicting How Objective, Clear, and Specific Were 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)	2243	6729 (2243)
Adjusted R-squared	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. Indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

Additional Analysis

In this section, we build on our pre-registered analysis with additional exploration of how application decisions relate to perceptions of where the bar is. In our Prolific study, we observe that men and women have similar perceptions of what the bar is, and yet we see 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-level job. We can compare these stated beliefs to their observed scores and ask whether participants whose scores exceed their believed bar indeed choose 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 know their score at the time of their application decision. This analysis asks how participants act on knowing that their score is above or below their perceived bar.

Table 6 provides these results. We see evidence of gender differences for individuals who believe 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 apply, compared to just 71% of women ($p<0.01$). This gender gap is largest in our Control treatment, where 75% of men who believe they have a score above the bar apply but just 57% of women do ($p=0.03$). In our view, this is a striking result that merits further investigation in future work: given that they believe they have the required qualification, why do nearly half of these women choose not to apply to the expert-level 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 sub-population.

Table 6. Beliefs and Behavior

	Application Rates to Expert Level Job			
Strictly Above Believed Bar	Overall	Control	Positive	Normative
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 v W	p=0.001	p=0.03	p=0.20	p=0.02
Score Equal to Believed Bar	Overall	Control	Positive	Normative
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 v W	p=0.07	p=0.10	p=0.80	p=0.07
Strictly Below Believed Bar	Overall	Control	Positive	Normative
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)	15.5% (N=187)	13.8% (N=209)
p-value M v 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.

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 apply to the expert-level 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-level job are low among this sub-population, but with persistent gender differences. Across each of our treatments, approximately 25% of men whose score does not meet what they believe the required qualification to be still choose to apply, while approximately 14% of women make 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.

IV. Discussion

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 to equally well-qualified men.

We show that exogenously reducing ambiguity about the required qualifications increases the rate at which qualified candidates apply to 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. Because of this, the treatments reduce the gender gap in qualified applicants in the field. Our results suggest that there may be soft touch employer interventions that can improve the breadth and diversity of 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.

While improving clarity around a job's qualification requirements may be a promising, feasible intervention for increasing the rate at which qualified women apply, one avenue for future work is understanding the persistence the gender gap in contexts like our controlled study. In particular, it seems worth delving deeper into why many women who 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. This is an interesting question for further study. 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.

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 or more clearly specifying desired years of experience, minimum GRE score, number of projects successfully completed in the past, etc. Assuming that indeed the employer has a bar in mind for these dimensions (that is, they only want 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. While 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.

While 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. While 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.

V. References

Abraham, Lisa and Alison Stein. 2020. "Words Matter: Experimental Evidence from Job Applications." Essays in Labor Economics, Doctoral dissertation, Harvard University, Graduate School of Arts and Sciences.

Altonji, Joseph G., and Rebecca M. Blank. 1999. "Chapter 48 Race and Gender in the Labor Market." In *Handbook of Labor Economics*, 3:3143–3259. Elsevier. https://doi.org/10.1016/S1573-4463(99)30039-0.

Bohnet, Iris, Alexandra van Geen, and Max Bazerman. 2016. "When Performance Trumps Gender Bias: Joint vs. Separate Evaluation." *Management Science* 62 (5): 1225–34. https://doi.org/10.1287/mnsc.2015.2186.

Bordalo, Pedro, Katherine B. Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review 109*(3), 739-73..

Bowles, H. R., Babcock, L., & McGinn, K. L. 2005. "Constraints and triggers: Situational mechanics of gender in negotiation." *Journal of personality and social psychology*, *89*(6), 951.

Coffman, Katherine B. 2014. "Evidence on Self-Stereotyping and the Contribution of Ideas." *The Quarterly Journal of Economics* 129 (4): 1625–60. https://doi.org/10.1093/qje/qju023.

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

Coffman, Katherine B, Christine L. Exley, and Muriel Niederle. 2021. "The Role of Beliefs in Driving Gender Discrimination." *Management Science*.

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

Del Carpio, Lucia and Maria Guadalupe. 2021. "More Women in Tech? Evidence from a Field Experiment addressing Social Identity." *Management Science.*

Dubey, Alpana, Kumar Abhinav, Mary Hamilton, and Alex Kass. 2017. "Analyzing Gender Pay Gap in Freelancing Marketplace." In *Proceedings of the 2017 ACM SIGMIS Conference on Computers and People Research*, 13–19. SIGMIS-CPR '17. New York, NY, USA: ACM. https://doi.org/10.1145/3084381.3084402.

Exley, Christine and Judd Kessler. 2020. "The Gender Gap in Self-Promotion."Forthcoming in *Quarterly Journal of Economics*.

Fernandez, Roberto M., and Marie Louise Mors. 2008. "Competing for Jobs: Labor Queues and Gender Sorting in the Hiring Process." *Social Science Research* 37 (4): 1061–80. https://doi.org/10.1016/j.ssresearch.2007.10.003.

Flory, Jeffrey A., Andreas Leibbrandt, and John A. List. 2015. "Do Competitive Workplaces Deter Female Workers? A Large-Scale Natural Field Experiment on Job Entry Decisions." *The Review of Economic Studies* 82 (1): 122–55. https://doi.org/10.1093/restud/rdu030.

Foong, Eureka, Nicholas Vincent, Brent Hecht, and Elizabeth M. Gerber. 2018. "Women (Still) Ask For Less: Gender Differences in Hourly Rate in an Online Labor Marketplace." *Proc. ACM Hum.-Comput. Interact.* 2 (CSCW): 53:1–53:21. https://doi.org/10.1145/3274322.

Gee, Laura K. 2018. "The More You Know: Information Effects on Job Application Rates in a Large Field Experiment." *Management Science* 65 (5): 2077–94. https://doi.org/10.1287/mnsc.2017.2994.

Ginther, Donna K., and Shulamit Kahn. 2009. "Does Science Promote Women? Evidence from Academia 1973 - 2001." In *Science and Engineering Careers in the United States: An Analysis of Markets and Employment*, edited by Richard B. Freeman and Daniel L. Goroff. A National Bureau of Economic Research Conference Report. Chicago: University of Chicago Press.

Heilman, Madeline E. 2001. "Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder." *Journal of Social Issues* 57 (4): 657–74. https://doi.org/10.1111/0022-4537.00234.

Ibarra, Herminia, Nancy M. Carter, and Christine Silva. 2010. "Why Men Still Get More Promotions Than Women." *Harvard Business Review*, September 1, 2010. https://hbr.org/2010/09/why-men-still-get-more-promotions-than-women.

Kuhn, Peter, Kailing Shen, and Shuo Zhang. 2020. "Gender-Targeted Job Ads in the Recruitment Process: Facts from a Chinese Job Board." *Journal of Development Economics* 147 (November): 102531.

Murciano-Goroff, Raviv. 2021. "Missing Women in Tech: The Labor Market for Highly Skilled Software Engineers." *Management Science*.

Niederle, M. 2016. "Gender," in *The Handbook of Experimental Economics 2*, Kagel John, Roth Alvin E., eds. (Princeton, NJ: Princeton University Press, 2016).

Niederle, Muriel, and Lise Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?" *The Quarterly Journal of Economics* 122 (3): 1067–1101. https://doi.org/10.1162/qjec.122.3.1067.

Niederle, Muriel and Alexandra H. Yestrumskas. 2008. "Gender Differences in Seeking Challenges: The Role of Institutions." *Working paper*.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2014. "How Stereotypes Impair Women's Careers in Science." *Proceedings of the National Academy of Sciences of the United States of America* 111 (12): 4403–8.

Riach, P. A., and J. Rich. 2002. "Field Experiments of Discrimination in the Market Place." *The Economic Journal* 112 (483): F480–518.

Samek, Anya Savikhin. 2019. "A University-Wide Field Experiment on Gender Differences in Job Entry Decisions." *Management Science* 65 (7): 3272-3281.

"Skills Tests." n.d. Upwork Help Center. Accessed February 4, 2019. http://support.upwork.com/hc/en-us/articles/211063198-Skills-Tests.

Upwork. n.d. "Upwork." Accessed December 5, 2018. https://www.upwork.com.

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, Saadia, and Herminia Ibarra. 2010. "The Corporate Gender Gap Report 2010." World Econonomic Forum.

Appendix

A. Experiment Materials

A.1. Experimental Instructions UpWork Field Experiment (under separate cover)

https://perma.cc/YTQ6-CYUC

A.2. Experimental Instructions Prolific Replication Study (under separate cover)

https://perma.cc/C5CF-P6XF

A.3. Experimental Instructions Real Job Ads Study (under separate cover)

https://perma.cc/2DZK-LHG8

A.4. Experimental Instructions MTurk Simulated Labor Market Study (under separate cover)

https://perma.cc/J33L-LWJM

B. Additional Tables

Table B1. Summary Statistics for Upwork Field Experiment

	All Freelancers in Dataset			Qualified Freelancers Only		
	Men	Women	p-value	Men	Women	p-value
Requested Hourly Rate	44.0	30.8	p<0.001	48.9	39.3	p<0.05
Hours Worked on Upwork	323	562	p=0.07	304	540	p=0.14
Jobs Worked on Upwork	13.7	15.6	p=0.49	11.8	22.0	p<0.05
Total Tests Displayed	6.18	7.32	p=0.002	6.62	7.62	p=0.29
Available Less than 30hrs/wk	0.18	0.22	p=0.12	0.14	0.20	p=0.26
Available More than 30hrs/wk	0.44	0.41	p=0.50	0.36	0.36	p=0.99
Available as Needed	0.37	0.34	p=0.36	0.49	0.40	p=0.18
College Degree	0.74	0.72	p=0.53	0.84	0.84	p=0.91
MBA Degree	0.14	0.08	p=0.001	0.21	0.16	p=0.25
Other Graduate Degree	0.20	0.21	p=0.47	0.28	0.28	p=0.93
Web/Mobile/Software Development	0.20	0.08	p<0.001	0.16	0.13	p=0.53
IT & Networking	0.08	0.005	p<0.001	0.07	0.010	p<0.05
Data Science & Analytics	0.18	0.11	p=0.001	0.21	0.19	p=0.67
Engineering & Architecture	0.04	0.01	p<0.001	0.07	0.02	p<0.10
Design & Creative	0.19	0.15	p=0.09	0.17	0.15	p=0.64
Writing	0.32	0.45	p<0.001	0.34	0.46	p<0.05
Translation	0.05	0.06	p=0.41	0.04	0.14	p<0.01
Legal	0.05	0.05	p=0.77	0.05	0.04	p=0.81
Administrative Support	0.25	0.48	p<0.001	0.22	0.47	p<0.001
Customer Service	0.04	0.11	p<0.001	0.02	0.06	p<0.10
Sales & Marketing	0.15	0.16	p=0.89	0.18	0.18	p=0.91
Accounting & Consulting	0.22	0.13	p<0.001	0.29	0.20	p<0.10
Analytical Skills Score	3.73	3.57	p<0.001	4.35	4.30	p=0.20
Time Taken on Analytical Test (minutes)	50.5	48.05	p=0.08	47.9	48.1	p=0.97
Management Skills Score	3.55	3.42	p<0.001	3.96	3.95	p=0.58
Time Taken on Management Test (minutes)	19.79	20.65	p=0.14	19.7	19.6	p=0.97
Proportion Qualified by Test Score	0.29	0.18	p<0.001			
Proportion in Analytical Skills Dataset	0.41	0.45	p=0.2	0.33	0.41	p=0.16
N	531	552		154	102	

Notes: p-values from binary variables are from two-tailed test of proportions. Continuous variables use two-tailed t-tests.

Table B2. Replication of UpWork Table 1 excluding the 14 observations who applied to both jobs initially

	OLS Predicting Decision to Apply to Expert-Level Job						
	All Participants			All Unqualified		All Qualified	
	I	II	III	III	V	VII	VIII
Positive	-0.025	-0.044*	-0.038	-0.047*	-0.069*	0.055	0.046
Treatment	(0.023)	(0.026)	(0.034)	(0.024)	(0.037)	(0.060)	(0.080)
Normative	-0.016	-0.053**	-0.015	-0.056**	-0.030	0.12**	0.027
Treatment	(0.023)	(0.026)	(0.032)	(0.024)	(0.036)	(0.059)	(0.072)
Female	-0.028	-0.025	-0.036	-0.0077	-0.0056	-0.070	-0.16*
	(0.021)	(0.021)	(0.034)	(0.022)	(0.036)	(0.053)	(0.085)
Qualified		-0.051					
		(0.045)					
Positive x		0.075					
Qualified		(0.055)					
Normative x		0.15***					
Qualified		(0.054)					
Female x			0.025		0.041		0.035
Positive			(0.047)		(0.050)		(0.12)
Female x Normative			-0.0025		-0.048		0.26**
			(0.046)		(0.049)		(0.12)
Controls	Y	Y	Y	Y	Y	Y	Y
Observations	1069	1069	1069	816	816	253	253
Adj. R-squared	0.037	0.042	0.035	0.039	0.041	0.020	0.034

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 (> than 30 hrs/wk, < than 30 hrs/wk, as needed), dummies for each self-reported skill, and a dummy for being in the second wave of experiment. * indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

Table B3. Summary Statistics for Prolific Replication Study

	Full Dataset			Qualified Participants Only		
	Men	Women	p-value	Men	Women	p-value
Year of Birth	1984	1987	p<0.001	1985	1990	p<0.001
Attended high school in US	0.97	0.98	p=0.72	0.98	0.98	p=0.73
White	0.79	0.82	p=0.16	0.80	0.84	p=0.07
Black or African American	0.068	0.081	p=0.23	0.044	0.052	p=0.49
Asian	0.12	0.099	p=0.11	0.14	0.12	p=0.25
Latino or Latina	0.069	0.071	p=0.87	0.064	0.056	p=0.51
Indigenous	0.020	0.010	p=0.04	0.013	0.007	p=0.25
No Formal Education	0.010	0.005	p=0.23	0.009	0.005	p=0.43
Secondary Education	0.015	0.016	p=0.84	0.012	0.008	p=0.56
High School Education	0.24	0.28	p=0.03	0.22	0.30	p<0.01
Community or Technical College Education	0.12	0.13	p=0.26	0.091	0.088	p=0.85
College Education	0.39	0.39	p=0.83	0.43	0.38	p=0.09
Graduate Degree	0.19	0.15	p=0.02	0.20	0.18	p=0.52
Education Missing	0.042	0.031	p=0.20	0.047	0.041	p=0.58
Approval Rate	0.996	0.996	p=0.67	0.997	0.997	p=0.72
Total Jobs	718	533	p<0.001	709	476	p<0.001
ASVAB Score	5.96	5.64	p<0.001	7.21	6.99	p<0.001
Positive Treatment	0.30	0.32	p=0.32	0.29	0.34	p=0.11
Normative Treatment	0.36	0.33	p=0.09	0.36	0.31	p=0.06

Notes: p-values from binary variables are from two-tailed test of proportions. Continuous variables use two-tailed t-tests.

Appendix C

Appendix C provides detailed report of the two additional experiments we ran. In the interest of space and cohesion, we omitted these studies from the main text. When we made the decision to run the pre-registered replication of the UpWork study in Fall 2021, we pre-committed to reporting that follow-up in the main text and moving these studies to the Appendix. Please see our pre-registration for more details (AEARCTR-0008223).

C1. Impressions of Real Job Advertisements

In this simple laboratory study, we seek to document the beliefs potential job-seekers' hold about how well-qualified they are for different openings. By varying the job advertisements, we ask how beliefs about qualification level vary with different features, including how objective, specific, and clear the desired qualifications are.

In April 2018, we constructed a random sample of real job postings from Indeed.com, representative of available online postings in the geographic area of our participants. We searched for full-time job postings in the Boston area that required a Bachelor's degree. We performed two searches: a search of entry-level jobs and a search of mid-level jobs. Within the entry-level search, we downloaded all ads that were returned from our search and randomly selected 20; within the mid-level search, we downloaded all ads that were returned and randomly selected 30. We then read each ad selected. In cases where the ad description did not appear to fulfill our search criteria (for example, not actually being full-time, despite being returned by Indeed.com), we eliminated the ad from our sample (13 cases). In addition, a coding error omitted 9 entry-level ads from the study (no participants were randomly assigned to view them). This left us with 28 job ads: 4 entry-level and 24 mid-level.

We use these ads to collect beliefs from participants at the Computer Lab for Experimental Research (CLER) at Harvard Business School. In the first part of the experiment, we collect participant impressions of how well-qualified they feel for a given job posting. The structure is as follows. Participants complete four rounds of job ad evaluation. Within each round, participants are given two minutes to view a randomly-selected ad from the set of 28.[16] We limit participants' time viewing the ads to ensure timely completion of the experiment, as completion in under 20 minutes was required for the laboratory format we took advantage of.

We ask three questions about perceived qualifications:

1. *On a scale of 1 (Extremely Poorly Qualified) - 10 (Extremely Well Qualified), how well-qualified do you feel you are for this job?*

2. *Thinking of the desired skills, characteristics, and qualifications stated in the advertisement, what percent of those skills, characteristics, and qualifications do you possess?*

[16] We create four non-overlapping subsets of the 38 ads. Each bucket contains 3 – 4 entry-level jobs and 6 – 7 mid-level jobs. Within each round of this experiment, participants view one randomly-selected ad from a given bucket, ensuring that no participant sees the same ad twice in one part of the experiment.

3. *More specifically, please list some of the desired skills, characteristics, and qualifications that you do possess, and some of the desired skills, characteristics, and qualifications that you do not possess.*

The first question gets at our core issue: how well-qualified an individual feels for a given position. The second question provides a more quantitative assessment of those beliefs. The third question encourages participants to reflect on the qualifications for this particular ad, and provides insight into the types of skills participants list.

We also ask two "decoy" questions of our participants:

4. *On a scale of 1 - 10, with 1 being not appealing at all and 10 being extremely appealing, how appealing is this job opening to you?*

5. *Approximately what salary would you expect this job to offer?*

We include these questions so that participants do not become solely focused on qualifications as they read additional ads in the experiment. We display the qualification questions first, followed by the decoy questions, in Rounds 1 and 3; we reverse the order in Rounds 2 and 4.

In the second part of this experiment, we collect additional data from *the same set of* participants about each of these 28 ads. In particular, we recognized that in analyzing the data on perceived qualifications, there were a number of ad characteristics that we would want to collect and use as controls in our analysis. Rather than code these characteristics ourselves, we chose to have participants code these characteristics, ensuring no researcher bias.

The format of the second experiment is nearly identical to the first. Participants complete four rounds. Within each round, they are given 2 minutes to view one randomly-selected ad from the 28. Note that this randomization operates independently from the randomization in the first part; thus, participants could be randomly assigned the same ad in both experiments, but this was not particularly likely. They are then asked four questions about the ad:

1. *In general do you think the stereotype associated with this job is more female-typed or more male-typed? Use the slider scale below to indicate your answer, where -1 indicates extremely female-typed and 1 indicates extremely male-typed.*

2. *How prestigious would you say this job is? Use the slider scale below to indicate your answer, where 1 indicates not prestigious at all and 7 indicates extremely prestigious.*

3. *Thinking of typical Harvard undergraduates, how well-qualified do you think the average Harvard undergraduate would be for this job? Use the slider scale below to indicate your answer, where 1 indicates not at all qualified and 10 indicates extremely well-qualified.*

4. *Thinking of how the qualifications in the job advertisement were described, how specific, clear, and objective were the stated qualifications? Use the slider scale below to indicate your answer, where 1 indicates not at all clear and 10 indicates extremely specific, clear, and objective.*

We hypothesized that each of these measures could be relevant in predicting participant beliefs about how well-qualified they were. The first gets at the gender-stereotype associated with the job, speaking to the mechanism of Coffman (2014), who finds that individual self-confidence and willingness to volunteer ideas is dependent on the gender congruence of the domain. If beliefs of own aptitude are a key driver in predicting beliefs of how well-qualified someone is, we would predict that as the maleness of the job posting increased, men would feel relatively more well-qualified while women would feel relatively less well-qualified. The second question allows us to try to separate out differences in the gender stereotype attached to the job from differences in the perceived prestige of the position.

The third question allows us to better account for variation across ads in how likely it is that any participant in our sample feels qualified for that particular ad. This is important given how heterogeneous the various postings are. Finally, the fourth question speaks to the hypothesis tested in the UpWork experiment: does the amount of ambiguity surrounding the desired qualifications matter for beliefs of how well-qualified individuals feel? In particular, does ambiguity contribute to a gender gap in these beliefs?

Results

In total 200 participants completed the two experiments as part of bundle sessions at the Computer Lab for Experimental Research at Harvard Business School, of which 197 provided information on their gender.[17] We provide summary statistics on our participants and our job ads in Tables C1 and C2, respectively. More than 80% of our sample identifies themselves as a current student.

Table C1. Summary Statistics for Laboratory Participants for Real Job Ads Study

	Men	**Women**	**P-value from test of proportions**
White	0.28	0.32	0.51
Black or African American	0.16	0.11	0.34
Asian	0.35	0.34	0.93
Latino or Latina	0.08	0.07	0.90
Multiracial	0.12	0.09	0.48

[17] In addition, one participant did not provide data on their age; thus they are excluded from analyses that control for age. We obtain answers to a standard bank of (non-mandatory) demographic questions that are asked of all participants in "bundle sessions" in the laboratory. This laboratory format we used bundles our project with short projects from other researchers. These bundle sessions are administered by the laboratory and target 200 participants in a single week. The placement of our experiments with respect to these other projects was varied across session, though note that the two parts of our project are always placed in the same order (Experiment 1 and then Experiment 2, as described above) and appear consecutively, with no other projects in between.

Middle East	0.01	0.01	0.89
Is a Student	0.80	0.87	0.22
Average Age	23.93	23.81	0.78
Highest obtained Education			
High School	0.12	0.05	0.05
Some College	0.30	0.29	0.80
Bachelor's Degree	0.36	0.49	0.06
Advanced Degree	0.21	0.18	0.51
Humanities Major	0.10	0.18	0.14
Social Science Major	0.25	0.29	0.53
STEM Major	0.50	0.44	0.41
Is fluent in English	0.99	0.99	0.89
Order of Experiment within Session	0.48	0.50	0.81

Notes: p-values from binary variables are from two-tailed test of proportions. Continuous variables use two-tailed t-tests.

Table C2. Summary Statistics on Job Ads

Panel A: Data on Bureau of Labor Statistics Sector for Ads	
BLS Sector	**Percent of Ads**
Educational Services	7
Financial Activities	7
Health Care and Social Assistance	21
Information	14
Leisure and Hospitality	11
Manufacturing	7
Professional and Business Services	25
State and Local Government	4
Transportation and Warehousing	4

Panel B: Summary of Participant Assessments of Job Ads					
Variable	**Min.Value**	**25th pctl**	**75th pctl**	**Max. Value**	**Mean**
Individual Level Well-Qualified	1	2	6	10	4.41
Ad Level Male Stereotype	-0.26	-0.12	0.2	0.5	0.070
Ad Level Prestige	2.86	3.44	4.14	4.82	3.856
Ad Level Objectivity	5.14	6.37	7.38	7.7	6.857
Ad Level Avg. Qualified	4.57	5.44	6.64	7.23	6.077

We find that, on average, men view themselves as marginally more well-qualified than women in our sample. Participants rate on a 1-10 scale how well-qualified they feel they are for each of four particular job ads. On average, men rate themselves a 4.65 (2.63 SD) while women rate themselves a 4.22 (2.51 SD). This gender gap is approximately 9% of the mean of how well-qualified individuals feel.

In Table C3, we present a regression that explores the determinants of these ratings. To increase interpretability, we create z-scores for the variables that were elicited with a scale. Controlling for ad fixed effects and demographics, we estimate that women rate themselves approximately 0.18 standard deviations less qualified than men ($p<0.10$, Column I). In Column II, we include more information about the ads. In particular, we use the assessments provided by our participants in the second part of the experiment, as to stereotype, prestige, believed qualifications of others, and objectivity of qualifications. For each ad, we take the average of the ratings provided by all raters who saw that ad for each characteristic.[18] Then, we take the z-score to capture where this particular ad falls relative to the full set of ads on that characteristic. We also include a dummy for whether the ad was for an entry-level position, and dummies for the major industry sector of the ad. Controlling for these ad characteristics does not have a large impact on our estimate of the gender gap (Column II).

Table C3. The Gender Gap in Perceived Qualification for Real Job Openings

	OLS Predicting How Well-Qualified an Individual Feels for Job Opening (z-score)			OLS Predicting What Percentage of Qualifications an Individual Believes She Possesses (0 – 100 scale)		
	I	II	III	IV	V	VI
Female	-0.18*	-0.21**	-0.22**	-5.52**	-6.59**	-6.70**
	(0.093)	(0.095)	(0.095)	(2.61)	(2.65)	(2.64)
Male Stereotype (z-score)		-0.25***	-0.23**		-8.78***	-8.58***
		(0.081)	(0.090)		(2.24)	(2.46)
Prestige (z-score)		0.090	0.15**		2.73*	4.21**
		(0.055)	(0.071)		(1.49)	(1.84)
Objectivity of Stated Qualifications (z-score)		-0.17***	-0.27***		-4.36***	-6.38***
		(0.051)	(0.069)		(1.51)	(2.05)
Average Belief of How Well-Qualified Average Undergrad Would be for this Ad (z-score)		0.048	0.028		1.10	-0.16
		(0.064)	(0.077)		(1.76)	(2.14)
Female x Male Stereotype			-0.034			-0.45

[18] We note that there are no significant differences in how men and women rate these ads on average in terms of stereotype, prestige, or objectivity of qualifications. And, across ads, the average male and average female ratings are highly correlated along each of these dimensions.

			(0.068)			(1.94)
Female x Prestige			-0.11			-2.53
			(0.072)			(1.99)
Female x Objectivity			0.17**			3.54
			(0.077)			(2.26)
Female x Avg. of Avg. Qualified			0.030			2.15
			(0.079)			(2.15)
Entry Level Dummy		0.34***	0.34***		6.86**	6.96**
		(0.12)	(0.12)		(3.24)	(3.24)
Ad Fixed Effects	Yes	No	No	Yes	No	No
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Order within Session	Yes	Yes	Yes	Yes	Yes	Yes
Observations (Clusters)	784 (196)	784 (196)	784 (196)	784 (196)	784 (196)	784 (196)
Adjusted R-squared	0.155	0.136	0.139	0.139	0.112	0.113

Notes: Controls are fixed effects for each ad in Columns I and IV, fixed effects for each race category, fixed effects for each education category, age, a dummy for fluent in English, and a dummy indicating where our pair of experiments fell within the session. In columns without ad fixed effects, we include dummies for major industry sector. * indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

In Column III, we interact the ad characteristics with the female indicator, to ask whether any of the characteristics of the job opening impact the gender gap in believed qualification level. We find a role for ambiguity in shaping the gender gap. We find that more objectively stated job qualifications, as rated by our participants, reduce the gender gap in perceived qualification level.

These results are very similar when we use the fraction of qualifications that a participant believes they possess (Columns IV – VI). On average, conditional on ad and individual characteristics, women believe they possess roughly 5pp fewer of the qualifications than men do (male average: 50%, female average: 45%, p<0.05). Again, as desired qualifications become more objective, men assess themselves to be less well-qualified on average in terms of the fraction of qualifications possessed; this effect is directionally weaker for women (p=0.12 on the interaction, Column VI). Conditional on believing they have the same fraction of qualifications, men and women rate themselves as equally well-qualified.

C2. MTurk Simulated Labor Market Study

In this online experiment, we examine both beliefs and application decisions in a simulated labor market. We exogenously vary the level of ambiguity in stated qualifications surrounding a promotion opportunity, and explore its impact on gender gaps in beliefs and behavior.

Overview of Design

We conduct our experiment on Amazon Mechanical Turk (MTurk).[19] The general structure of the experiment is as follows. Before we run the main experiment, which we call the "worker experiment," we recruit 10 MTurk employers to make contingent hiring decisions. We use these contingent decisions to later incentivize worker decisions, and to inform the crafting of qualifications in the main worker experiment. Next, we move to the "worker experiment." In Round 1 of the worker experiment, we collect data on participant aptitude in a diagnostic test and elicit participant beliefs about their aptitudes. Then, we confront workers with a decision; they are asked to decide whether to apply for a promotion for Round 2 of the experiment. We also directly elicit their perceived probability of promotion conditional on applying. Then, participants complete Round 2. Finally, on the back end, after the completion of the "worker experiment," the researchers use random matching to allocate workers from the main experiment to employers from the preliminary employer experiment to determine outcomes and payoffs. We provide a visual overview of the "worker experiment" in Figure C1. This is followed by a detailed description of each stage.

Round 1

In Round 1 of the worker experiment, participants take an assessment test that covers general science, arithmetic reasoning, math knowledge, mechanical comprehension, and assembling objects. We draw the questions from the Armed Services Vocational Aptitude Battery (ASVAB). These questions have a simple multiple-choice format, several of the categories we use are rather difficult to quickly "Google" answers to, and they cover stereotypically male-typed domains, matching our field setting. Participants have 5 minutes to answer up to 30 questions. All 30 questions appear on the same page and can be answered in any order. If Round 1 is chosen for payment, participants receive $0.20 per question answered correctly.

Beliefs of Round 1 Performance

After completing Round 1, each participant is asked to guess their score –how many problems she solved correctly in Round 1 – and how they rank relative to other MTurkers who are completing the HIT. They

[19] The study was restricted to workers with a United States based IP address who had completed at least 100 tasks (called Human Intelligence Tasks, or HITs) and had an approval rating by previous MTurk requesters of at least 95%. The study contains understanding questions and a participant must answer those understanding questions correctly in order to complete the study.

receive $0.10 if they guess their score correctly and $0.10 if they guess their bucket of rank correctly (bottom 5%, bottom 20%, bottom 40%, middle 20%, top 40%, top 20%, top 5%).

We note that all workers are then randomized into one of three feedback conditions: receiving either no feedback on Round 1 performance, a signal equal to their true score 60% of the time, or a signal equal to their true score 90% of the time. We then elicit posterior beliefs of Round 1 score from participants in each of the two noisy feedback conditions. This noisy feedback intervention and its impact on beliefs is the focus of a different paper, Coffman, Collis, and Kulkarni (2021). For our purposes, the only important thing to note that is that our measure of beliefs of Round 1 score in our analysis below will be the *posterior* beliefs of these participants: the beliefs they hold after receiving the information. In principle, this could work against us finding large gender gaps in beliefs or application decisions. But, Coffman, Collis, and Kulkarni (2021) show that gender gaps in prior and posterior beliefs in this setting are actually quite similar, suggesting the inclusion of this feedback treatment is not significantly impacting our conclusions in a meaningful way.

Application Decision

Participants are told that they will soon have a chance to participate in a second round of ASVAB problem-solving. Again, they will have 5 minutes to answer up to 30 ASVAB questions, but these questions will be more difficult on average than the questions in Round 1. In this way, Round 1 performance is predictive of Round 2 performance, but there is additional uncertainty due to the increased difficulty. If Round 2 is chosen for payment, the default option is that they will receive $0.20 per problem solved correctly.

Prior to completing Round 2, they have to make a decision about whether they want to apply for an "expert-level promotion". This "expert-level promotion" is an increase in compensation. Participants are presented with two options about how to be paid for Round 2 performance. Importantly, within each option, the problems faced in Round 2 are identical. Participants are explicitly told that the set of questions will be the same regardless of the option they chose. They simply choose how to be compensated:

"Option 1: Accept the novice job. If you choose this option and Round 2 is chosen for payment, you will get a Round 2 completion payment of $2 on top of the $0.20 per problem solved correctly in Round 2."

"Option 2: Apply for the expert-level job. If you choose this option **and** you are chosen to be promoted to the expert-level job, you will get a **promotion bonus** plus an extra $0.20 per problem solved correctly in Round 2, for a total of $0.40 per problem solved correctly. However, if you apply for the expert-level job and you are not promoted, you will only earn the $0.20 per problem solved correctly. You would not earn a Round 2 promotion bonus."

Participants complete a price list, choosing between Option 1 (accepting the novice job) and Option 2 (applying for the expert-level job) over a range of possible promotion bonuses. Within the price list, we vary the size of the promotion bonus from \$0 to \$6, in increments of \$0.50. Participants are aware that one row of the price list will be randomly-selected as the decision-that-counts, and that we will use their decision in that row to determine their application decision and associated payoffs. In Figure C2, we reproduce the price list used to elicit these decisions (full materials are available in Appendix A).

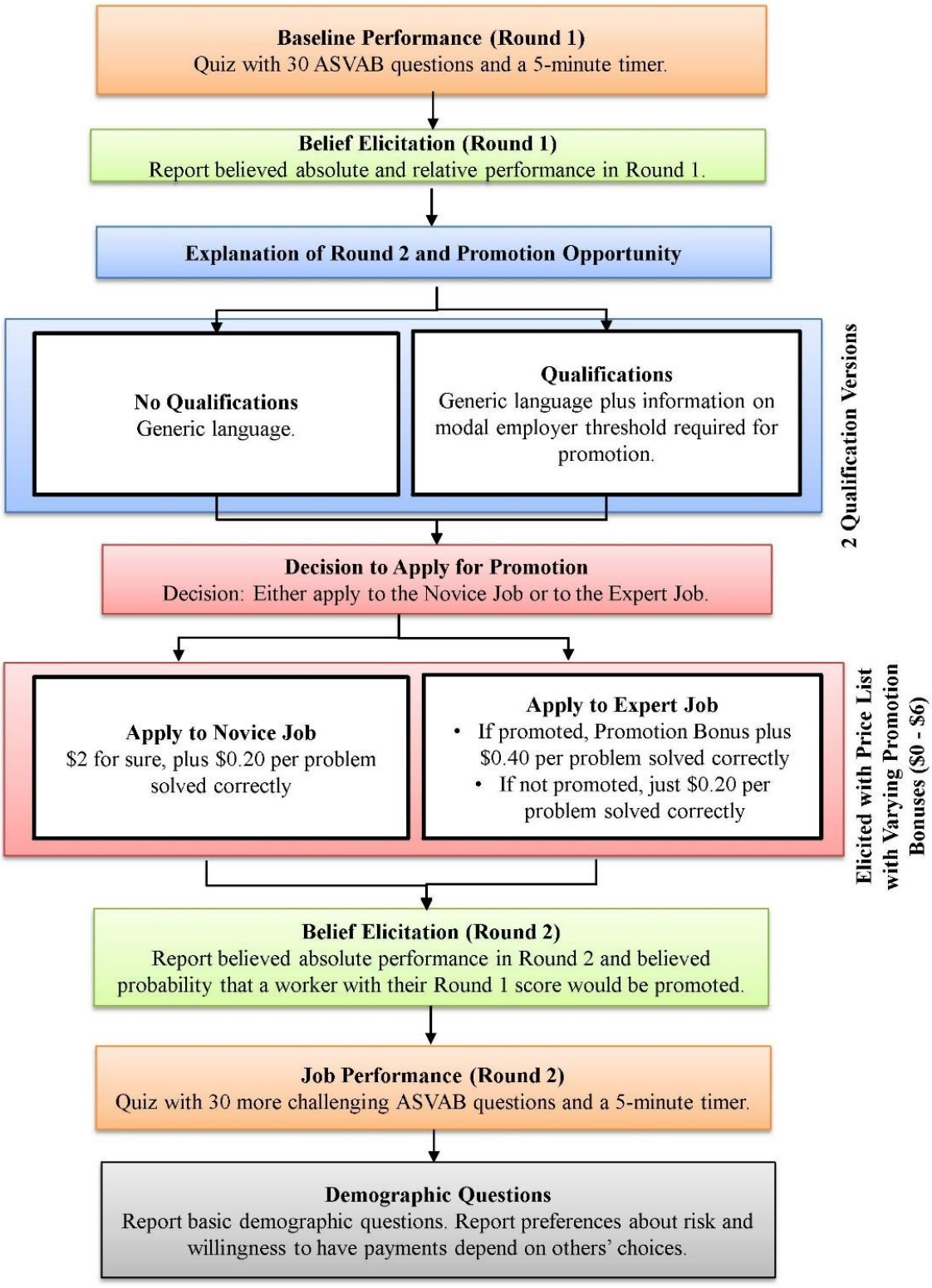

Figure C1. Design of Worker Experiment

While choosing to apply for promotion outside of our experiment typically entails applying both for higher compensation *and* different, more challenging work, we hold fixed the nature of the work. While this sacrifices some external validity, it comes with a number of advantages. First, by ensuring that all participants complete the same Round 2 problems, we can measure the returns to being promoted for each participant, absent any selection. Second, we can rule out explanations for not applying for promotion related to a distaste or disinterest in doing the work (i.e. if women apply for promotion less than men, it cannot be because they simply want to avoid doing harder problems). This way, we can better focus on our main channel of interest: beliefs.

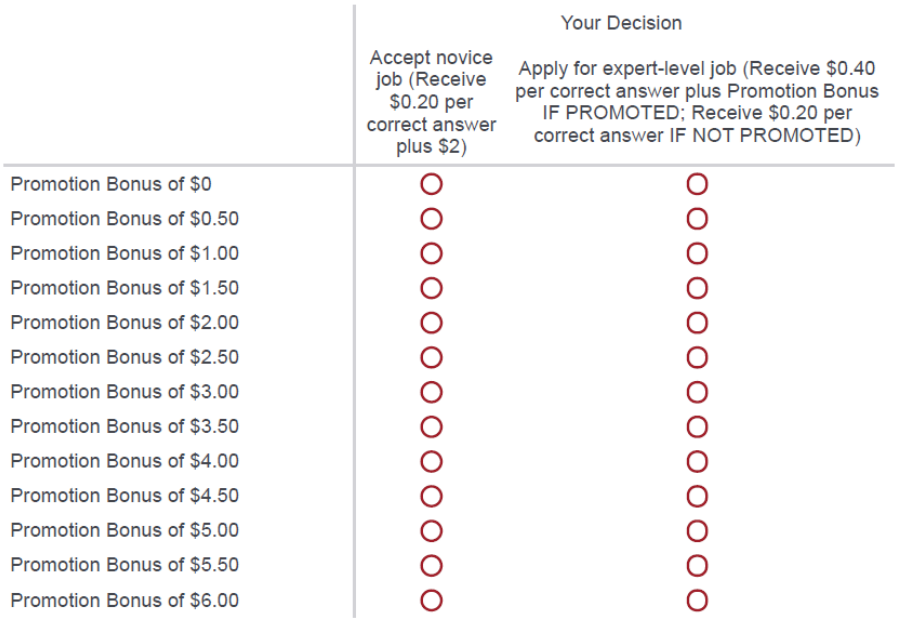

Figure C2. Price List Used to Elicit Application Decisions

Note that we build in an opportunity cost of applying for the promotion: a worker who chooses to apply for the expert-level job forgoes the $2 completion payment given to workers who choose Option 1, the novice job. Thus, a worker who applies for but does not receive the expert-level job earns less than a worker who simply accepts the novice job. This creates the incentive to apply for the expert-level job only if the worker believes she has sufficient probability of receiving it. In addition, because receiving the promotion entails a higher per-problem solved correctly wage ($0.40 versus $0.20), the returns to applying for and receiving the promotion are larger for more talented participants. We believe both of these features reflect many promotion opportunities outside of the laboratory.

We also structure the promotion opportunity in a way that mimics some of the uncertainty about the probability of receiving the promotion that would be present in the field. In particular, we wanted the probability of receiving the promotion to both be tied to individual performance in Round 1 (mimicking the role of resume, prior experience, and potential), but also dependent on the discretion of a potential "employer" with only imperfect knowledge of the candidate's true capacity for success. To achieve this, we recruited 10 other MTurk workers in a separate experiment to serve as employers. We recruit these employers in advance of running the worker experiment, and we ask them to make contingent decisions about what types of workers they would choose to "promote." We explain the worker experiment to them.

We told the employers that they would be randomly paired with *one worker who applied for the expert-level promotion*. If they chose to hire that worker, they would receive **$0.25 for each problem solved correctly by that worker in Round 2**. If not, they would receive a $1.25 fixed payment. Of course, at the time of the employer experiment, no workers had yet been recruited to complete the experiment. So, we ask employers to make contingent decisions, and we use these contingent decisions to execute their hiring decisions once the full worker experiment had been run. We show employers a series of possible Round 1 performances (i.e. 3 problems solved correctly, 4 problems solved correctly, etc.), and we ask whether they would want to hire a worker with that Round 1 performance. The employers are not provided with any other information such as gender. They make a series of binary decisions, covering all possible Round 1 scores.

We use these employer decisions to execute promotion decisions for all workers who apply to the expert-level promotion. Workers who apply are divided evenly and randomly among the 10 employers. Then, each employer's contingent decisions are used to determine whether each worker is promoted or not.[20]

Treatment Intervention

In our control treatment, "No Qualifications", we provide workers with a short information section entitled, "Should I apply?". We remind them of the details of Round 2 and we tell them about the incentives employers faced when making their hiring decisions. Employees are told that the only information provided to the employer is their Round 1 score.

[20] That is, suppose a worker has a Round 1 score of "7" and applies for the expert-level promotion. She would be randomly paired with one of the 10 employers and we would look at whether *that* employer was willing to hire a worker with a Round 1 score of "7". If the employer was willing, she would be hired for the promotion. If the employer was not willing, she would not be promoted. And, one of the matched workers for each employer is randomly selected to determine the employer's payoffs. Employers are aware of this payment rule. Both workers and employers have complete information on this process. See Appendix A for full instructions.

Participants in our "Qualifications" treatment receive the same language, but with one additional sentence that aims to reduce ambiguity about the bar: "While we can make no guarantees regarding your particular application, most employers indicated that they required a Round 1 score greater than 10 in order to be willing to promote a worker."[21] We argue that the key question workers must wrestle with is, "what test score do I need in order to get promoted?" Relative to the "No Qualifications" treatment, we argue that workers in the "Qualifications" treatment have a clearer, more specific, and more objective answer to this key question, reducing ambiguity as to where the bar is.

Beliefs about Promotion

After completing their application decisions, we ask participants how many problems they expect to solve correctly in Round 2, allowing us to calculate what their believed returns to promotion are conditional on being promoted. They receive $0.10 if they guess their Round 2 score exactly correctly. The second, unincentivized question asks participants what they believe the probability is that someone with their Round 1 score would be promoted, conditional on applying. This speaks directly to their beliefs of how well-qualified they are.

Round 2

Participants then complete the Round 2 problems. Following Round 2, they answer brief demographic questions about themselves: gender, education level, race, and whether they attended high school in the United States. They then complete a series of risk preference questions. Finally, they answer two questions about their decisions on MTurk in general, indicating whether they are reluctant to have their payments *on MTurk specifically* depend on chance or on the decisions of other MTurkers. This allows us to speak to whether their application decisions in our experiment might be distorted by an MTurk-specific skepticism about having payments be less transparent.

Results

The experiment was conducted in May 2018 with 1,502 workers. Table C4 provides summary statistics on the workers. We control for the demographic variables collected in the analysis that follows. Men outperform women on average in Round 1: 10.96 versus 9.65 problem solved correctly ($p<0.001$). Men on average rank in the 54th percentile, while women rank in the 46th percentile on average ($p<0.001$). Note that

[21] Indeed, this threshold is informed by employer decisions. A Round 1 score of 10 is the lowest score at which at least 5 of the 10 employers in our employer experiment were willing to hire a worker.

given the employer decisions and candidate performance, the average chance of receiving the promotion, conditional on applying, is 44%.

Table C4. Summary Statistics on Workers in the Online Simulated Labor Market

	All Participants			Qualified Participants Only		
	Men	Women	P-value	Men	Women	P-value
White	0.80	0.81	0.65	0.82	0.86	0.15
Black	0.06	0.09	0.08	0.04	0.05	0.44
Asian	0.10	0.06	0.01	0.11	0.07	0.03
Attended HS in US	0.98	0.96	0.05	0.98	0.97	0.16
HS Only	0.11	0.085	0.06	0.09	0.06	0.17
Some College/Assoc.	0.36	0.37	0.86	0.32	0.29	0.26
Bachelors	0.39	0.40	0.76	0.42	0.47	0.12
Advanced Degree	0.14	0.15	0.36	0.17	0.18	0.76
Rd. 1 Score	10.96	9.65	<0.001	14.0	13.2	<0.01
Rd. 2 Score	8.44	7.14	<0.001	9.73	8.55	<0.001
Prop. Assigned to Qualifications Treatment	0.49	0.50	0.78	0.50	0.53	0.44
N	798	704		460	336	

Notes: p-values from binary variables are from two-tailed test of proportions. Comparisons of Round 1 and Round 2 scores use two-tailed t-tests.

Beliefs and Decisions

We start by exploring our control treatment, where participants receive less guidance on where the bar is for promotion. We ask every participant to estimate their chances of being promoted, conditional on applying. We find that women's beliefs of their probability of being promoted are significantly lower than men's. Women believe they have a 39% chance of being promoted on average, while men believe they have a 48% chance of being promoted. In Table C5, we use regression analysis to probe this. When we condition on true aptitude, as measured by Round 1 performance, women believe they are significantly less likely to be promoted conditional on applying (Table 3, Column I, 7pp, $p<0.01$). Of course, conditional on Round 1 performance, *true* likelihood of being promoted is the same.

Table C5. Gender Differences in Believed Probability of Promotion

OLS Predicting Believed Probability of Promotion (0 – 100pp) *No Qualifications Treatment*			
	I	II	III
Female	-7.17***	-5.69***	-3.14**
	(1.64)	(1.60)	(1.41)

Round 1 Score	1.78***	0.31	0.55***
	(0.18)	(0.27)	(0.17)
Belief of Rd. 1 Score		2.02***	
		(0.28)	
Belief of Rd. 1 Rank			50.7***
			(2.96)
Controls	Y	Y	Y
Observations	759	759	759
Adjusted R-squared	0.175	0.226	0.408

Notes: Controls are fixed effects for each race category, fixed effects for each education category, and a dummy for attended high school in the US, as well as dummies for each feedback treatment (no signal, 60% signal, 90% signal). * indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

Why do women feel they are less likely to be promoted conditional on applying? Part of it seems related to beliefs about own aptitude. Indeed, we find gender differences in self-assessments. Conditional on Round 1 performance and demographic characteristics, women believe they performed significantly worse on average than men. A woman believes she scored 0.7 points worse than a man, conditional on having the same true score, ($p<0.01$), and believes she places 7.2pp worse in the distribution of performers ($p<0.01$) (see Appendix Table B7). In Table C5, Column II, we control for absolute beliefs of own ability and ask whether they explain the gender gap in believed probability of promotion. While beliefs of own aptitude are predictive of beliefs about promotion probability, they do not explain much of the gender gap, which remains 6pp ($p<0.01$).

In Column III, we ask whether beliefs of relative ability explain the gender gap. When we control for beliefs of relative ability, we see more explanatory power, suggesting they capture something important about this decision (Column III). This is worth noting, given that this is *not* a competitive environment (one candidate's decision to apply or ability to receive the promotion has no impact on another's). But, it seems reasonable that a person who believes their performance compares quite favorably to others will also view themselves as having a good chance of being promoted. In this way, beliefs of relative ability may reflect a mix of both beliefs about self and beliefs about what the bar is, hinting at whether the individual feels "good enough."

Next we consider our key behavioral outcome: willingness to apply for promotion at different wages. Conditional on applying, the average minimum promotion bonus at which men and women apply is nearly

identical: 264 cents for men and 260 cents for women. But, significantly more women than men choose to never apply (22% of women and 16% of men, p=0.04). From this point forward, we will code the decision to never apply as a minimum promotion bonus willingness to accept of 650 cents, 50 cents more than the maximum promotion bonus we offered. With this coding, the average min. promotion bonus required to induce a man to apply is 325 cents, while for women it is 344 cents (p=0.21).

Table C6 predicts the minimum promotion bonus at which someone was willing to apply for promotion for our No Qualifications treatment. Conditional on Round 1 performance, there are no significant gender differences in willingness to apply (Column I). We have focused on beliefs of how well-qualified one is, but conceptually, there are several other factors that could also influence willingness to apply in this setting. In particular, risk preferences and expected returns to promotion (i.e. beliefs of Round 2 score) may matter. In Column II, we add these factors to the regression to explore the role they play in shaping application decisions. As expected, each of these factors predict willingness to apply. Finally, in Column III, we include each of these factors and its interaction with the female indicator, asking whether any of these factors are more predictive for women than men. While we estimate that beliefs of Round 2 score and risk preferences are similarly important for men and women, we find that the effect of believed probability of promotion on the decision to apply is nearly twice as large for women as for men.

One question is why we observe a gender gap in believed probability of promotion but no corresponding gender gap in application rates. It could be that there are other meaningful factors that influence application decisions that we are not capturing, or that we are measuring these decisions with too much noise. We also collected data on a few other explanations. For instance, 29% of men and 25% of women reported that they would **never** choose to have their payment on MTurk depend upon the decisions of someone else if they had a choice, suggesting they would be highly reluctant to apply for promotion at any price, for reasons independent of our experiment. When asked on a 1-7 scale how reluctant they would be to have their payment depend upon chance or the decisions of others, with 7 being extremely reluctant, the average response is similar for men and women (4.4 and 4.5, respectively). Both of these measures are predictive of application decisions, although their inclusion does not change the estimated gender gap.

In sum, when there are no clearly stated qualifications for promotion, women believe they are significantly less likely to be promoted than men are, conditional on applying. This gender gap is partially explained by beliefs of own ability, and in particular relative ability. We estimate that, conditional on ability, women are directionally less willing to apply, but the gender gap is not significant.

Table C6. Willingness to Apply for Promotion

OLS Predicting Minimum Acceptable Promotion Bonus *No Qualifications Treatment*			
	I	II	III
Female	9.57	-12.3	73.3*
	(15.4)	(14.8)	(38.8)
Round 1 Score	-9.51***	-3.06	-3.04
	(1.71)	(2.01)	(2.01)
Believed Probability of Promotion		-2.40***	-1.68***
		(0.34)	(0.45)
Took Common		-76.5***	-67.0***
Risk Gamble		(14.2)	(19.4)
Beliefs of		-4.73*	-4.26
Round 2 Score		(2.53)	(2.94)
Female x			-1.53**
Believed Prob. of Prom.			(0.67)
Female x			-24.2
Risk Gamble			(28.7)
Female x Belief			-1.20
of Round 2 Score			(4.50)
Controls	Y	Y	Y
Observations	759	759	759
Adjusted R-squared	0.044	0.142	0.147

Notes: Controls are fixed effects for each race category, fixed effects for each education category, and a dummy for attended high school in the US, as well as dummies for each feedback treatment (no signal, 60% signal, 90% signal).
* indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

Does Reducing Ambiguity Narrow the Gender Gap?

In Table C7, we ask how more information on where the bar is impacts believed probability of promotion. We predict the worker's perceived chances of being promoted from her treatment assignment, controlling for her demographics, her true Round 1 score, her beliefs about her performance, and her risk preferences. We find that, on average, our treatment significantly reduces men's beliefs about their chances of being promoted (by approximately 3.5pp). For women, however, a clearer bar directionally increases their beliefs of being promoted (see Table 5, Column I). Note that this is true controlling for individual beliefs of own ability and risk preferences, suggesting that the mechanism is indeed operating through better communicating where the bar is (see Column II), not through changing beliefs of own ability. Thus, overall, more information significantly reduces the gender gap in believed probability of promotion.

Of course, we expect heterogeneous effects depending upon whether a participant actually possesses the desired qualification (or believes she possesses the desired qualification). So, in Columns II – V, we split the sample across unqualified and qualified participants. We do this both by actual Round 1 score (i.e. true score above the threshold, Columns II and IV), and believed Round 1 score (i.e. believed score above the threshold, Columns III and V).

As we expect, the qualifications treatment reduces the believed probability of being promoted significantly among the unqualified group. Qualified men's beliefs about their likelihood of receiving the promotion are directionally lower in the treatment than in the control, while qualified women's beliefs increase significantly, eliminating the gender gap.

But, once again, when we turn to the behavioral measure of the minimum promotion bonus at which a participant was willing to apply for promotion, these results do not hold. Table C8 replicates Table C7, but using the behavioral dependent variable. We estimate no significant impact of the qualifications on unqualified participants, nor any gender differences among these participants. Among qualified participants, the effects are also quite noisily estimated; if anything, it seems that the treatment directionally increases the gender gap in willingness to apply.

Table C7. The Impact of Clearly Stated Qualifications on Believed Probability of Promotion

OLS Predicting Believed Probability of Promotion					
	All Participants	Unqualified Participants (Round 1 score < 10)	Unqualified Participants (Believed Round 1 score < 10)	Qualified Participants (Round 1 score ≥10)	Qualified Participants (Believed Round 1 score ≥10)
	I	II	III	IV	V
Qualification	-3.55***	-4.47**	-5.54***	-2.09	-0.21
Treatment	(1.32)	(2.04)	(1.76)	(1.71)	(1.97)
Female	-2.31*	-2.25	-2.15	-2.52	-3.68
	(1.38)	(1.97)	(1.71)	(1.92)	(2.34)
Female x Qual.	3.88**	2.19	2.93	5.44**	6.12*
Treatment	(1.92)	(2.80)	(2.42)	(2.62)	(3.20)
Controls	Y	Y	Y	Y	Y
Observations	1502	706	920	796	582
Adjusted R-squared	0.429	0.285	0.306	0.408	0.340

Notes: Controls are Round 1 score, beliefs of Round 1 score – absolute and relative, beliefs of Round 2 score, risk preferences, fixed effects for each race category, fixed effects for each education category, and a dummy for attended high school in the US as well as dummies for each feedback treatment (no signal, 60% signal, 90% signal).
* indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

Table C8. The Impact of Clearly Stated Qualifications on Willingness to Apply

OLS Predicting Minimum Promotion Bonus at Which Applied					
	All Participants	Unqualified Participants (Round 1 score < 10)	Unqualified Participants (Believed Round 1 score < 10)	Qualified Participants (Round 1 score ≥10)	Qualified Participants (Believed Round 1 score ≥10)
	I	II	III	IV	V
Qualification	-8.70	-10.1	7.20	-10.4	-32.5*
Treatment	(14.2)	(24.4)	(20.8)	(16.7)	(18.2)
Female	-4.33	1.74	7.57	-11.1	-23.4
	(14.9)	(23.7)	(20.2)	(18.8)	(21.5)
Female x Qual.	12.7	20.7	-8.56	8.17	52.1*
Treatment	(20.7)	(33.6)	(28.6)	(25.6)	(29.4)
Controls	Y	Y	Y	Y	Y
Observations	1502	706	920	796	582
Adjusted R-squared	0.099	0.043	0.061	0.090	0.069

Notes: Controls are Round 1 score, beliefs of Round 1 score – absolute and relative, beliefs of Round 2 score, risk preferences, fixed effects for each race category, fixed effects for each education category, and a dummy for attended high school in the US as well as dummies for each feedback treatment (no signal, 60% signal, 90% signal). * indicates p<0.10, ** indicates p<0.05, ***indicates p<0.01, ****indicates p<0.001.

What can we make of these results? When no clearly stated qualifications are given for promotion, we find that women believe they have a significantly lower chance of being promoted than men. This is true even conditional on measured performance and measured beliefs about performance. Adding more information on the qualifications required for promotion helps to reduce this gap. However, this does not translate into significant differences in application behavior. Application decisions in our experiment, while correlated with believed probability of promotion, are also predicted by other factors. While some of these other factors might be externally relevant, such as risk preferences, others seem much less likely to be so, such as worries about having others' determine their payoffs on MTurk.