"Breaking Gender Barriers: Experimental Evidence on Men in Pink-Collar Jobs”

Alexia Delfino

11th Best Paper Award on Gender Economics

Working Paper Series

n. 154 ■ July 2021

Statement of Purpose

The Working Paper series of the UniCredit Foundation is designed to disseminate and to provide a platform for discussion of either work of the UniCredit economists and researchers or outside contributors (such as the UniCredit Foundation scholars and fellows) on topics which are of special interest to the UniCredit Group. To ensure the high quality of their content, the contributions are subjected to an international refereeing process conducted by the Scientific Committee members of the Foundation.

The opinions are strictly those of the authors and do in no way commit the Foundation and UniCredit.

These Working Papers often represent preliminary work. Citation and use of such a paper should take account of its provisional character.

The Working Papers are also available on our website (http://www.unicreditfoundation.org)

Breaking Gender Barriers: Experimental Evidence on Men in Pink-Collar Jobs

Alexia Delfino*

April 27, 2022

Abstract

Traditionally female-dominated sectors are growing and male-dominated ones shrinking, yet sectorial male shares are not changing. Why? I embed a field experiment within the UK national recruitment program for social workers to analyze barriers to men's entry into traditionally female-dominated sectors. I modify the content of recruitment messages to potential applicants to exogenously vary perceived gender shares and workers' effectiveness on the job. I find that perceived gender shares do not affect men's applications. In contrast, information that past workers' effectiveness was low encourages men to apply, improves the quality of male applicants and enables the employer to hire and retain more talented men. Survey evidence indicates that information of workers' effectiveness changes the gap in expected success on the job between talented and untalented applicants, improving selection. The net impact of the informational intervention is positive for the employer even after accounting for spillovers on female applicants.

*Bocconi University, alexia.delfino@unibocconi.it. I thank the Director of Selection and all the members of the Attraction and Selection teams in the partner organization. I thank Nava Ashraf, Oriana Bandiera, Matthew Levy and Erik Eyster for their priceless advice throughout the project. I also thank advice from Robin Burgess, Maitreesh Ghatak, Gharad Bryan, Edward Glaeser, Rachel Kranton, Marco Manacorda, Eliana La Ferrara, Thomas Le Barbanchon, Robert Akerlof, Steve Pischke, Florian Englmaier, Matthew Lowe, Johanna Rickne, Paul Gertler, Rocco Macchiavello and seminar participants at MIT, LMU, the Bank of Italy & CEPR labor market workshop and JILAEE. Francesco Sannino, Matteo Benetton, Michel Azulai, Nicola Limodio, Giulia Giupponi, Virginia Minni and Clement Minaudier provided helpful comments. I thank Miguel Espinosa for contributing to multiple rounds of revision. I thank Cristhian Acosta, Alex Blums, Andres Cordoba and Pedro Cabra for excellent research assistance. I acknowledge funding from ESRC, STICERD and the Unicredit Foundation. A special acknowledgement goes to focus group participants who gave invaluable inputs on the design. The experiment is registered in the AEA RCT Registry with ID AEARCTR-0002351 and was approved by the LSE Ethics Committee in August 2017.

1 Introduction

In the last half century, the shift from brawn-intensive to brain-intensive occupations has eroded the traditional advantage that men enjoyed in the labor market (Autor and Wasserman, 2013). A declining manufacturing share of employment in the US and other OECD countries increases the risk of men's involuntary displacement, long spells of unemployment and idleness (Ngai and Petrongolo, 2017; OECD, 2019). At the same time, the demand for female-dominated occupations, such as healthcare and education, is growing and commonly facing staff shortages (Blau and Kahn, 2017; Graves and Kuehn, 2021). And yet, organizations do not know how to attract more men into growing female-dominated occupations.[1]

Why don't men enter female-dominated jobs? The answer depends on two margins. First, we need to know whether men do not apply and identify potential frictions that can be cheaply addressed to increase male applications. Second, we need to understand whether firms can hire more men without lowering the quality of their workforce. More male applicants do not necessarily lead to more male workers, because employers will not hire untalented candidates.

This paper tackles both steps by means of a large-scale natural field experiment (List and Metcalfe, 2014) testing a low-cost real-world policy. First, I generate exogenous variation in the recruitment messages to potential applicants for a job in social work to back out supply-side barriers that hinder men from accessing female-dominated professions.[2] Second, I embed the experiment within the UK-wide recruitment of social workers, where I observe applications as well as hiring, on-the-job performance and retention for over two years for both genders.

Social work is a high-growth female-dominated occupation which, similarly to other caring professions, faces big challenges in recruiting and retaining talent, especially among men. Over the next decade, the growth rate of social workers is expected to be twice the average growth of other US occupations (Bureau of Labor Statistics, 2019), but the male share of social workers has not changed since 1970 (Blau et al., 1998). In the UK, turnover rates are as high as 30% yearly and 8% of posted jobs are unfilled (Skills for Care, 2021). Finding policies that bring talented men into social work will not only satisfy needs for diversity, but also for efficiency.

The experimental design unpacks the role of two factors commonly believed to influence the attractiveness of female-dominated jobs for men and which were top-of-mind for my partner organization: gender identity and expected effectiveness on the job.

The first factor, gender identity, embodies the idea that there is a wedge between men's and women's utility in female-dominated jobs. In this view, a man would avoid a career in social work because it is not appropriate for his gender, even if he could be very successful. Such prescription may come from the need to preserve masculinity (Akerlof and Kranton, 2000, 2005; Baranov et al., 2020) or internalized social norms (Bursztyn and Jensen, 2017; Oh, 2019;

[1] The shortage of programs targeted at men's entry into female-dominated roles contrasts with the abundance of interventions that encourage women into male-dominated sectors (see, for instance, Koch et al. (2014)). Lessons from these experiences are difficult to apply to the context of this paper if men and women face different constraints and perceptions related to female-dominated and male-dominated occupations differ.

[2] The experiment is carried out in a double-blind manner to reduce experimenter demand effects and to not interfere with the natural course of the hiring process. Compared to alternative sources of variation (e.g., monetary incentives), my design preserves the organizational systems in place and does not require hands-on administration, making it easy to scale (Al-Ubaydli et al., 2017).

Del Carpio and Guadalupe, 2021).[3] Many diversity-recruitment policies try precisely to address identity concerns by promoting images of male workers and manipulating the perceptions of gender shares in pink-collar jobs, such as nursing or teaching (The Economist, 2018).[4]

My first experimental treatment reproduces this policy benchmark by including photographs of real workers in the job advertisement and randomizing whether the portrayed worker was a man or a woman (Bertrand and Mullainathan, 2004; Benjamin et al., 2010). Auxiliary survey evidence that I gathered shows that the photographs have the intended impact of creating a wedge in perceived gender shares for both men and women (6 percentage points on average, 9% of the mean female share).

The second factor, expected effectiveness on the job, captures the idea that being a minority in a certain occupation may bring high uncertainty or even pessimism on the possibility to do a good job and be successful. In psychology and sociology, the need to feel competent and effective in the job are considered key determinants of occupational choices, especially in areas which are incongruent with own gender (Dweck and Elliot, 2005; Elliot et al., 2002; Skorikov and Vondracek, 2011; Hirschi, 2012). However, economics research on how beliefs about self-efficacy affect individual selection into jobs is scant. This is especially important in my setting, where the fear of not being competent, or perceived as such, may particularly discourage men's applications for activities which are stereotypically considered as a woman's domain (Goldin, 2006; Simpson, 2009).[5]

This is the motivation for my second experimental treatment, which aims to vary individual beliefs on the potential for being effective and perform well on the job. To create such variation, I reported in the job advert the aggregate performance in service delivery of a past cohort of workers from the organization. In a randomized way, half of the sample were informed that 66% or 89% of workers achieved the highest evaluation scores in interacting with their customers in a previous year.[6]. Auxiliary survey data I collected confirm that the statistics provided affect individual expectations of own effectiveness. In particular, I find that information of workers' effectiveness affects the gap in expected success on the job for talented vs untalented applicants. Respondents believe that talent can make a difference on the job when past workers' performance was low, but not when it was high.

The first result of the paper is that perceived gender shares do not affect men's applications, thus "a photograph is not worth a thousand words" in my setting.[7] This result is consistent

[3] Men may also have a distaste for working with a majority of women (Becker, 1957) or anticipate employers', clients' or coworkers' preferences for female workers (Folke and Rickne, 2020)

[4] The term "pink-collar" is commonly used for care-oriented jobs and fields historically considered to be women's work. For instance, portraying more male nurses is one of the pillars of the biggest recruitment drive in the history of the UK National Health System (see McGonagle, 2019).

[5] Attracting agents who are able to succeed in public service delivery is also an important and challenging objective for mission-oriented organizations and the public sector (Besley and Ghatak, 2005).

[6] I relied on the organization's historical performance records to communicate truthful, but partial information (Dal Bó et al., 2017)

[7] A potential concern is that the experimental sample may be well aware of the existing gender composition and thus does not react to the manipulation. I address this in two ways. First, I compare the reaction to treatment by respondents who are applicants for the partner organization's job to those of an online population of UK students and workers. Reassuringly, the difference in posteriors when receiving a female or male photograph is very similar in the two samples. Second, a different experiment that I ran with the same partner organization replicates the main finding that male photographs do not encourage men's applications.

with other studies which find little room for gender shares in explaining career choices (Wiswall and Zafar, 2018; Hsieh et al., 2019) and casts doubts on the usefulness of varying the display of workers' gender composition in recruitment campaigns to attract more men into female-dominated jobs. While organizations may want to showcase diversity for other reasons (e.g., reputation), their media content is unlikely to affect the sticky association between social work and women.[8]

In contrast, my second result is that informing male candidates of their potential for being effective in the job affects applications, the quality of the pool of applicants and the downstream performance that new male hires have on the job.

At the entry stage, men are more likely to apply when informed of low effectiveness among past workers. The increase in applications is 14% with respect to receiving information of outstanding past performance ($p<0.05$) and 9% with respect to a pure control group ($p=0.15$). In line with the psychology literature, finding out that workers achieved low success in the past increases individual expected effectiveness and encourages candidates to apply (Elliot et al., 2002). This effect is stronger among candidates with high predicted job-specific talent, a fact consistent with my survey evidence that the experimental information changes the gap in expected success between high and low talent people. High-talent applicants thus seem to be disproportionately attracted by the possibility of making a difference in the job.

The quality of the pool of applicants across treatments supports this interpretation. Applicants in the low past effectiveness group have better observable characteristics and receive more job offers than applicants in the high past effectiveness group. This gap in offer rates arises from two channels. Quality at the top of the distribution is highest in the low effectiveness treatment. At the same time, quality at the bottom is lowest in the high effectiveness treatment. This means that a workplace where everyone performs well attracts a tail of unfitted applicants who are seeking insurance from failure, but who are ultimately not hired by the employer.

Crucially, higher quality among applicants allows the employer to hire and retain better male workers. In the first two years on the job, men attracted by information of low workers' effectiveness show a consistently higher performance (by 0.25 SD), they are not more likely to leave the job and report higher levels of job satisfaction and intent to stay than workers in all other experimental groups.

The magnitude of the effect of providing information of low workers' effectiveness is large, which is noteworthy given that the treatment is cost-free for the employer. For a candidate with average job-specific talent, the increase in men's applications is comparable to the estimated effect of an 11% increase in the wage, which would cost the employer more than a million dollars a year to implement. How can information have such a big impact? One plausible explanation is that men have high uncertainty on their expected effectiveness, and thus react strongly to new information. This is supported by evidence that the informational treatment has the strongest impact on application rates among men with limited exposure to social work or female-dominated jobs in general (Jensen, 2010; Charles et al., 2018; Baranov et al., 2020).

[8] A possible reason for this stickiness may be that the association between social work and women is not linear in gender shares, and requires the male share to be above a certain threshold, like in models with tipping points (Pan, 2015).

The last result of the paper is that the benefit of providing information of low past effectiveness for the employer is positive even after taking into account spillovers on the selection of women. I find that women are insensitive to information provision on average, which could lead to the conclusion that changing expectations is a silver bullet for the employer: it achieves higher diversity, better quality among the gender minority and has no effect on the majority. Nevertheless, the very fact that men start entering social work might have an impact on women's behavior. If male shares in female-dominated jobs increased in the long run, what would be the impact on women's choices? I find that fewer women apply when they believe that there are more male social workers in the job (a difference of 7.5% between the male and female photograph arm). Moreover, female recipients of the male photograph who are hired, but under-perform in the first semester in the job, are more likely to quit the job compared to women with a similar performance, but who received a female photograph. Thus the male photograph creates a trade-off for the employer: on the one hand the quality of female workers who stay in the job is higher, but on the other hand turnover is larger. I conducted a survey with the recruitment personnel of my partner organization and found that most of them deem the increase in performance to be high enough to justify the higher turnover. This leads to a positive net effect, for the employer, of the informational policy that attracts more men in this job.

I use a simple selection model to go beyond the employer's perspective and think about economy-wide implications. The joint increase in the size and talent of the pool of applicants in the experiment is consistent with men's negative sorting in social work, with the marginal male applicant being more talented than the average one. Keeping fixed women's distribution across occupations, male sorting implies that bringing more men into female-dominated jobs can ameliorate talent allocation overall, by increasing average talent also in the outside option.

Taken together, my results suggest that breaking informational barriers to men's entry into female-dominated jobs might increase gender diversity and improve overall workforce quality in a gender-neutral way. This yields an optimistic message for policy. Both the stigma associated with working in a female occupation and men's perceptions of their expected success have been central in the debate around the conversion of unemployed men into female-dominated service jobs (Miller, 2017). The two have different policy implications. The femaleness associated with some occupations may be difficult to modify and changes in gender composition take time. While people can be monetarily compensated or compositional changes can be accelerated through quotas, uncertain or incorrect expectations can be more cheaply tackled through information provision and incentives, for example through low-cost organizational practices that recognize good performance. While increasing diversity, a focus on (cheap) ways to reward performance could also help public service jobs ameliorate the difficulty of finding and retaining talent (Delfgaauw and Dur, 2010; Finan et al., 2017).

This paper contributes to the growing empirical work on the impact of stereotypes and social identity on occupational choices (Hoff and Pandey, 2006; Bordalo et al., 2016; Glover et al., 2017; Alan et al., 2018; Carlana, 2019; Oh, 2019; Coffman et al., 2019; Del Carpio and Guadalupe, 2021). Previous work has analyzed the overall impact of identity on choices. However, group belonging (e.g., gender) and group-specific expected success jointly determine the identity-payoffs that a person enjoys in a certain task or occupation. We have little knowledge on how these two economic fundamentals compare in explaining behavioral differences between groups. My paper is a step in this direction and my design proposes a novel way to disentangle them.

Relatedly, while studies about gender differences across male-typical or female-typical domains abound in the lab (for a review see Azmat and Petrongolo (2014)), tasks in the latter tend to be highly stylized (e.g., knowledge of Disney movies). This is the first paper to study the determinants of men's choices in a real female-typical field setting. My setting is not only highly relevant from a policy perspective (Katz, 2014), but it is also a high-stakes, as my participants are taking a real-life choice with far-reaching implications for their careers.

The paper also speaks to the personnel economics literature on the effects of advertised job amenities or requirements on applications (Dal Bó et al., 2013; Marinescu and Wolthoff, 2020; Ashraf et al., 2020; Deserranno, 2019; Abebe et al., 2021; Flory et al., 2019; Coffman et al., 2021; Abraham and Stein, 2022). As in Flory et al. (2019) and Abraham and Stein (2022), my paper focuses on recruitment strategies aimed at increasing organizational diversity. I expand this work by following hires on the job, which allows me to quantify the net benefit of increasing diversity for the employer, as well as to uncover the existence of long-term effects of diversity recruitment policies on the majority in the job (Azmat and Boring, 2020). The importance of reducing uncertainty in men's expected effectiveness complements the results of Coffman et al. (2021), who show that reducing ambiguity related to their job-fit encourages women into male-typed domains.

2 Institutional context

There are several reasons why social work is a good setting for my research questions. Women historically represent more than 70% of social workers in the US and in the UK (Figure A.1a). Similarly to many other female-dominated service occupations, the demand for social workers is expected to grow (Figure A.1b) and to face looming staff shortages in the following decades (Lin et al., 2015). The growth rate of social workers is expected to be twice the average growth rate across all US occupations, and to be greater in areas of high male joblessness (Figure A.1c).

A consistently high female share of workers may create uncertainty about men's performance in social work. I asked students and candidates for a social work job what they think the performance of a male or female social worker is. Figure A.2 shows that both male and female respondents expect men to perform significantly worse than women in the job, but have a relatively low confidence level in their expectations.[9] Moreover, men are even more pessimistic and less confident than women. This evidence suggests that men may lack information to estimate their likelihood to succeed in social work.

During 2017, I collaborated with one of the main UK recruiters of public sector social workers. The organization offers a two-year on-the-job training position targeted at either final year students from all disciplines or current workers from any industry. Previous experience in social

[9] In my partner organization, female and male social workers perform similarly.

work is not required, a feature which allows me to study entry into this occupation broadly, not only looking at people with job-specific training. Moreover, heterogeneity in terms of background exposure to social work makes informational and psychological constraints particularly relevant for part of the sample. Every year, new hires are assigned to teams located in Local Authorities across England and earn a stipend which is comparable to the average UK annual entry salary in social services (26k GBP), primary school teaching (24k GBP) and nursing (22k GBP). The daily job involves both office tasks (e.g., case writing) and meetings with families in need and other stakeholders, such as lawyers, medical professionals and the police. After the program, the majority of workers stay in similar positions (between 60% and 70%). Among those who leave the job, many switch to policy-making positions in UK or international organizations.

The program is part of the movement towards the professionalization of the public sector. As such, the organization's key challenges are related to the attraction, retention and diversity of talent. Attracting gender diversity without sacrificing talent or retention was also the goal of the experiment we designed together.

Figure 1 illustrates the timeline of the organization's 2017 nationwide recruitment. The experiment happened from September to November, which is the application period. The hiring process consists of different stages (e.g., interviews), which are conducted in a centralized manner either online or at the organization's head office in London. The overall duration of the hiring process from application to job offer is around ten weeks. If a person was hired and accepted the job, work in local authorities started in July 2018. I followed workers for the entire duration of the program until July 2020.

3 Experimental design

Experimental participants are people with an interest in applying for the job offered by the partner organization. To express this interest, candidates fill in a short registration form on the organization's website which contains eligibility and demographic questions and takes between three and five minutes to be completed. If eligible to apply, respondents receive an invitation-to-apply email.[10] The need to register implies that the experimental sample is selected on a minimum interest for the job. From a policy perspective, it can be argued that this is exactly the relevant sample to be targeted. Application rates after registration are around 50% for men (60% for women), so there is substantial room for changing entry into the job within this group. I address external validity of the sample in Section 5.3.

I introduce exogenous variation in the content of the invitation-to-apply email along two dimensions: photographs and information. Two experimental conditions were cross-randomized in a fully nested design, leading to a total of four treatment emails. Participants could also be randomly assigned to receive a fifth control email containing no manipulation, which I used to compare the treatments with business-as-usual for the organization. Randomization was at the

[10] The email contains their candidate number, which is necessary to access the application process, and some basic information about the hiring process. Respondents who do not meet the eligibility requirements receive a standard rejection email. Eligible applicants should have a bachelor's degree with a certain minimum average grade and have obtained at least a C in Math and English pre-university qualifications. Social work students are not eligible.

individual level, with stratification by gender (man/woman), ethnicity (white/non-white) and whether a person registered during the week before the official opening date. The experiment was double-blinded: participants were not aware that the invitation-to-apply email was part of a research study and recruiters were not aware of candidates' treatment assignment. This design limits biases that arise from candidates' knowledge of being in a research study and prevents recruiters' assessment from being influenced by the candidates' treatment. I discuss each of the experimental manipulations below. Figure 2 shows an example of treatment email.[11]

Photograph manipulation - variation in perceived gender shares. The invitation-to-apply email contained a photograph of a real worker, who was randomized to be either a man or a woman. This experimental condition varies potential applicants' perceived gender shares if seeing a male photograph generates a perception of a higher male share than seeing a female photograph. While this is the main interpretation that I adopt in the paper, photographs may also vary the salience of the predominantly female composition of the job, making it more or less desirable for a man to apply.[12] These two interpretations are observationally equivalent and I do not aim to distinguish between them.

The reaction of participants to this manipulation identifies the utility given by the workplace gender composition (or related attributes) on their application decisions, if photographs affect choices mainly through changing perceived gender shares. The ethnicity of portrayed workers could confound this identification strategy: if white female candidates apply more after seeing an email portraying a white woman than they do after seeing an email with a non-white man, we would not know whether to attribute the effect to the gender or ethnicity match. Moreover, showing photographs of white people right before starting a selection process might cause a stereotype threat in non-white subjects (Steele and Aronson, 1995). For these reasons, I assigned different photographs to white or non-white people and matched the ethnicity of photographed workers with that of each candidate (randomizing gender).

The design of this manipulation also addresses other potential confounders. To attract the candidate's attention to the photograph, I added a short text where the photographed person addresses the candidate by name and recalls that she/he was also once an applicant. Drawing on studies on role models (Marx and Ko, 2012) and information retrieval (Schwarz et al., 1991), this message should facilitate the candidate's relatability to the portrayed person and the gender group she/he belongs to. The photographed people are real workers who did not feature in other advertising campaigns or multimedia content from the organization, in order to limit the impact of unobserved heterogeneity in candidates' exposure to the organization's media channels and recruitment materials. All photographs show the same background and are of the same size to limit visual differences. Other issues might arise if there is a systematic correlation between portrayed workers' characteristics and their gender. Table OA.1 shows that male and

[11] The pure control email did not contain any intervention box.

[12] My interpretation of the photograph manipulation aligns with audit studies (Bertrand and Mullainathan, 2004), where non-white sounding names increase the employer's expectations that the candidate is going to be non-white. Alternative interpretations based on salience align with priming studies (Benjamin et al., 2010) and could be accounted for in the model in Section 4 with a change in α_i. This manipulation is also inspired by common business practices. For instance, Glassdoor advices employers to "Include photographs of women and minority employees on your careers site, but don't use stock photography" as a strategy for diversity recruitment.

female photographed workers are deemed to be similar in characteristics such as friendliness, attractiveness, or work satisfaction by three different samples of survey respondents.

Information manipulation - variation in expected effectiveness. The goal of this manipulation was to convey to each person what their effectiveness on the job could be, given their talent. Facing constraints on the types of real information available, the organization and I decided to provide information about past workforce effectiveness, allowing each participant to infer their own potential for success. I communicated to subjects the outcome of a selected past cohort of workers, which had either low or high aggregate effectiveness in service delivery. The exact wording was the following:

> Did you know that in a past cohort X% of participants got commendable or excellent feedback to their interaction with families?

where X was equal to 66 or 89 in the two experimental treatments. Commendable or excellent are the highest grades that people can achieve in their performance assessments in the job, which are given by external reviewers who are experts in social work. The experimental information refers to the evaluation that workers got when interacting with their customers (i.e. families), thus it indicates the effectiveness in service delivery obtained by previous workers. Both statistics were computed using actual records of the organization. This enabled the communication of truthful but partial information, which on average creates a wedge in beliefs between experimental groups (Dal Bó et al., 2017). I label the treatment disclosing low past performance (66%) as "Low workers' effectiveness" and the one disclosing an outstanding past performance (89%) as "High workers' effectiveness". The reaction of participants to this manipulation is meant to identify the role of individual expectations of one's potential for being effective on the job, and not whether people care about workforce average effectiveness per sè. The next section shows manipulation checks and tackles alternative interpretations.

I reported information about on-the-job success in frontline interactions with clients - and not, for instance, academic grades - for several reasons, primarily to induce variation in people's beliefs concerning their effectiveness in generating output for the employer. Performance metrics on client service are also rarely collected and/or published in the industry, a fact which increases the likelihood that the provided information will affect a candidate's beliefs. Additionally, the quality of clients' interactions is one of the crucial objectives of the organization's mission and it is an important variable that candidates consider when applying (Besley and Ghatak, 2005).[13] Finally, effectiveness in service delivery also depends on clients' reactions. A low score can signal clients' hostility and/or discrimination towards the employees, which can disproportionately affect men's and other minorities' assessment of their potential job success.[14]

3.1 Main manipulation checks

I carried out auxiliary survey experiments to provide evidence on subjects' interpretation of the treatments. I chose not to conduct manipulation checks with participants from the field

[13] To make this even more salient, the box was positioned below a summary of the organization's mission, which is focused on the challenge of improving outcomes for disadvantaged communities. According to surveys run by the partner organization, 35% of applicants mention this challenge as their main motivation for applying.

[14] Information of low past performance could signal that clients are discriminating, which would make men and non-white people less likely to apply. The opposite effect - which is what I find - excludes this interpretation.

experiment because the survey could have interacted with their reaction to treatment (e.g., making gender salient) and interfered with the double-blind nature of the experiment. I administered the survey to two out-of-trial samples: the following year (2018) applicants for the partner organization's job and an online sample recruited through the platform "Prolific Academic". The former sample allows me to test the reaction to treatment in a subject pool as similar as possible to my field experimental sample. The latter group allows me to generalize the validity of my manipulation checks in a wider sample of students and workers matched on observables with the applicants. Appendix OE. describes the pre-registered survey implementation. I used a between-subject design and assigned each respondent to one of the treatment emails used in the field experiment. After mandatory comprehension checks, I elicited beliefs on a variety of characteristics of the job, applicants and workers.

Figure 3 shows the main manipulation checks for the photograph treatment. The first question asks: "100 people apply for the job after seeing the email ad that you have just seen. Of these 100 people that apply, how many do you think are women?". The mean female share is 73.8% in the female photograph group and 68.3% in the male photograph group ($p<0.001$). These percentages are almost identical between male (Panel 3a) and female respondents (Panel 3b). This means that participants predict that gender shares will change as an effect of the treatment. Because each participant is only aware of the existence of his/her treatment group, this is consistent with the interpretation of the photograph treatment in terms of a shock in perceived gender shares.

Two additional survey questions address whether the treatment changes the association between women and the job. On average, 61.4% of respondents think that the job is desirable for a man when seeing a female photograph, but this share jumps to 75.8% when respondents are shown a male photograph ($p<0.001$). Similarly, on average the male photograph decreases the share of people who think that the job is desirable for a woman. However, Figure 3 shows that these differences are almost entirely driven by female respondents (Panel 3b). Men and women share similar perceptions about the job's desirability for the two genders when seeing a female photograph, but only a limited share of men changes such perceptions when receiving a male photograph (Panel 3a).

The two samples of job applicants and online respondents react in the same way to the photograph manipulation (Table OA.3). Moreover, Figure 5a shows that photographs do not systematically convey other types of information about the job (e.g., difficulty, wage, promotions) and Table OA.1 shows that men and women portrayed in the photographs are deemed very similar on a variety of dimensions such as work satisfaction, happiness or attractiveness.[15]

Figure 4 shows the main manipulation checks for the information treatment, by gender. Two questions are used to check whether respondents believe the information provided and consequently update their beliefs on average workforce effectiveness (left-hand side Figures). Both male and female respondents reading the 89% statistic think that there is a higher share of successful workers in the job than when reading the 66% statistic (a gap of 6 and 9 percentage

[15] The statistical significance of differences in perceived gender shares is robust to corrections for Multiple Hypotheses Testing (MHT from here on), in the pooled data (Table A.2) and by sample or gender (Tables OA.3 and OA.5).

points, for men and women respectively, p<0.01 for both genders). Figure 4 also shows that people receiving information of outstanding past effectiveness think that 78.4% of applicants have the potential to get commendable or excellent feedback on the job (75% for men and 80% for women), while only 69% are deemed to have this potential in the group of people receiving information of low past performance (66% for men and 70% for women, p<0.01 for both genders).

This evidence indicates that respondents update their beliefs on the aggregate performance of workers and applicants, but we still don't know whether - and how - information affects *individual* expectations of effectiveness. On the one hand, having many successful peers may increase the likelihood that a person will become successful on the job as well (e.g., through better training). On the other hand, few past successful workers may signal that the employer is not selective in hiring or is lenient on the job, predicting that every person's expected success in this job will increase. Figure 5b rules out both these explanations by showing that respondents' estimates of the effort needed in the application process or of the quality standards required by the organization are the same in the two information treatments.

Alternatively, effects may be heterogeneous by job-specific talent. Few successful past workers may signal that only talented workers can make a positive difference in the job, implying that expected success may increase for talented people and decrease for untalented ones. Additional survey questions provide support for this interpretation. The last bars of Figure 4a show that self-reported ability in the two information treatments is the same for male respondents. I thus use this variable to classify people into high (above-mean) and low (below-mean) ability, and test whether expected success in the job changes differentially in these two groups.

I find that the information treatment affects the gap in expected job success between people of high and low ability.[16] This is shown in the right panel of Figure 4a, which reports mean answers to the question "Consider 100 people who are applying for this job. Based on the ad you just viewed, on a scale from 1 (worst) to 100 (best), how would you rank yourself for the job among them?", by information treatment and self-reported ability level. Low-ability people expect to do as well as high-ability people when receiving the 89% statistic (p>0.15). When receiving the 66% statistic, however, people believe that ability differences are going to map into different success on the job (p=0.01). This widening of the gap in expected success is driven by men below-mean ability, whose expected ranking significantly decreases when receiving the 66% statistic than the 89% statistics. All in all, this evidence indicates that information of past workers' effectiveness influences people's beliefs on the mapping between talent and success on the job, essentially affecting expectations of (non-monetary) rewards to talent. Figure 4b shows that this effect also holds for women.

Table OA.2 shows that the 2018 Applicants and the Online sample have similar beliefs in the two information treatments. Moreover, Figure 5b rules out alternative interpretations of the information provided, such as updating on job amenities (e.g., wages, promotions, training quality). There are few small differences in beliefs on other job characteristics for women (e.g.,

[16] Self-reported ability may be inflated because of demand bias or overconfidence. This is problematic only if it alters the ranking of abilities in the sample, but results on this possibility are mixed in the literature (see Moore and Healy, 2008; Coffman et al., 2019). Moreover, my manipulation checks are still valid even in the case of rank-reversals as long as the self-reported ability captures beliefs that drive people's choices.

social status and discrimination), but none of them survives MHT corrections (Tables A.2, OA.3 and OA.5).

4 Theoretical framework

I now present a theoretical framework to generate predictions on the size and quality of the applicants' pool in each treatment and for different parameter ranges.[17] The model formalizes the insight that information on workers' effectiveness affects the gap in expected job success between candidates with high and low talent, thus affecting expectations of rewards to talent.

Potential applicants are characterized by gender $g \in \{M, W\}$ and talent $a_i \sim U[.]$, which includes both skills and the motivation for doing a good job. Both g and a_i are known to each candidate.[18] At time 0, potential applicants assess whether to apply for a female-dominated job j. Their expected utility in j is:

$$U^j(a_i) = \alpha_i s_g + w + \theta_g(a_i - \hat{a}_g)$$

where s_g the share of workers of gender g in the job, w is the wage, θ_g is the marginal impact of talent and $\hat{a}_g$ is a minimum ability requirement (Lazear et al., 2018).

Agents care about the output generated for the employer through $\theta_g(a_i - \hat{a}_g)$. Everyone enjoys doing a better job than required (being above $\hat{a}_g$), but good performance has a disproportionate impact on utility when θ_g is high. This utility boost may come from expecting higher rewards to talent (e.g., career advancement), from internalizing the positive social impact of good performance (Besley and Ghatak, 2005) or from the satisfaction of feeling competent (Dweck and Elliot, 2005). If the employer's production function is linear in a_i, θ_g simply represents the marginal productivity of talent. Similarly, an untalented person ($a_i < \hat{a}_g$) suffers a greater disutility from feeling incompetent or worrying about not being useful for the clients when θ_g is high. When θ_g is low, having talent above or below a minimum requirement does not have a big impact on job utility, for instance because good performance is not rewarded by the employer or does not make a difference in the lives of the beneficiaries.

$\alpha_i s_g$ formalizes agents' utility from workplace gender composition, which is linear in the share of their own gender g with an individual positive weight $\alpha_i \in [0, 1]$. Different channels can rationalize this preference, such as gender identity (Akerlof and Kranton, 2000, 2005), anticipated harassment (Folke and Rickne, 2020) or image concerns (Bursztyn and Jensen, 2017).[19]

At time 0 agents know w and $\hat{a}_g$, but do not know the exact gender shares and how much utility they are going to get from being above the minimum ability requirement. Priors on both s_g and θ_g are normally distributed with gender-specific distributions: $s_g \sim N(\bar{s}_g, \sigma^2_{s_g})$ with $\bar{s}_W > 0.5$ and $s_M = 1 - s_W$ and $\theta_g \sim N(\bar{\theta}_g, \overline{\sigma}^2_g)$, with $\theta_W \perp \theta_M$. Priors on θ_g satisfy the following

[17] The framework draws on a two-occupations Roy model (1951) with perfect correlation between skills.

[18] Different transformations of a_i are possible (e.g., overconfidence) and do not affect the theoretical predictions as long as they do not alter the ranking of abilities in the sample.

[19] When it represents internalized social stigma, the individual component $\alpha_{iM} s_g$ can be micro-founded through a game between applicant i and his peers, where α_{iM} is the cost of social punishment for selecting a female job and $1 - s_g$ is the likelihood that the punishment will be enforced.

assumptions:[20]

Assumption 1. ***Gender differences in beliefs about θ_g***
On average, men's beliefs on the marginal impact of talent are lower than women's: $\bar{\theta}_M \leq \bar{\theta}_W$.

Assumption 2. ***Gender differences in uncertainty***
Men's priors on θ_g are noisier than women's: $\overline{\sigma}^2_M \geq \overline{\sigma}^2_W$.

At time 1 a non-strategic employer posts recruitment messages to potential applicants in order to change their expectations of effectiveness and beliefs on gender composition. Recruitment messages are vectors (P, S) which combine a photograph $p \in \{M, W\}$ and a signal of potential impact of talent $S \sim N(\theta, \sigma_s^2)$, where $\frac{1}{\sigma_s^2}$ is the signal precision. I denote the experimental realizations of the signal $s \in \{s_L, s_H\}$.

At time 2 each applicant sees one particular recruitment ad (p, s) and updates beliefs on s_g and θ_g. Seeing a photograph of gender g increases the perceived share of that gender in the job: $E[s_g|p = g] > E[s_g|p \neq g]$. Seeing that workers in the past were not effective induces people to think that a talented person can make a bigger contribution in the job, and thus increases the expected impact of talent in the job: $E[\theta_g|s = s_H] > E[\theta_g|s = s_L]$. In the experiment, s_H (s_L) corresponds to information that 66% (89%) of past workers were effective.

At time 3, potential applicants decide whether to apply to j given their posteriors on s_g and θ_g. Potential applicants apply for the female-job if: $U^j(a_i) - \frac{c}{p(a_i)} > U^o(a_i)$ where c is a small application cost and $p(a_i)$ is the probability of being hired, with $p(0) = 0$, $p'(a_i) > 0$ and $p''(a_i) < 0$. $U^o(a_i) = w^o_g + v_g a_i$ is utility in the outside option, where v_g is the marginal return of talent.[21]

4.1 Results

Define $\Delta U(a_i) = U^j(a_i) - U^o(a_i)$ and $C(a_i) = \frac{c}{p(a_i)}$. Assume that $\Delta U(0) < C(0)$, so that a person with no talent will not apply. Moreover, assume that there exists a level of talent a^* indifferent between the job j and the outside option o when there are no application costs, such that $\Delta U(a^*) = 0$. Whether the most talented or least talented people apply for j depends on the slope of utilities with respect to talent in the job ($U^{j'}(a_i) = \overline{\theta}_g$) and in the outside option ($U^{o'}(a_i) = v_g$). When $\overline{\theta}_g > v_g$, the most talented people select into the job and there is a unique threshold of ability $\overline{a}_g$ such that any $a_i > \overline{a}$ applies for the female-dominated job. When $\overline{\theta}_g < v_g$, people of either very low or very high ability do not apply for the job. The applicants' range of talent is defined by two thresholds of ability, $\underline{a}_g$ and $\overline{a}_g$, such that only levels of talent $\underline{a}_g < a_i < \overline{a}_g$ apply for the job.[22] From hereon I focus on the case in which marginal gains from ability in social work are lower than in the outside option $\overline{\theta}_g < v_g$, which is the empirically relevant case as long as private-sector options - which have steeper career progression or bonuses - contribute to candidates' outside options.

[20] Assumptions 1 and 2 are equivalent to assuming risk aversion in the utility function and keeping only the assumption of asymmetric uncertainty. Figure A.2 provides empirical evidence that men have lower and more dispersed expectations of performance in social work than women.

[21] To ease notation, utility in the outside option does not depend on gender shares. What matters is the difference in gender shares between job j and the outside option, thus s_g can be reinterpreted as such difference.

[22] See Appendix OF. for the formal proofs related to the results in this Section.

The first result states that more applications when people of gender g receive a same-gender ($p = g$) than other-gender ($p \neq g$) photograph identify the effect of gender shares on utility. Whether such increase in applications also leads to an increase in average quality depends on the relative change in marginal ability at the top ($\bar{a}$) or bottom ($\underline{a}$).

Result 1. ***The effect of a shock to perceived gender shares***
When $p = g$, the pool of applicants is larger than when $p \neq g$.
When $p = g$, marginal ability at the top ($\overline{a}$) is higher and marginal ability at the bottom ($\underline{a}$) is lower than when $p \neq g$, with an ambiguous comparison in average quality.

The difference in the size of the applicants' pool between the female and male photograph groups is increasing in the generated difference in perceived gender shares and the utility weight on gender composition α_i. Figure 6a shows Result 1 graphically.

The second result focuses on the effect of a change in expected impact of talent $\overline{\theta}_g$. Assume that the change in $\overline{\theta}_g$. is small enough not to invert the sign of the difference $\overline{\theta}_g - v_g$.[23]

Result 2. ***The effect of a shock to expected impact of talent***
If $\hat{a}_g \in [\underline{a}, \overline{a}]$ and $s = s_H$, the size of the pool of applicants can be larger or smaller than when $s = s_L$. Marginal abilities $\underline{a}_g$ and $\overline{a}_g$ are higher when $s = s_H$ than $s = s_L$, thus average ability is greater.
If $\hat{a}_g \notin [\underline{a}, \overline{a}]$ and $s = s_H$, the size of the pool of applicants is larger than when $s = s_L$. Marginal ability $\underline{a}_g$ decreases and $\overline{a}_g$ increases when $s = s_H$ than $s = s_L$, thus the comparison of average ability is ambiguous.

Result 2 shows the conditions under which raising expectations on the potential for being effective unambiguously improves the average quality of the pool of applicants. This happens when the minimum ability requirement $\hat{a}$ takes intermediate values, such that talents above and below $\hat{a}$ apply for the job. Quality improves by discouraging applications at the bottom of the ability distribution ($\underline{a}$ improves) and, at the same time, increasing them at the top ($\overline{a}$ improves). The discouragement effect at the bottom makes predictions on the size of the applicants pool ambiguous.

When the minimum ability requirement $\hat{a}$ is low (high), only talents above (below) $\hat{a}$ apply for the job. The effect of an increase in θ_g is now similar to the effect of increasing beliefs on gender shares: the pool of applicants increases by attracting more applications either at the top or the bottom. Figure 6b shows Result 2 graphically. The difference in utility between the treatment providing $s = s_H$ and $s = s_L$ is proportional to the change in beliefs between the two conditions and the distance between individual talent a_i and $\hat{a}$. Because updating is stronger for candidates with weaker priors, Assumption 2 implies that men will update more than women when receiving the same signal: $\Delta\theta_M > \Delta\theta_W$. The effect of the shock to expected θ_g on the applicants' pool will thus be greater for men.

In sum, if people care about workplace gender composition, an increase in the perceived share of own gender in the job can increase applications. However, the ability level of the pool of applicants depends on the shape of the cost function, which determines whether low or high ability react the most to the treatment. Increasing the expected effectiveness of talent can

[23] The posterior expected impact of talent when receiving s_H could be higher than v_g,thus violating condition $\theta_g < v_g$. I only consider the case in which both priors and posteriors when receiving s_H are lower than v_g.

unambiguously improve the quality of applicants by discouraging low ability candidates and encouraging high ability ones. This prediction is consistent with the differential updating of expected ranking by ability seen in Figure 4a.

The model can also accommodate alternative interpretations of the treatments. For instance, a change in the salience of gender caused by the photograph treatment could be accounted for through an increase in α_i, which would generate the same predictions of result 1. Information of low past effectiveness could decrease beliefs on the minimum ability requirement $\hat{a}$, causing an upward parallel shift in utility on the job across levels of ability. Result 2 would still go through.

5 Empirical strategy, sample and balance

5.1 Empirical strategy

My main empirical strategy relies on the independent random assignment of the photograph and information manipulations and compares the impact of one type of photograph (or type of information) against the other in affecting men's entry and selection into social work. I estimate the following regression, separately for men and women:

$$y_i = c + \beta_1 Pic_i^M + \beta_2 Effectiveness_i^L + X_i'\lambda + \epsilon_i \tag{1}$$

where Pic_i^M is equal to one if potential applicant i was assigned to receive a male photograph (zero for female photograph) and $Effectiveness_i^L$ is an indicator variable for information of low workers' effectiveness (zero for high). The vector of controls X_i contains stratification variables: indicators for non-white ethnicity and for whether the person registered before the official opening date. Randomization was at the individual level, so I use Eicker-Huber-White robust standard errors. For robustness, I use randomization inference (Young, 2018) and, when appropriate, I adjust standard errors for multiple-hypotheses testing (Benjamini et al., 2006; List et al., 2019).

The coefficient β_1 tests the null hypothesis of no effect of perceived gender shares on outcome y_i, conditional on also getting information of workers' effectiveness. The coefficient β_2 measures how outcome y_i changes in the low effectiveness vis-á-vis high effectiveness treatment, conditional on getting also a worker's photograph. Coefficients β_1 and β_2 identify the causal effect of gender shares and expectations on y_i, respectively, under the assumption of no interaction between the two manipulations. The no-interaction assumption is empirically valid for men (Table A.4 and Figure OA.1). Appendix B shows alternative specifications which include the pure control. However, the interpretation of the comparison with the pure control is not straightforward because each treatment email simultaneously changes information and photographs.

I consider two primary sets of variables y_i, which were pre-registered: applicants' outcomes (entry and quality) and workers' outcomes (performance and retention).

Applicants' outcomes: While candidates have to pass through different stages of selection, empirically I am only interested in the cumulative effect of the treatment on the candidates' decision to enter the job over the entire process. Accordingly, my outcome variable for application takes the value of one if a candidate decides to submit the application form and shows up at any later stage of the selection process, conditional on being admitted to that stage. To assess whether the treatment attracts better-skilled or worse-skilled applicants, I consider whether i receives a job offer (conditional on application) and whether s/he accepts the offer (conditional on receiving it).[24]

Workers' outcomes: new hires are continuously evaluated throughout the duration of the two-year program. I measure performance in the job for worker i as the weighted average test score, where weights are given by the credits that the organization assigns to each test. Performing well on these tests is important to be able to qualify as a social worker as well as for future career opportunities.[25] To measure retention, I define an indicator for quitting the program before the end.

Identification assumptions: my empirical strategy aims to show that recruitment ads affect selection into social work, which in turn affects the performance of people who enter the job. To be able to interpret differences in outcomes as the causal effect of the treatment on the selection of the pool of applicants, the identification assumption is that the individual probability of being successful from one stage to the next is independent of treatment assignment. This assumption could be violated if treatment assignment influences the employer's screening criteria or candidates' effort. While the double-blind design mitigates these concerns, Section 9.1 provides empirical support.

5.2 Sample

The experimental sample consists of 5417 candidates, of whom 1013 are men and 807 of them are assigned to one of the treatment groups. Table 1 presents summary statistics by gender and balance checks for the overall experimental sample. Candidates' average age is 27 and 3 out of 10 are ethnically non-white. Approximately 32% of the candidates studied in a top-tier UK university, but at the same time the share of people from lower socio-economic backgrounds is substantial, with 19% of subjects coming from families where parents have an unskilled occupation, 27% of subjects receiving economic support in school and 2% being looked after by a social worker as a child.

Men and women have similar socio-economic backgrounds and exposure to the organization, but differ in terms of demographics, education and employment. Men tend to be older and, therefore, more likely to have graduated before 2016 or to be in fulltime employment (FTE). The same share of men and women attended a top UK university or got a first grade, but men are more likely to have studied scientific subjects and, if working, to be in corporate, scientific or business jobs.

Treatment assignment is balanced on observables. Columns 7 and 8 of Table 1 report the F-statistics and the related p-value of a regression for each of the row-variables on the set of

[24] I show results on unconditional variables in Table A.3.

[25] Assessments include both theory (e.g., case studies, essays) and practice tests. The theory assessments are evaluated by experts in social work in anonymous form. Anonymity is not possible in the first month performance review and the practice assessment. The former is a score given by teachers at the end of the mandatory classroom-based training phase which evaluates the fitness and potential of each worker to do a good job in interacting with families. The practice score is given through direct observation of the way in which a worker interacts with customers. Evaluators were not aware of candidates' treatment or of the experiment.

four treatment indicators. The last column of Table 1 reports the minimum p-value of pairwise t-tests for the difference in means between each pair of treatments along the variables reported. I also fail to reject the null hypothesis of zero effect of all the variables reported in Table 1 in a joint test of orthogonality on assignment to any treatment group (F(23, 4865)=0.67).

5.3 External validity

Participants in my study have expressed an interest in the job, a fact which could limit the external validity of the results. Here I discuss how representative the study group is in terms of preferences, beliefs or constraints (List, 2020).

Men in my sample may care less about the job gender composition (low α_i in the model) compared to the male population. If so, they will be less sensitive to the photograph manipulation and my estimate of β_1 will be biased downward. My male sample may also be more informed about social work than average, implying more precise priors of expected success (low σ^2). If this is the case, they will react less to the information manipulation, thus the estimate of β_2 will be biased downward.

Differences in both preferences and beliefs predict that my estimates of the effect of photographs as well as information should be biased downward in my sample compared to the overall male population. As long as the sample is equally selected on attitudes towards gender composition and information on expected success, sample selection should not introduce a differential bias between treatment groups. I maintain this assumption throughout the paper.

Table A.1 provides empirical support by showing that applicants have indeed more precise priors on both men's and women's performance in social work and gender shares than an online population of workers and students matched on observables.[26]

Finally, participants in the experiment may also have different constraints and outside options than average men (differing in parameters such as v_g or w_g^o). This implies that selection on talent could be different in other samples facing different structural parameters. I perform heterogeneity analysis which is portable across contexts to address this point.

6 Results

This section contains the main results of the paper. First, I find that gender shares do not affect men's applications (Figure 7a). In contrast, information of low workers' effectiveness encourages more men to apply and attracts more qualified male applicants, who get more job offers, perform consistently better on the job for two years and are not more likely to leave compared to men who believe that workers' effectiveness is high.

[26] Appendix Table OB.1 compares the experimental sample with a random subsample from the UK Labour Force Survey (LFS) and shows that men and women in my experiment are more likely to have experience in public sector or healthcare jobs, which reinforces that the sample can be equally selected on preferences and beliefs.

6.1 Application and selection process

The first three columns of Table 2 summarize men's journey in the application and selection process using empirical specification (1). The dependent variables in columns (1), (2) and (3) are indicator variables for applying, receiving a job offer (conditional on applying) and accepting the job offer (conditional on receiving the offer), as defined in Section 5.1.

The first result is that the workplace gender composition does not affect men's applications. Receiving an email featuring a male worker reduces men's applications by 1.7 percentage points with respect to an email featuring a female worker (Column (1), Table 2), but this coefficient is imprecisely estimated and I cannot reject the null hypothesis of no difference between the two photographs. The male photograph has no impact also when compared to the pure control group (Panel C, Table B.1).

Nevertheless, men react to the expectations manipulation (bottom row of Column (1), Table 2). The coefficient on $Effectiveness_i^L$ shows an increase in applications of 7 percentage points in the treatment with low as compared to high workers' effectiveness (p=0.04). This effect represents 14% of the mean application rate in the high effectiveness treatment (52%) and 13% of the pure control mean (53%).

This effect could be caused by men applying more when knowing that workers are not effective or applying less when workers are highly effective. I distinguish these cases by comparing application rates in each information group with the pure control email (Panel B, Table B.1). Men's application rate increases by 4.9 percentage points in the low effectiveness treatment (p=0.2) and decreases by 2.1 percentage points in the high effectiveness treatment (p=0.6) with respect to the pure control. This suggests that, on average, people's baseline beliefs on workers' effectiveness are closer to the 89% statistics and that the low effectiveness treatment comes as a surprise.[27]

Table A.4 shows that the difference in men's applications between the two information treatments is nearly the same when combined with a male or a female photograph. The additivity assumption thus seems appropriate, but a caveat of this exercise is a limited statistical power.

I now turn to hiring outcomes. Column (2) of Table 2 shows that male applicants in the low workers' effectiveness treatment get more job offers than applicants in the high effectiveness treatment. The male offer rate in the former group is 16%, which is 6 percentage points higher than in the group of past outstanding performance. When looking at the unconditional offer rate (Column (6) of Table A.3), the gap is still statistically significant, but is reduced to 4.2 percentage points, indicating that there is selection on quality at the application stage.

The difference in offer rates can come from an improvement in the quality of applicants in the low effectiveness group or a decrease in the quality in the high effectiveness group. I find evidence of both.

First, applicants in the low effectiveness treatment have better observables which are correlated with receiving a job offer. Figure A.3 shows that there is a bigger mass of applicants in the

[27] Table B.1 Panel A shows that men's application rates are significantly higher in the low effectiveness treatment compared to a pooled control group which includes either high effectiveness information or pure control emails (by 7 percentage points, p=0.035).

low effectiveness treatment with above-median "desirable skills", which is an index including having a first grade in university, being from a top tier university, having volunteered frequently in the past, cognitive skills above the median and maximum scores in English pre-university qualifications.

Moreover, quality at the top of the applicants' distribution also differs across information treatments, a fact consistent with changes in marginal applications. To identify top applicants, I follow the employer's screening rules, which were blind to treatment status and created a unique ranking of candidates independently of gender. I take the distribution of average test scores during the selection process for both men and women and focus on the top quintile, whose size corresponds to the number of slots available for job offers that year.This group identifies applicants who would have been hired by the employer based on their performance in the selection process. I measure quality in four different ways: predicted on-the-job performance based on observables, the index of "desirable skills" defined above, the average score in the selection process and an index which combines the three measures.[28] Figure A.4 shows that the highest quality - for any of the four measures - is achieved in the low effectiveness information treatment. Higher quality is correlated with more job offers at the top: 46% in the low effectiveness treatment compared to 37% in the high effectiveness treatment or pure control (a difference which is not statistically significant given N=128).

Information of high past effectiveness attracts instead lower quality. The rate of job offers in the treatment providing information of high past effectiveness is significantly lower than the pure control group by 9 percentage points (Column (2), Panel A of Table B.1). This means that there is a tail of low-talent candidates who apply for the job when they receive information that almost anyone can be effective. This is consistent with evidence of Section 3.1, where low-talent candidates have expectations of success as positive as high-talent candidates in this treatment.

To sum up, relative to information of high effectiveness, information of low workers' effectiveness attracts more and better male applicants, who are consequently also more likely to be hired. The joint increase in the size of the pool of applicants, the exit of low-skilled applicants and the entry of high-skilled ones are consistent with Result 2 of the model and the case of negative sorting of men in social work ($\theta_g < v_g$).

There are some differences in applicants' quality also between photograph treatments. The average applicant who received a male photograph has better observable skills (Figure A.3) and is more likely to be offered the job (by 5.5 pp, p=0.10). However, the gap in offer rates - and skills - between the two photograph treatments is driven by the female photograph group getting significantly fewer job offers than the pure control, while the male photograph has not impact on selection (Panel C, Table B.1).[29]

Offering the job to more talented men exposes the firm to the risk that they would reject the offer because they have better outside options. Column (3) of Table 2 shows indeed that the

[28] Section 7.1 describes the construction of predicted on-the-job performance. All the indexes are standardized to be mean zero and unitary standard deviation in the pooled sample of men (Kling et al., 2007). To define cognitive skills, I coded the most recent job (reported in the application) into standardized SOC4 categories and followed Acemoglu and Autor (2011).

[29] Differences in quality without changes in pool size are possible in the model when θ_g or $\hat{a}$ change.

acceptance rate for men conditional on receiving an offer is weakly lower in the low effectiveness treatment than in the alternative information group (by 4 pp over a mean of 70%, $p>0.70$), but the estimates are not statistically significant. Moreover, this gap is reversed when looking at acceptance conditional on application, so before selection by the employer (Table A.3). This reversal suggests that applicants' motivation to accept the job is not differential across information groups.

I now look at the outcomes of male newcomers in the job.

6.2 On-the-job outcomes

The last three Columns of Table 2 show the effect of the treatments on on-the-job outcomes for new male hires, using my main empirical specification (1). The sample in this set of regressions is the subset of job offerees who accepted the offer (43 out of 67 offerees). After a first month of training, new hires are sent to their allocated team in different UK regions.[30]

Columns (4) and (5) of Table 2 show that information of low workers' effectiveness allows the employer to select male workers who perform consistently better on the job. The average difference in average test scores between the two information treatments is 4.8 percentage points over two years, which corresponds to 8.6% of the mean average score in the high effectiveness treatment. The average performance of men who perceive workers to not be effective is 9.5% percent higher than the performance of men with higher expectations during the first semester (Column (4) of Table 2). This gap slightly decreases to 8% in the last year of assessments, but the average level of performance also drops and becomes more heterogeneous.

Male workers attracted to apply by low past job effectiveness perform better along the entire distribution of on-the-job performance. Figure 8 shows the distribution of men's residualized average test scores by experimental treatment, after controlling for strata controls. The distribution of average test scores is completely shifted to the right for men who received information of low vis-à-vis high past effectiveness (right-hand side figure, KS test $p < 0.05$). Not only workers are better in the low effectiveness treatment compared to the high effectiveness group, but also with respect to the pure control group (Table B.2, Panel B), even if this difference is estimated with noise ($p=0.15$).

Good performance on-the-job benefits both the employer and service beneficiaries, but might come at the expenses of workers' job-satisfaction. Newspaper headlines often report high levels of burnout, stress and, consequently, turnover among social workers. In the UK, as many as 50% of social workers plan to stay less than two years in the occupation and work overload is one of the most common reasons for leaving (Ravalier, 2019). One might worry that newly hired high-achievers would soon become overloaded with cases and/or be more likely to leave. Moreover, studies have shown that men are disproportionately likely to exit fields which are atypical for their gender (Torre, 2018).

Against these concerns, Column (6) of Table 2 shows that men in the low effectiveness

[30] Regional and team assignment is orthogonal to expected performance and based on candidate's regional preferences, diversity and slot availability. Out of the ones who accepted the offer, 70% were allocated to the first ranked region. There are a total of 52 communities in my sample and the average team size is 4 people.

treatment are not more likely to drop out of the program. Furthermore, male hires attracted by low past effectiveness are more likely to state that they want to keep working in social work after the program and have more positive attitudes towards their work (Table 4). They are more likely to say that they are aware of the social impact of their work and that they are confident of their skills in interacting with families. At the same time, Columns (1) to (3) of Table 4 show that men in the low past effectiveness group are not more likely to be concerned with the work environment and do not have a higher perceived or actual workload.

To show the robustness of the performance results and overcome empirical issues related to the small sample, I exploit the availability of repeated tests for each worker by estimating the model:

$$score_{ia} = \alpha + \beta_1 Pic_i^M + \beta_2 Effectiveness_i^L + X_i'\lambda + \epsilon_{ia} \tag{2}$$

where $score_{ia}$ is worker's i grade in assessment a normalized by the mean and standard deviation of male workers' grades.The vector X_i includes the strata controls of specification (1) and dummies for the region of work, for whether the worker has been allocated to his preferred region and for whether the worker had already applied for the job in the past. I also estimate the same model controlling for the index of desirable skills introduced in section 6.1 and an index of difficulty of the local community where a worker is allocated to.[31] Standard errors are clustered at the worker level.

This specification compares the score achieved in a given test by two men working in the same (preferred) region - and assigned to a local community of similar difficulty - who were originally assigned to receive different photographs or pieces of information at recruitment. Coefficients β_1 and β_2 measure the causal effect of the experimental manipulations on the selection of more talented workers under the identifying assumption that the treatments do not have a direct impact on workers' effort and/or motivation on the job.

The bottom row of Table 3 confirms that men receiving information of low past effectiveness perform significantly better: their scores are 0.25 standard deviations higher than men who received information of high past effectiveness ($p < 0.10$) and the effect slightly increases when controlling for the difficulty level of local communities (Column (3)). Assuming no spillovers on inframarginal workers, the performance of those hires induced to apply by the low effectiveness treatment is two thirds of a standard deviation higher than hires in the alternative treatment. Once I control for workers' observable skillset, the coefficients on the low effectiveness treatment dummy decrease by two percentage points and become marginally insignificant (F-stat = 1.56). This means that the observable measures of talent used by the employer to make job offers partially account for the difference in performance between the two treatment groups, which is consistent with my interpretation of the treatment affecting job performance through selection.

[31] I use data from the Department for Education on the Children and Family Social Work Workforce (2017) in England and data from the 2016/17 report of Her Majesty's Chief Inspector of Education, Children's Services and Skills (by Ofsted). For each local authority, I compute an index of "difficulty" by averaging the score in social workers' caseload, turnover, absenteeism and Ofsted's scores on helping children, childcare, leadership effectiveness. Specification 2 controls for candidates who applied for the job in the past, because they represent more than 10% in the final cohort of workers (vs 5% overall) and are more likely to be allocated to a team in their preferred region.

The residual difference can be due to a wider range of skills which are not observable to the researcher (but which the employer is able to pick-up in hiring) or to a direct impact of the treatment on effort on the job. I provide evidence against the latter channel in section 9.1.2. Results on performance on the job are even stronger when adding the pure control group to the estimation (Table B.2), as men attracted to apply by past low effectiveness perform also better than the pure control male hires.

The top row of Table 3 also shows that men attracted through a male photograph perform worse than men attracted through a female photograph. Men in the former treatment show 0.28 standard deviations lower scores than men in the latter treatment ($p>0.10$), an effect which is enhanced when controlling for ability measures ($p=0.10$). However, this gap is entirely driven by men in the female photograph group performing better than both the pure control and the male photograph group (Table B.2). Overall, thus, the male photograph has no impact on men's entry, the quality of the pool of applicants or job performance. The female photograph, instead, attracts on average worse applicants who are substantially screened-out by the employer, so that the residual ones who get hired end up being better.

All in all, the evidence of this section is consistent with a story in which disclosing information of low workers' effectiveness improves the male workforce by improving the expected impact of talent in the job. A higher expected effectiveness of skills, in turn, attract a better pool of male applicants, from which the employer selects talented hires who also perform better once in local communities.

Discussion

The null effect of the photograph manipulation on men's applications speaks to many policies advocating that media campaigns portraying people of the same gender will help attract men to teaching or nursing (Flynn, 2006; The Economist, 2018). For instance, in 2002 the Oregon Center for Nursing tried to appeal to young men by launching the notorious "Are you man enough to be a nurse?" recruitment campaign, which portray a line-up of masculine men engaged in a variety of extreme sports.[32] A more realistic representation of male nurses is also one of the pillars of the biggest recruitment drive in the history of the UK National Health System (NHS from here on, see McGonagle, 2019). In light of my results, it's surprising that the campaign has been considered the cause of the 9% increase in men's enrollment in nursing school between 2018 and 2019 (News, 2019).[33]

One way to reconcile my results with current policies is sample selection. Even if men in my sample update perceived gender shares similarly to the general population (discussed in Section 3.1), they may care less about workplace gender composition than the average man (low α_i). If so, the estimated effect of the male photograph provides a lower bound. To address this point, I can use evidence from an additional experiment that I ran with the same partner organization.[34] The goal of this trial was to explore whether adverts portraying men could be

[32] See the webpage: http://www.oregoncenterfornursing.org/.

[33] See more at these links: www.england.nhs.uk/ and www.nhsemployers.org/your-workforce/recruit.

[34] The experiment was run at the same time of the main trial presented in the paper, and was indeed part of the same pre-registration.

effective in encouraging a wider population of male students - not yet interested or aware - to apply for the job (see Appendix C). I find no evidence that male photographs encourage this pool into the job. This extends the external validity of the null result in my main experiment and suggests that the gender composition among campaigns' actors has limited impact on men's recruitment.

My survey evidence uncovers a plausible explanation for the null effect of photographs: male photographs are not powerful enough to change the association between social work and women. In my auxiliary survey data I asked respondents the extent to which they think that the job is desirable for a man or for a woman. Figure 3 shows that both men and women change their expectations of gender shares in the two photograph treatments, but men do not update their beliefs on the desirability of the job for a man or a woman when seeing a male photograph.[35] This provides direct evidence of the stickiness of the association between female gender and social work.

My second result demonstrates that the informational content of campaigns is what really matters for men's selection into female-dominated jobs. Information provision has an economically substantial effect on men's applications. Previous field recruitment experiments which exogenously varied on-the-job incentives find effects of higher wages on application rates to be between 18% and 26% (Abebe et al., 2021; Dal Bó et al., 2013). In my experiment, the effect of providing information of low workers' effectiveness is between two thirds and half of the one obtained in the aforementioned papers, but it's also nearly costless for the employer.[36]

How can information have such a big impact? According to Bayesian updating, the magnitude of the coefficient on the $Effectiveness_i^L$ dummy is proportional to men's uncertainty in priors. From the model, random assignment of the two information treatments guarantees that the difference in posteriors $\Delta\theta_g$ is decreasing in priors' precision and independent of priors' levels according to: $\Delta\theta_g = E[\theta_g|s = s_H] - E[\theta_g|s = s_L] = \frac{\overline{\sigma}_g^2}{\overline{\sigma}_g^2+\sigma_s^2} \cdot (s_H - s_L)$. Thus the large gap in application rates between the two information arms uncovers an important barrier to men's entry in female-dominated jobs: informational constraints. While it is surprising that limited information plays a role in my context, where one could assume there are nearly unlimited opportunities for learning and experimentation, the willingness to experiment is itself a function of the expected usefulness of information. The sheer fact that some occupations are almost exclusively done by women can impair men's inclination to gather information on careers that are uncommon for their gender.[37] The reluctance to get informed can be especially detrimental to people with high job-specific talent and a valuable outside option.[38]

[35] Notice that the lack of updating by men should not be taken as an indication of a low powered intervention, because female respondents react to the photograph manipulation and substantially change their gender-social work association.

[36] In a similar light-touch intervention, Del Carpio and Guadalupe (2021) get around 28% higher applications by women in the treatment group. However, application costs are very different between our settings and the level of application rates in the control group is 53% in my context versus only 7% in their setting.

[37] Consistently with this view, the experiment described in Appendix C shows that men are more likely to gather information about social work when they think there are more men in the sector.

[38] Some authors find evidence of employer's discrimination against men in female-dominated jobs (Booth and Leigh, 2010; Rich, 2014; Hangartner et al., 2021). Receiving information that talent matters and can make a difference may weaken concerns that the employer will be unfair, a channel which can enhance the impact of my information treatment.

To the best of my knowledge, this is also the first paper showing that disclosure of a low past success rates can be motivating for new applicants. This evidence is in contrast with the basic idea of role model interventions, which provide statistics of high success to minority members to increase their perceived chances to succeed in jobs which are uncommon for their group (Porter and Serra, 2020; Breda et al., 2020; Del Carpio and Guadalupe, 2021). These two perspectives are not in conflict if we recognize that the interpretation of "high success" depends on people's priors. In sectors where people expect low rewards to talent, such as social work, high success may reinforce these perceptions and encourage only people of low ability to apply (as in my high effectiveness treatment). Previous research, however, could have not uncovered this effect given its focus on entry in male-dominated occupations, where priors are substantially different.

A final point is that my information treatment is not related to incentive schemes or earnings, but about the extent to which talent can be impactful and effective in the job. Highlighting the valuable contribution that men can make in female-dominated jobs is a common policy content (Flynn, 2006), but this factor has not been considered in the rich literature on the determinants of occupational choices by gender, ethnicity or socio-economic background (Wiswall and Zafar, 2015, 2018; Boneva and Rauh, 2017).

7 Theory-driven heterogeneous treatment effects

7.1 Heterogeneity by job-specific talent (a_i)

In this section I test the interpretation of the low effectiveness information treatment as an increase in the slope of expected utility with respect to talent. I use a discrete-choice framework to quantify the change in application probability with respect to job-specific talent generated by the information manipulation and to benchmark this effect against a change in wage.

Consider the individual decision of whether to apply to the job or not: $Pr(apply = 1) = Pr(U^j(\alpha_i, s_g, \theta_g, a_i, \hat{a}) + \xi_j > U^o(v_g, a_i, \overline{w}_g) + \xi_o)$, where ξ_j and ξ_o are errors with type I generalized extreme value distributions and the cost of application is assumed to be zero. I use Maximum Likelihood to estimate the following logit model:

$$log\frac{Pr(apply)}{(1-Pr(apply))} = \beta_1\overline{w}_g + \beta_2 OwnGender_i + \beta_3 a_i + \beta_4 Effectiveness_i^L + \beta_5 Effectiveness_i^L * a_i$$

where $OwnGender_i$ is a dummy for a same-gender photograph, $Effectiveness_i^L$ is a dummy for receiving information of low workers' effectiveness, $\overline{w}_g$ is the de-meaned difference between the log-wage in the job and in the outside option and a_i is a de-meaned proxy of job-specific talent. This proxy is the predicted on-the-job performance score based on observables and constructed through a linear truncated regression in the pure control group, under the assumption that these observables affect job performance independently of being hired and treatment.[39]

The coefficient of interest in the model is β_5, which identifies the difference in slopes with

[39] The truncated regression uses the following variables: ranking and completion rate of the candidate's university, subject studied, obtaining a first grade, whether the grade is confirmed, age, age squared and full-time employment. Data on universities are from the 2015-2016 University and Subject League Tables. Construction of the outside option is described in Appendix OB. and the logit estimates are reported in Appendix OC..

respect to job-specific talent between the two information treatments $\Delta\theta_g = \theta_H - \theta_L$. The estimated β_5 for men is 0.02, which implies that the odds of applying increase by 20% for a one unit increase in job-specific ability. As a benchmark, just above mean talent this coefficient is comparable to an 11% increase in the wage in the job (an increase in the hourly wage from 20 to 22.2 GBP). Consistently, Figure 10 shows that the slope of the predicted probability of applying with respect to job-specific talent is higher in the low workers' effectiveness vis-á-vis the alternative information group. This supports the idea that the treatment changes the slope of expected utility with respect to job-specific ability.

7.2 Heterogeneity by priors' uncertainty (σ_g^2)

Do expectations of effectiveness matter relatively more for men with less exposure to social work? Theoretically, the impact of new information should be increasing in initial uncertainty on the potential for making a difference in the job ($\overline{\sigma}_g^2$). I build two empirical proxies for men's uncertainty: exposure to labor market gender segregation *during teenage years* and an index of familiarity with the position in social work offered by the partner organization.[40]

To construct the former proxy, I use microdata from the 2011 U.K. Census and construct the Duncan index of occupational segregation (Duncan and Duncan, 1955) in the local area where a candidate went to secondary school and/or lived during teenage years.[41] The first two columns of Table 5 estimate heterogeneous treatment effects on applications by splitting the sample between subjects exposed to higher-than-median (Column 1) and lower-than-median (Column 2) occupational gender segregation. The bottom row of Panel A shows that men exposed to higher-than-median occupational gender segregation react significantly more to the information manipulation. Their applications increase by 16.7 pp when receiving information of low vs high past effectiveness, which represents 35% of the mean in latter group.[42]

Appendix OD.2 further motivates the use of occupational gender segregation as a proxy of men's uncertainty. Through an Implicit Association Test (IAT), I show that exposure to segregation increases the implicit association between social work and women. Moreover, men coming from areas of with a high Duncan index tend to have higher uncertainty in beliefs about men and women's performance in female-jobs (Table OD.1).

Columns (3) and (4) of Table 5 estimate heterogeneous treatment effects on applications by splitting the sample between subjects with higher-than-median (Column 1) and lower-than-

[40] Using exposure to labor market gender segregation as a proxy for men's uncertainty is motivated by a rich literature on the association between segregation and gender norms (Goldin, 2014; Cortes and Pan, 2018; Baranov et al., 2020). Moreover, exposure to gender segregation has been shown to affect the persistence of biased beliefs on group ability (Guryan and Charles, 2013), an insight used by Arrow (1998) to explain long-term statistical discrimination.

[41] The Duncan index identifies the percentage of women (or men) that would have to change occupations for the occupational distribution of the two genders to be equal and is computed using the following formula: $\frac{1}{2}\sum_{i=1}^{N}|\frac{m_i}{M} - \frac{f_i}{F}|$, where m_i and f_i are the male and female population, respectively, in occupation i and M and F are the total working population of the local labor market. It takes values between 0 (complete integration) and 1 (complete segregation). I compute the index through the subjects' secondary school postcode (40% of sample), or home postcode (62% of cases).

[42] The estimates are robust to the inclusion of controls for observable differences between men coming from areas with high versus low gender segregation and for the ratio of male to female unemployment in the local area, to account for gender differences in working opportunities.

median (Column 2) familiarity with the job. The index averages baseline data on the participants' exposure to social work and/or the job offered by the partner organization: whether the person studied a common university subject for social workers, worked in healthcare or the public sector, attended any information event about the job or applied in the past. The estimates indicate that the increase in applications in the low effectiveness treatment is entirely driven by men with limited information about the job, a fact consistent with the theoretical predictions. Neither exposure to occupational gender segregation or the job seem to mediate reaction to photographs instead.[43]

7.3 Heterogeneity by outside option returns to talent (v_M)

The joint increase in male applications and hires is consistent with a model where potential applicants face steeper returns to talent in the outside option than in the job (case $\theta_M < v_M$). In this section, I test this theoretical prediction empirically.

Using data from the 2017-2018 Labour Force Survey, I proxy returns to talent in the outside option with the average wage dispersion faced in the UK labor market. For a candidate who studied subject s, wage dispersion is computed as the weighted average of the 75/25 interquartile range of the distribution of hourly wages across industries, where weights are given by the proportion of graduates of subject s working in each industry. Panel B of Table 5 shows heterogeneous treatment effects by splitting the sample of candidates by above/below median wage dispersion. As hypothesized, Columns (5) and (6) of Table 5 show that the increase in application rates in the low effectiveness treatment is stronger among men facing above-median wage dispersion. This suggests that the marginal applicant induced to apply by the information treatment faces a steeper outside option, a fact which also contributes to explain the subsequent increase in quality among top-ranked applicants.

8 Net effect of policies to attract men in pink-collar jobs

Is it desirable to attract men into pink-collar occupations by changing applicants' expected effectiveness? As any policy can rarely target only one particular gender, we need to know whether attracting more men with information of low workers' effectiveness also affects women. Knowledge of the sign and magnitude of these spillovers on women's selection will allow me to assess the net impact of the policy for the employer. I conclude by discussing implications for the economy overall.

8.1 Spillovers on women's selection

Women are insensitive to the experimentally-provided information on workers' effectiveness on average (Figure 7b). The second row of Table 6 shows that the coefficient on the $Effectiveness_i^L$ dummy on women's applications is -0.015 and statistically insignificant, while Column (1) of

[43] I define "common university subjects" those belonging to knowledge areas listed for social work by O*Net.

Table A.5 confirms that the two genders react differently to the information treatment.[44] This evidence could lead to the conclusion that improving candidates' expected effectiveness is a silver bullet for the employer: it achieves higher diversity, better quality among the gender minority and has no effect on the majority.

Nevertheless, we can go a step further. While the suggested informational policy does not affect women's selection, the very fact that men start entering social work might have an impact on women's behavior. Coding and hospital administration are historical examples of female-dominated occupations where the gender composition tipped in favor of men (Arndt and Bigelow, 2005; Ensmenger, 2012). If male shares in female-dominated jobs increased, what would be the impact on the number and quality of female applications and hires? I provide an answer to this question using my photograph manipulation. Showing a male photograph allows me to simulate - in people's minds - a counterfactual world in which the share of men in the job is higher. Women's behavior when receiving a male photograph thus tells us how women would react if more men became social workers.

I find that increasing social workers' male shares discourages women from applying for the job, but without negative consequences on the quality of the pool of female applicants. Column (1) of Table 6 shows that there are 7.5% fewer women's applications in the male vis-à-vis female photograph treatment, but offer rates in the two arms are similar (Column (2), Table 6).[45]

The male photograph affects women's decisions in two other stages of the process: offer acceptance and job retention. Women who applied despite seeing the male photograph are more likely to accept the job offer vis-á-vis women who received the female photograph (Column (3) of Table 6), especially those with lower observable ability (Figure A.8b). However, this greater motivation fades when women find it hard to keep-up with the requirements of the job. Colum (6) of Table 6 shows that women in the male photograph are more likely to leave the program before completing it. This effect comes entirely from women with low performance in the first six months in the job (Figure 9), of whom 42% belong to the group of women with below-median observable skills. These results indicate women with high job-specific talent are relatively unaffected by changing perceived male shares in the job. A higher male share instead affects the decisions of lower ability women, who seem to become more sensitive to either positive or negative signals about their fit with the role either at recruitment or in the job (Coffman et al., 2019). While it is beyond the scope of the paper to fully explain women's behavior, Section 8.3 briefly discusses potential reasons.

8.2 Net impact for the employer and the economy

The evidence on women's behavior allows me to answer the question of whether attracting more men by raising expectations of success - for instance by disclosing information of low workers'

[44] The null effect of information provision on women's applications seems in contrast with the fact that they update their beliefs in the two treatments (section 3.1). Even in the surveys, however, women's updating is partial: their expected performance react to information when combined with a female photograph, but not with a male photograph.

[45] Women may also think that the employer is looking for men, a belief which may still anticipate a future change in gender composition. Independently of the exact interpretation, the main result is that employers' active policies to attract more men in female-dominated jobs might discourage women from applying.

effectiveness - is desirable for the employer.

First, if we only look at recruitment, the answer is positive. Information on low workers' effectiveness attracts more and better men, and does not affect women's applications on average. In turn, a higher perceived male share does not affect the quality of new female hires and increases their acceptance.

The desirability of a policy which attracts more men by changing expected effectiveness is instead ambiguous when considering the outcomes of the incoming pool of workers. The reason is that a higher male share combined with information of low past effectiveness discourage low-performing women from staying in the job, creating a trade-off between performance and retention for the employer. Is the increase in performance among stayers high enough to justify the loss of female workers in the program?

I conducted a survey with the recruitment personnel of my partner organization to provide an answer to this question. Respondents choose between two hypothetical cohorts of workers: one with no drop-out, but lower average performance and one with greater drop-out, but a resulting higher performance.[46] Results show that 58% of respondents prefer a cohort with higher performance rather than longer retention, and this percentage increases to 71% if higher performance is associated with greater gender diversity. All in all, this evidence indicates that the net effect of disclosing low workers' effectiveness is positive, even after taking into account the cost of female turnover for the employer. Compared to the other experimental conditions, providing information of low workers' effectiveness indeed achieves the highest overall performance and diversity in the workforce (Table A.12).

What do my results imply for talent allocation in the aggregate economy? In a world with two sectors (e.g., female and male), the answer depends on the nature of men and women's sorting in each. If men's sorting in female-dominated jobs is negative - as my results suggest - their reallocation will improve average skills in both sectors as switchers have the lowest male-sector ability. Predictions are less straightforward if we consider the crowd-out of women from the female sector. There will be aggregate gains from talent reallocation if women are positively sorted in the female sector and negatively sorted in the outside option. However, the net effect of both women and men's reallocation is ambiguous if women are positively sorted in both sectors. Predictions for the economy are ultimately difficult to draw if we consider further effects which are not included in a standard selection model, such as tipping-point reallocation across occupations (Pan, 2015) or backlash by women (Rudman and Fairchild, 2004).

8.3 Do women care about the workplace gender composition?

Women's negative reaction to the male photograph could suggest that women value co-workers gender more than men do (Haile, 2012). Yet, I cannot reject the null hypothesis that men and women react in the same way to the photograph manipulation (Table 6). Moreover, women's results are entirely driven by the interaction between the photograph and information manipulation (see Table A.4). The male photograph has no impact on women's decisions when

[46] To construct the two scenarios, I used data about female workers' outcomes in treatment (Female Photo, 89%) and in treatment (Male Photo, 66%) respectively. Details of the survey are reported in Appendix OG..

combined with information of high past effectiveness, but becomes demotivating when combined with information of low past effectiveness. This interaction suggests that women's behavior does not stem from a generalized preference for working with their own gender.

I consider two hypotheses to explain women's behavior. First, women may dislike working with men in more challenging environments (Niederle and Vesterlund, 2007). Women indeed believe that the job will be more difficult when seeing information of low past effectiveness (Figure 5). Nevertheless, this explanation cannot explain the observed differences in behavior by low-performing vs high-performing women, unless preferences are correlated with ability.

A second hypothesis is that gender shares affect women's inference of their expected success on the job, as suggested by previous work (Croson and Gneezy, 2009; Dreber et al., 2014; Coffman et al., 2019). Low past performance in combination with a male photo may signal that the employer is trying to hire men in order to make up for a bad female performance. This is bad news for all women, but especially for those of lower ability who believed their gender could favor their hiring. I find support for this interpretation: turnover is concentrated among women working in teams with a high male share and who perform worse than their peers (Figure A.6).[47]

9 Robustness and alternative mechanisms

9.1 Identification checks

9.1.1 Does the treatment affect employer's screening?

To attribute the change in offer rates to the causal effect of the treatment on applicants' composition one needs to exclude the possibility that the treatment affects the employer's screening criteria (Ashraf et al., 2020). I test this assumption using the following specification:

$$o_i = \sum_j \alpha_j^{T^1} T_i^1 X_i^j + \sum_j \alpha_j^{T^2} T_i^2 X_i^j + S_i' \lambda + \epsilon_i$$

where o_i is equal to one if i received a job offer (conditional on applying), T_i^1 and T_i^2 are indicator variables for one of the two treatments for each condition (e.g., male and female photograph respectively) and S_i are strata controls. X_i^j are indicator variables equal to one if candidate i has a certain qualification, such as receiving a first grade, studying a common university subject for social workers, having received the maximum grade in Maths or English pre-university qualifications.

Columns (1) and (3) of Table A.6 report the coefficients $\alpha_j^{T^1}$ and $\alpha_j^{T^2}$ for the information and photograph conditions, respectively. Columns (2) and (4) report the p-value of tests of equality of coefficients $\alpha_j^{T^1} = \alpha_j^{T^2}$. The tables shows that, first, the employer finds some qualifications more desirable than others. For instance, candidates who received a first grade in university are 11 percentage points more likely to receive an offer, while receiving a high Math score does not matter. Secondly, I cannot reject the null hypothesis of equality of the employer's selection criteria across treatments, as coefficients $\alpha_j^{T^1}$ and $\alpha_j^{T^2}$ are statistically indistinguishable in most

[47] This mechanisms can be modeled by adding beliefs based on stereotypes to the theory (Bordalo et al., 2019).

the cases.

9.1.2 Does the treatment affect candidates' effort?

The higher quality of male job offerees and workers is consistent with better selection generated by the treatment reporting low workers' effectiveness. An alternative explanation of this effect is a self-fulfilling prophecy: believing in higher chances to be successful might encourage men to put more effort into the hiring process or in the job, with a subsequent higher offer rate and performance at work. There are three main pieces of evidence against this explanation.

First, any motivating effect of the treatment should be stronger right after receiving the invitation-to-apply email. In contrast, Table A.7 shows that men in the two information treatments do not differ in the effort put into application completion, as measured by the percentage of fields filled-in and the number of characters used to answer the application questions. Second, male workers across treatments have on average the same actual and perceived workload in the job, which also excludes differential effort in the workplace (Columns (1) and (2) of Table 4). Third, we should expect higher effort to be correlated with higher likelihood of job acceptance, perhaps through a sunk cost fallacy. Evidence reported in Table 2 contradicts this hypothesis.

A related concern is that performance effects are an artefact coming from a "surprise" once people compare expected and actual potential for being successful on the job. According to this hypothesis, performance effects should be waning over time. Figure A.5 shows that there is no decreasing trend in the coefficients on the treatment indicator variable in separate regressions for each on-the-job assessment, with the caveat that estimates become noisier over time.

9.2 Alternative mechanisms

9.2.1 Social comparison and employer's selectivity

The information manipulation could make participants update their beliefs about people competing for the same role or about the selectivity of the employer. According to models of tournament entry (Niederle and Vesterlund, 2007; Lazear et al., 2018) and directed search (Wright et al., 2019), we should expect low ability people to shun away from high competition or a selective employer, predicting an increase in average quality in the treatment featuring the 89% statistic. This is in contrast with the evidence shown in the paper. Survey respondents also do not update beliefs on the effort required to complete the application or the quality standards required by the organization (Figure 5b).

9.2.2 On-the-job dating market

Experimental photographs may be a signal of dating opportunities. Under this hypothesis, we expect heterosexual (non-heterosexual) men to apply more when seen a female (male) photograph and weaker effects of photographs on married people. Table A.10 tests for differential treatment effects of the male photograph on applications by sexual orientation (in odd Columns) and by marital status (in even Columns). I do not find support for the on-the-job dating channel.

There are no significant differences in the effect of photographs by marital status. Moreover, non-heterosexual men and women both react positively to the male photograph, implying that this reaction is not related to dating.

9.2.3 Gender differences in preferences

Different variance in past success among workers could affect the perceived riskiness of performance in the job, leading to different reactions by gender as men tend to be less risk averse than women (Holt and Laury, 2002; Eckel and Grossman, 2008). There are two main comments against this interpretation. First, more variance in past success does not necessarily imply higher uncertainty if people know their own ability. A high aggregate past performance could even be associated with higher uncertainty if people think that success is determined by other (unclear) factors rather than ability. Secondly, even allowing low workers' effectiveness to be associated with higher risk, we should expect women to apply less in this condition on average.

Another rich literature shows that men tend to be more overconfident than women (Niederle and Vesterlund, 2007; Dreber et al., 2014). The main concern is that information of low past workers' effectiveness could attract overconfident male applicants who incorrectly believe they will improve service delivery. First, results on actual job performance contrast with this view. Second, Table A.11 shows that men in my sample tend to be less overconfident than women, especially in job-specific skills, which is aligned with evidence that the gender gap in overconfidence is context-dependent and even reverses in female domains (Coffman et al., 2019). Third, Appendix D shows that the increase in men's applications is driven by men with low confidence in their estimates of people's performance in female-dominated jobs. As long as confidence about others is correlated with confidence in one's own ability, my effects are driven by the least confident men (Moore and Healy, 2008).

Finally, low workers' effectiveness might signal that the workplace environment is competitive. Well-known results are that men are more likely to select into competitive environments than women (Niederle and Vesterlund, 2007), especially in tasks perceived as "masculine" (Dreber et al., 2014; Flory et al., 2015). I check whether reaction to the treatment differs by participants' competitive background. I identify two types of candidates, who have the same on-the-job average performance: those used to competition, who studied a male-dominated subject in a top-tier university, and those not used to competition, who studied a female-dominated subject in lower-tier universities. Qualitatively, Figure A.9 shows that both men and women react similarly to the expectations treatment independently of this proxy of competitiveness, but statistical precision in this exercise is limited.

10 Concluding remarks

Blue-collar employment is shrinking across the developed world (Autor et al., 2013). These trends challenge the traditional role of men in society and within households by creating male idleness and financial insecurity, especially on the left tail of the ability distribution (Dorn et al., 2019; Coile and Duggan, 2019). At the same time, female-dominated sectors such as healthcare and education are growing and face relatively little risk of automation in the future (Nedelkoska and Quintini, 2018). And yet, male labor supply is still relatively untapped as a resource for addressing the shortage of teachers and nurses in many industrialized economies. Understanding the interaction between traditional gender norms and gender-specific information in rapidly changing labor markets is crucial to allowing men in declining industries achieve new opportunities (Binder and Bound, 2019).

In this paper, I provided evidence that the limited entry of men into female-dominated jobs can be explained by limited information on expected success rather than job-gender composition. I show that providing information on the chances of being effective increases men's applications in social work, especially when their exposure to the sector is limited. Men who expect higher chances to make a difference in the job are more likely to be hired, perform better and are happier on the job vis-á-vis men with lower expected chances. At the same time, information which is motivating for men can be discouraging for women, indicating that the way in which men are attracted to the job might affect the majority.

Historically, job advertisement has been a common strategy to change the demographic composition of male-dominated occupations. Rosie the Riveter is a long-lived testimonial of the crucial role of advertising in recruiting women to supply-short male jobs during WWII (Honey, 1985). This legacy inspired recent attempts to attract men to female-dominated sectors portraying masculine men working as nurses or teachers. My results are a cautionary tale against strategies designed to promote a new male identity in these roles without addressing informational constraints.

Informational asymmetries between men and women are central in this paper, which assumes that men and women only differ in terms of their informational endowments. This assumption contrasts with many studies on gender differences in preferences and moves the focus of research from natural to nurturing differences, which emerge as a result of being the minority in a certain occupation. In particular, my paper highlights that more research is needed in understanding how the basic need to feel competent and have a positive impact in the workplace can affect gender or ethnic minorities' career choices.

Many questions are left for future research. How can we encourage men's (early) exposure to female-dominated jobs in order to reduce their uncertainty? How do supply-side and demand-side factors interact in determining whether men apply and get hired in female-dominated jobs? Hopefully answering these questions may prevent communities such as the ones in the Rust Belt or the North of England from being left behind by a rapidly evolving economy.

References

Abebe, G., Caria, A. S., and Ortiz-Ospina, E. (2021). The selection of talent: Experimental and structural evidence from ethiopia. *American Economic Review*, 111(6).

Abraham, L. and Stein, A. (2022). Words matter: Experimental evidence from job applications. *Job Marker Paper*.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

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

Akerlof, G. A. and Kranton, R. E. (2005). Identity and the economics of organizations. *Journal of Economic Perspectives*, 19(1):9–32.

Al-Ubaydli, O., List, J. A., and Suskind, D. L. (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–86.

Alan, S., Ertac, S., and Mumcu, I. (2018). Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics*, 100(5):876–890.

Arndt, M. and Bigelow, B. (2005). Professionalizing and masculinizing a female occupation: The reconceptualization of hospital administration in the early 1900s. *Administrative Science Quarterly*, 50(2):233–261.

Arrow, K. J. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, 12(2):91–100.

Ashraf, N., Bandiera, O., Davenport, E., and Lee, S. S. (2020). Losing prosociality in the quest for talent? sorting, selection, and productivity in the delivery of public services. *American Economic Review*, 110(5):1355–94.

Autor, D. and Wasserman, M. (2013). Wayward sons: The emerging gender gap in labor markets and education. *Third Way Report*, 20013.

Autor, D. H., Dorn, D., and Hanson, G. H. (2013). The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, 103(6):2121–68.

Azmat, G. and Boring, A. (2020). Gender diversity in firms. *Oxford Review of Economic Policy*, 36(4):760–782.

Azmat, G. and Petrongolo, B. (2014). Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics*, 30:32–40.

Baranov, V., De Haas, R., and Grosjean, P. A. (2020). Men. roots and consequences of masculinity norms. *UNSW Business School Research Paper*.

Becker, G. (1957). The economics of discrimination. *Chicago: University of Chicago Press*.

Benjamin, D. J., Choi, J. J., and Strickland, A. J. (2010). Social identity and preferences. *American Economic Review*, 100(4):1913–28.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Besley, T. and Ghatak, M. (2005). Competition and incentives with motivated agents. *American Economic Review*, 95(3):616–636.

Binder, A. J. and Bound, J. (2019). The declining labor market prospects of less-educated men. *Journal of Economic Perspectives*, 33(2):163–190.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Boneva, T. and Rauh, C. (2017). Socio-economic gaps in university enrollment: The role of perceived pecuniary and non-pecuniary returns. *CESifo Working Paper*.

Booth, A. and Leigh, A. (2010). Do employers discriminate by gender? a field experiment in female-dominated occupations. *Economics Letters*, 107(2):236–238.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2019). Beliefs about gender. *American Economic Review*, 109(3):739–73.

Breda, T., Grenet, J., Monnet, M., and Van Effenterre, C. (2020). Do female role models reduce the gender gap in science? evidence from french high schools. Technical report, Institute of Labor Economics (IZA).

Bursztyn, L. and Jensen, R. (2017). Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9:131–153.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Charles, K. K., Guryan, J., and Pan, J. (2018). The effects of sexism on american women: The role of norms vs. discrimination. *NBER Working Paper*, (w24904).

Coffman, K. B., Collis, M., and Kulkarni, L. (2019). *Stereotypes and Belief Updating*. Number 19-068.

Coffman, K. B., Collis, M., and Kulkarni, L. (2021). *When to Apply?* Harvard Business School.

Coile, C. C. and Duggan, M. G. (2019). When labor's lost: Health, family life, incarceration, and education in a time of declining economic opportunity for low-skilled men. *Journal of Economic Perspectives*, 33(2):191–210.

Cortes, P. and Pan, J. (2018). Occupation and gender. *The Oxford Handbook of Women and the Economy*, pages 425–452.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–74.

Dal Bó, E., Dal Bó, P., and Eyster, E. (2017). The demand for bad policy when voters underappreciate equilibrium effects. *The Review of Economic Studies*, 85(2):964–998.

Dal Bó, E., Finan, F., and Rossi, M. A. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *The Quarterly Journal of Economics*, 128(3):1169–1218.

Del Carpio, L. and Guadalupe, M. (2021). More women in tech? evidence from a field experiment addressing social identity. *Management Science*.

Delfgaauw, J. and Dur, R. (2010). Managerial talent, motivation, and self-selection into public management. *Journal of Public Economics*, 94(9):654–660.

Deserranno, E. (2019). Financial incentives as signals: Experimental evidence from the recruitment of village promoters in uganda. *American Economic Journal: Applied Economics*, 11(1):277–317.

Dorn, D., Hanson, G., et al. (2019). When work disappears: Manufacturing decline and the falling marriage market value of young men. *American Economic Review: Insights*, 1(2):161–78.

Dreber, A., von Essen, E., and Ranehill, E. (2014). Gender and competition in adolescence: Task matters. *Experimental Economics*, 17(1):154–172.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217.

Dweck, C. S. and Elliot, A. J. (2005). *Handbook of Competence and Motivation*. Guilford Press New York.

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of Experimental Economics Results*, 1:1061–1073.

Elliot, A. J., McGregor, H. A., and Thrash, T. M. (2002). The need for competence. *Handbook of self-determination research*.

Ensmenger, N. (2012). *The Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise*. History of Computing. MIT Press.

Finan, F., Olken, B. A., and Pande, R. (2017). The personnel economics of the developing state. *Handbook of economic field experiments*, 2:467–514.

Flory, J. A., Leibbrandt, A., and List, J. A. (2015). Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1):122–155.

Flory, J. A., Leibbrandt, A., Rott, C., and Stoddard, O. (2019). Increasing workplace diversity: Evidence from a recruiting experiment at a fortune 500 company. *Journal of Human Resources*, pages 0518–9489R1.

Flynn, S. (2006). Campaign to attract more men into teaching. *https://www.irishtimes.com/news/campaign-to-attract-more-men-into-teaching-1.1005448*.

Folke, O. and Rickne, J. (2020). Sexual harassment and gender inequality in the labor market. Technical report, CEPR Discussion Papers.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

Goldin, C. (2006). The quiet revolution that transformed women's employment, education, and family. *American Economic Review*, 96(2):1–21.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119.

Graves, J. and Kuehn, Z. (2021). Specializing in growing sectors: Wage returns and gender differences. *Labour Economics*, 70:101994.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Guryan, J. and Charles, K. K. (2013). Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal*, 123(572):F417–F432.

Haile, G. A. (2012). Unhappy working with men? workplace gender diversity and job-related well-being in britain. *Labour Economics*, 19(3):329–350.

Hangartner, D., Kopp, D., and Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *Nature*, 589(7843):572–576.

Hirschi, A. (2012). Callings and work engagement: moderated mediation model of work meaningfulness, occupational identity, and occupational self-efficacy. *Journal of counseling psychology*, 59(3):479.

Hoff, K. and Pandey, P. (2006). Discrimination, social identity, and durable inequalities. *American Economic Review*, 96(2):206–211.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.

Honey, M. (1985). Creating rosie the riveter: Class, gender, and propaganda during world war ii. *Minerva*, 3(2):65.

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The allocation of talent and u.s. economic growth. *Econometrica*, 87(5):1439–1474.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2):515–548.

Katz, L. F. (2014). America's jobs challenges and the continuing role of the us department of labor. *ILR Review*, 67(3_suppl):578–583.

Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119.

Koch, J., Irby, B., and Polnick, B. (2014). *Girls and women in STEM: A never ending story*. IAP.

Lazear, E. P., Shaw, K. L., and Stanton, C. T. (2018). Who gets hired? the importance of competition among applicants. *Journal of Labor Economics*, 36(S1):S133–S181.

Lin, V. W., Lin, J., and Zhang, X. (2015). Us social worker workforce report card: Forecasting nationwide shortages. *Social Work*, 61(1):7–15.

List, J. (2020). Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial. Technical report.

List, J. A. and Metcalfe, R. (2014). Field experiments in the developed world: An introduction. *Oxford Review of Economic Policy*, 30(4):585–596.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

Marinescu, I. and Wolthoff, R. (2020). Opening the black box of the matching function: The power of words. *Journal of Labor Economics*, 38(2):535–568.

Marx, D. M. and Ko, S. J. (2012). Superstars "like" me: The effect of role model similarity on performance under threat. *European Journal of Social Psychology*, 42(7):807–812.

McGonagle, E. (2019). Nhs celebrates male nurses in mission to overcome gender divide. *https://www.campaignlive.co.uk/article/nhs-celebrates-male-nurses-mission-overcome-gender-divide/1661044*.

Miller, C. C. (2017). Why men don't want the jobs done mostly by women. *The New York Times*.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2):502.

Nedelkoska, L. and Quintini, G. (2018). Automation, skills use and training. *OECD Social, Employment, and Migration Working Papers*, (202):0_1–119.

News, N. (2019). Young male nursing applicants surge after 'we are the nhs' recruitment campaign.

Ngai, L. R. and Petrongolo, B. (2017). Gender gaps and the rise of the service economy. *American Economic Journal: Macroeconomics*, 9(4):1–44.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

OECD (2019). Oecd employment outlook 2019 the future of work. *OECD Publishing, Paris*.

Oh, S. (2019). Does identity affect labor supply? *Job Marker Paper, Columbia University*.

Pan, J. (2015). Gender segregation in occupations: The role of tipping and social interactions. *Journal of Labor Economics*, 33(2):365–408.

Porter, C. and Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*, 12(3):226–54.

Ravalier, J. M. (2019). Psycho-social working conditions and stress in uk social workers. *The British Journal of Social Work*, 49(2):371–390.

Rich, J. (2014). What do field experiments of discrimination in markets tell us? a meta analysis of studies conducted since 2000. *Working Paper*, (8584).

Rudman, L. A. and Fairchild, K. (2004). Reactions to counterstereotypic behavior: The role of backlash in cultural stereotype maintenance. *Journal of Personality and Social Psychology*, 87(2):157.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., and Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology*, 61(2):195.

Simpson, R. (2009). *Men in caring occupations: Doing gender differently*. Springer.

Skorikov, V. B. and Vondracek, F. W. (2011). Occupational identity. In *Handbook of Identity Theory and Research*, pages 693–714. Springer New York.

Steele, C. M. and Aronson, J. (1995). Stereotype threat and the intellectual test performance of african americans. *Journal of Personality and Social Psychology*, 69(5):797.

The Economist (2018). A shortage of nurses calls for the recruiting of more men. but that is a tricky task. *The Economist*.

Torre, M. (2018). Stopgappers? the occupational trajectories of men in female-dominated occupations. *Work and Occupations*, 45(3):283–312.

Wiswall, M. and Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *Review of Economic Studies*, 82(2):791–824.

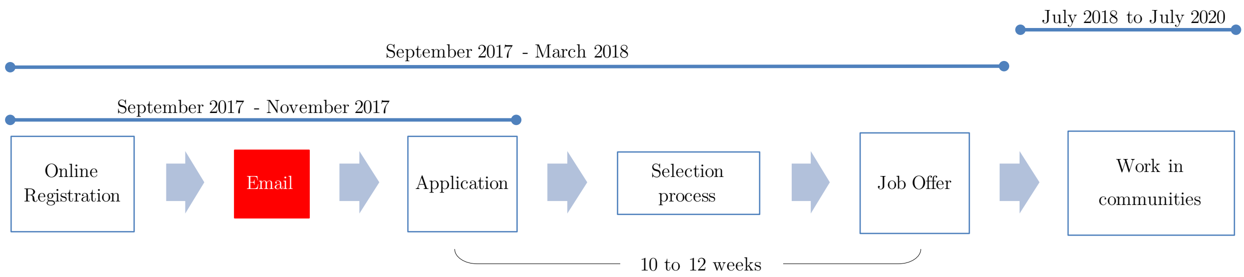
Wiswall, M. and Zafar, B. (2018). Preference for the workplace, investment in human capital, and gender. *The Quarterly Journal of Economics*, 133(1):457–507.

Wright, R., Kircher, P., Julien, B., and Guerrieri, V. (2019). Directed search and competitive search: A guided tour. *Journal of Economic Literature*.

Young, A. (2018). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

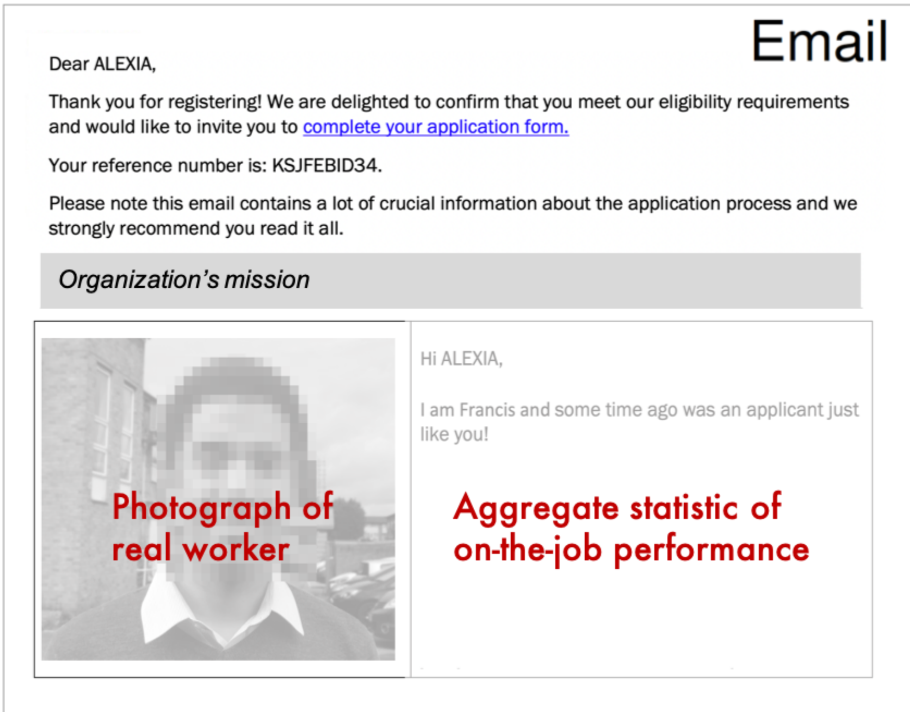
11 Figures

Figure 1. Recruitment timeline



Note. The Figure shows the recruitment timeline of the partner organization. Applications were open from September until November 2017. A given candidate was randomized in one of the different invitation-to-apply emails between his/her online registration and application. The organization completed the recruitment process in March 2018. For a successful applicant, it usually took between ten and twelve weeks between application and job offer. If a person was hired and accepted the job, actual work in local communities started in July 2018. Performance data are collected for the whole duration of the two-year program between July 2018 and July 2020.

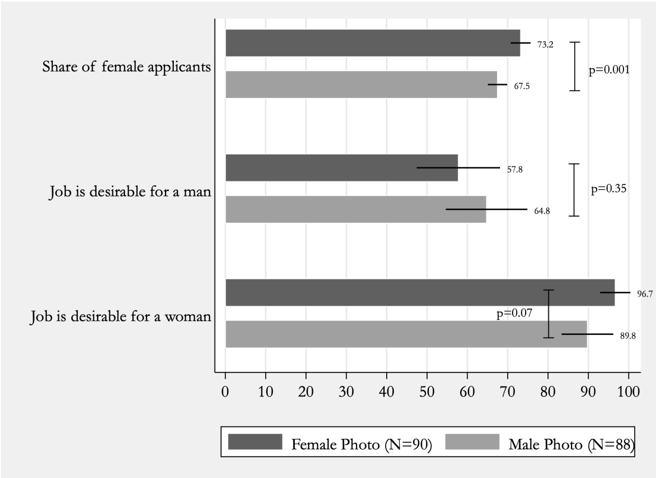
Figure 2. Intervention email template



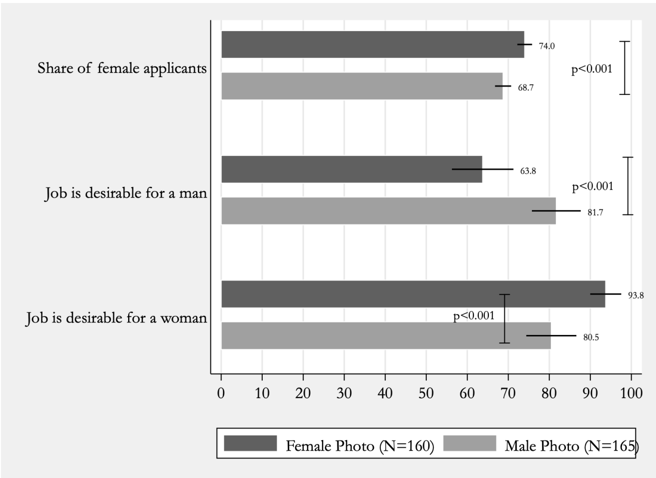
Note. The Figure shows a stylized example of one of the email templates used in the intervention. The intervention box containing both the photograph and information manipulations was located in the top quarter of the email and could be visualized in the email preview in any smartphone or tablet. To further increase attention to it, the box was positioned below the candidate number, necessary to access the application portal, and the text on the right of the picture addressed the candidates by name. The footer of the email was signed by the Director of Selection, contained the organization's logo and a disclaimer of confidentiality. The pure control email did not contain any intervention box.

Figure 3. Photograph treatment: manipulation checks by gender

(a) Male respondents



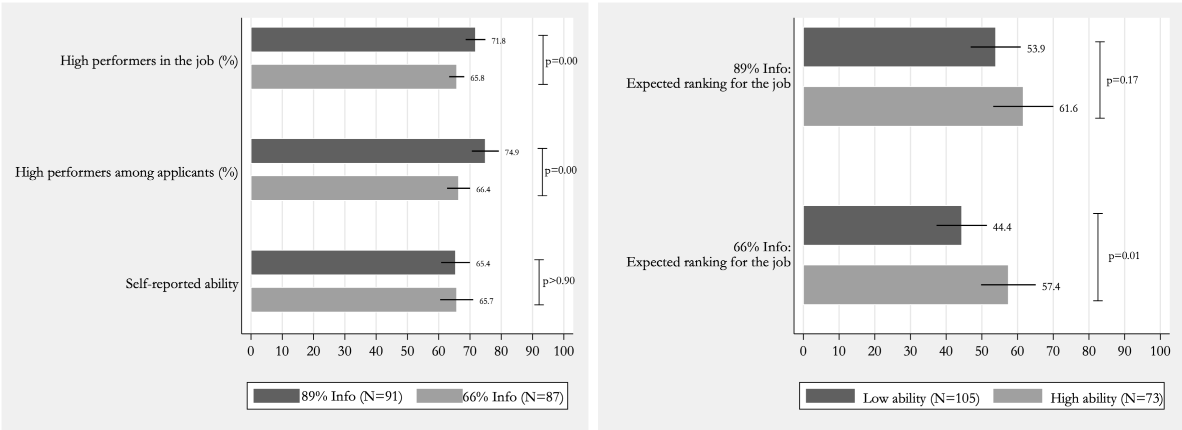
(b) Female respondents



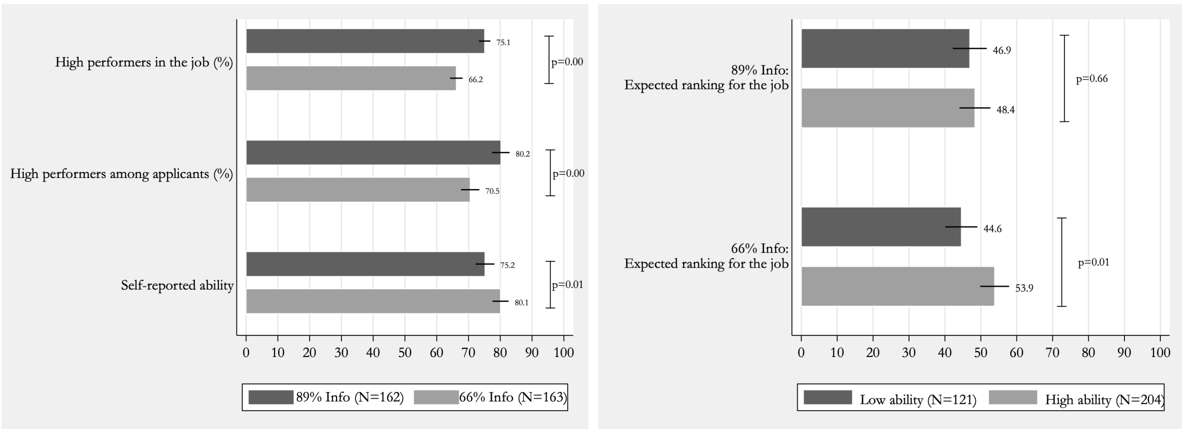
Note. Figures (a) and (b) show the average answer to the following questions (from top to bottom): “Consider the following situation. 100 people apply for the job after seeing the email ad that you have just seen. Of these 100 people that apply, how many do you think are women?”, “By looking at this ad, to what extent do you agree with the following statements? The job is desirable for a man / The job is desirable for a woman”. For the latter two questions, answers were on a 6-points scale from “Strongly Agree” to “Strongly Disagree” and I created dummy variables equal to one for the three highest options. Averages are computed separately for respondents assigned to the email with a female or male photograph, only for male respondents in (a) and for female respondents in panel (b). Error bars show 95% confidence intervals. Data are from the auxiliary online surveys. The number of respondents is 504, of whom 262 are from the Prolific Academic sample and 242 from the applicants’ sample, 178 are men and 325 are women (and one observation has missing gender). Differences between treatment groups in expected gender shares are robust to multiple-hypotheses testing corrections for both male and female respondents (see Table OA.5). For female respondents, all the differences shown in the figure survive multiple-hypotheses testing corrections. See also Table OA.3 for the comparison between the Prolific Academic sample and the 2018 Applicants’ sample and Table A.2 for manipulation checks in the pooled sample of all respondents.

Figure 4. Expectations shock: manipulation checks by gender

(a) Male respondents



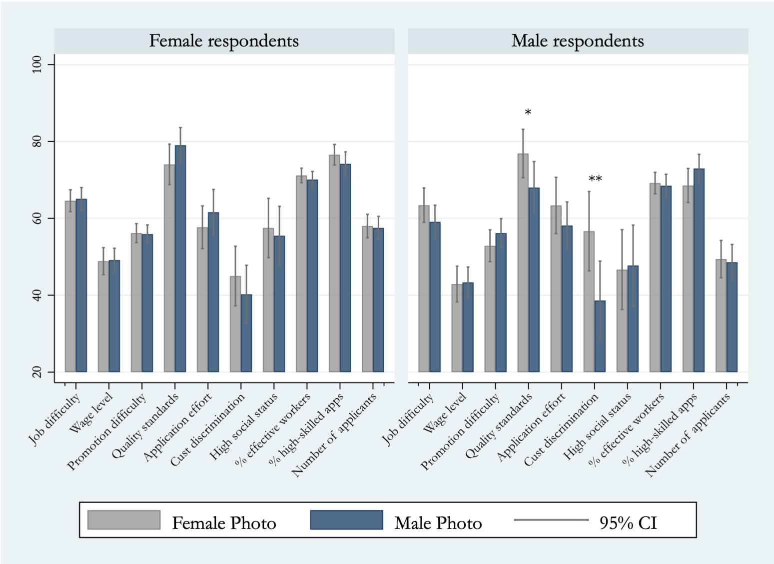
(b) Female respondents



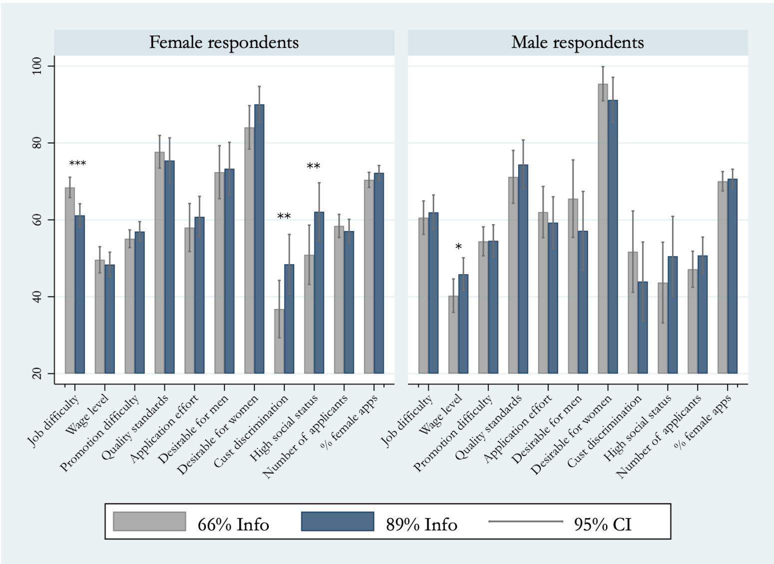
Note. In both Figures (a) and (b), "High-performers in the job (%)" is the weighted average of answers to the questions "Now that you have seen the email ad, indicate the proportion of [women/men] that you think are successful on-the-job", with weights given by actual gender shares. "High performers among applicants (%)" is the average of answers to the questions "Out of 100 [women/men] that apply for this job after seeing the email ad, how many do you think that have the potential to get commendable or excellent feedback on the job?". "Self-reported ability" is the expected performance of the respondent in interacting with families in need. Answers in the last question are given on a scale from 1 (min) to 10 (max), and are multiplied by 10 in the graph above for comparability. The left panels of Figures (a) and (b) show the average answers to the listed questions, separately for respondents assigned to the email with a statistic of 66% (lighter bars) or 89% (darker bars) past effectiveness. The right panel of Figures (a) and (b) shows the average "Expected ranking for the job", which is calculated from the average of the questions: "Consider 100 [men/women] who are applying for this job. Based on the ad you viewed, on a scale from 100 (best) to 1 (worst), how would you rank yourself for the job among them?". The top two bars are for respondents who saw the 89% statistic and the bottom bars for respondents who saw the 66% statistic. The sample is also split by high or low ability, where each level is defined as above or below mean answers to the variable "Self-reported ability". Error bars show 95% confidence intervals. Data are from the auxiliary online surveys. The number of respondents is 504, of whom 262 are from the Prolific Academic sample and 242 from the applicants' sample, 178 are men and 325 are women (and one observation has missing gender). See Table OA.4 for multiple-hypotheses testing corrections in the male and female samples, Table OA.2 for the comparison between the Prolific Academic sample and the 2018 Applicants' sample and Table A.2 for manipulation checks in the pooled sample of all respondents.

Figure 5. Further manipulation checks from auxiliary online surveys by gender

(a) Photograph treatment



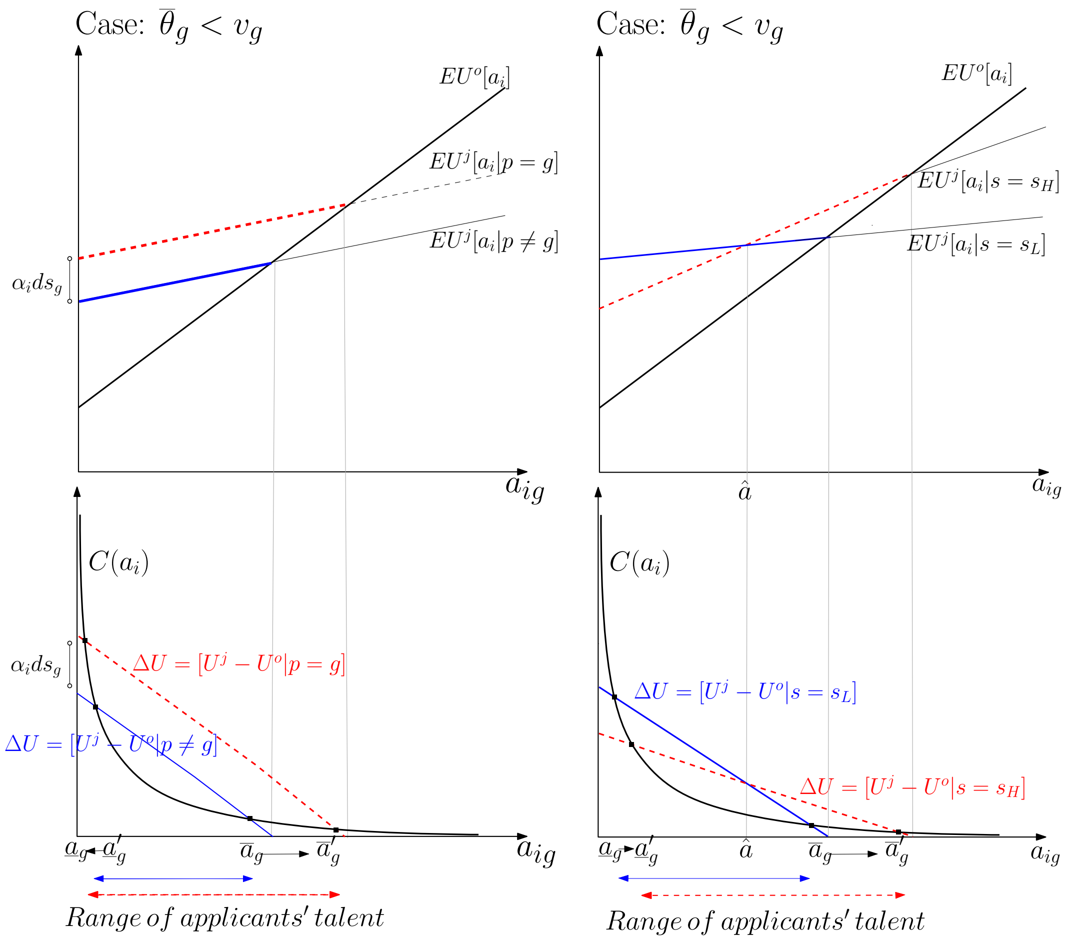
(b) Information treatment



Note. Figure (a) shows the average answers to additional survey questions about the job and its workers, separately for respondents assigned to the email with a male (darker blue bars) or female photograph (lighter gray bars), by gender. Figure (b) shows the average answers separately for respondents assigned to the email with a statistic of 66% (lighter gray bars) or 89% (darker blue bars) past effectiveness, by gender. Variables "Cust discrimination" and "High social status" show the share of people that agree with the statements "Customers discriminate workers" and "The job has a high social status". Answers were on a 6-points scale from "Strongly Agree" to "Strongly Disagree" and I created dummy variables equal to one for the three highest options. Variables "Job difficulty", "Wage level", "Quality standards", "Promotion difficulty" and "Application effort" asked participants to rate the given dimension of the job on a scale from 0 to 100. "Number of applicants" is the believed number of people that apply out of 100 who are considering whether or not to apply for the job. Figure (a) also shows mean answers to the questions used for the information manipulation checks, defined in the note to Figure 4. Figure (b) also shows mean answers to the questions used for the photographs manipulation checks, defined in the note to Figure 3. Data are from the auxiliary online surveys. The number of respondents is 504, of whom 262 are from the Prolific Academic sample and 242 from the applicants' sample, 178 are men and 325 are women (and one observation has missing gender). See Table OA.4 for multiple-hypotheses testing corrections in the male and female samples, Tables OA.2 and OA.3 for the comparison between the Prolific Academic sample and the 2018 Applicants' sample and Table A.2 for manipulation checks in the pooled sample of all respondents.

Figure 6. Theory predictions in the case of negative sorting in social work

(a) Shock in gender shares s_g (b) Shock in expectations θ_g



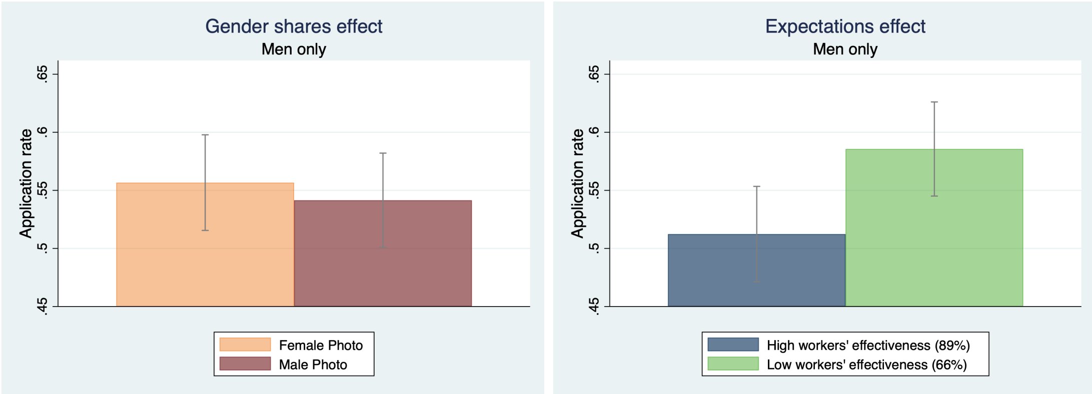
Note. The figure plots the application decision for potential applicants of gender g. Both panels consider the case ${U^j}'(a_i) < {U^o}'(a_i)$. Panel (a) shows the effect of a shock to perceived gender shares and Panel (b) to expectations of job effectiveness. The solid thick line shows expected utility in the outside option.

Panel (a): the top figure shows the expected job utility when receiving a photo of the same ($p = g$) gender (dashed red line) or a different gender ($p \neq g$) (solid blue line), and utility in the outside option. The vertical distance between these two lines comes from the assumption of the model $E[s_g|p = g] > E[s_g|p \neq g]$. The distance between each of the job expected utility lines and the outside option is then plotted in the graph at the bottom, which shows the comparison of $\Delta U(a_i)$ with $C(a_i)$. Two thresholds of ability for the marginal applicants $\underline{a}_g$ and $\overline{a}_g$ are determined from the intersection of $\Delta U(a_i)$ with $C(a_i)$. From Result 1, the size of the applicants' pool is greater when $p = g$ than $p \neq g$ and the marginal applicant $\overline{a}'_g$ is more skilled than $\overline{a}_g$, while $\underline{a}'_g$ is less skilled than $\underline{a}_g$.

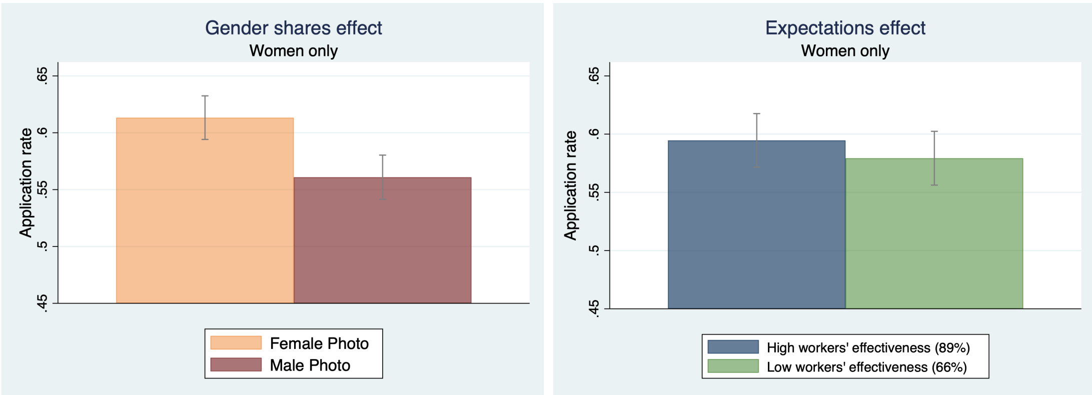
Panel (b): the top figure shows the expected job utility when receiving information of high marginal impact of talent ($s = s_H$) (dashed red line) or low marginal impact of talent ($s = s_L$) (solid blue line), and utility in the outside option. The different slope of the two expected job utility lines is explained by $E[\theta|s = s_H] > E[\theta|s = s_L]$. The distance between each of the job expected utility lines and the outside option is then plotted in the graph at the bottom, which shows the comparison of $\Delta U(a_i)$ with $C(a_i)$. Two thresholds of ability for the marginal applicants $\underline{a}_g$ and $\overline{a}_g$ are determined from the intersection of $\Delta U(a_i)$ with $C(a_i)$. From Result 2, the size of the applicants' pool may be larger or smaller when $s = s_H$ compared to $s = s_L$ as long as $\hat{a} \in [\underline{a}, \overline{a}]$. The marginal applicant $\overline{a}'_g$ is more skilled than $\overline{a}_g$ and $\underline{a}'_g$ is more skilled than $\underline{a}_g$.

Figure 7. Application rates by treatment and gender

(a) Men

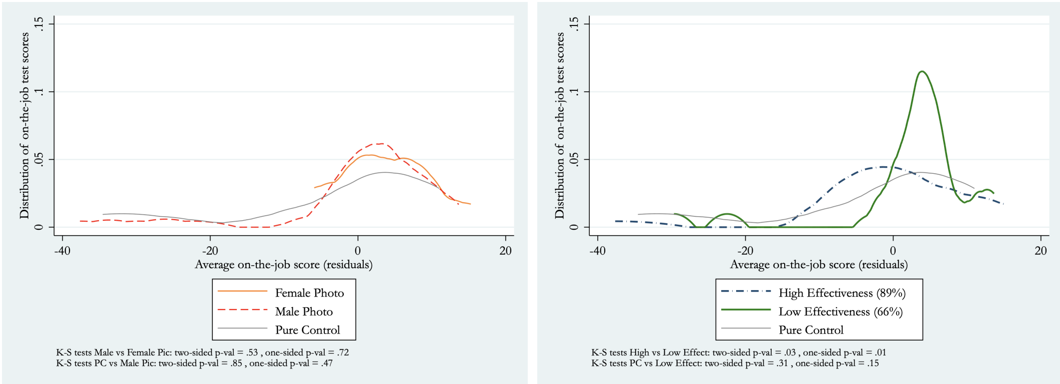


(b) Women



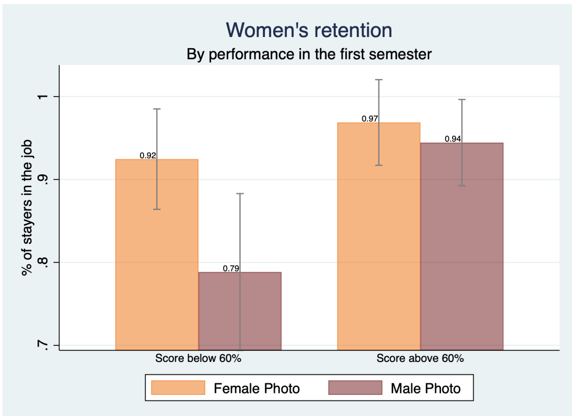
Note. Panel A shows application rates for men by photograph treatment (left-hand side) and information treatment (right-hand side). Panel B shows application rates for women by photograph treatment (left-hand side) and information treatment (right-hand side). Error bars show 95% confidence intervals.

Figure 8. Men's average on-the-job test scores by treatment



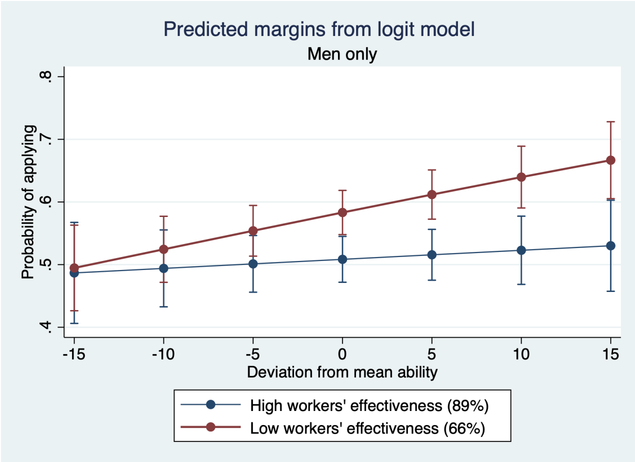
Note. The figure shows the distribution of men's average on-the-job test scores. The average on-the-job test score is the weighted average of the scores in the nine performance assessments required in the job, where weights are given by the credits assigned to each test by the organization and residualized after controlling for stratification variables (ethnicity and early registration). The Figure on the left-hand side shows the distributions by photograph treatment: the dashed line is for the male photograph and the red solid line for the female photograph. The Figure on the right-hand side shows the distributions by information treatment: the dashed line is for high workers' effectiveness (89%) and the green solid line is for the low workers' effectiveness (66%) treatment. In both graphs, the grey thin solid line shows the distribution for the pure control.

Figure 9. Women's retention by performance



Note. The figure shows female workers' retention by their performance in the first semester in the job and photograph treatment. I categorize on-the-job performance into two levels: above or below 60% average score in the first semester. Retention is defined as being still in the program by the end of the second year. Error bars show 95% confidence intervals.

Figure 10. Predicted application probability: logit estimation



Note. The figure shows predictive margins from a logit choice model for men only. The variable on the x-axis is the de-meaned predicted on-the-job performance. Predicted on-the-job performance is calculated using a truncated linear regression in the pure control group only with the following independent variables: ranking and average completion rate of the university attended by the candidate, subject studied, obtaining a first grade, whether the grade is expected or obtained, age, age squared and whether the person is in FTE. Error bars show 90% confidence intervals.

12 Tables

Table 1. Balance checks and summary statistics

	Men			Women			Joint		Pairwise
VARIABLES	*N*	*Mean*	*SD*	*N*	*Mean*	*SD*	*F-stat*	*p-val*	*min p-val*
Demographics									
Male	1013	1.00	0.00	4404	0.00	0.00	0.04	1.0	0.72
Non-white	1013	0.28	0.45	4404	0.27	0.45	0.08	1.0	0.60
Age	1013	28.7	9.2	4404	26.4	7.9	0.29	0.88	0.42
Married	995	0.19	0.4	4331	0.12	0.33	0.19	0.95	0.47
Caring duties	1013	0.16	0.36	4404	0.16	0.37	0.96	0.43	0.11
Non heterosexual	959	0.13	0.34	4131	0.07	0.26	0.36	0.84	0.33
Education and employment									
Top UK university	1013	0.33	0.47	4404	0.32	0.47	0.205	0.936	0.38
First grade	1013	0.2	0.4	4404	0.18	0.39	0.697	0.594	0.13
Graduate	1013	0.46	0.5	4404	0.35	0.48	0.473	0.756	0.19
Scientific subject	1013	0.09	0.28	4404	0.05	0.21	0.496	0.738	0.18
FTE	1013	0.49	0.5	4404	0.42	0.49	0.25	0.911	0.41
in: public sector	500	0.46	0.5	1840	0.56	0.5	1.06	0.373	**0.05**
in: healthcare	500	0.16	0.36	1840	0.17	0.37	0.87	0.483	0.11
in: corporate/business	500	0.32	0.47	1840	.22	.41	1.17	0.324	**0.05**
Registration									
Past application	1013	0.07	0.26	4404	0.06	0.24	0.08	0.99	0.61
Pre-submission call	1013	0.11	0.32	4404	0.08	0.28	0.48	0.75	0.27
Early registration	1013	0.04	0.2	4404	0.05	0.21	0.31	0.87	0.40
Registration before November	1013	0.53	0.5	4404	0.57	0.5	0.02	1.00	0.83
Any event	1013	0.00	0.05	4404	0.01	0.11	0.13	0.97	0.52
Socio-economic background									
Economic school support	1013	0.27	0.44	4404	0.27	0.45	0.62	0.65	0.15
Low socio-econ status	1013	0.60	0.49	4404	0.62	0.49	1.23	0.30	**0.08**
Young carer	999	.04	.2	4339	.04	.2	0.62	0.15	**0.02**
Care leaver	1006	.03	.17	4369	.02	.15	0.46	0.76	0.26

Note. The Table shows summary statistics for the overall experimental sample. “Caring Duties” is a dummy equal to one if the respondent is a primary or secondary carer of children. “Top UK” universities belong to the Russell Group (http://russellgroup.ac.uk). “First Grade” is a dummy for achieving the highest grade in university. “Graduate” indicates whether the candidate graduated in 2016 or before. “Scientific Subject” is equal to one if the person studied Engineering, IT/Computer Science, Maths or Natural Sciences. “Past application” is equal to one if the candidate applied already in the past for the same job. “Pre-submission call” indicates whether the candidate received a call from a recruitment officer to encourage application. “Early registration” is a dummy equal to one if the person had access to an early opening of the application. “Registration before November” is a dummy for whether the person registered online before the 1^{st} of November. “Any event” is a dummy equal to one if the candidate participated in any of the organization’s career events. “Economic school support” equals one if the candidate received free school meals or any other type of economic support (e.g., scholarship) during school. “Low socio-econ status” equals one if the occupation of the household’s highest earner in the candidate’s family was unemployment, routine manual or routine semi-manual or for parents with no degree. “Care leaver” is a dummy equal to one if the person spent some time looked after by a social worker before the age of 14. Columns 7 and 8 (under “Joint”) report the F-statistic and p-value from a joint test of the significance of the set of treatment dummies in explaining each row variable in a regression with pooled genders and with robust standard errors. The last Column reports the minimum p-value from the associated t-test between pairs of treatment groups with robust standard errors and with pooled genders. I also fail to reject the null hypothesis of zero effect of all the variables reported in the Table in a joint test of orthogonality on assignment to any treatment group (F(23, 4865)=0.67).

Table 2. Men's results

	(1)	(2)	(3)	(4)	(5)	(6)
	Application and selection process			On-the-job outcomes		
VARIABLES	Applied & no drop-out	Received Offer	Accepted Offer	Mean on-the-job score Semester 1	Semesters 2/3	Early Exit
Male Photo	-0.016	0.055	0.089	-3.489	-5.824	0.055
	(0.035)	(0.034)	(0.123)	(2.678)	(4.942)	(0.091)
	[0.62]	[0.10]	[0.48]	[0.20]	[0.22]	[0.32]
Low Past Effectiveness	0.071**	0.061*	-0.039	5.004*	3.512	-0.062
	(0.035)	(0.033)	(0.127)	(2.944)	(5.699)	(0.104)
	[0.04]	[0.06]	[0.75]	[0.09]	[0.51]	[0.29]
Summary Stats						
Observations	807	440	67	43	43	43
Mean Dep Var PC	0.53	0.21	0.83	56.16	45.42	0.32
Mean Dep Var	0.52	0.10	0.70	58.07	53.27	0.14
Male Ph = Low Past Effect						
p-value	0.08	0.90	0.47	0.04	0.18	0.30

Robust standard errors in parentheses
Randomization inference p-values in square brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. The table reports results of six different regressions. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). The dependent variables are indicators for applying and never dropping out of the process, receiving a job offer (conditional on applying) and accepting the job offer (conditional on receiving the offer) in Columns (1), (2) and (3). The dependent variable in column (4) is the average on-the-job test score achieved in the first five assessments during the first semester on the job and in column (5) is the average on-the-job test score achieved in the four additional assessments during the second and third semester on the job. The score goes from 0 to 100 in each test, the average is weighted by the credits assigned to each exam by the organization. The dependent variable in column (6) is a dummy equal to one if the person left the program before completing it. All the regressions control for stratification variables: access to early registration and non-white ethnicity. Square brackets contain the p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions. "Mean Dep Var" is the mean of the outcome variable in the omitted category and "Mean Dep Var PC" is the mean of the outcome variable in the pure control group. See Table A.3 for results on unconditional variables and Table B.1 for the comparison with the pure control group.

Table 3. New male hires: on-the-job performance

	DV: On-the-Job Std. Scores			
	(1)	(2)	(3)	(4)
Male Photo	-0.284	-0.300*	-0.293	-0.305*
	(0.194)	(0.166)	(0.194)	(0.168)
Low Past Effectiveness	0.248*	0.225	0.258*	0.234
	(0.141)	(0.144)	(0.143)	(0.147)
Controls				
Strata Controls	✓	✓	✓	✓
Exam FE	✓	✓	✓	✓
Quality Controls	×	✓	×	✓
Location Difficulty Controls	×	×	✓	✓
Summary Stats				
Observations	387	387	387	387
Mean Dep Var	0.19	0.19	0.19	0.19
Mean Dep Var in PC	-0.12	-0.12	-0.12	-0.12
Test Male Ph = Low Past Effect				
p-value	0.03	0.03	0.03	0.03

Clustered s.e. in parentheses (worker level)

*** p<0.01, ** p<0.05, * p<0.1

Note. OLS panel estimates for men only. The table reports results of four different regressions. The dependent variable is the on-the-job test score achieved in nine different assessments, standardized to be mean zero and unitary standard deviation in the full sample of male workers. The score goes from 0 to 100 in each test, and each test is weighted by the credits assigned to it by the organization. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). Columns (3) and (4) additionally control for an index of "difficulty" of the community where the worker is allocated to, using data collected by the UK Statistical Office. For each local authority, the index is obtained by averaging the score in these variables: social workers' caseload, turnover, absenteeism and scores on helping children, child care, leadership effectiveness. All the regressions control for the basic set of controls X_i made of the following dummies: access to early registration, non-white ethnicity, past application, workplace region fixed effect and a dummy for being allocated to the preferred region. Columns (2) and (4) control for the index of observable qualifications which are positively correlated with receiving a job offer. "Mean Dep Var" is the mean of the outcome variable in the omitted category and "Mean Dep Var PC" is the mean of the outcome variable in the pure control group. Standard errors are clustered at the worker level.

Table 4. New male hires: attitudes on the job

DV:	(1) Workload Actual	(2) Workload Perceived	(3) Concerned	(4) Satisfied	(5) Perceived impact at work	(6) Perceived impact outside	(7) Confidence in practice	(8) Intent to stay
Male Photo	0.178	-0.005	0.096	0.355***	0.126	0.111	0.153	0.113
	(0.227)	(0.133)	(0.178)	(0.121)	(0.101)	(0.124)	(0.112)	(0.080)
	{0.91}	{0.94}	{0.82}	{0.00}	{0.72}	{0.75}	{0.66}	{0.69}
Low Past Effectiveness	0.199	0.028	-0.043	0.195**	0.358***	0.121	0.287**	0.258**
	(0.213)	(0.123)	(0.169)	(0.090)	(0.102)	(0.119)	(0.113)	(0.108)
	{0.73}	{0.81}	{0.61}	{0.35}	{0.02}	{0.42}	{0.02}	{0.03}
Controls								
Strata Controls	✓	✓	✓	✓	✓	✓	✓	✓
Survey Wave FE	✓	✓	✓	✓	✓	✓	✓	✓
Summary Stats								
Observations	22	83	89	89	67	66	89	89
Mean Dep Var	0.33	0.31	0.50	0.36	0.55	0.09	0.43	0.57
Mean Dep Var in PC	0.50 0.44	0.59	0.63	0.67	0.25	0.67	0.89	
Test Male Ph = Low Past Effect								
p-value	0.93	0.86	0.57	0.31	0.06	0.96	0.36	0.30

Clustered s.e. in parentheses (worker level)
FWER corrected p-values in curly brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. The table reports results of eight different regressions. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). "Actual workload" is an indicator variable equal to one if a worker has a workload of 17 or more cases (Column 1). "Perceived workload" is an indicator variable equal to one if a worker perceives their workload to be too high or much too high (Column 2). "Concerned" is an indicator variable equal to one if a worker feels any personal, financial, work, well-being or health-related concern (Column 3). "Satisfied" is a dummy for whether the person feels satisfied with their work (Column 4). "Perceived impact" is an indicator equal to one if a worker feels that he is having positive social impact in his work (Column 5) or outside work (Column 6). "Confidence in practice" is equal to one if the worker feels confident in interacting with families in need (Column 7). "Intent to stay" is an indicator equal to one if the worker says he is moderately or very likely to stay in the same community or in the same job (Column 8). All the regressions control for strata variables: access to early registration and non-white ethnicity, plus an additional dummy for the survey wave. Square brackets report FWER corrected p-values using the procedure in List et al. (2019). Standard errors are clustered at the worker level.

Table 5. Heterogeneous treatment effects by background and outside option

	DV: Applied and never DO = 1					
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Background				*Panel B: Outside option*	
	Job Genderization		Familiarity with Job		Wage dispersion	
	High	Low	Low	High	Low	High
Male Photo	-0.026	-0.013	-0.050	0.002	-0.015	-0.023
	(0.050)	(0.050)	(0.057)	(0.044)	(0.046)	(0.055)
	[0.62]	[0.79]	[0.48]	[0.76]	[0.73]	[0.69]
Low Past Effectiveness	0.167***	-0.022	0.170***	0.011	0.035	0.121**
	(0.050)	(0.050)	(0.057)	(0.044)	(0.046)	(0.054)
	[0.00]	[0.66]	[0.02]	[0.52]	[0.42]	[0.03]
Summary Stats						
Observations	390	402	297	510	477	330
Mean Dep Var in PC	0.55	0.52	0.46	0.57	0.56	0.45
Mean Dep Var	0.48	0.57	0.44	0.57	0.56	0.45
Test Male Ph = Low Past Effect						
p-value	0.01	0.90	0.01	0.89	0.44	0.06

Robust standard errors in parentheses

Randomization inference p-values in square brackets

*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). All the regressions control for stratification variables: access to early registration and non-white ethnicity. Square brackets report p-values of the coefficients on the treatment dummies from randomization inference (randomization-t) with 1000 repetitions.

Panel A reports results of four different regressions. The variable "Job Genderization" is the Duncan index of occupational segregation by gender computed at the local area level (MSOA) where the subject went to secondary school or lives, using data from the 2011 U.K. Census. The level "high" or "low" is defined for values of the index respectively above or below the gender-specific median in the experimental sample. The variable "Familiarity with Job" is an index which averages baseline data on the participants' exposure to social work and/or the job offered by the partner organization: whether the person studied a common university subject for social workers, worked in healthcare or the public sector, attended any information event about the job or applied in the past. The level "high" or "low" is defined for values of the index respectively above or below the gender-specific median in the experimental sample.

Panel B reports results of two regressions. In Columns (5) and (6), wage dispersion for individual studying subject s is defined as the weighted average of the 75/25 interquartile range of the distribution of hourly wages across industries in the UK labor market, with weights are given by the proportion of graduates of subject s working in each industry.The level "high" or "low" is defined for values of the index respectively above or below the gender-specific median in the experimental sample. Data are from the 2017 and 2018 UK Labour Force Survey.

Table 6. Women's results

	(1)	(2)	(3)	(4)	(5)	(6)
	Application and selection process			On-the-job outcomes		
VARIABLES	Applied & no drop-out	Received Offer	Accepted Offer	Mean on-the-job score Semester 1	Semesters 2/3	Early Exit
Male Photo	-0.051***	0.013	0.131**	1.572	-3.063	0.076*
	(0.017)	(0.015)	(0.055)	(1.144)	(2.051)	(0.042)
	[0.00]	[0.43]	[0.03]	[0.15]	[0.17]	[0.07]
Low Past Effectiveness	-0.015	0.004	-0.000	-0.469	-2.432	0.095**
	(0.017)	(0.015)	(0.055)	(1.118)	(2.103)	(0.043)
	[0.36]	[0.83]	[0.96]	[0.88]	[0.19]	[0.03]
Summary Stats						
Observations	3,513	2,062	301	191	191	191
Mean Dep Var	0.60	0.14	0.55	57.3	60.1	0.02
Mean Dep Var in PC	0.59	0.15	0.68	59.5	57.8	0.07
Test Male Ph = Low Past Effect						
p-value	0.12	0.66	0.09	0.21	0.81	0.73

Robust standard errors in parentheses
Randomization inference p-values in square brackets
*** p<0.01, ** p<0.05, * p<0.1

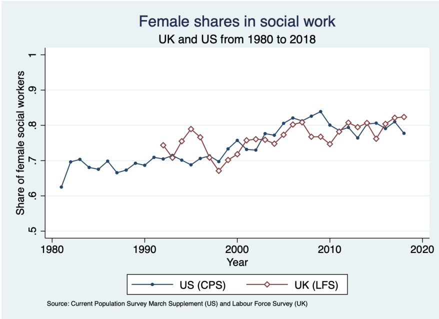
Note. OLS estimates for women only. The table reports results of six different regressions. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). The dependent variables are indicators for applying and never dropping out of the process, receiving a job offer (conditional on applying) and accepting the job offer (conditional on receiving the offer) in Columns (1), (2) and (3). The dependent variable in column (4) is the average on-the-job test score achieved in the first five assessments during the first semester on the job and in column (5) is the average on-the-job test score achieved in the four additional assessments during the second and third semester on the job. The score goes from 0 to 100 in each test, the average is weighted by the credits assigned to each exam by the organization. The dependent variable in column (6) is a dummy equal to one if the person left the program before completing it. All the regressions control for stratification variables: access to early registration and non-white ethnicity. Square brackets contain the p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions. See Table A.9 for results on unconditional variables and Table B.3 for the comparison with the pure control group.

Appendices

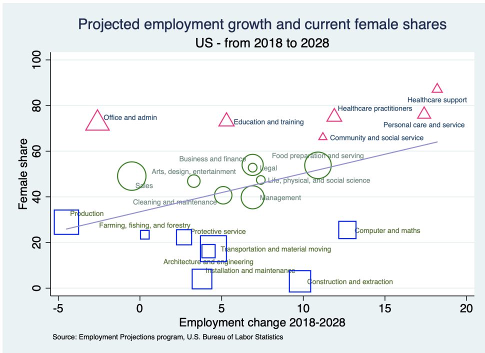
A Appendix figures and tables

Figure A.1. Three facts about social work

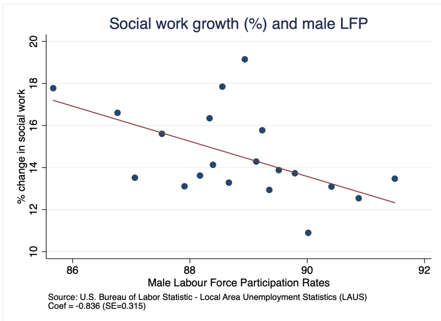
(a) Historical female shares in social work



(b) Predicted growth in occupations by current female shares (2018-2028)



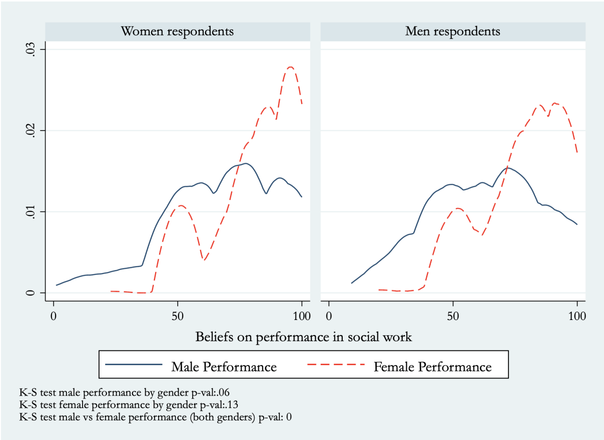
(c) Social work predicted growth and male labor force participation



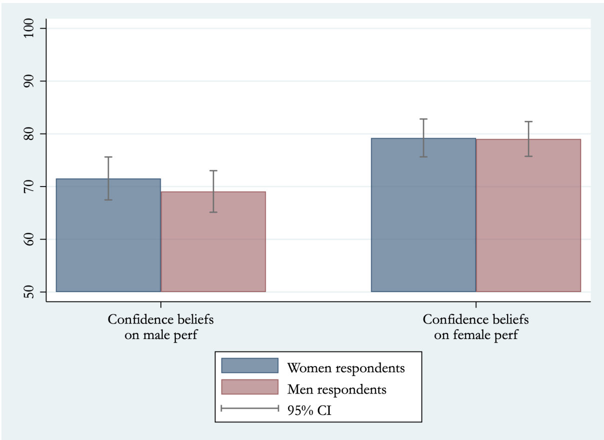
Note. Figure (a) shows the female share of social workers from 1980 to 2018 in the US (CPS data, March Supplement) and UK (LFS data). Figure (b) shows the correlation between predicted percentage change in employment between 2018 and 2028 for major occupational groups in the US (on the x-axis) and the 2018 female share (on the y-axis). Data are from the US Bureau of Labor Statistic and the Employment Projections program. Figure (c) shows a binned scatterplot between the 2018 male labor force participation (on the x-axis) and percentage change in employment in social work between 2018 and 2028 (on the y-axis) across US states. The graph controls for the overall growth rate across occupations and the state-level female labor force participation. Data are from the US Bureau of Labor Statistic Local Area Unemployment Statistics (LAUS) and the Employment Projections program.

Figure A.2. Beliefs about men's and women's performance in social work

(a) Beliefs by gender



(b) Confidence in beliefs by gender



Note. Figure (a) shows densities of answers to the following question: "On a scale from 0 (minimum) to 100 (maximum), what do you think is the performance of a [WOMAN/MAN] in social work?", where zero is for extremely bad performance, fifty for neither bad nor good performance and a hundred for extremely good performance. The graph on the left-hand side shows the distribution of women's beliefs and the one on the right of men's beliefs. Dashed lines are for beliefs about women's performance and solid blue lines for beliefs about men's performance. Figure (b) shows means answers to the following question: "On a scale from 0 (minimum) to 100 (maximum), how confident are you of your answer [about the performance of a man/woman in social work]?". The bars on the left shows the mean answers regarding beliefs on men's performance in social work and bars on the right beliefs on female performance. Dark blue bars show answers for respondents who are women and lighter red bars for men. Both men and women think that men have on average a lower performance in social work (Figure (a)), is consistent with the assumption $\theta_M < \theta_W$. The variance of the distribution of beliefs about men's performance is larger than the one of the distribution of beliefs about women, which supports the assumption $\overline{\sigma}^2_M > \overline{\sigma}^2_W$. Indeed both men and women are less confident of their predictions regarding a male social worker's performance than a female social worker's performance.

Table A.1. Comparison of prior beliefs: job applicants versus Prolific Academic sample

	2018 Applicants	Prolific Academic (AC)	(1)-(2)
	Mean	Mean	Difference
	(SD)	(SD)	(SE)
Priors on Female Share in Female jobs	77.57	74.76	2.81***
	(10.01)	(11.37)	(0.96)
Priors on Female Perf in Female jobs	81.18	81.18	0.00
	(17.66)	(15.54)	(1.48)
Priors on Male Perf in Female jobs	69.94	62.60	7.35***
	(22.86)	(23.34)	(2.06)
Priors on Female Share in Male jobs	22.82	21.83	0.99
	(12.25)	(12.51)	(1.10)
Priors on Male Perf in Male jobs	79.28	78.32	0.96
	(18.27)	(16.04)	(1.53)
Priors on Female Perf in Male jobs	69.73	67.36	2.37
	(24.30)	(21.89)	(2.06)
Confidence in Perf in Social Work	81.61	74.44	7.17***
	(22.48)	(20.17)	(1.92)
Think about promotion	0.79	0.81	-0.02
	(0.41)	(0.40)	(0.04)
N	242	262	

Note. The Table shows differences in prior beliefs about female-dominated and male-dominated jobs between the two samples of respondents to the auxiliary online surveys: 2018 applicants for the partner organization's job and respondents recruited online through the platform "Prolific Academic" and matched on observables with the applicants' sample. Questions in the survey asked about four different occupations: social work and primary school teaching (female-dominated), repair/maintenance work and computer specialist (male-dominated). For each of them, I asked three questions: "Out of 100 workers, what is the number of female workers in your opinion?", "On a scale from 0 (minimum) to 100 (maximum), what do you think is the performance of a man/woman in the following jobs?". Variables in the Table report mean answers to each of these questions, averaging across female-dominated and male-dominated jobs. Variable "Confidence in Perf in Social Work" is the answer to the question "On a scale from 0 (minimum) to 100 (maximum), how confident are you of your answer [about the performance of a man/woman in social work]?". The variable "Think about promotion" takes value of one if the respondents state that they think about future promotion possibilities when applying for a new job.

Table A.2. Manipulation checks: pooled samples with corrections for multiple hypotheses testing

	Female - Male Photo				High (89%) - Low (66%) Info			
	Diff	P-values			Diff	P-values		
			FDR	FWER			FDR	FWER
% female applicants	5.47	0.00	0.00	0.00	1.44	0.18	0.40	0.82
Job desirable for men	-0.14	0.00	0.00	0.01	-0.03	0.51	0.69	0.94
Job desirable for women	0.11	0.00	0.00	0.00	0.03	0.36	0.66	0.95
% of high-skilled applicants	0.017	0.99	1.00	0.99	9.29	0.00	0.00	0.00
% effective workers	0.93	0.44	1.00	0.99	7.86	0.00	0.00	0.00
Own expected performance	-0.20	0.30	1.00	0.95	-0.34	0.07	0.16	0.52
Job difficulty	1.18	0.50	1.00	0.99	-4.25	0.02	0.07	0.22
Wage level	-0.47	0.81	1.00	1.00	1.49	0.439	0.66	0.94
Quality standards	0.33	0.90	1.00	0.99	-0.18	0.96	1.00	0.96
Promotion difficulty	1.02	0.50	1.00	0.99	1.25	0.40	0.66	0.95
Application effort	-0.50	0.86	1.00	1.00	0.63	0.85	1.00	1.00
Customers discriminate workers	0.09	0.04	0.11	.332	0.05	0.30	0.61	0.94
Job has high social status	0.01	0.82	1.00	1.00	0.10	0.03	0.09	0.27
Number of applicants	0.55	0.78	1.00	1.00	0.33	0.86	1.00	0.98

Note. Variables are defined as follows. "% of high-skilled applicants" is the average of answers to the questions "Out of 100 [women/men] that apply for this job after seeing the email ad, how many do you think that have the potential to get commendable or excellent feedback on the job?". "% of high-performers in the job" is the weighted average of answers to the questions "Now that you have seen the email ad, indicate the proportion of [women/men] that you think are successful on-the-job". "Own expected performance" is the expected on-the-job performance on a scale from 1 (min) to 10 (max). "Own expected ranking for the job" is the average expected ranking among 100 [men/women] applying for this job. Variables "customers discriminate workers" and "job has high social status" show the share of people that agree with the given statement. Answers were on a 6-points scale from "Strongly Agree" to "Strongly Disagree" and I created dummy variables equal to one for the three highest options. "Job difficulty", "Wage level", "Quality standards", "Promotion difficulty" and "Application effort" asked participants to rate the given dimension of the job on a scale from 0 to 100. "Number of applicants" is the believed number of people that apply out of 100 who are considering whether or not to apply for the job. Columns denoted with "FDR" and "FWER" report multiplicity-adjusted p-values following Benjamini et al. (2006) and List et al. (2019).

Table A.3. Men's results: unconditional outcomes

VARIABLES	(1) Applied & no drop-out	(2) Applied	(3) Drop-Out Cond on App	(4) Drop-Out Uncond	(5) Received offer Cond on App	(6) Received offer Uncond	(7) Accepted offer Cond on Offer	(8) Accepted offer Cond on App	(9) Accepted offer Uncond
Male Photo	-0.016	-0.030	-0.021	-0.014	0.055	0.030	0.089	0.047*	0.026*
	(0.035)	(0.034)	(0.026)	(0.016)	(0.034)	(0.019)	(0.123)	(0.028)	(0.015)
	[0.62]	[0.34]	[0.43]	[0.44]	[0.11]	[0.13]	[0.45]	[0.10]	[0.12]
	{0.61}	{0.76}	{0.85}	{0.77}	{0.36}	{0.46}	{0.75}	{0.35}	{0.42}
Low Past Effectiveness	0.071**	0.049	-0.042	-0.022	0.061*	0.042**	-0.039	0.030	0.022
	(0.035)	(0.034)	(0.026)	(0.016)	(0.033)	(0.019)	(0.127)	(0.027)	(0.016)
	[0.03]	[0.15]	[0.09]	[0.20]	[0.07]	[0.03]	[0.75]	[0.26]	[0.16]
	{0.19}	{0.46}	{0.33}	{0.41}	{0.33}	{0.14}	{0.61}	{0.47}	{0.40}
Summary Stats									
Observations	807	807	485	807	440	807	67	440	807
Mean Dep Var	0.52	0.59	0.13	0.07	0.10	0.05	0.70	0.07	0.03
Mean Dep Var in PC	0.53	0.58	0.08	0.05	0.21	0.11	0.830	0.17	0.09
Test Male Ph = Low Past Effectiveness									
p-value	0.08	0.10	0.57	0.73	0.90	0.66	0.47	0.66	0.86

Robust standard errors in parentheses
Randomization inference p-values in square brackets
FWER corrected p-values in curly brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. The table reports results of nine different regressions. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). Columns (1), (5) and (7) report the same regressions of Columns (1) to (3) of Table 2. In Column (2), the dependent variable is an indicator for applying at the first stage of the section process, independently of whether the applicant keeps applying at later stages of the process or not. The outcome variable in Columns (3) and (4) is an indicator equal to one if a person drops out of the hiring process at any later stage after applying in stage one. Column (3) only considers applicants and Column (4) reports results on unconditional drop-out. Column (6) reports results on receiving a job offer, not conditional on applying. Columns (8) and (9) report results on offer acceptance, respectively conditional on applying and unconditionally. All the regressions control for stratification variables: access to early registration and non-white ethnicity. "Mean Dep Var" is the mean of the outcome variable in the omitted category and "Mean Dep Var PC" is the mean of the outcome variable in the pure control group. Square brackets report p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions and curly brackets report multiplicity-adjusted p-values following List et al. (2019).

Table A.4. Application rates by treatment and gender: interacted treatments

	(1)	(2)
	Applied and never drop out	
Sample:	Men	Women
Female Ph + High Past Effect	-0.015	0.011
	(0.049)	(0.023)
	[0.75]	[0.63]
	{0.80}	{0.86}
Female Ph + Low Past Effect	0.060	0.037
	(0.049)	(0.023)
	[0.20]	[0.11]
	{0.66}	{0.34}
Male Ph + High Past Effect	-0.027	0.001
	(0.049)	(0.023)
	[0.55]	[0.97]
	{0.93}	{0.96}
Male Ph + Low Past Effect	0.039	-0.055**
	(0.049)	(0.023)
	[0.41]	[0.02]
	{0.85}	{0.10}
Summary Stats		
Treated obs	807	3513
PC obs	206	891
Mean Dep Var in PC	0.53	0.59
Pairwise Coeff Comparisons p-values		
Male = Female | High Past Effect	0.81	0.65
	{0.93}	{0.85}
Male = Female | Low Past Effect	0.67	0.00
	{0.95}	{0.00}
High = Low | Female Ph	0.13	0.27
	{0.55}	{0.55}
High = Low | Male Ph	0.18	0.02
	{0.62}	{0.09}

Robust standard errors in parentheses
Randomization inference p-values in square brackets
FWER corrected p-values in curly brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates run separately for men (Column 1) and women (Column 2). The outcome variable is a dummy equal to one if a person applied for the job and never dropped-out of the hiring process at later stages. For each gender g, the omitted category is the pure control group. Each regressor is a treatment dummy for the combination of a male (Male Ph) or female (Female Ph) photograph and high (High Past Effect) or low (Low Past Effect) past effectiveness information. All the regressions control for the basic set of controls X_i for access to early registration and non-white ethnicity. "Mean Dep Var PC" is the mean of the outcome variable in the pure control group. Square brackets report p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions and curly brackets report multiplicity-adjusted p-values following List et al. (2019).

Table A.5. Do women and men react differently to treatments?

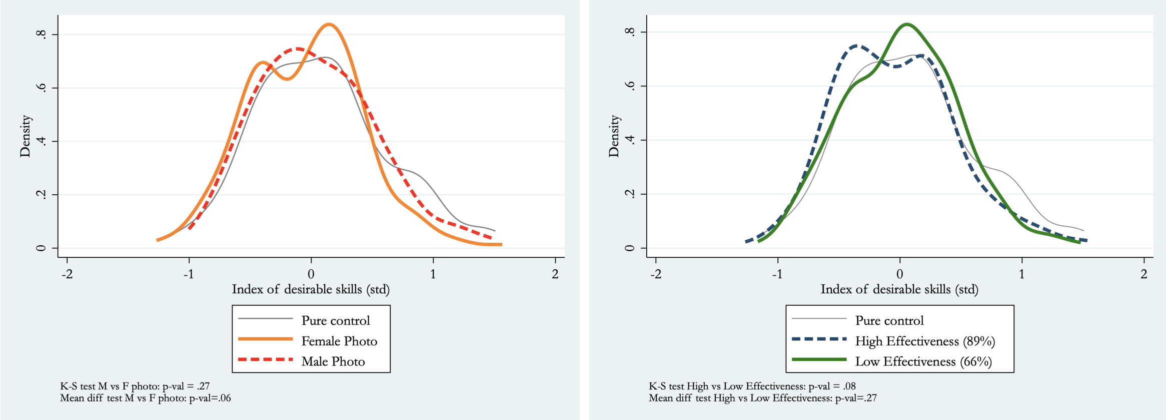
	(1)	(2)	(3)
VARIABLES	Applied and never drop out	Received Offer	Accepted Offer
Male Candidate	-0.102***	-0.043	0.060
	(0.033)	(0.029)	(0.127)
Male Photo	-0.051***	0.014	0.132**
	(0.017)	(0.015)	(0.055)
Male Photo x Male Candidate	0.035	0.041	-0.049
	(0.039)	(0.037)	(0.133)
Low Past Effect	-0.015	0.004	0.000
	(0.017)	(0.015)	(0.055)
Low Past Effect x Male Candidate	0.086**	0.056	-0.064
	(0.039)	(0.036)	(0.132)
Observations	4,320	2,502	368
Mean Dep Var	0.60	0.14	0.55

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

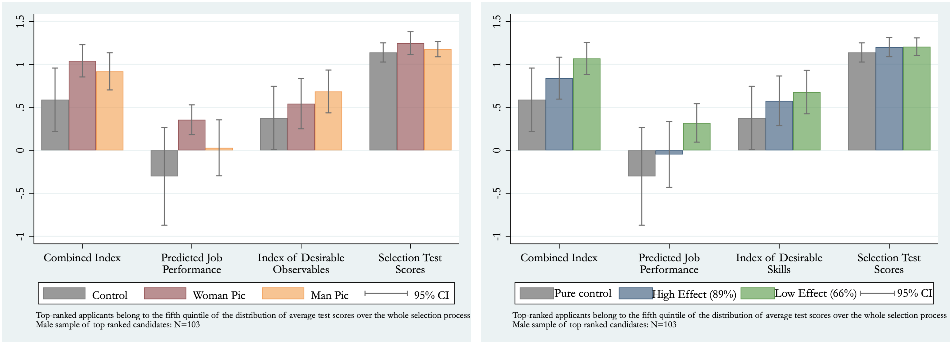
Note. OLS estimates for the pooled sample of men and women. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effect" is a dummy equal to one for receiving information on low past effectiveness (specification (1) of Section 5.1). The dependent variables are indicator dummies for application and never dropping out, receiving a job offer (conditional on applying) and accepting the job offer (conditional on receiving the offer) in Columns (1), (2) and (3). All the regressions control for the basic set of controls X_i made of the following dummies: past application, access to early registration, non-white ethnicity.

Figure A.3. Male applicants' qualifications



Note. The figure shows the distribution of male applicants' observable skills. The "desirable skills index" is computed as the mean of the following standardized variables: receiving a first grade, being from a top tier university, frequent past volunteering, high cognitive skills and score in English pre-university tests. The Figure on the left-hand side shows the distributions by photograph treatment and the dashed red line is for the male photograph. The Figure on the right-hand side shows the distributions by information treatment and the solid green line is for information of low past workers' effectiveness. In both graphs, the solid grey line is for the pure control group. The index of desirable skills shown is residualized on stratification variables first.

Figure A.4. Qualifications among top-ranked applicants by information treatment



Note. The figure shows quality among top-ranked applicants, by photograph treatment (on the left-hand side) and information treatment (on the right-hand side). Top-ranked applicants are defined as those belonging to the fifth quintile of the distribution of average test-scores across the whole hiring process and for both genders. There are three main proxies for quality in the figure. “Predicted job performance” is calculated using a truncated linear regression in the pure control group only with the following independent variables: ranking and average completion rate of the university attended by the candidate, subject studied, obtaining a first grade, whether the grade is expected or obtained, age, age squared and whether the person is in FTE. The “index of desirable skills” is computed as the mean of all the variables which are correlated with receiving a job offer: receiving a first grade, being from a top tier university, frequent past volunteering, high cognitive skills and score in English pre-university tests. “Selection test scores” is the average obtained by the candidate across the tests given by the employer in the hiring process. The combined index is obtained by averaging the three indexes (all standardized to be mean 0 and unitary standard deviation) and re-standardizing the resulting average.

Table A.6. Employer's hiring criteria

	Information		Photographs	
	(1)		(2)	
DV:	Offer	p-val	Offer	p-val
Top University * T^1	0.055*		0.068**	
	(0.029)	0.76	(0.028)	0.7
Top University * T^2	0.067**		0.053*	
	(0.029)		(0.029)	
First Grade * T^1	0.102***		0.063**	
	(0.032)	0.21	(0.030)	0.06
First Grade * T^2	0.109***		0.152***	
	(0.032)		(0.034)	
Aligned Subject * T^1	-0.007		0.008	
	(0.019)	0.09	(0.019)	0.77
Aligned Subject * T^2	0.028		0.014	
	(0.020)		(0.020)	
Past Volunteering * T^1	0.040**		0.050**	
	(0.020)	0.59	(0.020)	0.83
Past Volunteering * T^2	0.054***		0.044**	
	(0.021)		(0.020)	
Math Pre-Uni Score * T^1	0.016		-0.017	
	(0.028)	0.31	(0.025)	0.45
Math Pre-Uni Score * T^2	-0.023		0.012	
	(0.027)		(0.029)	
English Pre-Uni Score * T^1	0.080***		0.083***	
	(0.025)	0.66	(0.024)	0.5
English Pre-Uni Score * T^2	0.065**		0.060**	
	(0.025)		(0.026)	
Observations	2,271		2,271	
R-squared	0.069		0.069	
Strata Controls	✓		✓	

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Notes: OLS estimates. In Column (1), T^2 indicates information on low past effectiveness (and T^1 the alternative information). In Column (2), T^2 indicates a male photograph (and T^1 a female photograph). All regressions include controls for gender and ethnicity (stratification variables). Independent variables are interacted with the treatment and control dummies. "Top University" is equal to one if the candidate attended a top tier university in the U.K. "First Grade" is equal to one if the candidate got a first grade in university. "Past Volunteering" is equal to one if the candidate volunteered frequently in the past. "Maths Pre-Uni Score" and "English Pre-Uni Score" are equal to one if the candidate took the highest grade in Maths and English pre-university qualifications.

Table A.7. Effort in application completion

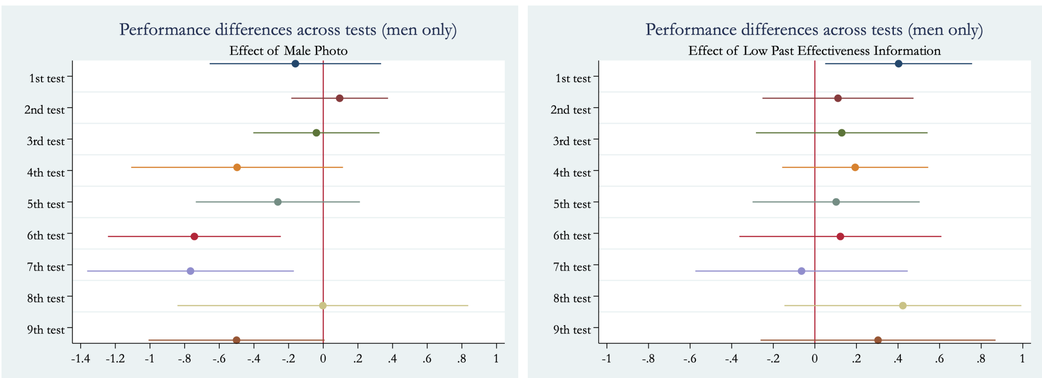
VARIABLES	(1) Access to portal	(2) # edits	(3) % completed	(4) Qst 1 length	(5) Qst 2 length
Male Photo	0.029	1.522	-0.021	-23.168	-35.240
	(0.025)	(2.125)	(0.023)	(55.645)	(46.005)
Low Past Effect	0.008	4.553**	0.026	32.439	41.648
	(0.025)	(2.185)	(0.023)	(55.686)	(46.057)
Observations	804	687	807	807	807
Strata Controls	✓	✓	✓	✓	✓
Week dummies	✓	✓	✓	✓	✓

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. The omitted category is the treatment group that received information on high past effectiveness. The variable "Access to portal" is a dummy for whether the person ever accessed the application portal to make changes to the application. The variable "# edits" counts how many times a candidate logged-in to make changes to the application form before submitting it. "% completed" is percentage of fields filled-in (not blank) in the application form. The variables "Qst 1 length" and "Qst 2 length" count number of characters used in each of the two motivational questions contained in the application form. All the regressions contain dummies for the week in which the candidate registered. The regressor "Low Past Effect" is a dummy equal to one for information on low past effectiveness (specification (1) of Section 5.1). All the regressions control for the basic set of controls X_i: access to early registration and non-white ethnicity.

Figure A.5. On-the-job test scores differences by treatment over time



Note. The figure reports the coefficients from a regression of each of the nine on-the-job assessment scores on the treatment dummy for receiving a male photograph (on the left) and the low past effectiveness information (on the right) for men only. Scores have been standardized by subtracting the gender-specific mean and dividing by the standard deviation. Coefficients are reported in chronological order from the top (first assessment) to the bottom (most recent assessment). All the regressions control for the basic set of controls X_i made of the following dummies: access to early registration, non-white ethnicity, past application, workplace region and allocation to preferred region.

Table A.8. New female hires: on-the-job performance

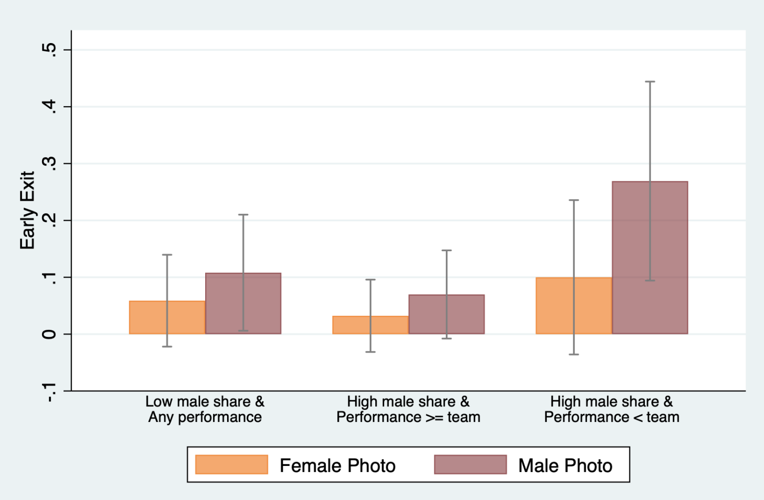
DV: On-the-Job Std. Scores				
	(1)	(2)	(3)	(4)
Male Photo	-0.006	-0.024	-0.009	-0.025
	(0.094)	(0.087)	(0.096)	(0.089)
Low Past Effect	-0.118	-0.081	-0.117	-0.081
	(0.087)	(0.078)	(0.086)	(0.078)
Controls				
Basic Controls	✓	✓	✓	✓
Exam FE	✓	✓	✓	✓
Quality Controls	×	✓	×	✓
Location Difficulty Controls	×	×	✓	✓
Summary Stats				
Observations	1,716	1,716	1,716	1,716
Mean Dep Var	0.03	0.03	0.03	0.03
Mean Dep Var in Pure C	0.03	0.03	0.03	0.03
Test Male Ph = Low Past Effect				
p-value	0.32	0.59	0.33	0.33

Clustered s.e. in parentheses (worker level)

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

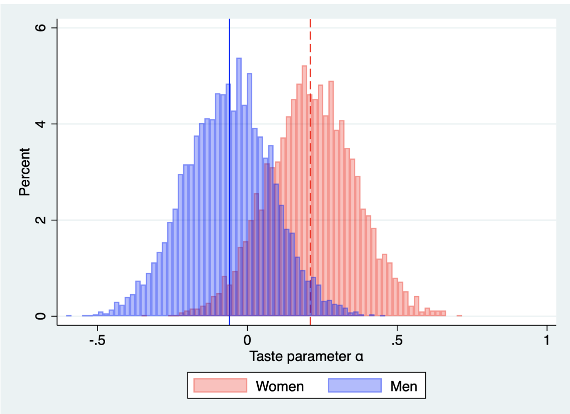
Note. OLS panel estimates for women only. The table reports results of four different regressions. The dependent variable is the on-the-job test score achieved in nine different assessments, standardized to be mean zero and unitary standard deviation in the full sample of male workers. The score goes from 0 to 100 in each test, and each test is weighted by the credits assigned to it by the organization. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). Columns (3) and (4) additionally control for an index of "difficulty" of the community where the worker is allocated to, using data collected by the UK Statistical Office. For each local authority, the index is obtained by averaging the score in these variables: social workers' caseload, turnover, absenteeism and scores on helping children, child care, leadership effectiveness. All the regressions control for the basic set of controls X_i made of the following dummies: access to early registration, non-white ethnicity, past application, workplace region fixed effect and a dummy for being allocated to the preferred region. Columns (2) and (4) control for the index of observable qualifications which are positively correlated with receiving a job offer. "Mean Dep Var" is the mean of the outcome variable in the omitted category and "Mean Dep Var PC" is the mean of the outcome variable in the pure control group. Standard errors are clustered at the worker level.

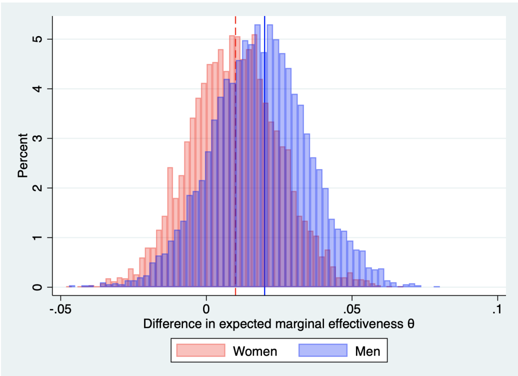
Figure A.6. Women's early exit from job by team male share and relative performance



Note. The Figure shows the average rate of turnover among women, splitting the sample in three categories: i) women allocated in teams with median or below median male share ($\leq 20\%$) and any relative performance, ii) women allocated in teams with higher than median male share ($> 20\%$) and with individual performance which is better than the leave-one-out team average and iii) women allocated in teams with higher than median male share ($> 20\%$) and with individual performance which is worse than the leave-one-out team average.

Figure A.7. Estimated weight on workplace gender shares and change in expected impact of talent

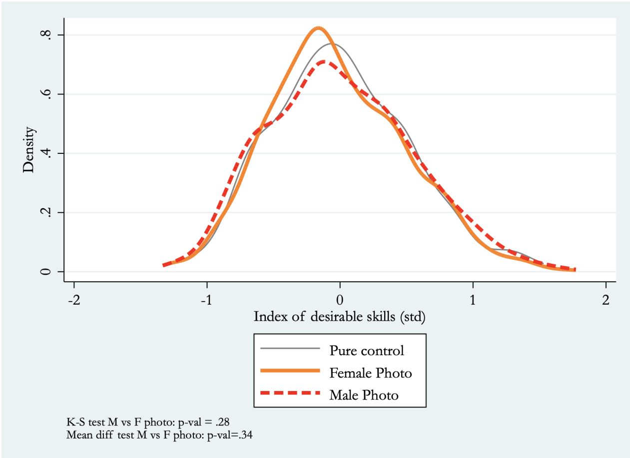




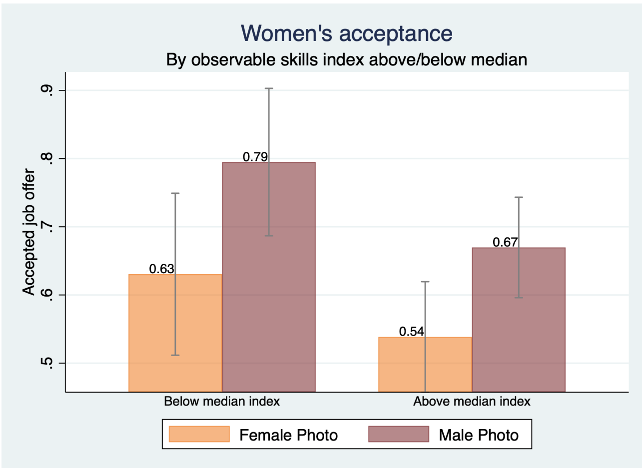
Note. The figure on the left-hand side shows distributions of the estimated weight on workplace gender shares α. The figure on the right-hand side shows distributions of the estimated difference in expected impact of talent $\Delta\theta$. Both graphs use the discrete-choice framework of section 7.1. Darker (blue) bars are for men and lighter (red) bars are for women. Solid lines are the mean value of the parameters for men and dashed lines are the mean value of the parameters for women. Multiple estimations are obtained through 5000 bootstrap replications of the logit model described in the main body of the paper with equal sample size (N=800) for the two genders.

Figure A.8. Female applicants' skills and offer acceptance by skill level and photo treatment

(a) Female applicants' skills



(b) Women's offer acceptance by skills



Note. The figure shows the distribution of female applicants' observable skills (a) and female applicants' offer acceptance by their level of observable skills (b). The index of observable skills ("index of desirable skills") is computed as the mean of the variables which are correlated with receiving an offer: receiving a first grade, being from a top tier university, frequent past volunteering, high cognitive skills and score in English pre-university tests. The figure in panel (a) shows the distribution of this index by photograph treatment among women who apply for the position. The figure in panel (b) shows the acceptance rate for women by above/below median "desirable skills index" and by photograph treatment, conditional on receiving a job offer.

Table A.9. Women's results: unconditional outcomes

VARIABLES	(1) Applied & no drop-out	(2) Applied	(3) Drop-Out Cond on App	(4) Drop-Out Uncond	(5) Received offer Cond on App	(6) Received offer Uncond	(7) Accepted offer Cond on Offer	(8) Accepted offer Cond on App	(9) Accepted offer Uncond
Male Photo	-0.051***	-0.051***	0.007	0.000	0.013	0.000	0.131**	0.028**	0.012
	(0.017)	(0.016)	(0.011)	(0.007)	(0.015)	(0.009)	(0.055)	(0.013)	(0.008)
	[0.00]	[0.00]	[0.54]	[0.98]	[0.39]	[0.98]	[0.03]	[0.03]	[0.13]
	{0.01}	{0.01}	{0.84}	{0.99}	{0.70}	{1}	{0.05}	{0.08}	{0.35}
Low Past Effectiveness	-0.015	-0.010	0.009	0.005	0.004	-0.000	-0.000	0.003	-0.001
	(0.017)	(0.016)	(0.011)	(0.007)	(0.015)	(0.009)	(0.055)	(0.013)	(0.008)
	[0.36]	[0.55]	[0.39]	[0.51]	[0.78]	[0.97]	[0.99]	[0.81]	[0.93]
	{0.85}	{0.90}	{0.89}	{0.94}	{0.99}	{1.00}	{0.97}	{1.00}	{1.00}
Summary Stats									
Observations	3,513	3,513	2,229	3,513	2,062	3,513	301	2,062	3,513
Mean Dep Var	0.60	0.64	0.06	0.04	0.14	0.09	0.55	0.08	0.05
Mean Dep Var in PC	0.59	0.62	0.05	0.03	0.15	0.09	0.68	0.10	0.06
Test Male Ph = Low Past Effectiveness									
p-value	0.120	0.07	0.86	0.64	0.66	0.95	0.09	0.15	0.25

Robust standard errors in parentheses
Randomization inference p-values in square brackets
FWER corrected p-values in curly brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. The table reports results of nine different regressions. The omitted category is the treatment group which received the female photograph and information on high past effectiveness. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment and the regressor "Low Past Effectiveness" is a dummy equal to one for receiving information on low workers' effectiveness (specification (1) of Section 5.1). Columns (1), (5) and (7) report the same regressions of Columns (1) to (3) of Table 2. In Column (2), the dependent variable is an indicator for applying at the first stage of the section process, independently of whether the applicant keeps applying at later stages of the process or not. The outcome variable in Columns (3) and (4) is an indicator equal to one if a person drops out of the hiring process at any later stage after applying in stage one. Column (3) only considers applicants and Column (4) reports results on unconditional drop-out. Column (6) reports results on receiving a job offer, not conditional on applying. Columns (8) and (9) report results on offer acceptance, respectively conditional on applying and unconditionally. All the regressions control for stratification variables: access to early registration and non-white ethnicity. "Mean Dep Var" is the mean of the outcome variable in the omitted category and "Mean Dep Var PC" is the mean of the outcome variable in the pure control group. Square brackets report p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions and curly brackets report multiplicity-adjusted p-values following List et al. (2019).

Table A.10. Treatment effects by sexuality and marital status

DV: Applied and never DO				
	(1)	(2)	(3)	(4)
	Women		Men	
Male Photo	-0.060***	-0.058***	-0.056	-0.030
	(0.018)	(0.018)	(0.039)	(0.039)
Non Hetero	0.009		-0.197***	
	(0.050)		(0.069)	
Male Photo * Non Hetero	0.051		0.199*	
	(0.066)		(0.106)	
Married		-0.048		-0.029
		(0.036)		(0.066)
Male Photo * Married		0.035		0.023
		(0.052)		(0.089)
Observations	3,294	3,455	757	793
Mean Dep Var	0.60	0.61	0.53	0.54

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

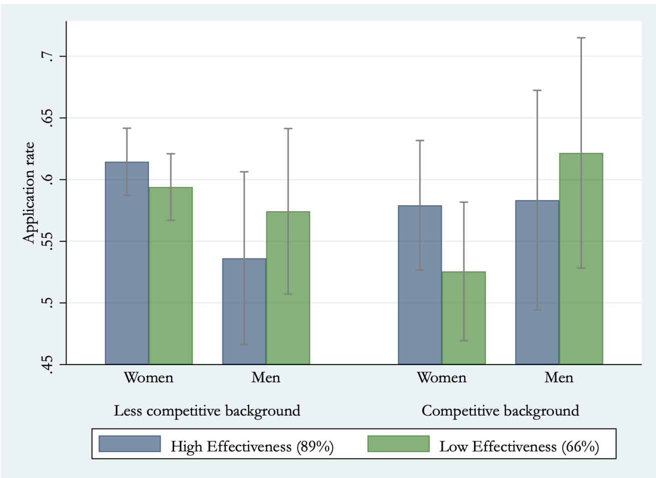
Note. OLS estimates. The regressor "Male Photo" is a dummy equal to one for the male photograph treatment."Non hetero" is a dummy equal to one if the person stated to be non-heterosexual and missing for refusing to answer the question on sexuality. "Married" is a dummy for being married or in a civil partnership and missing for refusing to answer the question on marital status. All the regressions control for the stratification controls X_i: access to early registration and non-white ethnicity.

Table A.11. A measure of overconfidence by gender

Overconfidence: self-reported number of skills above the mean							
	Women			Men			
	Mean	SD	N	Mean	SD	N	p-val
			Overall				
General	5.63	2.84	548	5.36	2.96	85	.43
Job specific	2.92	1.63	548	2.49	1.7	85	.03**
			Pure Control only				
General	5.5	2.73	123	5.63	2.95	19	.85
Job specific	2.82	1.55	123	2.53	1.84	19	.45

Note. The measure of overconfidence is defined in the following way. I asked survey respondents (N=633) to rate themselves in ten skills on a scale from 1 (min) to 10 (max). The skills are both general (i.e. complex problem solving, finance management, critical thinking, creativity, adaptability) and job specific (active listening, effective communication, leadership, empathy, client support). For each person, I construct a measure of overconfidence by counting the number of skills rated above the sample mean. The Table shows the mean measure of overconfidence by gender across treatments (in the first two rows) and in the pure control only (last two rows). Survey respondents are the subset of field participants who responded to a survey invitation (11.4%) sent to everyone in the invitation-to-apply email and subsequently encouraged through an ad-hoc email adding monetary incentives. The survey sample is representative of the overall pool of candidates (e.g., balanced on gender, treatment assignment, FTE status).

Figure A.9. Shock to expectations and competitiveness



Note. The graph shows raw differences in application rates in the high and low past effectiveness treatments by gender and a proxy of competitive attitudes. The proxy of competitive attitudes is built using information on the candidates' educational background. "Competitive background" is defined as having studied a male-dominated subject (e.g., engineering, business, math) in a top tier university in the UK. "Less competitive background" is defined as having studied a female-dominated subject (e.g., psychology, languages, humanities) in a non top tier university. "Women" and "Men" indicate the candidates' gender.

Table A.12. Overall gender shares and performance

	Male Share			**Share of Top Performers**		
	Applicants	Offerees	Workers	Women	Men	Overall
			Photograph			
Female Photo	18%	15%	16%	54%	67%	56%
Male Photo	17%	21%	21%	57%	42%	54%
			Information			
High Past Effectiveness	16%	14%	14%	54%	40%	52%
Low Past Effectiveness	19%	22%	**23%**	57%	58%	**57%**

Note. The first three columns of this table show the male share among applicants (Column 1), people who received a job offer (Column 2) and workers (Column 3), excluding people who drop out before completing the program. The last three columns show the share of top performers (people with an average score above 60%) by women (Column 4), men (Column 5) and for the pooled sample, excluding people who do not complete the program.

B Appendix tables: comparison with pure control

Table B.1. Men's results: including pure control

	(1)	(2)	(3)	(4)	(5)	(6)
	Application and selection process			On-the-job outcomes		
VARIABLES	Applied	Received Offer	Accepted Offer	Mean on-the-job score Semester 1	Mean on-the-job score Semesters 2/3	Early Exit
Panel A: Pooled Comparison Treatments + PC						
Male Photo	-0.020	0.023	0.015	-3.033	-1.237	-0.044
	(0.032)	(0.032)	(0.107)	(2.487)	(5.107)	(0.088)
	[0.526]	[0.518]	[0.868]	[0.205]	[0.808]	[0.648]
Low Past Effectiveness	0.067**	0.027	-0.118	5.153**	6.954	-0.146*
	(0.032)	(0.032)	(0.108)	(2.311)	(4.773)	(0.084)
	[0.037]	[0.404]	[0.30]	[0.039]	[0.153]	[0.087]
Panel B: Information vs PC						
High Past Effectiveness	-0.021	-0.089**	-0.163	-1.029	6.828	-0.190
	(0.043)	(0.044)	(0.123)	(3.309)	(7.532)	(0.138)
	[0.606]	[0.037]	[0.20]	[0.765]	[0.354]	[0.184]
Low Past Effectiveness	0.049	-0.028	-0.197*	3.584	9.633	-0.249**
	(0.042)	(0.045)	(0.111)	(2.604)	(6.952)	(0.124)
	[0.221]	[0.513]	[0.086]	[0.152]	[0.148]	[0.054]
Panel C: Photographs vs PC						
Female Photo	0.022	-0.084*	-0.239*	3.840	11.901*	-0.254*
	(0.043)	(0.043)	(0.121)	(2.744)	(6.779)	(0.128)
	[0.628]	[0.057]	[0.065]	[0.203]	[0.098]	[0.202]
Male Photo	0.006	-0.029	-0.147	0.504	6.509	-0.209
	(0.043)	(0.046)	(0.111)	(2.885)	(7.214)	(0.128)
	[0.885]	[0.551]	[0.186]	[0.85]	[0.396]	[0.085]
Summary Stats						
Treated obs	807	440	67	43	43	43
PC obs	206	109	23	19	19	19
Mean Dep Var PC	0.53	0.21	0.83	56.16	45.42	0.32
Equality of Coefficients p-values						
Panel A: Male Ph = Low Effect	0.08	0.93	0.45	0.04	0.24	0.39
Panel B: Low Effect = High Effect	0.04	0.06	0.79	0.13	0.62	0.56
Panel C: Male Ph = Female Ph	0.65	0.10	0.45	0.22	0.29	0.62

Robust standard errors in parentheses
Randomization inference p-values in square brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. Each panel reports the results of six different regressions. Panel A shows the main specification of the paper, but now the omitted category includes both the treatment group which received the female photograph and information on high past effectiveness as well as the pure control group. In Panels B and C, the omitted category is the pure control group which received no photograph and no information. In panel B, each column is a regression of the outcome variable on two dummies, one for receiving information on high and one for low past workers' effectiveness, respectively. In panel C, each column is a regression of the outcome variable on two dummies, one for receiving a male photograph and one for receiving a female photograph, respectively. The dependent variables are indicator dummies for application and never dropping out, receiving a job offer (conditional on applying) and accepting the job offer (conditional on receiving the offer) in Columns (1), (2) and (3). The dependent variable in column (4) is the average on-the-job test score achieved in the first five assessments during the first semester on the job and in column (5) is the average on-the-job test score achieved in the four additional assessments during the second and third semester on the job. The score goes from 0 to 100 in each test, the average is weighted by the credits assigned to each exam by the organization. The dependent variable in column (6) is a dummy equal to one if the person left the program before completing it. All the regressions control for stratification variables: access to early registration and non-white ethnicity. Round brackets report robust standard errors and square brackets report the p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions.

Table B.2. New male hires: on-the-job performance including PC

	DV: On-the-Job Std. Scores			
	(1)	(2)	(3)	(4)
Panel A: Pooled Comparison Treatments + PC				
Male Photo	-0.180	-0.183	-0.167	-0.174
	(0.165)	(0.161)	(0.170)	(0.165)
Low Past Effectiveness	0.293**	0.284**	0.302**	0.291**
	(0.128)	(0.128)	(0.130)	(0.130)
Panel B: Information vs PC				
High Past Effectiveness	0.047	0.055	0.075	0.075
	(0.223)	(0.215)	(0.228)	(0.221)
Low Past Effectiveness	0.259	0.253	0.288	0.275
	(0.178)	(0.172)	(0.184)	(0.178)
Panel C: Photographs vs PC				
Female Photo	0.303	0.307*	0.327*	0.324*
	(0.188)	(0.174)	(0.190)	(0.178)
Male Photo	0.073	0.068	0.102	0.089
	(0.207)	(0.201)	(0.216)	(0.211)
Controls				
Strata Controls	✓	✓	✓	✓
Exam FE	✓	✓	✓	✓
Quality Controls	×	✓	×	✓
Location Difficulty Controls	×	×	✓	✓
Summary Stats				
Treated obs	387	387	387	387
PC obs	171	171	171	171
Mean Dep Var in PC	-0.12	-0.12	-0.12	-0.12
Photo = Past Effect p-values				
Panel A	0.03	0.03	0.03	0.03
Panel B	0.19	0.21	0.19	0.21
Panel C	0.18	0.13	0.20	0.14

Clustered s.e. in parentheses (worker level)

*** p<0.01, ** p<0.05, * p<0.1

Note. OLS panel estimates for men only. Each panel reports results of four different regressions. Panel A shows the main specification of the paper, but now the omitted category includes both the treatment group which received the female photograph and information on high past effectiveness as well as the pure control group. In Panels B and C, the omitted category is the pure control group which received no photograph and no information. In panel B, each column is a regression of the outcome variable on two dummies, one for receiving information on high and one for low past workers' effectiveness, respectively. In panel C, each column is a regression of the outcome variable on two dummies, one for receiving a male photograph and one for receiving a female photograph, respectively. The dependent variable is the on-the-job test score achieved in nine different assessments, standardized to be mean zero and unitary standard deviation in the full sample of male workers. The score goes from 0 to 100 in each test, and each test is weighted by the credits assigned to it by the organization. All the regressions control for the basic set of controls X_i made of the following dummies: access to early registration, non-white ethnicity, past application, workplace region and a dummy for being allocated to the preferred region. Columns (3) and (4) additionally control for an index of "difficulty" of the community where the worker is allocated to, using data collected by the UK Statistical Office. For each local authority, I compute an index of "difficulty" by averaging the score in these variables: social workers' caseload, turnover, absenteeism and scores on helping children, child care, leadership effectiveness. Columns (2) and (4) control for the index of observable qualifications which are positively correlated with receiving a job offer. Standard errors are clustered at the worker level.

Table B.3. Women's results: including pure control

	(1)	(2)	(3)	(4)	(5)	(6)
	Application and selection process			On-the-job outcomes		
VARIABLES	Applied	Received Offer	Accepted Offer	Mean on-the-job score Semester 1	Semesters 2/3	Early Exit
Panel A: Pooled Comparison Treatments + PC						
Male Photo	-0.042***	0.009	0.095*	1.060	-2.094	0.053
	(0.015)	(0.014)	(0.050)	(0.993)	(1.895)	(0.039)
	[0.01]	[0.55]	[0.06]	[0.30]	[0.30]	[0.17]
Low Past Effectiveness	-0.006	-0.000	-0.034	-0.932	-1.526	0.074*
	(0.015)	(0.014)	(0.051)	(1.019)	(1.881)	(0.041)
	[0.67]	[1.00]	[0.50]	[0.35]	[0.43]	[0.07]
Panel B: Information vs PC						
High Past Effectiveness	0.006	-0.007	-0.051	-0.615	1.097	-0.026
	(0.020)	(0.019)	(0.064)	(1.158)	(2.117)	(0.042)
	[0.75]	[0.71]	[0.39]	[0.61]	[0.64]	[0.56]
Low Past Effectiveness	-0.009	-0.003	-0.051	-1.149	-1.175	0.066
	(0.020)	(0.019)	(0.065)	(1.153)	(2.284)	(0.051)
	[0.63]	[0.84]	[0.42]	[0.34]	[0.62]	[0.21]
Panel C: Photographs vs PC						
Female Photo	0.024	-0.012	-0.116*	-1.759	1.607	-0.020
	(0.020)	(0.019)	(0.065)	(1.219)	(1.995)	(0.043)
	[0.21]	[0.57]	[0.08]	[0.17]	[0.58]	[0.30]
Male Photo	-0.027	0.002	0.013	-0.173	-1.354	0.052
	(0.020)	(0.019)	(0.063)	(1.106)	(2.322)	(0.049)
	[0.17]	[0.93]	[0.84]	[0.88]	[0.46]	[0.66]
Summary Stats						
Treated obs	3,513	2,062	301	191	191	191
PC obs	891	524	79	54	54	54
Mean Dep Var PC	0.59	0.15	0.68	59.48	57.76	0.07
Equality of Coefficients p-values						
Panel A: Male Ph = Low Effect	0.12	0.66	0.10	0.22	0.83	0.71
Panel B: Low Effect = High Effect	0.35	0.82	1	0.64	0.28	0.03
Panel C: Male Ph = Female Ph	0	0.39	0.02	0.17	0.15	0.08

Robust standard errors in parentheses
Randomization inference p-values in square brackets
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for women only. Each panel reports the results of six different regressions. Panel A shows the main specification of the paper, but now the omitted category includes both the treatment group which received the female photograph and information on high past effectiveness as well as the pure control group. In Panels B and C, the omitted category is the pure control group which received no photograph and no information. In panel B, each column is a regression of the outcome variable on two dummies, one for receiving information on high and one for low past workers' effectiveness, respectively. In panel C, each column is a regression of the outcome variable on two dummies, one for receiving a male photograph and one for receiving a female photograph, respectively. The dependent variables are indicator dummies for application and never dropping out, receiving a job offer (conditional on applying) and accepting the job offer (conditional on receiving the offer) in Columns (1), (2) and (3). The dependent variable in column (4) is the average on-the-job test score achieved in the first five assessments during the first semester on the job and in column (5) is the average on-the-job test score achieved in the four additional assessments during the second and third semester on the job. The score goes from 0 to 100 in each test, the average is weighted by the credits assigned to each exam by the organization. The dependent variable in column (6) is a dummy equal to one if the person left the program before completing it. All the regressions control for stratification variables: access to early registration and non-white ethnicity. Round brackets report robust standard errors and square brackets report the p-values of the coefficients on the indicated treatment dummies from randomization inference (randomization-t) with 1000 repetitions.

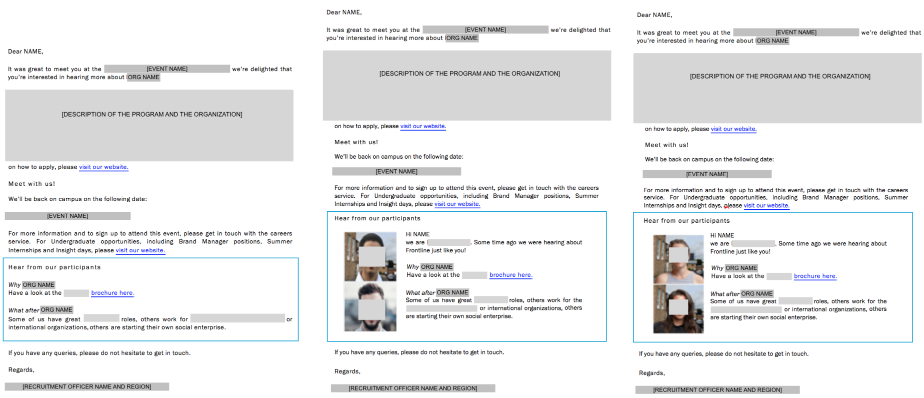
C External validity of the gender shares effect: evidence from a complementary experiment.

During 2017 I conducted a second field experiment with the same partner organization in order to try and increase the pool of candidates interested in the job. This second experiment was pre-registered together with the main field experiment presented in the paper and focused on testing whether manipulating perceptions of gender shares in social work could encourage students to get informed about the job and eventually apply for it. The results of this second experiment can thus be used to provide evidence on whether the null effect of gender shares on men's applications generalizes in a sample of young men who are not yet selected on interest in the job.

Setting: In the fall of 2017, the partner organization visited 52 universities across the country to promote its program and encourage applications through a variety of events (e.g., job fairs, workshops). On average, each university was visited three times (max 6). Each university is assigned to a Recruitment Officer (RO) in charge of organizing and conducting the events, collecting email addresses of event participants and sending a follow-up email with further information about the job.

Experiment: People who took part to career events and left their email address in a mailing list were randomly assigned to three groups, which differed in the format of the follow-up email received. The text content of these three emails was exactly the same, but they might show i) no picture, ii) a picture of previous female workers, ii) or a picture of previous male workers. The three email templates are shown in Figure C.1. Assignment to treatment was stratified by university, event and gender.

Figure C.1. Experiment in universities: treatments



Outcomes and balance: Each email contains links to the organization's website which are trackable at the level of stratification and treatment. This allows me to know the number of participants of gender g in event e in university u that clicked on any email link, whether they are first time users and some metrics of their online behavior for each treatment group. The main outcome of this experiment is whether people click on "Apply" on the organization's website. Each event had an average number of 30 sign-ups, for a total of 2877 unique participants. Table

C.1 presents summary statistics of the experimental sample and balance checks.

Table C.1. Experiment in universities: balance and summary statistics

	Overall			Joint test		Pairwise tests	
	N	*mean*	*sd*	*F stat*	*p-value*	*min p value*	*max diff*
Male	2877	0.22	0.42	1.397	0.248	*0.095	0.032
Last year	2500	0.58	0.49	0.120	0.887	0.662	-0.011
Graduates	2500	0.10	0.30	0.298	0.742	0.453	0.011
First/second year	2500	0.32	0.47	0.067	0.935	0.739	-0.008
Science or business	2334	0.21	0.41	1.230	0.292	0.168	0.028
Heard about the job	2334	0.29	0.45	0.863	0.422	0.245	0.027
- on campus	1221	0.21	0.41	1.411	0.244	0.125	0.043
- in news/ads	1221	0.55	0.50	1.492	0.225	*0.091	-0.058
- from friends	1221	0.07	0.26	0.090	0.914	0.680	0.008
- online	1221	0.17	0.37	0.317	0.729	0.454	-0.020

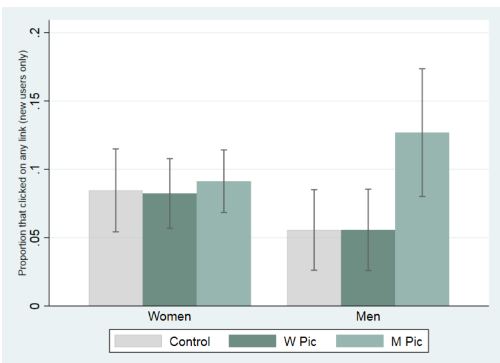
Note. “Last year” and “First/second year” are indicators for the year of enrolment in university. “Science or business” is an indicator for studying a scientific or economics/business subject. “Heard about the job” is equal to one if the person heard of the organization before attending the event. Columns 4 and 5 report the F-statistic and p-value from a joint test of the significance of the treatment dummies in explaining each row variable. The last two Columns report the minimum p-value and maximum difference from t-tests between pairs of treatment groups.

Results: Men are twice as likely to access the organization’s website when they receive an email containing male photographs as compared to both the control group and the treatment with female photographs (Figure C.2 shows the number of clicks by gender and treatment). However, men’s behavior on the website does not lead to more applications. Table C.2 estimates the effect of treatment emails on application for a person of gender g, event e and university u using:

$$y_{geu} = c + \beta_1 MPic_{geu} + \beta_2 WPic_{geu} + X'_{eu}\beta_3 + \delta_u + \epsilon_{geu}$$

where δ_u are university fixed effects and X_{eu} are event controls (type of event, month, number of participants, gender of RO). I use robust standard errors (randomization was at individual level) and add analytical weights by treatment group size. Table C.2 shows that the male photograph treatment doesn’t increase men’s applications, which reinforces the external validity of the null effect of the male photograph from my main field experiment.

Figure C.2. Experiment in universities: effect on information acquisition



Note: The bar chart shows the proportion of clicks by new users in the different treatment groups of the experiment.

Table C.2. Experiment in universities: effects on applications

DV: Event participant registered to apply		
	(1)	(2)
VARIABLES	M	W
Women's Pic	-0.046	0.017
	(0.038)	(0.028)
Men's Pic	0.008	0.007
	(0.054)	(0.029)
Scientific Subject	-0.075**	-0.107***
	(0.029)	(0.024)
Observations	337	1,259
R-squared	0.148	0.109
Mean Dep Var	0.082	0.17

Clustered standard errors in parentheses (uni level)
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS regressions for men and women separately. The dependent variable is equal to one of the participant filled-in the online registration form necessary to apply for the job. The omitted category is the group receiving emails with no workers' photographs. "Women's Pic" and "Men's Pic" are indicator variables for the two experimental treatments. The regression includes university fixed effects and event controls X_{eu} for event type, month, number of participants and gender of RO. I add analytical weights by treatment group size. The table limits the sample to last year students or graduates.

D Can overconfidence explain the results?

In this section I further explore whether the effect of the information treatment can be explained by (gender differences in) overconfidence. I use the survey questions defined at the end of Section OD.2 to construct a proxy of individual over-precision in their priors on men and women's performance in female jobs. I select the most important observable predictors of this measure using Lasso regression and impute the coefficients to my experimental sample. This provides a measure of "predicted confidence" (overprecision) in others' performance in social work and teaching.

Table E.1 shows the treatment effect on men's application likelihood depending on their predicted confidence. The increase in application rates is driven by men with over-precision below the median. As long as this is correlated with a higher likelihood of under-placement of own ability with respect to others, it suggests that the effects are actually driven by the least confident men. Moore and Healy (2008) show that lack of precision on beliefs about others is positively correlated with overplacement in easy tasks, but also positively correlated with under-placement in hard tasks. In other words, unprecise estimates of others' performance increase people's tendency to under-place one's own performance in hard tasks. This seems the relevant case in my context, as it is consistent with the hypothesis that information provision benefits the most men who start off with greater uncertainty about own success in female-dominated jobs.

Table E.1. Treatment effects by predicted priors' uncertainty

	DV: Applied and never drop-out			
	(1)	(2)	(3)	(4)
	Confidence in women's ability		Confidence in men's ability	
	$<$ med	$>$ med	$<$ med	$>$ med
Low Past Effectiveness	0.108**	0.041	0.137***	0.017
	(0.048)	(0.049)	(0.049)	(0.051)
Observations	394	398	386	406
R-squared	0.024	0.016	0.025	0.014
Mean Dep Var	0.53	0.56	0.48	0.62

Bootstrapped se in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note. OLS estimates for men only. Columns (1) and (2) split the sample at the median level of predicted confidence in priors about women's performance in social work and primary school teaching. Columns (3) and (4) do the same for priors about men. The variables used to predict confidence are age, whether the person studied in a top university, non-white ethnicity, whether the person studied a common university subject for social workers, exposure to occupational gender segregation and gender. The omitted category is the group that received information on high past effectiveness.

[For Online Publication]

OA. Appendix tables: manipulation checks

Table OA.1. Workers' photographs: robustness checks

	Female Photo			Male Photo			Diff means
	Mean	SD	N	Mean	SD	N	P-val
Panel A: 2018 Applicants							
White pictures							
Friendliness	.79	.41	92	.63	.48	95	.01
Work satisfaction	.91	.28	92	.84	.37	95	.14
Non-white pictures							
Friendliness	.86	.36	28	.82	.39	28	.72
Work satisfaction	.82	.39	28	.93	.26	28	.23
Panel B: Prolific Ac sample							
White pictures							
Friendliness	.87	.34	98	.74	.44	95	.02
Work satisfaction	.81	.4	98	.76	.43	95	.42
Non-white pictures							
Friendliness	.97	.17	33	.92	.28	36	.35
Work satisfaction	.97	.17	33	.92	.28	36	.35
Panel C: Amazon Turk sample							
White pictures							
Happy feeling	.79	.41	39	.66	.48	38	.18
Friendliness	.9	.31	39	.74	.45	38	.07
Work satisfaction	.87	.34	39	.76	.43	38	.22
Trust	.85	.37	39	.82	.39	38	.73
Attractiveness	.72	.46	39	.76	.43	38	.66
Professional clothing	.38	.49	39	.87	.34	38	0
Non-white pictures							
Happy feeling	.9	.3	42	.9	.3	42	1
Friendliness	.98	.15	42	.95	.22	42	.56
Work satisfaction	.95	.22	42	.88	.33	42	.24
Trust	.93	.26	42	.88	.33	42	.46
Attractiveness	.95	.22	42	.74	.45	42	.01
Professional clothing	.93	.26	42	.9	.3	42	.7

Note. Friendliness of the person in the picture was rated answering the question: "How does the person in the photograph appear to you?" on a 5-points scale. The variable "Friendliness" is a dummy equal to 1 if the person replied Friendly or Very Friendly and 0 otherwise. Work satisfaction was rated answering: "In your opinion, how satisfied is this person in his/her work?" on a 5-points scale. The variable "Work Satisfaction" is a dummy equal to 1 if the person replied Satisfied or Very Satisfied and 0 otherwise. The question "To what extent does this image make you feel happy?" assessed emotional reaction to the picture on a 7-points scale. The variable "Happy feeling" takes values between -3 ("Extremely unhappy") and 3 ("Extremely happy"). The variable for trust is defined from answers to the question "If this person was giving you some information about her job, would you trust him/her?", to which people answered on a 5-points scale; the variable has values between -2 ("Definitely not") and 2 ("Definitely yes"). The variable attractiveness is defined from answers to the question "In your opinion, how does this person look like?", to which people answered on a 5-points scale; the variable has values between -2 ("Not attractive") and 2 ("Attractive"). The variable professional clothing is a dummy equal to one if the respondent would describe the clothes of the portrayed person as "professional" and 0 if "unprofessional". In the Amazon Mechanical Turk sample the number of respondents for each question may vary by design: the more sensitive questions on clothing, ethnicity, attractiveness and trust were asked only on a subset of respondents. Data for Panels A and B come from the auxiliary surveys conducted with 2018 Applicants to the partner organization and Prolific Academic respondents. In these surveys, I asked respondents to categorize people portrayed in the intervention photographs along two characteristics: friendliness and work satisfaction. Each respondent was asked about only one photograph, which was the same used afterwards in the survey in displaying the full intervention. Data for Panel C come from an additional survey I conducted with Amazon Turk workers. See Appendix OE. for more details on the sampling.

Table OA.2. Manipulation checks by sample: information

	66% Info Mean (SD)	89% Info Mean (SD)	Difference L-H (SE)	P-values Uncorrected	FDR	FWER
Panel A: 2018 Applicants sample (N=242)						
Main manipulation checks						
% of high-performers on the job	68.20	73.72	-5.52	0.00	0.02	0.04
	(11.95)	(14.08)	(1.68)			
% of high-skilled applicants	72.63	80.27	-7.64	0.00	0.02	0.07
	(19.62)	(20.05)	(2.55)			
Own expected performance	8.28	8.01	0.28	0.12	0.21	0.91
	(1.29)	(1.42)	(0.17)			
Alternative mechanisms						
% female applicants	69.17	70.49	-1.32	0.43	0.39	1.00
	(13.60)	(12.73)	(1.69)			
Job desirable for men	0.71	0.74	-0.03	0.64	0.47	0.99
	(0.46)	(0.44)	(0.06)			
Job desirable for women	0.81	0.88	-0.07	0.16	0.22	0.96
	(0.40)	(0.33)	(0.05)			
Customers discriminate workers	0.39	0.53	-0.14	0.04	0.08	0.54
	(0.49)	(0.50)	(0.06)			
Job has high social status	0.51	0.68	-0.17	0.01	0.03	0.16
	(0.50)	(0.47)	(0.06)			
Job difficulty	65.81	60.31	5.49	.028	0.08	0.47
	(17.69)	(21.25)	(2.52)			
Wage level (N=84)	51.14	51.32	-0.18	0.96	0.69	0.96
	(12.88)	(15.76)	(3.13)			
Quality standards (N=78)	77.55	73.09	4.46	0.37	0.39	0.99
	(17.54)	(23.96)	(4.70)			
Promotion difficulty	55.46	55.98	-0.52	0.81	0.60	0.99
	(15.77)	(18.04)	(2.19)			
Application effort (N=78)	59.36	62.13	-2.77	0.63	0.47	1.00
	(25.91)	(23.62)	(5.64)			
Number of applicants	61.72	58.36	3.36	0.17	0.22	0.96
	(17.74)	(19.26)	(2.38)			
Panel B: Prolific Ac sample (N=262)						
Main manipulation checks						
% of high-performers on the job	64.02	74.03	-10.01	0.00	0.00	0.00
	(12.42)	(12.26)	(1.53)			
% of high-skilled applicants	65.81	76.62	-10.81	0.00	0.00	0.00
	(16.43)	(17.99)	(2.13)			
Own expected performance	6.79	6.39	0.40	0.18	0.69	0.96
	(2.43)	(2.28)	(0.29)			
Alternative mechanisms						
% female applicants	71.32	72.87	-1.56	0.27	0.86	0.99
	(11.09)	(11.62)	(1.40)			
Job desirable for men	0.69	0.61	0.08	0.19	0.69	0.96
	(0.46)	(0.49)	(0.06)			
Job desirable for women	0.95	0.93	0.01	0.63	0.91	0.99
	(0.23)	(0.25)	(0.03)			
Customers discriminate workers	0.45	0.41	0.04	0.55	0.91	1.00
	(0.50)	(0.49)	(0.06)			
Job has high social status	0.46	0.49	-0.03	0.60	0.91	1.00
	(0.50)	(0.50)	(0.06)			
Job difficulty	65.61	62.51	3.10	0.21	0.69	0.96
	(19.82)	(19.56)	(2.43)			
Wage level	43.95	45.95	-2.00	0.37	0.91	0.99
	(19.64)	(17.59)	(2.30)			
Quality standards (N=132)	73.90	76.28	-2.38	0.52	0.91	1.00
	(21.57)	(19.25)	(3.61)			
Promotion difficulty	54.29	56.20	-1.91	0.35	0.91	0.99
	(16.30)	(17.77)	(2.11)			
Application effort (N=132)	59.73	59.04	0.69	0.85	0.96	0.97
	(20.59)	(22.41)	(3.78)			
Number of applicants	47.85	51.58	-3.73	0.17	0.69	0.96
	(21.83)	(22.49)	(2.74)			

Note. Variables under "Alternative mechanisms" are defined as follows."% female applicants" is the perceived female share among 100 applicants. Variables "job desirable for men", "job desirable for women", "customers discriminate workers", "job has high social status" show the share of people that agree with the given statement. Answers were on a 6-points scale from "Strongly Agree" to "Strongly Disagree" and I created dummy variables equal to one for the three highest options. Variables "Job difficulty", "Wage level", "Quality standards", "Promotion difficulty" and "Application effort" asked participants to rate the given dimension of the job on a scale from 0 to 100. "Number of applicants" is the believed number of people that apply out of 100 who are considering whether or not to apply for the job. Variables in the section "Main manipulation checks" are defined in Section 3.1. Some questions were shown to subsamples only, implying differences in the number of respondents. Columns denoted with "FDR" and "FWER" report multiplicity-adjusted p-values following Benjamini et al. (2006) and List et al. (2019).

Table OA.3. Manipulation checks by sample: photographs

	Female Ph. Mean (SD)	Male Ph. Mean (SD)	Diff L-H (SE)	P-values Uncorrected	FDR	FWER
Panel A: 2018 Applicants sample (N=242)						
Main manipulation checks						
% female applicants	72.50	67.22	-5.28***	0.00	0.01	0.01
	(12.60)	(13.22)	(1.66)			
Job desirable for men	0.62	0.82	0.19***	0.00	0.00	0.03
	(0.49)	(0.39)	(0.06)			
Job desirable for women	0.96	0.73	-0.23***	0.00	0.00	0.00
	(0.20)	(0.45)	(0.04)			
Alternative mechanisms						
% of high-skilled applicants	77.35	75.63	-1.73	0.51	0.84	1.00
	(19.11)	(21.20)	(2.60)			
% of high-performers on the job	72.60	69.39	-3.21*	0.06	0.20	0.76
	(11.99)	(14.40)	(1.70)			
Own expected performance	8.08	8.20	0.12	0.47	0.84	1.00
	(1.31)	(1.41)	(0.18)			
Customers discriminate workers	0.51	0.41	-0.10	0.15	0.39	0.95
	(0.50)	(0.49)	(0.06)			
Job high social status	0.63	0.55	-0.08	0.21	0.54	0.98
	(0.48)	(0.50)	(0.06)			
Job difficulty	63.09	63.01	-0.08	0.98	1.00	1.00
	(20.34)	(19.16)	(2.55)			
Wage level (N=84)	52.09	50.27	-1.82	0.57	0.86	1.00
	(16.10)	(12.09)	(3.13)			
Quality standards (N=78)	73.21	77.36	4.14	0.39	0.84	1.00
	(20.25)	(20.84)	(4.72)			
Promotion difficulty	56.52	54.93	-1.59	0.47	0.84	1.00
	(17.22)	(16.63)	(2.19)			
Application effort (N=78)	60.21	61.83	1.62	0.78	0.96	1.00
	(25.34)	(23.79)	(5.60)			
Number of applicants	60.61	59.45	-1.16	0.63	0.89	1.00
	(18.56)	(18.62)	(2.39)			
Panel B: Prolific Ac sample (N=262)						
Main manipulation checks						
% female applicants	74.91	69.29	-5.62***	0.00	0.00	0.00
	(10.62)	(11.43)	(1.36)			
Job desirable for men	0.60	0.70	0.10*	0.10	0.78	0.90
	(0.49)	(0.46)	(0.06)			
Job desirable for women	0.94	0.94	0.00	1.00	1.00	1.00
	(0.24)	(0.24)	(0.03)			
Alternative mechanisms						
% of high-skilled applicants	70.50	72.02	1.52	0.49	0.99	1.00
	(18.71)	(17.36)	(2.23)			
% of high-performers on the job	68.48	69.64	1.16	0.48	0.99	1.00
	(12.96)	(13.66)	(1.65)			
Own expected performance	6.47	6.72	0.25	0.39	0.99	1.00
	(2.51)	(2.20)	(0.29)			
Customers discriminate workers	0.47	0.38	-0.09	0.14	0.78	0.94
	(0.50)	(0.49)	(0.06)			
Job high social status	0.45	0.50	0.05	0.38	0.99	1.00
	(0.50)	(0.50)	(0.06)			
Job difficulty	65.13	62.96	-2.17	0.39	0.99	1.00
	(18.67)	(20.71)	(2.44)			
Wage level	44.28	45.63	1.34	0.56	1.00	1.00
	(19.73)	(17.51)	(2.30)			
Quality standards (N=132)	76.43	73.38	-3.06	0.40	0.99	1.00
	(20.67)	(20.32)	(3.60)			
Promotion difficulty	53.56	56.95	3.40	0.11	0.78	0.91
	(17.62)	(16.36)	(2.10)			
Application effort (N=132)	59.35	59.36	0.00	1.00	1.00	1.00
	(21.32)	(21.86)	(3.77)			
Number of applicants	49.76	49.70	-0.06	0.98	1.00	1.00
	(22.59)	(21.89)	(2.75)			

Note. Variables under "Alternative mechanisms" are defined as follows. "% of high-skilled applicants" is the average answer to the questions "Out of 100 [women/men] that apply for this job after seeing the email ad, how many do you think that have the potential to get commendable or excellent feedback on the job?". "% of high-performers in the job" is the average of answers to the questions "Now that you have seen the email ad, indicate the proportion of [women/men] that you think are successful on-the-job". "Own expected performance" is the expected on-the-job performance on a scale from 1 (min) to 10 (max). Variables "customers discriminate workers" and "job has high social status" show the share of people that agree with the given statement. Answers were on a 6-points scale from "Strongly Agree" to "Strongly Disagree" and I created dummy variables equal to one for the three highest options. "Job difficulty", "Wage level", "Quality standards", "Promotion difficulty" and "Application effort" asked participants to rate the given dimension of the job on a scale from 0 to 100. "Number of applicants" is the believed number of people that apply out of 100 who are considering whether or not to apply for the job. Variables in the section "Main manipulation checks" are defined in Section 3.1. Some questions were shown to subsamples only, implying differences in the number of respondents. Columns denoted with "FDR" and "FWER" report multiplicity-adjusted p-values following Benjamini et al. (2006) and List et al. (2019).

Table OA.4. Manipulation checks by gender: information

	66% Info Mean (SD)	89% Info Mean (SD)	Diff L-H (SE)	P-values Uncorrected	FDR	FWER
Panel A: Men (N=178)						
Main manipulation checks						
% of high-performers on the job	65.76	71.77	-6.01	0.00	0.03	0.06
	(11.43)	(15.26)	(2.03)			
% of high-skilled applicants	66.36	74.91	-8.55	0.00	0.03	0.03
	(17.41)	(20.87)	(2.89)			
Own expected performance	65.75	65.38	0.36	0.92	1.00	1.00
	(24.99)	(22.28)	(3.55)			
Alternative mechanisms						
% female applicants	70.05	70.69	-0.65	0.72	1.00	1.00
	(11.94)	(12.04)	(1.80)			
Job desirable for men	65.52	57.14	8.37	0.25	0.91	1.00
	(47.81)	(49.76)	(7.32)			
Job desirable for women	95.40	91.21	4.19	0.27	0.91	1.00
	(21.06)	(28.47)	(3.77)			
Discrimination by customers	0.52	0.44	0.08	0.30	0.91	1.00
	(0.50)	(0.50)	(0.08)			
Job high social status	0.44	0.51	-0.07	0.36	0.99	1.00
	(0.50)	(0.50)	(0.08)			
Job difficulty	60.57	61.95	-1.37	0.67	1.00	1.00
	(20.63)	(21.93)	(3.19)			
Wage level	40.28	45.87	-5.59	0.07	0.41	0.76
	(18.71)	(18.76)	(3.09)			
Quality standards	71.19	74.40	-3.21	0.50	1.00	1.00
	(22.56)	(20.47)	(4.76)			
Promotion difficulty	54.41	54.57	-0.16	0.96	1.00	0.96
	(17.92)	(20.10)	(2.86)			
Application effort	62.03	59.28	2.75	0.57	1.00	1.00
	(20.54)	(22.36)	(4.83)			
Number of applicants	47.17	50.74	-3.56	0.30	0.91	0.99
	(22.20)	(23.25)	(3.41)			
Panel B: Women (N=325)						
Main manipulation checks						
% of high-performers on the job	66.17	75.07	-8.91	0.00	0.00	0.00
	(12.85)	(11.71)	(1.36)			
% of high-skilled applicants	70.54	80.19	-9.65	0.00	0.00	0.00
	(18.67)	(17.73)	(2.02)			
Own expected performance	80.06	75.19	4.88	0.01	0.04	0.21
	(16.54)	(19.12)	(1.98)			
Alternative mechanisms						
% female applicants	70.42	72.23	-1.82	0.19	0.23	0.97
	(12.64)	(12.29)	(1.38)			
Job desirable for men	72.39	73.29	-0.90	0.86	0.58	1.00
	(44.84)	(44.38)	(4.96)			
Job desirable for women	84.05	90.06	-6.01	0.11	0.14	0.87
	(36.73)	(30.01)	(3.73)			
Discrimination by customers	0.37	0.48	-0.12	0.03	0.07	0.52
	(0.48)	(0.50)	(0.05)			
Job high social status	0.51	0.62	-0.11	0.04	0.08	0.57
	(0.50)	(0.49)	(0.05)			
Job difficulty	68.44	61.17	7.27	0.00	0.00	0.00
	(17.18)	(19.57)	(2.04)			
Wage level	49.62	48.37	1.25	0.60	0.48	1.00
	(17.29)	(16.13)	(2.38)			
Quality standards	77.71	75.43	2.27	0.53	0.46	1.00
	(18.27)	(21.71)	(3.58)			
Promotion difficulty	55.09	56.98	-1.89	0.28	0.28	1.00
	(14.97)	(16.52)	(1.75)			
Application effort	58.00	60.78	-2.78	0.50	0.46	1.00
	(23.73)	(23.24)	(4.15)			
Number of applicants	58.42	57.09	1.33	0.54	0.46	1.00
	(19.46)	(19.78)	(2.18)			

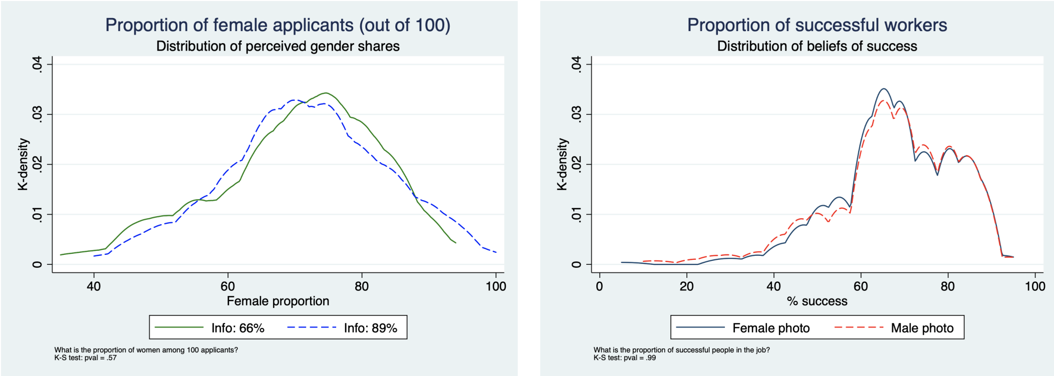
Note. For variable definitions, see the notes to Figure OA.2. Variables in the section "Main manipulation checks" are defined in Section 3.1. Some questions were shown to subsamples only, implying differences in the number of respondents. Columns denoted with "FDR" and "FWER" report multiplicity-adjusted p-values following Benjamini et al. (2006) and List et al. (2019).

Table OA.5. Manipulation checks by gender: photographs

	Female Ph. Mean (SD)	Male Ph. Mean (SD)	Diff L-H (SE)	P-values Uncorrected	FDR	FWER
Panel A: Men (N=178)						
Main manipulation checks						
% female applicants	73.21	67.48	-5.73	0.00	0.02	0.03
	(11.90)	(11.37)	(1.75)			
Job desirable for men	57.78	64.77	6.99	0.35	0.61	1.00
	(49.67)	(48.04)	(7.33)			
Job desirable for women	96.67	89.77	-6.89	0.08	0.25	0.81
	(18.05)	(30.47)	(3.74)			
Alternative mechanisms						
% of high-skilled applicants	68.55	72.96	4.41	0.14	0.37	0.95
	(21.38)	(17.61)	(2.94)			
% of high-performers on the job	69.18	68.48	-0.70	0.74	0.97	1.00
	(13.46)	(14.24)	(2.08)			
Own expected performance	65.00	66.14	1.14	0.76	0.97	1.00
	(25.05)	(22.10)	(3.54)			
Job difficulty	63.42	59.08	-4.34	0.18	0.41	0.97
	(21.61)	(20.79)	(3.18)			
Wage level	42.91	43.36	0.45	0.87	0.97	1.00
	(20.49)	(17.19)	(3.13)			
Quality standards	76.86	68.00	-8.86	0.06	0.25	0.72
	(21.17)	(21.14)	(4.69)			
Promotion difficulty	52.87	56.16	3.29	0.25	0.54	0.99
	(19.96)	(17.95)	(2.85)			
Application effort	63.35	58.14	-5.21	0.30	0.54	0.99
	(22.57)	(20.38)	(4.80)			
Discrimination by customers	0.57	0.39	-0.18	0.02	0.12	0.30
	(0.50)	(0.49)	(0.07)			
Job high social status	0.47	0.48	0.01	0.90	0.97	0.99
	(0.50)	(0.50)	(0.08)			
Number of applicants	49.40	48.58	-0.82	0.81	0.97	1.00
	(23.49)	(22.09)	(3.42)			
Panel B: Women (N=325)						
Main manipulation checks						
% female applicants	74.00	68.73	-5.27	0.00	0.00	0.00
	(11.54)	(12.84)	(1.36)			
Job desirable for men	63.75	81.71	17.96	0.00	0.00	0.00
	(48.22)	(38.78)	(4.86)			
Job desirable for women	93.75	80.49	-13.26	0.00	0.00	0.00
	(24.28)	(39.75)	(3.67)			
Alternative mechanisms						
% of high-skilled applicants	76.55	74.18	-2.37	0.26	0.95	0.99
	(17.17)	(20.26)	(2.09)			
% of high-performers on the job	71.15	70.08	-1.07	0.45	1.00	1.00
	(12.20)	(13.87)	(1.45)			
Own expected performance	76.50	78.73	2.23	0.25	0.95	0.99
	(18.67)	(17.33)	(2.00)			
Job difficulty	64.58	65.07	0.49	0.82	1.00	1.00
	(18.27)	(19.22)	(2.08)			
Wage level	48.87	49.15	0.28	0.90	1.00	1.00
	(17.82)	(15.59)	(2.38)			
Quality standards	74.04	79.01	4.98	0.16	0.80	0.96
	(20.17)	(19.24)	(3.53)			
Promotion difficulty	56.16	55.89	-0.27	0.88	1.00	1.00
	(15.87)	(15.71)	(1.76)			
Application effort	57.69	61.60	3.92	0.34	1.00	0.99
	(23.01)	(23.82)	(4.11)			
Discrimination by customers	0.45	0.40	-0.05	0.38	1.00	1.00
	(0.50)	(0.49)	(0.06)			
Job high social status	0.57	0.55	-0.02	0.02	1.00	0.72
	(0.50)	(0.50)	(0.06)			
Number of applicants	58.02	57.51	-0.51	0.82	1.00	1.00
	(19.62)	(19.63)	(2.18)			

Note. For variable definitions, see the notes to Figure OA.2. Variables in the section "Main manipulation checks" are defined in Section 3.1. Some questions were shown to subsamples only, implying differences in the number of respondents. Columns denoted with "FDR" and "FWER" report multiplicity-adjusted p-values following Benjamini et al. (2006) and List et al. (2019).

Figure OA.1. Further manipulation checks: interaction between photographs and information



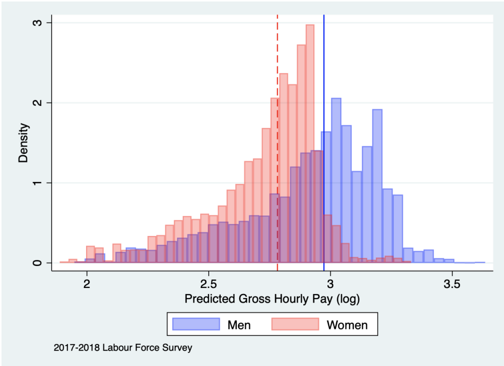
Note. The left panel shows the distribution of answers to the question "Consider 100 people who apply for this job. How many do you think are women?", separately for respondents assigned to the email with a high or low information on returns to talent. The right panel shows the distribution of answers to the question "After seeing the email ad, please indicate below the proportion of [WOMEN/MEN] that you think are successful on-the-job", separately for respondents assigned to the email with a female or male photograph. Data are from the auxiliary online surveys. The number of respondents is 504: 262 are from the Prolific Academic sample and 242 from the organization's sample.

OB. Outside option: methodology

I compute the individual expected hourly wage in the UK as a measure of the individual outside option (w^o), which I use in the logit model described in Section 7.1 of the paper. Using the Labour Force Survey (LFS) quarterly data between January 2017 and December 2018, I estimate a Mincerian regression of the log-hourly wage on a set of observables which are available both in the LFS and my experimental dataset. I then impute the coefficients of the Mincerian regression to my data to predict an individual-level expected wage.

I limit the sample to men and women between 16 and 64 years old and, to match the eligibility criteria of my organization, I include only people who have at least a bachelor degree or, if students, who are currently studying towards a bachelor degree or higher university title. Following the LFS guidance, the variable for the hourly pay has been truncated between 0 and 99 (variable called HOURPAY) and is computed for all respondents who are employees and those on a government scheme. I estimate a Mincerian regression of the log-hourly wage on the following set of dummies: university subject (16 categories), age, age squared, British nationality, gender, marital status, non-white ethnicity, first grade in university. The omitted category are non-married white women who studied Arts. Figure OB.1 shows the distribution of the computed outside option by gender. Table OB.1 compares my experimental sample with a random subsample from the LFS, which I generated to reproduce the same age distribution of candidates in my experiment.

Figure OB.1. Outside option distribution by gender



Note. The figure shows the distribution of outside option for men (in blue) and women (in red). The red dashed (blue solid) line is the women's (men's) median.

Table OB.1. Labour Force Survey and experimental sample comparison

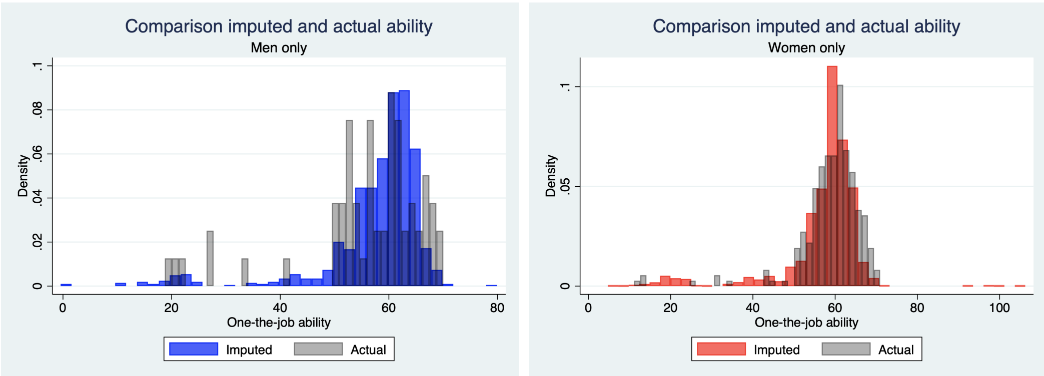
	Labour Force Survey					Experiment	
	Women		Men		Diff (1)-(2)	W	M
	Mean	SD	Mean	SD	p-val	Mean	Mean
Non-white	.12	.33	.14	.34	.07	0.27	0.28
Age	28.77	8.36	29.3	8.73	.01	26.35	28.68
Married	.28	.45	.27	.44	.51	0.12	0.19
First Grade	.15	.35	.14	.35	.3	0.18	0.20
Graduated before 2016	.73	.44	.75	.44	.19	0.34	0.45
FTE in Public Sector	.49	.5	.27	.44	0	0.71	0.60
Scientific Subject	.15	.36	.32	.47	0	0.05	0.09
Aligned Subject	.44	.5	.27	.45	0	0.70	0.48

Note. The first five Columns of the table show summary statistics from a random sample of the LFS which I generated to reproduce the same age distribution of the experimental sample. The last two Columns show the shares of women and men with the given row characteristic in the field experiment. Column "Diff (1)-(2)" contains the difference in the proportions of women and men that have the characteristic of the corresponding row in the LFS sample. "FTE in Public Sector" is an indicator variable for working in the government and includes jobs in healthcare.

OC. Logit model: methodology

Section "Heterogeneity by job-specific talent" in the paper uses a proxy for job-specific ability which is the predicted on-the-job performance score, obtained from the pure control group through a linear truncated regression described in the main body of the paper. The following figure shows the distribution of imputed ability against the distribution of actual test scores in the job from the raw data, by gender. The following table shows the coefficients of the logit estimation.

Figure OC.1. Comparison of imputed and actual on-the-job performance



Note. The figure shows the comparison of imputed and actual on-the job performance distributions. The histograms on the left-hand side are for men and on the right-hand side for women. Ability is on a scale from 0 (min) to 100 (max).

Table OC.1. Logit estimation: output by gender

DV:	Applied and never DO	
	(1)	(2)
	Men	Women
$w - \overline{w}$	0.171	0.736**
	(0.253)	(0.319)
a_i	0.006	-0.011
	(0.013)	(0.011)
Treat θ_H	0.303**	-0.070
	(0.144)	(0.148)
Treat $\theta_H \times a_i$	0.018	0.009
	(0.017)	(0.015)
Treat $p = g$	-0.059	0.215
	(0.146)	(0.147)
Constant	-0.010	0.229*
	(0.133)	(0.134)
Observations	807	3,513

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

OD. Exposure to occupational gender segregation

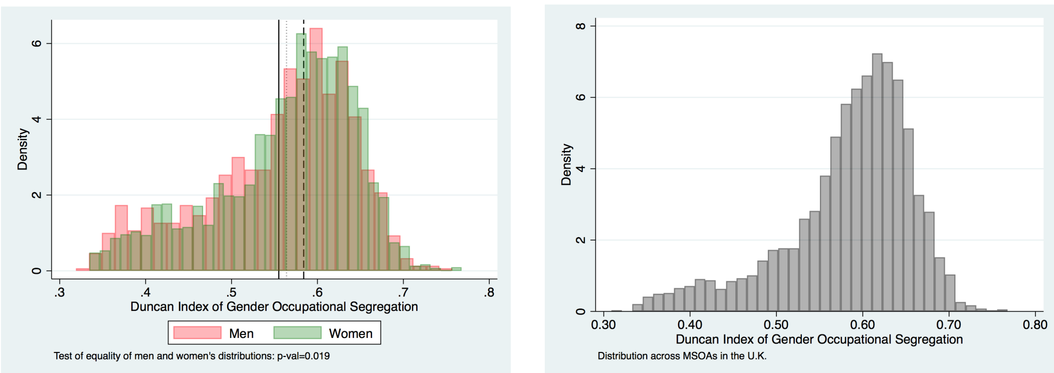
OD.1 Methodology

I use microdata on the local occupational structure by gender from the 2011 U.K. Census (10% random sample) to construct the Duncan index of occupational segregation (Duncan, 1955). The dataset contains the distribution of workers across 362 detailed SOC4 occupational categories at the Medium Layer Super Output Areas (MSOA) level. The data contain 7201 MSOA and the median MSOA comprised 188 8-digits postcodes, with a minimum of 89 postcodes to a maximum of 1033.

The Duncan index is computed using the following formula: $\frac{1}{2}\sum_{i=1}^{N}|\frac{m_i}{M}-\frac{f_i}{F}|$, where m_i and f_i are the male and female population, respectively, in occupation i and M and F are the total working population in the local labor market.

Using a bridge between the Census local area codes and 7-digit postcodes, I merged the indexes with my experimental data through the subjects' secondary school postcode and, when missing (for 62% of subjects), home postcode. The subsample with only home postcode available is made of 50% students and 50% workers. For students, home postcode is mostly the postcode of their parents' home, which is likely to be where they grew up. For workers, it is instead the current domicile. The distribution of the Duncan index in my experimental sample is representative of the overall Country, as shown in Figure OD.1. The U.K. average Duncan Index across MSOAs is 0.5839 and the average in my sample is 0.563.

Figure OD.1. Duncan Index in the experimental sample and in the UK



The figure on the left shows the distribution of the Duncan Index in the experimental sample by gender (postcode level). We can see that men's distribution is shifted to the left of women's distribution (Kolmogorov-Smirnov test of equality of distributions: p-val=0.019). The vertical black line shows the mean for men (0.554) and the vertical dashed line shows the mean for women (0.564). The distribution for the whole U.K is showed in the figure on the right (MSOA level).

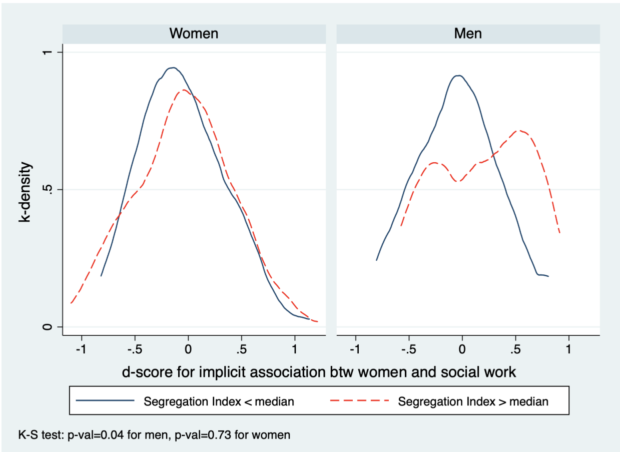
I use the Duncan Index as an individual measure of exposure to gender-segregated labor markets in the previous decade before the current job application. One shortcoming of this method is that it does not equalize the age of exposure to local labor markets across candidates, but timing of exposure has been shown to be a crucial variable for norms internalization (Heckman and Kautz, 2012). Reassuringly, the correlation in my experimental data between the 2001 and 2011 Duncan index is 0.70 (p-val = 0.000).

OD.2 Empirical relationship with beliefs and IAT

Section 7.2 uses exposure to occupational gender segregation as a proxy for uncertainty in men's beliefs in female-dominated jobs. I present two exercises to validate the proxy used.

First, I show that men who come from areas with above-median occupational gender segregation display a higher implicit association between social work and women. In the invitation-to-apply email, all the experimental subjects were invited to participate in a complementary research survey, which included a Single-Target Implicit Association test (Greenwald et al., 1998). Response rate was 12.5% for the main survey and 6% to the IAT (604 and 300 respondents respectively).

Figure OD.2. Implicit Association Test and exposure to gender occupational segregation



Note. The figure shows kernel density estimates of the d-score computed from an Implicit Association Test (IAT) I administered to the job candidates as part of a research survey (12% response rate). Respondents to the IAT count 337 women and 52 men (61% of the survey respondents). The d-score measures the degree of implicit association between female gender and social work: the higher and positive, the greater the implicit association. The d-score is the standardized mean difference score of the "hypothesis-inconsistent" rounds and "hypothesis-consistent" rounds. In the former type of rounds, individuals are instructed to categorize to one side of the screen female names and to the opposite side of the screen male names and social work activities ("hypothesis-inconsistent pairings"). The latter are rounds in which individuals must categorize to one side of the screen female names and social work activities and to the opposite side of the screen male names only ("hypothesis-consistent pairings").

Subjects are presented with two sets of stimuli. The first set of stimuli are typical English female (e.g., Rebecca) and male names (e.g., Josh), and the second set are words related to social work (e.g., family assistance). Subjects are required to categorize the words as quickly as possible for four rounds. In "hypothesis-inconsistent" ("hypothesis-consistent") rounds individuals categorize to one side of the screen female names (male names) and to the opposite side of the screen male names and social work activities (female names and social work). The measure of implicit association between female gender and social work is given by the standardized mean difference score in the two types of rounds. A high and positive d-score indicates a strong association between the two concepts. The order of the two types of blocks was randomized at the individual level.

Figure OD.2 shows the distribution of d-score for women (left panel) and men (right panel), splitting the sample according to exposure to different levels of the Duncan Index. The distribution of d-score values for men exposed to higher-than-median gender segregation is strikingly shifted to the right of the distribution of men from lower-than-median gender segregation (Kolgorov-Smirnov test: p-val=0.043). A similar pattern is observed for women, but the difference is smaller and I cannot reject the null hypothesis of equal distribution between the groups (Kolgorov-Smirnov test: p-val=0.73).

I then show that people exposed to high gender occupational segregation display higher uncertainty in their beliefs on men's and women's performance in female occupations. In the surveys, I asked people the following questions: "On a scale from 0 (min) to 100 (max), what do you think is the performance of a [woman/man] in social work?" and "On a scale from 0 (min) to 100 (max), how confident are you of your answer?". I use answers to the former question as a proxy for the priors on male and female performance in social work and to the latter as a proxy of priors' precision. The proxy for precision is the dependent variable in Table OD.1, computed as an average of the precision levels stated in the two questions (one about a man and one about a woman). The independent variable is an indicator variable for a higher than median Duncan index of the postcode where a respondent was living when she/he was 14 years old.

Table OD.1. Correlation between gender occupational segregation and beliefs

DV: Confidence in beliefs of performance in social work		
	(1)	(2)
Online sample:	M	W
Exposure to high gender segregation	-7.149**	2.641
	(3.319)	(3.927)
Observations	110	116
R-squared	0.268	0.169
Mean Dep Var	74.66	80.18

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Note. The dependent variable is the average of answers to the questions "On a scale from 0 (minimum) to 100 (maximum), how confident are you of your answer [about the performance of a man/woman in social work]?". "Exposure to high gender segregation" is equal to one if the Duncan index of occupational gender segregation in the postcode where a respondent was living when she/he was 14 years old is above the median of the sample. The regression controls for ethnicity, survey wave and the average of the answers to the questions "On a scale from 0 (minimum) to 100 (maximum), what do you think is the performance of a [woman/man] in the social work?". Data are from the auxiliary online surveys and the sample size is determined by the number of people who answered to the postcode question and whose postcode could be matched with the 2011 Census.

OE. Auxiliary online surveys: procedures

I have collected auxiliary surveys on three different samples between July and December 2018 to address potential confounders in the interpretation of the experimental manipulations. Pre-analysis plans of these data have been submitted to the AEA RCT Registry and are available upon request. The following paragraphs describe each sample.

Survey 1 - Amazon Turk. This survey tests for differences between photographs on a sample of 161 Amazon Mechanical Turk workers. This allows me to understand whether images differed in some important dimensions other than gender, but correlated with it. Respondents were online workers who had not participated in any of the researchers' previous experiments conducted on the same platform and who had been granted the "Master" qualification on the website. The survey was conducted with a pool of workers all around the world. The survey was run in different waves between May and July 2018. A total of 188 answers were collected (on average 47 per photograph) and I excluded answers which were only partial (with less than 95% completed). The final sample is made up of 161 answers, of which 39 were for the white woman, 38 for the white man and 42 for the non white photographs. The survey took an average of 2 minutes and was rewarded with 20 cents.

Survey 2 - 2018 Applicants. At the beginning of November 2018, one year after the field experiment, I collaborated with the partner organization to invite current candidates to participate in my online survey. Invitations were sent to 4500 people over two days. The sample comprises candidates at different stages of the selection process who registered between the beginning of September and the beginning of November. As incentive for participation I compensated the first 300 respondents with $5, which they could keep for themselves or donate to a UK social work charity of their choice (out of two listed charities). All the participants were also automatically enrolled into a raffle for a $150 Amazon voucher. A total of 303 people fully completed the survey, which corresponds to a response rate of around 7%. While men's proportion corresponds to the population mean - less than 20% - their number is too small to allow analyses by gender in this sample.

Survey 3 - Prolific Academic. Respondents in this sample are Prolific Academic workers who i) have not participated in any of the researchers' previous surveys conducted on the same platform, ii) are of British nationality, iii) have an approval rate between 75 and 100 percent, iv) are between 18 and 64 years old and v) have at least a bachelor degree. The final sample is made up of 130 women and 131 men, selected through independent survey postings on the website. I collected answers in different waves to match the composition of the field sample on the following observable criteria: gender, ethnicity, student status, university subject, employment status, job sector. Fee was $1.50.

Both surveys 2 and 3 were conducted with a between-subject design. I first showed respondents a photograph and asked two short questions about the portrayed worker (from survey 1). Then participants looked at one intervention email for at least 30 seconds (see intervention table in Figure 3). After mandatory understanding checks, I elicited beliefs on a variety of dimensions about the job and its applicants (e.g., wage, difficulty). Average completion time was 15 minutes.

OF. Theory proofs

Proof. **Existence of threshold of ability a_i^***

Define $\Delta U(a_i) = U^j(a_i) - U^o(a_i) == U^j(a_i, \hat{a}, s_g, \alpha_i, \theta_g) - U^o(a_i, c, v_g, \bar{w})$ and $C(a_i) = \frac{c}{p(a_i)}$ Consider a closed intervals of ability a_i: $[a_1, a_2]$, with a_1 and a_2 bounded away from 0 and infinite. Assume that $\Delta U(a_i)$ and $C(a_i)$ satisfy the following conditions:

a0. They are both continuous in the interval $[a_1, a_2]$

a1. $\Delta U(a_1) < C(a_1)$

a2. $\Delta U(a_2) > C(a_2)$

Define the function $H(a_i) = \Delta U(a_i) - C(a_i)$, which is continuous as well in $[a_1, a_2]$. Then:

$H(a_1) = \Delta U(a_1) - C(a_1) < 0$ from a1

$H(a_1) = \Delta U(a_2) - C(a_2) > 0$ from a2

Since $H(.)$ is continuous, by the Intermediate Value Theorem (IVT) there must be a value $a_i^* \in [a_1, a_2]$ such that $H(a_i^*) = 0$. Thus the two functions $\Delta U(a_i)$ and $C(a_i)$ must intersect in a_i^*. □

Proof. **Result 1**

We need to consider how the change in own gender proportion s_g affects the marginal applicant's ability. Define $G(a_i, \hat{a}, s_g, \alpha_i, \theta_g, c, v_g, \bar{w}, p) = \Delta U(a_i) - C(a_i) = U^j(a_i) - U^o(a_i) - C(a_i)$, where $U^j(a_i)$ and $U^o(a_i)$ are as defined in the previous proof. Consider the vector $\overline{x}_0 = (a_{i0}, \hat{a}_0, \alpha_{i0}, s_{g0}, \theta_{g0}, \bar{w}_0, c_0, v_{g0})$ such that $G(\overline{x}_0) = 0$. Assume that $\frac{\partial G(\overline{x}_0)}{\partial a_i} \neq 0$ and that $p(a_i) = pa_i$, with $p \in [0, 1]$. By the Implicit Function Theorem (IFT):

$$\frac{\partial a_i}{\partial s_g} = -\frac{\frac{\partial G(.)}{\partial s_g}}{\frac{\partial G(.)}{\partial a_i}}$$

From the definition of $G(.)$:

- $\frac{\partial G(.)}{\partial s_g} = \frac{\partial U^j(.)}{\partial s_g} = \alpha_i$. Thus $sign\left(\frac{\partial G(.)}{\partial s_g}\right) = sign(\alpha_i) > 0$ under the assumptions of the model.
- $\frac{\partial G(.)}{\partial a_i} = \frac{\partial \Delta U(.)}{\partial a_i} - \frac{\partial C(.)}{\partial a_i} = (\theta_g - v_g) + \frac{cp}{(pa_i)^2}$. The sign of this difference depends on i) the relative slope of the on-the-job expected utility and the outside option and ii) cost c and the derivative of the probability of being hired with respect to talent.

It follows that $sign\left(\frac{\partial a_i}{\partial s_g}\right) = -sign\left(\frac{\alpha_i}{\theta_g - v_g + \frac{cp}{(pa_i)^2}}\right)$. Consider $(\theta_g - v_g) < 0$, we have two cases:

- If $|\theta_g - v_g| < \frac{cp}{(pa_i)^2}$, then we are at the lower marginal threshold $\underline{a}$ for selection into the job. Here an increase in s_g will decrease the threshold $\underline{a}$, lowering ability at the bottom.
- If $|\theta_g - v_g| > \frac{cp}{(pa_i)^2}$, then we are at the higher marginal threshold $\overline{a}$ for selection into the job. Here an increase in s_g will increase the threshold $\overline{a}$, increasing ability at the bottom.

If instead $(\theta_g - v_g) > 0$, then $sign\left(\frac{\partial a_i}{\partial s_g}\right)$¡0 unambiguously. In any of these cases, there is an increase in the mass of people applying to the job. The magnitude of the change in a^* is increasing in α_i and decreasing in $v_g - \theta_g$.

□

Proof. **Result 2**

We need to consider how the change in expected job impact of talent θ_g affects the marginal applicant's ability. Consider $G(a_i, \hat{a}, s_g, \alpha_i, \theta_g, c, v_g, \bar{w}, p) = \Delta U(a_i) - C(a_i) = U^j(a_i) - U^o(a_i) -$

$C(a_i)$ as defined above. Consider the vector $\overline{x}_0 = (a_{i0}, \hat{a}_0, \alpha_{i0}, s_{g0}, \theta_{g0}, \bar{w}_0, c_0, v_{g0})$ such that $G(\overline{x}_0) = 0$. Assume that $\frac{\partial G(\overline{x}_0)}{\partial a_i} \neq 0$. By the Implicit Function Theorem (IFT): $\frac{\partial a_i}{\partial \theta_g} = -\frac{\frac{\partial G(.)}{\partial \theta_g}}{\frac{\partial G(.)}{\partial a_i}}$.

From the definition of $G(.)$:

- $\frac{\partial G(.)}{\partial \theta_g} = \frac{\partial U^j(.)}{\partial \theta_g} = a_i - \hat{a}$. Thus $sign\left(\frac{\partial G(.)}{\partial s_g}\right)\Big|_{a_i^*} = sign(a_i^* - \hat{a})$.
- $\frac{\partial G(.)}{\partial a_i} = \frac{\partial U^j(.)}{\partial a_i} - \frac{\partial U^o(.)}{\partial a_i} - \frac{\partial C(.)}{\partial a_i} = (\theta_g - v_g) - C'(a_i)$. The sign of this difference depends on the relative slope of the on-the-job expected utility and the outside option, as well as of the cost function.

It follows that there are four possible cases for $sign\left(\frac{\partial a_i}{\partial \theta_g}\right)$, given by the combination of one level of a_i^* - above or below $\hat{a}$ - and the relationship between on-the-job and outside option returns to talent with respect to the slope of the cost function. The Table below summarises the possible combinations. A positive sign of the derivative of a_i with respect to θ_g means that we expect an increase in the quality of the marginal applicant when on-the-job marginal returns to talent increase. For instance, the case of $a_i^* - \hat{a} < 0$ and $(\theta_g - v_g) > C'(a_i)$ corresponds to what happens at the marginal threshold $\underline{a}$ in Result (2) and in Figure 6. Here an increase in θ_g increases marginal ability.

	$\theta_g - v_g > C'(a_i)$		$\theta_g - v_g < C'(a_i)$	
	$a_i^* > \hat{a}$	$a_i^* < \hat{a}$	$a_i^* > \hat{a}$	$a_i^* < \hat{a}$
$\frac{\partial a_i}{\partial \theta_g}$	$-$	$+$	$+$	$-$

□

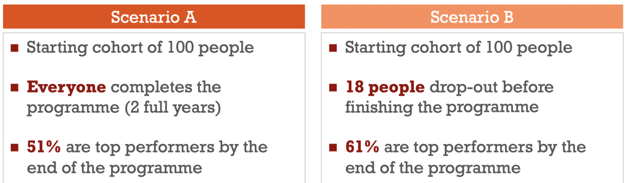
OG. The trade-off between retention and performance: a survey

In October 2020, I sent a survey to the management team of the partner organization to ask about their opinion regarding the trade-off between performance and retention in the program (N=31). Half of the respondents are supervisors of social workers in local communities and half belong to recruitment functions. 55% of the respondents have been working in the organization for less than two years and 45% of them have managerial responsibilities. I used two vignette-type questions (see Figure OG.1). For both vignettes, one scenario is such that all the people hired stay for the full two-year program, but average performance in practice tests is moderate. In the alternative scenario, 18% of people exit before finishing the program, but there is a higher average performance net of leavers. I ask respondents to choose their favourite scenario i) with just this basic information (question 1) and ii) adding information about gender diversity in the two scenarios (question 2). I also elicited the performance level that would make them indifferent between the two scenarios after their choice. The statistics used in the two scenarios come from data on women from treatment groups “Woman Photo + High Past Effectiveness” and “Male Photo + Low Past Effectiveness”.

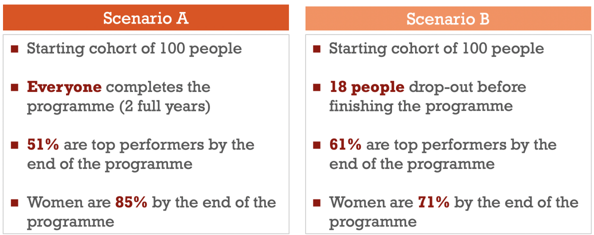
Consider the following two scenarios. Consider them identical besides the characteristics given below. Which of these scenarios would you prefer for the local communities where the organization operates? In both, consider as top performer somebody with commendable or excellent average scores ($\geq 60\%$) in the practice tests of the program.

Figure OG.1. Survey questions

(a) Vignette question 1



(b) Vignette question 2



Note. The figure shows the two main questions asked in the survey with the organization about the trade-off between retention and performance.

UniCredit Foundation

Piazza Gae Aulenti, 3

UniCredit Tower A

20154 Milan

Italy

Giannantonio De Roni – *Secretary General*

e-mail: giannantonio.deroni@unicredit.eu

Annalisa Aleati - *Scientific Director*

e-mail: annalisa.aleati@unicredit.eu

Info at:

www.unicreditfoundation.org

UniCredit Foundation