GRAPE Working Paper #94

Affirmative action and gender-neutral hiring of junior scholars

Magdalena Smyk, Joanna Tyrowicz, Lucas van der Velde

FAME | GRAPE, 2024

Foundation of Admirers and Mavens of Economics
Group for Research in Applied Economics

| GRAPE Working Paper | #94

Affirmative action and gender-neutral hiring of junior scholars

Magdalena Smyk
FAME|GRAPE, and
Warsaw School of
Economics

Joanna Tyrowicz
University of Warsaw,
University of Regensburg and
IZA

Lucas van der Velde
FAME|GRAPE, and
Warsaw School of
Economics

Abstract

We examine the role for external experts in providing unbiased evaluations of candidates in the contexts of affirmative action. Affirmative action policies can promote the advancement of minority candidates, but the empirical results have been inconclusive. We conduct a field experiment with Polish academics, asking them to assess the quality of job candidates and decide which candidates should be invited for interviews. We implement two treatments: a binding and a non-binding equal opportunity clause. Additionally, we vary the gender composition of the candidates being evaluated. Our findings show no evidence of bias against women, either in quality assessments or in subsequent interview invitations. Under the binding equal opportunity clause, external evaluators tend to favor women, suggesting alignment between external experts and institutional objectives.

Keywords:

affirmative action, gender discrimination, academia, experiment

JEL Classification

J71, J16, C93

Corresponding author

Lucas van der Velde,, l.vandervelde@grape.org.pl

Acknowledgements

The experiment was approved by the GRAPE Ethics Board on 03.04.2024 (decision #5/2024). The experiment was pre-registered in the Open Science Framework Registry on 03.04.2024 (https://doi.org/10.17605/OSF.IO/UQW9C). The authors express gratitude for insightful comments and suggestions from Adrian Chadi, Klarita Gerxhani, Julia McQuillan, Nuria Rodriguez-Planas, Shelly Lundberg, Andrea Weber, and participants of EEA 2020, GGAPS 2023, Gender Workshop in Prague 2019 as well as seminars at GRAPE, WSE and University of Warsaw. Aleksandra Makarska provided wonderful research assistance. This paper was written within the scope of EARHART project funded through Norwegian Financial Mechanism 2014-2021, project number # 2019/34/H/HS4/00481, whose support is gratefully acknowledged. Bram Timmermans, our project partner from NHH provided valuable insights at various stages of its development.

Published by: FAME | GRAPE
ISSN: 2544-2473

Foundation of Admirers and Mavens of Economics
ul. Koszykowa 59/7
00-660 Warszaw
Poland

W	grape.org.pl
E	grape@grape.org.pl
TT	GRAPE_ORG
FB	GRAPE.ORG
PH	+48 799 012 202

1 Introduction

Despite the commitment of many institutions, the academic sector remains biased against women.[1] Across disciplines and institutions, gender inequality prevails in hiring (Rivera 2017, Antecol et al. 2018, Sarsons et al. 2021), publications (Symonds et al. 2006, Van Arensbergen et al. 2012, Card et al. 2020) and access to research funds (Lincoln et al. 2011, Witteman et al. 2019). Within Poland, the country on which we conduct our study, universities have implemented Gender Equality Plans to address persistent gender differences in academia. Among the promoted actions, those focused on recruitment stand out. Some institutions, such as the University of Warsaw (the largest university in the country), will use gender (and other minority traits) to break ties between candidates who are perceived to have similar competences.[2]

Were talent directly observable, the mechanism would automatically promote more gender equality. However, talent is never directly observable, and recruitment packages provide a blurred picture of the underlying ability. Recruitment committees could require inputs from experts to assess the quality of the resume and the fit to the specific position for which they are recruiting. These assessments can be biased, either due to unconscious heuristics held by experts, or due to intentionally biased assessments. These (intentionally) biased reviews could occur if experts perceive affirmative actions as unfair and actively overturn this clause (Crosby et al. 2006). Moreover, one cannot rule out that existing (male) academics have an interest in preventing the entry of women into the field, which occurs if male dominance is associated with other factors, such as the prestige of the said field (see Goldin 2014, for more on gender seggregation in occupations). Recruitment committees are in a position where they need to rely on inputs from potentially biased experts, which reduces the reliability of their assessments. In other words, well-intended affirmative actions might not have an impact, or be even prejudicial for women, due to the pre-existing gender biases among current faculty.

We set up an experiment to test whether external experts provide biased assessments of resumes for an entry position in academia when the hiring institution promotes the use of gender markers to break ties between candidates.[3] We focus on hiring decisions for young faculty for two reasons. First, and as documented in (Lundberg and Stearns 2019), it is an area where the progress of women has stalled (and probably declined) in the 2010s. The lack of female representation at the lower echelons of academia might fuel greater biases in the upper echelons in the years to come. Second, among junior faculty, there is greater scope for subjectivity in recruitment processes. For many young scholars, resumes provide a relatively weak signal of their abilities due to early stage of their careers, and thus typically few achievements to report. Given the imperfect knowledge on candidates' ability, faculty entrusted to make the hiring decisions are bound to take decisions under uncertainty, and face a higher risk of falling into non-neutral gender heuristics (Leslie et al. 2015).

Our experimental design introduces another innovation. We acknowledge that the progress made by women differs across academic disciplines. Even within economics, there are substantial differences across fields of study (Lundberg and Stearns 2019). The perception that some fields are more female friendly might attenuate biases against women (Heilman et al. 2004, Goldin 2014). By contrast, in male-dominated fields the entry bar would be higher for women, as a way to maintain the purity of the discipline. In other words, affirmative action policies might be less effective in places where they are more required. We manipulate the perception of how female/male-dominated a field is by asking participants to review more candidates of a given gender.

The results show that evaluations by external experts did not differ much across gender of candidates. Moreover, and against the initial intuition, experts' hiring recommendations when an affirmative action clause

[1] As well as members of diverse minorities. Within economics, both the American and European Economic Associations have standing committees that monitor and promote the participation of women in Economics. These are the Committee on the Status of Women in the Economics Profession (CSWEP), which dates back to 1972, and the Women in Economics (WinE), which started its activities in 2003.

[2] Universities in Germany are subject to similar rules.

[3] The experiment was pre-registered at the LINK [anonymized for refereeing purposes].

is in place tend to disfavor male candidates. In other words, the evaluators produced recommendations that align well with policy goals. Our study extends previous literature findings to a new context (Williams and Ceci 2015).

The article is organized as follows. Section 2 discusses previous empirical findings. Section 3 provides further details on the experimental design. Section 4 presents the analyzes that were pre-registered. Then, Section 5 presents an additional exploratory analysis that we developed in response to the collected data. Finally, we discuss the implications of our analysis in Section 6.

2 Literature review

Affirmative action policies are those taken with the aim of increasing the employment of underrepresented minorities and groups (Holzer and Neumark 2006, Crosby et al. 2006). According to its critics, these policies involve a form of reverse discrimination, as characteristics usually irrelevant for performing at a job, such as gender or race, are taken into account when deciding hiring or salaries.

The range of possible affirmative action policies is vast, since organizations can take different steps to improve representation of their workforce. In this paper we focus on two affirmative action policies. The first type corresponds to affirmative actions implemented during the search for potential candidates. For example, institutions might pledge to interview at least one candidate from a disfavoured group. More often, institutions would simply encourage participants from disfavoured groups to apply, without listing any specific reasons on why it would be beneficial. These type of policies could also involve efforts to craft job offers that avoid gender connotations. Oppenheimer (1988) bundles these activities into the weakest form of affirmative action, as it simply tries to avoid discrimination.

A second type of affirmative action corresponds to commitments at the hiring stage. At this stage, institutions can pledge (or be obliged) to recruit candidates from disadvantaged group (i.e. there are recruitment quotas) (see Oppenheimer 1988). A somehow weaker version of this type of policies emerges when institutions use ascription to a disadvantaged group to break ties between candidates who present otherwise identical qualifications.

A trade-off usually emerges between conceptions of equality and efficiency, as the policy is often perceived to promote the former at the expense of the second. From a theoretical perspective it is not clear whether implementing these policies improves long-term outcomes of the disadvantaged group. In the model of Coate and Loury (1993), differences between groups arise as a self-fulfilled prophecy. Groups that expect lower returns, for example due to statistical discrimination, invest less in their skills. In this case, affirmative action, particularly quotas for members of the disadvantaged groups, can act either as a catalyst that increases investment among that group or it can lead to a patronizing equilibrium. In the second equilibrium case, having secured certain positions reduces the incentives to compete with members from the advantaged group, and further reduces investment in the disadvantage group. The resulting equilibrium depends on how ambitious the quota is in relation to the size of the disadvantaged group .

As shown by empirical evidence from different domains, the perception that equality comes at the expense of efficiency does not match reality. Holzer and Neumark (1999) shows that establishments that implement affirmative actions recruited more often candidates with a stigmatized background (either based on gender or racial minorities), while the average performance level was at par with non-implementing establishments. Similar results were obtained by Baltrunaite et al. (2014) in their analysis of gender quotas in the election of local politicians. They found that in municipalities adopting the quotas, the increase of women was accompanied by an increase in the average level of education of politicians. In other words, more educated women replaced less qualified men. These findings are by no means unique. Affirmative action policies have

been found to improve equality and efficiency (see also Balafoutas and Sutter 2012, Niederle et al. 2013).
Hypothesis 1 When equal opportunity clauses are binding, female candidates received lower scores than male candidates.

The literature also demonstrated that subjective judgements depend on whether candidates adjust to the perceived gender stereotypes. Women are more penalized when they succeed in positions that are traditionally dominated by men (Heilman et al. 2004, Heilman and Okimoto 2007). Successful women in those fields are considered less likeable than men with an identical record. These penalties are lower when women express adherence to gender stereotypes, for example by being mothers (Heilman and Okimoto 2007) or when transgression was perceived to be unintentional, e.g. when women were demanded to take leadership roles by someone with higher authority (Toneva et al. 2020). These findings suggest that equal opportunity clauses might be less effective there where they are most needed, i.e. when there is greater gender disparity. We propose a third hypotheses:
Hypothesis 2: the extent of gender inequality is greater when woman are minority candidates.

Henningsen et al. (2021) studied if the evaluation of resumes was linked to the university's standards concerning affirmative action in recruitment (the comparison group has a pledge to excellence). They did not find evidence of a preferential treatment against women when an equal opportunity clause was in place. Our study differs in two regards. First, their study focuses on professorship positions, i.e. the upper echelons of the academic hierarchy. At that level, resumes might be more informative about candidates ability, reducing uncertainty and the scope for subjective judgement. Moreover, the labor market for professors might be thinner than among entrants. As shown by Kuhn and Shen (2012), a thin labor market reduces the scope for discrimination.
Hypothesis 3: the extent of gender inequality decreases when candidates present better qualifications.

3 Study design

We design the experiment such that in a fair and unbiased judgement there should be no systematic differences in evaluation of the candidates due to gender. In our experiment, we asked external experts to evaluate whether and to what extent our concluded recruitment processes gave all candidates a fair chance. We invited scholars from Poland, from institutions not related to ours, to review the candidates so that we could compare our actual recruitment decisions with their recommendations. We describe the design in steps.

The task Each expert evaluated two sets of candidates, with three candidates in each set. To make the task manageable for experts, the candidates were presented through short biographical profiles, a so-called descriptive approach (Williams and Ceci 2015). Ours was a deception-free experiment: All the candidates were actual applicants in recruitment processes in our institutions. From roughly 400 candidates over the course of a few recruitment processes, we selected pairs of one man and one woman whose professional achievement could be adequately described in the same words, a biographical profile.

In addition to being truthful, the added advantage of this approach is that the advice of external experts was given on real cases of candidates rather than artificially constructed, and thus possibly nonexistent ones. Once we constructed distinct biographical profiles, each of these actual individuals could be truthfully described as a man and a woman. Consequently, our design yields a clear prediction for a fair and unbiased evaluation of candidates: for a given biographical profile, there should be no differences between genders.

After observing the biographical profiles, subjects are asked two questions about each candidate and one summarizing question for each set. First, they should rate the competences of each candidate on a scale of 1 to 100 using a slider. Second, we ask whether it would be a mistake *not to* invite each of the candidates to an interview. Finally, we asked them to sort the candidates from the most qualified to the least qualified.

Once participants evaluated all candidates, they were presented with a short survey. We asked about their academic background: the year of Ph.D. completion, current academic field, and whether they are tenured. The survey also asks whether the respondents had practical experience in recruiting junior candidates.

Constructing the biographical profiles We construct seven biographical profiles. This number was dictated by the design of our experiment and it will become clear once we describe the contextual factors considered in our experiment. This part of the preparations was entirely qualitative. Thus, the three authors of this paper independently performed this part of the preparations, and the seven final pairs selected were consistently chosen by each of the authors under the veil of ignorance. First, we reread all the applications. Then, we developed criteria for classifying the candidates' applications as excellent quality (E) or high quality (H). Next, we grouped the candidates whose applications displayed similar traits according to these prespecified criteria. In some cases, the applications were richer than necessary to satisfy the criteria, but in no single case were they poorer. Eventually, we matched them in pairs. As a final step, we wrote the biographical profiles based on the contend of the applications. The profiles were written using one candidate in pairs as a starting point and then adjusted to adequately fit both candidates in each pair. The seven biographical profiles are reported in the Appendix B.[4]

Conveying gender to the participants Information about the gender of the candidate was conveyed several times throughout the biographical profiles. Polish is a highly gendered language: it exhibits a gender-specific conjugation of verbs and a declination of nouns and adjectives, as well as pronouns. Furthermore, candidates were given fictitious names, which also unequivocally signalled their gender.[5]

In the experiment, we manipulate one central dimension and two contextual factors. The experimental design included variation between subjects and within subjects. Each expert was randomly assigned to a treatment condition or a control condition. Treatment concerns whether the institution has implemented a strict hiring commitment. The two contextual factors refer to the gender composition of each set and the quality of the profiles. Each expert evaluated two sets of candidates, with three candidates in each set. Via purposefully manipulating the candidates included in each set, we introduce both the within-subject treatment contextualization and the between-subject treatment contextualization.

Treatment conditions The central dimension refers to the type of affirmative action announced by our institution at recruitment. The participants in our experiment were informed before the experiment that our institution implements an equal opportunity policy. There were two specific policies, and they were randomized between participants.[6] Treatment varies between subjects.

In the control treatment, the institution wants to ensure that the institution supports equal treatment and particularly encourages women to participate in the recruitment process.[7] This statement corresponds to soft affirmative action at the recruitment stage, as the statement does not involve any specific commitment on the

[4] We report a translation from Polish. Note that allowing external review of our procedure would necessitate disclosing personal information of candidates in our recruitment to a third party. In the interest of the candidates, we did not ask for external assistance in this task. Balancing between assuring the no-deception design and privacy protection of the candidates, we decided not to ask an external evaluation on our work.

[5] To protect the anonymity of the candidates in our real recruitment processes, we replace real, international names with random Polish names. Participants in our experiment were explicitly informed that the candidates were real, but the names were fictitious.

[6] There was no deception in this description either, as our recruitments were performed at two independent institutions, differing with respect to the EO policy.

[7] Specifically, the instruction read: "Our institution values equality, we encourage especially women to apply" [in Polish: reads "*Instytucja wspiera rownosc i zacheca w szczegolnosci kobiety do udzialu w rekrutacji*"].

side of the recruiting institution. We term this treatment a *no hiring commitment* or NHC. In the experimental treatment, the recruitment institution pledges that in case of equal qualifications, the female candidate will be preferred.[8] This statement involves strict and specific commitment on the side of the recruiting institution. We refer to this treatment as *hiring commitment* or HC.

Contextual factors In this study we shed light on two important contexts of fair and objective hiring. The first contextual factor refers to the gender composition of the candidates in each set. Intuitively, positions men are considered to perform better in male dominated tasks/fields, which would give an edge to male candidates (see meta-analyses by Davison and Burke 2000, Koch et al. 2015). Moreover, if women *pollute* occupations or disciplines, then incumbents will exert greater effort to prevent the entry of women (Goldin 2014). The second contextual factor refers to the quality of the candidates. In positions requiring more specific skills, competence is fiercer, leaving less room to promote candidates based on ascriptive characteristics (Kuhn and Shen 2012).

For the gender composition, we design the set of biographical profiles to contain three applicants. Thus, in our sets, a given female candidate will be in either a gender minority or a gender majority. Each academic participating in our study evaluated two sets of candidates: one with minority women and one with majority women. Consequently, this contextual factor varies within-subject. We can thus study both within-subject and between-subject responses to the gender composition. However, notice that the interaction between the gender composition of the set and the affirmative action policy enforced by the institution varies only between subjects.

To test the fairness of judgment vis-a-vis objectively weaker candidates, we also purposely construct sets of biographical profiles. We construct either exceptional or of high quality biographical profiles.[9] We distinguish three settings: (i) when the man's and the woman's profiles are both exceptional, and the third profile is high quality; (ii) when the man's and the woman's profiles are high quality, and the remaining one is exceptional; (iii) and when the man's and the woman's profiles are both high quality, and the remaining one is of high quality as well. The quality composition varies across sets and hence presents within-subject variation. However, like in the previous contextual factor, the differences between affirmative action policies enforced by the institution only vary between subjects.

Note that each participant was offered to review six different biographical profiles, and it was our responsibility to ensure the distinction between E profiles and H profiles, as well as to make H profiles sufficiently similar to one another and E profiles similar to one another.

Of the 64 possible combinations of the two contexts, we selected the eight sets that allow us to test our hypotheses. Table 1 presents the allocation of biographical profiles to the two recruitment processes, evaluated by the participant in our experiment. Each participant is randomly assigned one of the four sets for recruitment processes I and II.

Hypotheses Recall that our experimental design is such that under fair and objective evaluation we expect no differences across genders, conditional on biographical profile. We identify the overall effect of experimental treatment from the comparison of each candidate evaluation between the participants assigned to the NHC and the HC conditions. Take, for example, set 1a. Each of the six candidates in this set received a score from a given participant. A different participant also received set 1a, but was randomized into a different experimental condition. As we average over participants, the treatment effect will be measured as a difference

[8] Specifically, the instruction read: "In the case of scores being equal among the top two candidates, we are committed to hiring a woman" [in Polish: "*Jesli kandydaci reprezentuj takie same kwalifikacje zatrudniona zostanie kobieta*"].

[9] The terms *exceptional* and *high* are relative. *Exceptional* candidates are such in comparison to the other candidates in the set.

Table 1: Distribution of biographical profiles across recruitment processes

Recruitment I		Recruitment II	
Set 1	EW(1), EM (2) , HW (4)	Set a	EW (3), HM (5), HM (6)
Set 2	EW(1), EM (2) , HM (4)	Set b	EW (3), HM (5), HW (6)
Set 3	EM(1), EW (2) , HW (4)	Set c	HW (5), HW (6), HM (7)
Set 4	EM(1), EW (2) , HM (4)	Set d	HW (5), HM (6), HM (7)

Notes: The table presents the distribution of profiles across recruitment sets. E and H represent Exceptional and High quality, and W and M signify Women and Men. Numbers in parentheses identify profiles. See Appendix B for detailed profiles.

in average scores for and all other biographical profiles in this set. In other words:

$$H_0 : \quad \forall_{p\in\{1,\ldots,7\}} \quad E(response_p \mid g = M) = E(response_p \mid g = W), \tag{1}$$

where the average $response$ in our experiment denotes the answer of the participant to both the scoring question and the invitation question. We denote gender by $g \in \{M, W\}$ and biographical profiles by p.

We identify the treatment effect on female candidates by comparing the score given to each biographical profile between treatment conditions and gender. This is the key coefficient of interest in our experiment. The null condition of objective an unbiased evaluation regardless of the treatment can be written down as:

$$\begin{aligned} H_0 : \quad \forall_{p\in\{1,\ldots,7\}} \qquad & E(response_p \mid g = M, t = NHC) = E(response_p \mid g = W, t = NHC) = \\ & E(response_p \mid g = M, t = HC) \quad = E(response_p \mid g = W, t = HC), \end{aligned} \tag{2}$$

where $t \in \{NHC, HC\}$ denotes treatment. In other words. Our estimated effect comes from the difference between the evaluation of the biographical profiles for both genders and between respondents who were assigned to different treatment conditions. This hypothesis is similar to a difference in the differences.

Recovering the role of contextual factors involves triple difference. To understand whether the quality of the applications matters, we compare the treatment effect in Recruitment I (always two excellent biographical profiles) to the treatment effect in Recruitment II (always one excellent biographical profile). The natural comparison in Recruitment II is between the third candidate in sets (a) and (b), and the second candidate in sets (c) and (d).

The second contextual factor is the gender composition of the set. This can be estimated from a triple difference of the profiles in the Recruitment I. For example, a comparison of the first candidate of Recruitment I in sets 1 and 2 indicates whether women benefit from being a minority candidate. A comparison of this effect with that obtained for the first candidates in sets 3 and 4 indicates whether women benefit more than men from being a minority candidate. A comparison of these effects across treatment conditions serves to estimate whether the effect of being in a minority group for women is greater when the institution commits to hiring a woman.

Implementation We administered our experiment through an online survey. We contacted all faculty in all registered higher education institutions in Poland.[10] The invitations to participate in the survey were sent out by email. We explained that we are evaluating the fairness of two recruitment processes and we ask the participant to give us their evaluation of the candidates. The participants were truthfully informed that the recruitment processes were closed and the only purpose of the survey was to collect their insights. Participation

[10] We constructed a database with names, institutional email addresses, and the field of research of all Polish-based researchers.

was rewarded with an entry into a lottery, with twenty smart watches as a reward.[11]

The survey was administered anonymously. We created two separate links for men and women among Polish faculty, so that we could control for the gender of the participants while preserving the anonymity of the survey. A priori, the gender of the participant variable is an important moderator, as on average women might be more aware of gender inequality in academia and be more likely to promote other women.

Our database of contacts contained 61 281 academics with valid email addresses. Approximately 70% of these emails reached the mailboxes (that is, they did not bounce due to typo, no longer existing email or out-of-office note). We sent the invitation email on April 9th, and around 450 complete surveys were collected over the period of 8 days. Automated email open monitoring reveals that on average 10,8 percent of women and 19,7 percent of men opened the email. This substantially reduces the sample size from the initial 61,000 to 6 863 potential respondents. A reminder email was scheduled for April 18th.[12] We closed the experiment a week after the reminder email. During this week, 570 additional complete questionnaires were collected.

Sample In total, 1 026 academics participated in the study. We interpret the response rate to be approximately 14,9 percent.[13] This places our paper at a regular spectrum of response rates in this literature.[14] Our final sample is relatively large by the standards of this field: more than a thousand evaluators and more than 6,000 evaluations. Although response rates might adversely affect the external validity of our findings, they do not bias the estimated parameters.

Note that our sample is unique in some respects. First, our invitation email was widely distributed across many disciplines. Some of the faculty contacted us to explain that they do not feel qualified to evaluate candidates in economic sciences. Next, it is not very common in Poland to recruit candidates in international markets. On the one hand, wages are not competitive, so few candidates are interested in applying. On the other hand, there are strong traditions in hiring Ph.D. program graduates in their *alma mater*. Furthermore, asserting fairness and objectivity in hiring has not been an important objective in many institutions. It is fairly recent and fairly rare that academic institutions in Poland develop gender equality plans and equal opportunity policies. Hence, a part of the academic faculty may have found our invitation unusual and thus abstained from participating.

Table C1 reports the composition of our sample. We have fewer women than men among the respondents (this characteristic is common between treatments due to randomization), reflecting the skewed gender proportions in Polish academia.[15] Around two thirds of the participants had been involved in previous recruitment processes. This statistic raises confidence in our results for two reasons. First, it indicates that participants were familiar with the task of evaluating resumes. Second, it increases the external validity of our study, as these are the same individuals who would be contacted to evaluate candidates in the real world. Despite overall successful randomization, respondents in the HC treatment were less likely to have obtained full professorship and more likely to be on the lower rungs of the academic ladder (p-value for an independence test 0.066).

The last panel of Table C1 shows three proxies for the quality of responses. The first is the time to complete the survey. The median time to complete the survey is around seven to eight minutes in both conditions. However, a few outliers implies that under the HC the average time required to complete the survey was

[11] The value of the reward in monetary terms was approximately 120 EUR. The participants had the choice to leave any preferred email address to enter the lottery. Some participants decided to answer our questions for the benefit of the science and did not leave their email address.

[12] Due to the anonymous character of the survey, we cannot tell how many *new* respondents were reached with the reminder email.

[13] The effectively read 6 863 emails refer to the original invitation, we cannot tell how many *new* respondents were reached through the reminder email.

[14] In Gërxhani et al. (2023) response rate was around 20 percent, while in Williams and Ceci (2015) it was close to 35%.

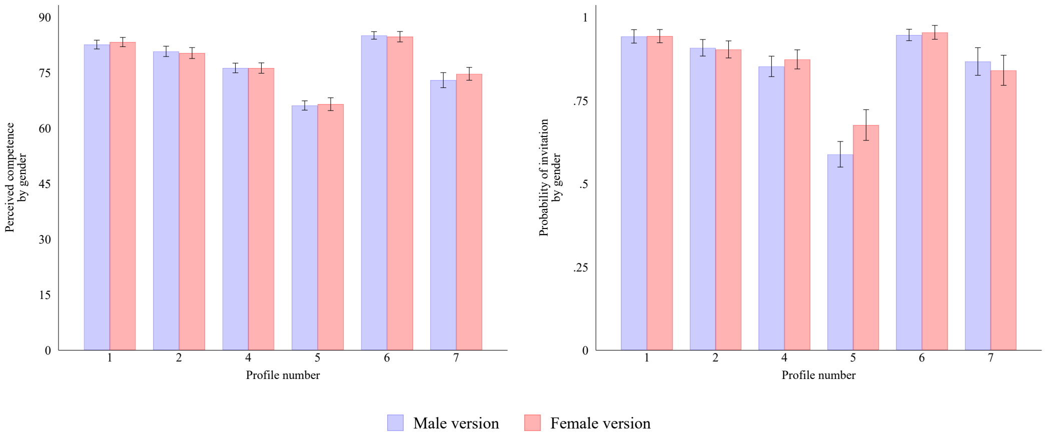
[15] The response rate was slightly higher for men due to a higher open rate in this group.

10 minutes longer. This difference appears to be driven by outliers. The remaining two proxy variables are intended to capture the lining up behavior on the side of respondents. In an effort to fill the survey faster, the respondents may have not reflected on the characteristics of each biographical profile and may have provided the same (or a very similar) evaluation to all candidates. The two measures presented indicate whether the participants suggested that all candidates should be invited for an interview and whether all candidates were assessed to have identical competences.

4 Pre-registered analysis

Although we find differences in the evaluation of biographical profiles, we do not find differences across the gender of the candidate described. We report these results in Figure 1. In the left panel, we present average perceived competences, whereas in the right panel we portray the proportion of participants in our study, who argue that not inviting a given candidate would be a mistake. The bars represent the averages for subgroups (as indicated by colors). We report each biographical profile separately.

Figure 1: Perceived competences and invitation to interview across biographical profiles

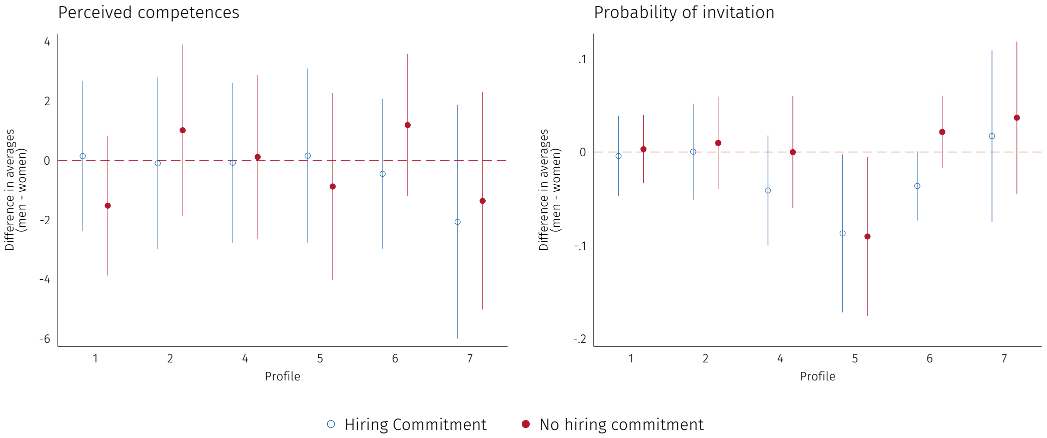


Notes: The figure portrays average perceived competences (left) and the probability of stating that not inviting the candidate would be a mistake (right) for men and women for each profile. Vertical lines represent 95% confidence intervals. We omit biographical profile #3 because it was only used in the female version.

Several observations stand out. First, the evaluation of competence was consistent. The male and female versions of the biographical profiles were evaluated to have similar abilities and are invited to participate in interviews at the same rate. The only exception seems to be on profile five, where women have an edge of around fiver percentage points. Second, there is variation between profiles; that is, the evaluators did not randomly pick values. The first three profiles, excellent, tend to score higher than profiles four, five, and seven, which were classified as high. We also observe that profile six, which was high quality, is evaluated at levels similar (or slightly higher) to excellent profiles. Overall, these results are consistent with the unbiased evaluations by the external reviewers. Next, we study differences across treatment conditions.

The external evaluators provided the same unbiased judgment, regardless of the treatment condition. In Figure 2 we report the estimated differences between the average score for the male variant of the biographical profile and the female variant. We do that for both treatment conditions. The tests are reported separately for each biographical profile. In all cases, the confidence intervals overlap. Finally, we do not observe systematic differences based on the quality of the candidates (first two profiles compared to last four). Only in the case of the biographical profile # 5 five, the probability of invitation differs by gender of the profile. The only profile for which differences appear to be statistically significant is the biographical profile # 6, where a small disadvantage for women under NHC condition is transformed into an advantage in the HC condition.

Figure 2: Gender gap in perceived competence and invitation to interview across biographical profiles



Notes: The figure portrays differences in average perceived competences (left) and the probability of stating that not inviting the candidate would be a mistake (right) for men and women under different treatment conditions. Vertical lines represent 95% confidence intervals from t-tests assuming unequal variances.

These descriptive statistics speak against gender bias among external evaluators. Indeed, women are evaluated no worse than men, and differences between treatment conditions appear to be very minor. Emerging minor differences tend to favor female candidates. In other words, we do not find evidence that referees are biased against women under any of the treatment conditions.

4.1 Regression analysis

Given that the tests presented in Figures 1-2 may have low power, we extend the analysis through regression models. In these models, we aggregate the differences across biographical profiles to tease out the effect of candidate gender and treatment effect, as well as the interaction of these two variables. The model looks as follows:

$$y_{i,p,t} = \beta_0 + \beta_F Female + \beta_T Treatment + \beta_{F,T} Female \times Treatment + \gamma_i + x_i\beta + e_{i,p,t} \qquad (3)$$

where $y_{i,p,t}$ is the evaluation made by participant i, of biographical profile p, in the treatment condition t. We consider two outcome variables: the perceived competence of the candidate, whether participants considered it would be a mistake not to interview the candidate. The parameter of interest is $\beta_{F,T}$, which shows the differential effect of being a female candidate applying to an institution with a hiring commitment in place. The term γ_p identifies biographical profile fixed effects. The inclusion of this term ensures that only variation within profiles is used to identify $\beta_{F,T}$. [16] Finally, $x_i\beta$ identify respondents characteristics (gender, graduation year, previous experience).

In order to test Hypothesis 2, we expand the previous specification by including additional interactions with an indicator for whether the biographical profile is of the underrepresented gender in the current set. The variable equals to one for female (male) profiles in sets composed of two male (female) profiles and one female

[16] We also exclude biographical profile number 3 from the sample, as this profile was only distributed in its female variant.

(male) profile. The regression is of the following form:

$$\begin{aligned} y_{i,p,t} &= \beta_0 + \beta_F Female + \beta_T Treatment + \beta_M Minority \\ &+ \beta_{F,T} Female \times Treatment + \beta_{F,M} Female \times Minority + \beta_{T,M} Treatment \times Minority \\ &+ \beta_{F,T,M} Female \times Treatment \times Minority + \gamma_i + x_i\beta + e_{i,p,t} \end{aligned} \tag{4}$$

The regression includes additional terms and parameters, which are related to hypothesis two. The parameters $\beta_{F,M}$ and $\beta_{F,T,M}$ indicates whether outcome variables are smaller for female biographical profiles, and whether the relationship is different when institution announces a hiring commitment. The hypothesis two states that $\beta_{F,M} < 0$, i.e. profiles of women are perceived as less competent in male dominated positions, and $\beta_{F,T,M} < 0$, i.e. differences in evaluation are more negative when the institution announces a hiring commitment.

Finally, we test Hypothesis 3 by interacting the treatment variables with an indicator on whether the biographical profile corresponded to the excellent- or the high-quality type. The regression is almost identical to the previous one, except for the interaction terms, which now refer to quality of the biographical profile and not to the minority status.

$$\begin{aligned} y_{i,p,t} &= \beta_0 + \beta_F Female + \beta_T Treatment + \beta_Q Quality \\ &+ \beta_{F,T} Female \times Treatment + \beta_{F,Q} Female \times Quality + \beta_{T,Q} Treatment \times Quality \\ &+ \beta_{F,T,Q} Female \times Treatment \times Quality + \gamma_i + x_i\beta + e_{i,p,t} \end{aligned} \tag{5}$$

Table 2 presents the results of estimating specifications 3-5 We print only the estimates related to treatment and its interactions with gender. The full set of coefficients is available in Table C2 in the Appendix. The results are consistent with the descriptive statistics. When looking at gender, $\hat{\beta}_F$, the coefficients are all very close to zero, and not statistically significant. In the first specification, women received 0.154 fewer points in the assessment of competences than men. This result suggests that participants evaluated candidates similarly regardless of their gender. Most of the interactions are not statistically significant either. The largest coefficient corresponds to the three-way interaction between gender, affirmative clause, and minority condition. In this case, the coefficient suggests that female profiles, when affirmative action clause at hiring, in a pool dominated by male candidates, receive 2.5 fewer points, around 3% of the mean number of points.

When it comes to whether it is mistake not to invite a candidate, we observe that gender of the candidate is more salient. When institution announces a hiring commitment (HC), respondents were more likely to considered that not inviting a women was a mistake: the difference between men and women under HC was around three percentage points. This difference results mostly from the underperformance of male profiles in the hiring condition, and not from an increase support for female candidates. These results contrast with those presented in Williams and Ceci (2015), who found a strong preference for women in the United States.

The only statistically significant coefficients in Table 2 correspond to the effect of announcing a hiring commitment on the probability of inviting candidates to an interview. Participants were less likely to state that not inviting a (male) candidate is a mistake when the institution commits to hiring women in the case of ties. Point estimates for the interaction between hiring commitment and the female profile are positive, and we cannot reject the null hypothesis of no treatment effect for women. These results are consistent with previous findings (Williams and Ceci 2015, Henningsen et al. 2021).

Table C2, in the Appendix, presents the full set of coefficients. Some respondent characteristics appear to play a role in scoring biographical profiles. Women evaluated competences somehow higher than men. However, men and women did not differ in the likelihood of inviting candidates for an interview. We also find differences between people at various stages of their academic careers. Tenured professors were the most likely

Table 2: Regression results

	Competence			Invitation		
Specification	(1)	(2)	(3)	(1)	(2)	(3)
Female CV	0.110	-0.473	0.00318	0.00409	-0.00859	0.0103
	(0.471)	(0.866)	(0.743)	(0.0103)	(0.0156)	(0.0171)
HC	-0.550	-0.987	-0.507	-0.0385**	-0.0441**	-0.0427***
	(0.817)	(0.969)	(0.668)	(0.0153)	(0.0183)	(0.0163)
Female CV × HC	0.168	0.977	0.480	0.0253	0.0427*	0.0355
	(0.671)	(1.216)	(1.033)	(0.0156)	(0.0230)	(0.0242)
Minority=1		-0.626			-0.00438	
		(0.856)			(0.0183)	
Female CV × Minority=1		1.885			0.0427	
		(1.675)			(0.0293)	
HC × Minority=1		1.280			0.0161	
		(1.208)			(0.0255)	
Female CV × HC × Minority=1		-2.542			-0.0561	
		(2.345)			(0.0429)	
Female CV × High quality=1			0.251			-0.0166
			(1.200)			(0.0231)
HC × High quality=1			-0.130			0.0126
			(1.143)			(0.0229)
Female CV × HC × High quality=1			-0.758			-0.0273
			(1.697)			(0.0333)
Resume FE	Yes	Yes	No	Yes	Yes	No
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5639	5639	5639	5639	5639	5639
R-squared	0.171	0.171	0.171	0.128	0.129	0.129

Notes Estimates obtained using linear regression. Column title indicate specifications. Standard errors in parentheses. Columns 1 and 2 cluster standard errors at the individual level, while in Column 3 we use heteroscedasticity consistent standard errors. ***, **. * indicate p-values lower than 0.01, 0.05 and 0.1 .

to invite candidates for interviews. Finally, the coefficients for the field of studies indicate the differences in expectations that exist between disciplines. Respondents in STEM disciplines were considerably less likely to invite candidates to participate in the interview.

5 Exploratory analysis

The results from the previous section indicate that Polish academics evaluate competences of male and female candidates equally. When it comes to inviting candidates for interviews, differences are small and tend to be against male candidates when the hiring institution has a preference for a minority candidates. The results are against the initial hypotheses. Under a positive light, these results suggest that institutions can rely on reports prepared by external referees to enforce affirmative action policies.

While a null result is promising, estimates presented in Table 2 might be biased towards zero for a number of reasons rendering our conclusions invalid. In this subsection, we explore four alternative explanations. The first explanation is that respondents were inattentive, which would suggest that the answers were randomly given, and coefficients underestimate the true effects. The second explanation is that respondents failed to perceived differences among candidates, for example in terms of academic achievement. Third, it is possible that effects are heterogeneous across respondents, which might cancel each other out on average. Finally, it is possible that the linear regression models lack power to detect deviations.

5.1 Were respondents attentive?

The initial design of the survey did not include any manipulation checks to avoid making the affirmative action and gender of the candidates too salient. A natural substitute is to check whether answers to the outcome variables are consistent. If a candidate is attentive, we would expect that profiles that possess greater competencies will also be invited more frequently for an invitation. This expectation can be violated in two ways. First, candidates with same perceived competences have different recommendations, which occurs thirty times in our sample (around 1,5% of recruitment sets). Second, it is possible that candidates with lower perceived competences are invited to an interview, while candidates with higher competences are not. This situation occurred five times, i.e. 0.2% of of recruitment sets. Given how unfrequent these discrepancies are, we cannot conduct a meaningful analysis of their prevalence in terms of gender of the profiles or across treatments.

Besides discrepancies, we can proxy the attention paid by the respondents using the time to complete the survey. In this subsection, we exclude respondents whose response time was among the lowest 5 percent (who answered the survey in less than 3 minutes) or among the highest 5% (who answered in over thirty-two minutes). The decision to exclude the fastest respondents reflects the concern that they might have not read the descriptions carefully, whereas in the case of slowest respondents, we were concern about simultaneous engagement in other tasks, which could have distracted them.

The results from this re-estimation are present in Table C3 in the Appendix. Following our expectations, point estimates tend to diverge more from zero than those presented in Table 2. However, differences in the estimated coefficients tend to be smaller than 1 percentage point of the dependent variable. Moreover, confidence intervals overlap, suggesting that differences with the main sample are not statistically significant either.

5.2 Top coding

As stated in our Hypothesis 3, we expect differences in evaluation of competences to be lower when candidates present high qualifications. Based on that hypothesis, we selected resumes from a pool of candidates who applied to a position in Warsaw. While resumes were representative of participants in the job market for economists, they might not be representative of candidates who apply to the institutions for which participants worked. It is possible that respondents did not judge biographical profiles against each other, but in the wider context of all recruitment processes in which they were involved. If that were the case, and assuming that candidates from other recruitment processes presented lower qualifications, we might expect that candidates in our experiment are evaluated with the highest possible mark, and that not inviting any of the candidates would be a mistake.[17]

We address this possibility by estimating our main regression on the sub-sample of participants whose answers differed across biographical profiles. We keep only participants who did not recommend inviting all candidates, and those who had a difference of at least one point between the highest and the lowest rated candidate. These two restrictions reduced the number of observations, as around 60% of participants indicated that all candidates should be invited for an interview. The second restriction was less binding. Only two participants provided the same rating to all candidates, while at the same time suggesting that only some should be invited for an interview.

The results are presented in Table C4. When it comes to perceived competences, biographical profiles of men and women are still evaluated identically: the largest coefficients suggest a difference of around one

[17] One can speculate about other reasons to invite all candidates. If relevant information such as social skills or intrinsic motivation cannot be gathered from the biographical, then not inviting any candidate, regardless of the perceived competences, could be seen as a mistake.

percentage point. However, for the second outcome differences across treatments become more apparent. The pattern that we observed in the previous section comes to the forefront. Men are less likely to be invited for an interview when the institution has a hiring clause. The effect of the hiring clauses among women is null or slightly positive depending on the specification.

The most important difference between Tables 2 and C4 corresponds to the second specification. The coefficient $\hat{\beta}_{H,M}$ is positive and large. This result indicates that the penalty against men that we observed when the institution has an affirmative clause corresponds only to situations where there were more male than female candidates. The coefficient $\hat{\beta}_{F,M}$ is also positive, women also benefit from a boost in the probability of being invited when they are the minority candidates. Finally, the three-way interaction between the affirmative action clause, female profile, and being a minority candidate is negative and large (the probability of inviting in the reduced sample was 0.7) . It represents a decline of 9 percentage points in the probability of being called for an interview when compared to minority women applying to an institution without a hiring clause $(0,122 - 0,210 = -0,088$, with a p-value of $0,1056)$.

Taken together, the findings provide some tentative support to Hypothesis 2, the implementation of hiring clauses can be detrimental to women when they are the minority candidate.

5.3 Possible sources of heterogeneity treatment effects

An alternative explanation for the lack of average treatment effects is effect heterogeneity across groups. In this subsection we refer to two possible sources of heterogeneity: gender and field of studies. If participants expressed preferences for candidates of the same gender, then, on average these preferences might cancel each other out. The second dimensions refers to how male or female dominated a given field is. We can expect two phenomena. First, if a field is male dominated, then the affirmative action clause will tend to favor women, and might be resisted by men. The opposite might hold in female dominated fields. Second, following Goldin (2014) (male) insiders might have a preference to prevent the entry of women, which is more acute when the field is male dominated.

The results of estimating these heterogeneous treatment effects are presented in Tables C5 and C6. For the sake of brevity, we only present the results from estimating Model 3. Table C5 shows that some differences exist between men and women, which are particularly visible in the case of point estimates for perceived competence. Estimates suggest some degree of homophily: men appear to perceive profiles of male candidates as more competent, while women perceive profiles of female candidates as more competent. However, neither for men nor for women are these differences statistically different from zero. For invitations, the level and interaction effects have the same sign for men and women. When the institution announces a hiring constraint, the probability of contact for male candidates fall, while that of women remains unchanged.

The estimation by fields of study produces patterns that stand against our hypotheses. Female candidates received higher scores in Natural Sciences (which comprise STEM disciplines), and lower scores in Social Sciences (Sociology, Economics, etc.). Experts from male dominated fields do not undervalue biographical profiles of women. In Panel 2, we present the estimates for whether not inviting candidates is a mistake. The announcement of a hiring constraint by the institution has a stronger effect on profiles of men in male dominated disciplines (Engineering and agricultural sciences) and is less pronounced in some female dominated disciplines, e.g. Social sciences. However, the number of observations varies greatly across disciplines, and these comparisons should be taken with a grain of salt.

The regressions show that there is indeed some degree of heterogeneity, which is more pronounced in the case of disciplines than across gender. Point estimates remain on the low side, indicating in most cases on or two percentage point difference between men and women. When differences emerge, they suggest a less favourable evaluation of profiles of men in the hiring condition, when compared to lack of treatment.

5.4 Alternative estimation procedures

The pre-registered analysis concluded that perceived competences of male and female biographical profiles do not differ across treatments. While our preferred conclusion is that participants were not biased, and evaluated similarly male and female candidates, this is not the only possibility. Here, we explore several alternative estimation procedures to deal with other potential sources of lack of statistical (and economical) significance. We estimate different versions of Equation 3. We restrict the analysis both to keep exposition short, and because we would not expect complex interaction to play a role if the main effects are not there. The estimated coefficients are presented in C7.

The main specification does not include individual fixed effects, as the institutional commitment varies only across subjects. This introduces additional noise to our regressions, since respondents can have different views on how successful candidates should look like. While estimates from these regressions remain unbiased, the standard errors can be too conservative, which could explain the null results. Hence we expand the initial regression by including participant fixed effects. Since this approach drops all variation between subjects, it effectively prevents the estimation of the effect of the affirmative action. However, we are still able to test the interaction with gender. Column 1 of Table C7 contains the results of this estimation for perceived competences. The point estimates remain close to zero, and not statistically significant. Confidence intervals suggest that differences are up to 1.5 points in favor of women. Given an average of almost 78, the effect of being a woman and the affirmative action appears negligible.

An alternative to individual fixed effects consists on focusing on the comparisons that are more relevant for the study. As we stated, in the first recruitment the third profile acted as a signal of whether the position was more or less female oriented. We will now exclude this candidate, and focus on comparisons between Anna/Adam and Barbara/Bartosz. The dependent variable is the difference in scores between these resumes, and the independent variables indicate the gender of the first profile (which by construction determines the gender of the second profile), the presence of a hiring commitment, and the interaction between the previous two. These estimates, shown in Column 2 of Table C7 are less precise, as can be observed from the standard errors. As a result, one cannot rule out differences of up to 3 points between men and women under each treatment condition.

Besides differences between subjects, the estimate can be not statistically significant in the presence of ceiling effects. Perceived competences could only be rated between 1 and 100. This creates artificial censoring, as some participants might want to rate candidates above this threshold.[18] When this happens, the dependent variable is measured with an error and the estimated coefficients are biased towards zero. We address this concern by fitting a Tobit model to the data. The point estimates from a Tobit model are presented in Table column 3, and are not statistically different from zero.

5.5 Are differences really negligible?

In the previous subsections, we discussed the statistical significance of the coefficients and considered the possibility of them being downward biased. In this subsection, we address a related question: are the effects economically meaningful? Table 2 shows that women receive a boost of 0.168 in the evaluation of competences when the institution announces a hiring commitment. In this regression, we could not reject the null hypothesis that the true coefficient is zero. However, we also could not reject the null hypotheses that the coefficient

[18] One can imagine a participant who believes that the first profile is rated at 50, the second profile should receive twice as many points, and the third resume thrice as many points. In this case, the actual perceived competence of the third profile is censored at 100.

is 1 or -1.[19] Assuming that the parameter equals 1, does this value represent a *strong* advantage in favor of women?

To answer this question, we study how differences between candidates are distributed under a variety of assumptions. These distributions are presented in Table 3. In the first column, we consider the differences between the top two candidates in each recruitment process. As differences are obtained within evaluator, and among two profiles considered similar, the distribution corresponds to a lower bound of what can be expected in real scenarios. The second column presents the distribution of differences in average scores between two randomly selected profiles from each recruitment process. To avoid negative numbers, we compute the absolute value of the difference. Finally, the last column presents a distribution of differences between two randomly selected evaluations. This is an upper bound, as these differences come from different evaluators who evaluated randomly selected profiles.[20]

Table 3: Distribution of differences

	Top 2 candidates	2 candidates	Random
0	0.18	0.12	0.04
1	0.05	0.04	0.03
2	0.04	0.03	0.03
3	0.04	0.04	0.03
4	0.04	0.02	0.03
5	0.15	0.11	0.07
6-10	0.21	0.20	0.18
11-20	0.20	0.23	0.25
More than 20	0.10	0.21	0.35
Mean	8.75	13.03	18.06
Median	6.00	10.00	15.00

Notes Table presents possible distributions of differences across candidates. First columns includes differences between the top two candidates in each recruitment process for each evaluator. Column 2 presents differences between two randomly selected candidates in each recruitment process for each evaluator. Column three presents differences between two randomly selected evaluations for different resumes.

We see zero figures as a prominent value in that between 14 and 18 percent of differences within the same external reviewer take this value. When we compare evaluations for different candidates from different evaluators, just four percent are identical. This is the proportion of cases where bias evaluations can give an advantage to a given candidate. If we consider an advantage of one point based on gender alone, this bias will be sufficient to close the gap in around 5 percent of differences (second row).

Table 3 also presents the average and median differences. As expected, these values increase as one moves from left (more similar candidates and evaluations) to the right (candidates being more dissimilar). In the latter case, the average and median differences are twice as large as in the former. The impact of an additional point is null.

In addition to an analysis of the raw differences in perceived competences, we can also study how higher values of perceived competences translate onto whether a candidate should be invited for an interview. For this, we reestimate Equation 3 but including perceived competences as an additional covariate. We present this estimates in Table C8.

The coefficient on perceived competence indicates that increases of this variable by one point raises the probability of being invited for an interview by 0,86 percentage points (95% CI = $\{0.0079, 0{,}0093\}$)[21]. To grasp the magnitude of this effect, it suffices to remember that the probability of being invited in the sample is 85 percentage points, i.e. the effect is around 1% of the mean.

Table C8 also includes separate regressions for subsamples of male and female profiles. The resulting

[19] In fact, we would not be able to reject the null hypothesis for any value in the confidence interval.

[20] We restrict comparisons to cases when the resumes are different.

[21] We estimate this coefficients using a linear probability model, which contains clustered standard errors at the level of recruitment and candidate.

coefficients are virtually identical. This result suggests that there is no gender heterogeneity when mapping perceived competences to probability of invitation. Having similar coefficients further reinforces the result that evaluators are not gender biased.

6 Discussion and conclusions

Higher education institutions have begun discussing measures that promote gender equality in academia. Prominent among them are tools aimed at levelling the field in recruitment, be it through the introduction of gender quotas, or by providing a favourable treatment of minority candidates when competences are equal. Unlike quotas, the favourable treatment requires an objective assessment of candidates' qualifications, which is impossible to contract, i.e. one cannot design a template of desirable attributes for all possible positions to be created in the future. Instead, recruitment committees could turn to specialists to evaluate the content of the resumes and provide insights on which candidate provides a better fit for a given position. For this system to work, experts should be free from biases: gender (minority status more generally) should not be an important factor when making their decisions. If expert report reproduces prevailing gender stereotypes, the introduction of an affirmative action could actually hamper women, as it might legitimize discriminatory practices.

In this study, we design a correspondence experiment to investigate whether expert recommendations are biased against women. We consider three different hypothesis: 1) that differences in scores between men and women are larger when the institution expresses a preference for hiring women; 2) that differences are smaller for more qualified candidates; and 3) that differences are larger when women are perceived to be in the minority. We only find tentative support for hypothesis (2). In the remaining cases, reports were unbiased, and if anything, they tend to be more favourable towards women.

Even though proving the null is ultimately impossible, we do provide additional evidence that the lack of a relationship might be attributed to the lack of an effect, and not to deficiencies in the survey design. Our additional regressions in different subsamples provided similarly small point estimates. Moreover, additional methods that improve precision do not produce significantly higher estimates. We conclude that either expert reports are unbiased, or that the bias is unlikely to make a difference in the final evaluation.

While the use of expert inputs seems a promising way to parse relevant information from the resumes, the generalizability of our results beyond the experimental setting depends on a number of factors. First, the survey was ran among external, as opposed to internal, experts. BY being external, these experts have less "skin in the game," which reduces the motivation to produce biased reports. We can expect that if a referee has a strong preference for working with someone of the same gender, then the reports of (potential) co-workers will be less reliable. Second, one can ponder about the sample composition. Given the framing of the experiment, individuals who took part in the survey might have been those who are more receptive to gender equality goals. On the bright side, these are the same individuals who would participate in similar "voluntary" evaluations. Yet, if such an evaluation policy is institutionalized, and experts are "required" to write reports, then our results might not apply.

One final caveat concerns how information was presented. In our design, resumes were presented using stories, which indeed represented candidates of different genders. These stories effectively ironed out some systematic differences between male and female candidates, such as the content of recommendations letters. In actual recruitment processes, this additional information could be available, leading to reports that would favor candidates of a given gender (i.e. experts evaluation reproduce the biased heuristics implicit in recommendation letters). A further research question is how to transmit the information to external experts to minimize the impact of differences in source material (see also Amer et al. 2024, for an analysis of information transmission and its impact on assessment bias by gender).

Our research provides ground from some cautious optimism concerning the role of affirmative action policies in recruitment. Differences in evaluation of candidates were marginal, and do not seem to backlash against minority candidates, in this case, women.

References

Amer, A., Craig, A. and Van Effenterre, C.: 2024, Decoding gender bias: The role of personal interaction.

Antecol, H., Bedard, K. and Stearns, J.: 2018, Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies?, *American Economic Review* **108**(9), 2420–41.

Balafoutas, L. and Sutter, M.: 2012, Affirmative action policies promote women and do not harm efficiency in the laboratory, *Science* **335**(6068), 579–582.

Baltrunaite, A., Bello, P., Casarico, A. and Profeta, P.: 2014, Gender quotas and the quality of politicians, *Journal of Public Economics* **118**, 62–74.

Card, D., DellaVigna, S., Funk, P. and Iriberri, N.: 2020, Are referees and editors in economics gender neutral?, *Quarterly Journal of Economics* **135**(1), 269–327.

Coate, S. and Loury, G. C.: 1993, Will affirmative-action policies eliminate negative stereotypes?, *The American Economic Review* **83**(5), 1220–1240.

Crosby, F. J., Iyer, A. and Sincharoen, S.: 2006, Understanding affirmative action, *Annual Review of Psychology* **57**(1), 585–611.

Davison, H. K. and Burke, M. J.: 2000, Sex discrimination in simulated employment contexts: A meta-analytic investigation, *Journal of Vocational Behavior* **56**(2), 225–248.

Goldin, C.: 2014, *Human Capital in History: The American Record*, University of Chicago Press, Chicago, chapter 9. A Pollution Theory of Discrimination: Male and Female DiVerences in Occupations and Earnings, pp. 313–354.

Gërxhani, K., Kulic, N., Ruscon, A. and Solga, H.: 2023, Gender bias in evaluating assistant professorship applicants?evidence from harmonized survey experiments in germany and italy, *Manuscript* .

Heilman, M. E. and Okimoto, T. G.: 2007, Why are women penalized for success at male tasks?: The implied communality deficit., *Journal of Applied Psychology* **92**(1), 81–92.

Heilman, M. E., Wallen, A. S., Fuchs, D. and Tamkins, M. M.: 2004, Penalties for success: Reactions to women who succeed at male gender-typed tasks., *Journal of Applied Psychology* **89**(3), 416–427.

Henningsen, L., Horvath, L. K. and Jonas, K.: 2021, Affirmative action policies in academic job advertisements: Do they facilitate or hinder gender discrimination in hiring processes for professorships?, *Sex Roles* **86**(12), 34–48.

Holzer, H. J. and Neumark, D.: 2006, Affirmative action: What do we know?, *Journal of Policy Analysis and Management* **25**(2), 463–490.

Holzer, H. and Neumark, D.: 1999, Are affirmative action hires less qualified? evidence from employeremployee data on new hires, *Journal of Labor Economics* **17**(3), 534–569.

Koch, A. J., DMello, S. D. and Sackett, P. R.: 2015, A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making., *Journal of Applied Psychology* **100**(1), 128–161.

Kuhn, P. and Shen, K.: 2012, Gender discrimination in job ads: Evidence from China, *The Quarterly Journal of Economics* **128**(1), 287–336.

Leslie, S.-J., Cimpian, A., Meyer, M. and Freeland, E.: 2015, Expectations of brilliance underlie gender distributions across academic disciplines, *Science* **347**(6219), 262–265.

Lincoln, A. E., Pincus, S. H. and Leboy, P. S.: 2011, Scholars' awards go mainly to men, *Nature* **469**(7331), 472–472.

Lundberg, S. and Stearns, J.: 2019, Women in economics: Stalled progress, *Journal of Economic Perspectives* **33**(1), 3–22.

Niederle, M., Segal, C. and Vesterlund, L.: 2013, How costly is diversity? affirmative action in light of gender differences in competitiveness, *Management Science* **59**(1), 1–16.

Oppenheimer, D. B.: 1988, Distinguishing five models of affirmative action, *Berkeley Women's LJ* **4**, 42.

Rivera, L. A.: 2017, When two bodies are (not) a problem: Gender and relationship status discrimination in academic hiring, *American Sociological Review* **82**(6), 1111–1138.

Sarsons, H., Gerxhani, K., Reuben, E. and Schram, A.: 2021, Gender differences in recognition for group work, *Journal of Political Economy* **129**(1), 000–000.

Symonds, M. R., Gemmell, N. J., Braisher, T. L., Gorringe, K. L. and Elgar, M. A.: 2006, Gender differences in publication output: towards an unbiased metric of research performance, *PloS one* **1**(1), e127.

Toneva, Y., Heilman, M. E. and Pierre, G.: 2020, Choice or circumstance: When are women penalized for their success?, *Journal of Applied Social Psychology* **50**(11), 651–659.

Van Arensbergen, P., Van der Weijden, I. and Van den Besselaar, P.: 2012, Gender differences in scientific productivity: a persisting phenomenon?, *Scientometrics* **93**(3), 857–868.

Williams, W. M. and Ceci, S. J.: 2015, National hiring experiments reveal 2:1 faculty preference for women on stem tenure track, *Proceedings of the National Academy of Sciences* **112**(17), 5360–5365.

Witteman, H. O., Hendricks, M., Straus, S. and Tannenbaum, C.: 2019, Are gender gaps due to evaluations of the applicant or the science? a natural experiment at a national funding agency, *The Lancet* **393**(10171), 531–540.

A Additional experimental materials

B Biographical profiles

Below we list all the different biographical profiles presented in a narrative form. The first three resumes correspond to high quality candidates, while the remaining four correspond to candidates who were in the middle of the ranking. The biographical profiles are presented in female version. For the male version, all words denoting gender[22] were changed to its correct form.

Profile #1: Anna (Adam) is currently a PhD candidate at a top10 US university. Before the PhD program, she has studied in her country of origin. Her research falls at the intersection of public economics and focuses on quantifying the effects of government policies on individuals outcomes and welfare.

She has already published a paper in a top general interest journal and has a portfolio of a job market paper (coauthored) and three (coauthored) working papers.

She received a number of fellowships and awards for work as a graduate student. She has taught tutorials with her supervisor during her PhD studies.

She provided three references, from PhD advisors and coauthors

Profile #2: Barbara (Bartosz) is currently a PhD candidate at a top European university, previously graduating from an MA program from a top national university from another European country.

Her research interest concern political economy and inequality.

In addition to the job market paper, she has two revise & resubmit decisions at top field journals (one coauthored with supervisors and one single-authored) and two more coauthored papers submitted to a journal.

Her work was presented in numerous prestigious general interest and field conferences and workshops. The job market paper has received the Best Paper Award from a professional association in her field.

She has taught tutorials with her supervisor during her PhD studies.

She provided four reference letters. This list includes scholars from his/her alma mater and from visited institutions, including a Noble prize winner and a foreign coauthor.

Profile #3: Natalia is currently a post-doctoral research fellow at top Asian university, having graduated from one of the best Chinese universities a year ago.

Her research interests concern asset pricing, both on a theoretical and empirical side. In addition to a job market paper, she has two revise & resubmit decisions on coauthored papers, both from top field journals, and two more coauthored complete papers. These papers were presented in top field and general interest conferences. In addition, Natalia works in two additional studies (one single authored and one coauthored).

During the PhD program, she was a teaching assistant and after graduation she was invited with guest PhD lectures.

She provided three references, from her past and current Chinese institutions

Profile #4: Justyna (Jan) will graduate on time from a good US university, previously studying in Europe.

Her work concerns monetary economics with particular focus on the link between firm financing and macroeconomic fluctuations. She studies the degree to which the response of firms to economic conditions can be independent drivers of economic fluctuations.

[22] As Polish language includes gender distinctions in pronouns, nouns, adjectives, and verbs.

In addition to a single-authored job market paper, She has developed (coauthored) working papers which are already submitted to journals.

She has spent some time visiting the research departments of central banks. Her research was supported by several fellowships, She also received awards for excellence in teaching .

She provided references from her current academic institution.

Profile #5: Marta (Marek) is currently a PhD candidate at a mid-range US university, previously graduating from an MA program in her country of origin.

Her research interest concerns the effects of public policies on human capital and labor market.

In addition to the job market paper she has developed one more single-authored study. The job market paper was presented in a prestigious general interest conferences.

She has taught tutorials with her supervisor during her PhD studies.

She provided one reference letter, from her advisor.

Profile #6: Paulina (Piotr) is currently a post-doctoral research fellow at good US university. She holds a PhD from a top Spanish university, and has previously graduated from a top university in her home country.

Her work is interdisciplinary. She works on policy-relevant questions using historical evidence to answer important questions about economic and social policy.

Besides her job market paper, She has two other studies submitted to journals and three more papers in progress. Her dedication to academic excellence is evidenced by being granted the Best Paper Award from a professional association in her field.

She has taught tutorials at her alma mater (both quantitative and theoretical).

Paulina provided four references. This list includes scholars from all her academic institutions (MA, PhD, current position) as well as a foreign coauthor.

Profile #7: Katarzyna (Karol) is currently graduating from her PhD program at a top European university. She is a dedicated PhD candidate, graduating on time. Moreover, she has spent a year visiting at a top US university.

Her work combines trade theory with environmental economics to address current challenges of productivity slowdown and the implementation of eco-friendly policies.

In addition to the job market paper, she has one more study in her portfolio.

She has taught tutorials with her supervisor.

She provided three references. This list includes scholars from her current institution.

C Additional tables and figures

Table C1: Randomization: do participants characteristics differ across treatment conditions

	HC		NHC	
	mean	sd	mean	sd
Female respondent	0.3	0.47	0.3	0.47
Year completed PhD studies	2001.040	10.78	2000.384	11.10
Degree				
PhD	0.300	0.46	0.266	0.44
Tenured	0.141	0.35	0.113	0.32
University professor	0.323	0.47	0.306	0.46
Full professor	0.220	0.41	0.290	0.45
No answer	0.015	0.12	0.024	0.15
Field of study				
Humanities	0.141	0.35	0.115	0.32
Social sciences	0.312	0.46	0.270	0.44
Exact sciences	0.141	0.35	0.149	0.36
Life sciences	0.109	0.31	0.109	0.31
Technical siences	0.164	0.37	0.225	0.42
Agricultural sciences	0.036	0.19	0.040	0.20
Medical sciences	0.080	0.27	0.083	0.28
Art	0.015	0.12	0.008	0.09
Experience in recruitment				
Yes	0.683	0.47	0.692	0.46
No	0.264	0.44	0.262	0.44
No answer	0.054	0.23	0.046	0.21
Quality of answers				
Time to complete survey (minutes)	24.085	120.73	15.754	67.99
Invited all candidates	0.549	0.50	0.579	0.49
No differences between candidates	0.033	0.18	0.024	0.15
Observations	523		503	

Notes Characteristics of participants in the sample across affirmative action clauses.

Table C2: Full set of coefficients from Table 2

	Competences			Invitation		
	(1)	(2)	(3)	(1)	(2)	(3)
Female CV	0.110	-0.473	0.00318	0.00409	-0.00859	0.0103
	(0.471)	(0.866)	(0.743)	(0.0103)	(0.0156)	(0.0171)
Female CV × HC	0.168	0.977	0.480	0.0253	0.0427*	0.0355
	(0.671)	(1.216)	(1.033)	(0.0156)	(0.0230)	(0.0242)
Female CV × Minority=1		1.885			0.0427	
		(1.675)			(0.0293)	
Female CV × HC × Minority=1		-2.542			-0.0561	
		(2.345)			(0.0429)	
Female CV × High quality=1			0.251			-0.0166
			(1.200)			(0.0231)
Female CV × HC × High quality=1			-0.758			-0.0273
			(1.697)			(0.0333)
HC	-0.550	-0.987	-0.507	-0.0385**	-0.0441**	-0.0427***
	(0.817)	(0.969)	(0.668)	(0.0153)	(0.0183)	(0.0163)
HC × Minority=1		1.280			0.0161	
		(1.208)			(0.0255)	
HC × High quality=1			-0.130			0.0126
			(1.143)			(0.0229)
Minority=1		-0.626			-0.00438	
		(0.856)			(0.0183)	
Profile fixed effects						

Table C2, continued

	Competences			Invitation		
	(1)	(2)	(3)	(1)	(2)	(3)
2	-2.409***	-2.396***		-0.0370***	-0.0367***	
	(0.469)	(0.470)		(0.00902)	(0.00905)	
4	-6.684***	-6.521***		-0.0798***	-0.0743***	
	(0.410)	(0.461)		(0.00975)	(0.0110)	
5	-16.66***	-16.61***		-0.319***	-0.317***	
	(0.465)	(0.474)		(0.0151)	(0.0153)	
6	2.029***	2.157***		0.00921	0.0135*	
	(0.374)	(0.408)		(0.00659)	(0.00747)	
7	-9.085***	-9.086***		-0.0890***	-0.0891***	
	(0.618)	(0.618)		(0.0148)	(0.0148)	
Respondent characteristics						
Female respondent=1	2.301***	2.275***	2.301***	0.0191	0.0185	0.0192**
	(0.817)	(0.816)	(0.439)	(0.0141)	(0.0141)	(0.00943)
Year completed PhD studies	0.0414	0.0413	0.0415*	0.000518	0.000516	0.000527
	(0.0495)	(0.0496)	(0.0251)	(0.000911)	(0.000912)	(0.000556)
Academic degree						
Tenured	1.088	1.077	1.089	0.0525**	0.0522**	0.0524***
	(1.440)	(1.440)	(0.750)	(0.0216)	(0.0215)	(0.0143)
University professor	-0.532	-0.534	-0.530	-0.00575	-0.00584	-0.00568
	(1.043)	(1.044)	(0.548)	(0.0186)	(0.0186)	(0.0120)
Full professor	-1.128	-1.135	-1.123	-0.0213	-0.0215	-0.0211
	(1.437)	(1.439)	(0.756)	(0.0256)	(0.0256)	(0.0166)
No answer	-11.60**	-11.67**	-11.61***	-0.145*	-0.147*	-0.146***
	(5.640)	(5.653)	(2.494)	(0.0770)	(0.0773)	(0.0429)
Experience in recruitment						
No	0.741	0.771	0.740	0.00954	0.0103	0.00933
	(0.972)	(0.973)	(0.506)	(0.0161)	(0.0162)	(0.0104)
No answer	-1.400	-1.340	-1.394	0.0214	0.0227	0.0215
	(2.224)	(2.226)	(1.071)	(0.0284)	(0.0284)	(0.0186)
Field of studies						
Social sciences	-0.818	-0.873	-0.825	-0.00679	-0.00802	-0.00693
	(1.206)	(1.208)	(0.664)	(0.0205)	(0.0205)	(0.0140)
Exact sciences	-0.550	-0.571	-0.554	-0.0481*	-0.0486*	-0.0480***
	(1.382)	(1.382)	(0.748)	(0.0256)	(0.0256)	(0.0168)
Life sciences	0.183	0.165	0.183	0.0240	0.0236	0.0240
	(1.526)	(1.527)	(0.824)	(0.0223)	(0.0223)	(0.0159)
Technical siences	-1.823	-1.874	-1.827**	-0.0657**	-0.0669***	-0.0657***
	(1.436)	(1.435)	(0.755)	(0.0259)	(0.0259)	(0.0160)
Agricultural sciences	-1.874	-1.908	-1.878	-0.00939	-0.0102	-0.00935
	(2.630)	(2.636)	(1.328)	(0.0352)	(0.0353)	(0.0238)
Medical sciences	-1.021	-1.045	-1.025	-0.0413	-0.0418	-0.0413**
	(1.771)	(1.773)	(0.933)	(0.0304)	(0.0305)	(0.0200)
Art	2.495	2.518	2.504	0.0520	0.0525	0.0529
	(3.521)	(3.507)	(2.031)	(0.0400)	(0.0401)	(0.0338)
Constant	0.837	1.242	-4.537	-0.0651	-0.0617	-0.169
	(99.43)	(99.48)	(50.44)	(1.829)	(1.831)	(1.116)
Observations	5639	5639	5639	5639	5639	5639
R-squared	0.171	0.171	0.171	0.128	0.129	0.129

Table C3: Excluding fast and slow respondents

	Competence			Invitation		
	(1)	(2)	(3)	(1)	(2)	(3)
Female CV	0.104	-0.564	0.00522	0.00256	-0.0105	0.0109
	(0.492)	(0.915)	(0.772)	(0.0107)	(0.0164)	(0.0178)
HC	-0.908	-1.448	-0.950	-0.0419***	-0.0478**	-0.0451***
	(0.844)	(0.993)	(0.694)	(0.0158)	(0.0188)	(0.0170)
Female CV × HC	0.192	1.202	0.411	0.0266*	0.0439*	0.0332
	(0.701)	(1.276)	(1.074)	(0.0161)	(0.0238)	(0.0252)
Minority		-0.833			-0.00641	
		(0.906)			(0.0193)	
Female CV × Minority		2.131			0.0433	
		(1.773)			(0.0312)	
HC × Minority		1.595			0.0173	
		(1.269)			(0.0266)	
Female CV × HC × Minority		-3.164			-0.0556	
		(2.460)			(0.0450)	
Female CV × Excellent			0.258			-0.0224
			(1.256)			(0.0239)
HC × Excellent			0.123			0.00977
			(1.195)			(0.0237)
Female CV × HC × Excellent			-0.568			-0.0180
			(1.771)			(0.0346)
Resume FE	Yes	Yes	No	Yes	Yes	No
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5172	5172	5172	5172	5172	5172
R-squared	0.177	0.178	0.177	0.137	0.137	0.137

Notes Estimates from equation 3 on a subsample that excludes the fastest and slowest 5% of respondents. All estimations include resume fixed effects and respondent characteristics. In Columns 1 and 2, standard errors are clustered at the individual level, in Column 3, heteroskedasticity consistent standard errors are used. In both cases, standard errors are reported in parentheses. ***, **, * indicate p-values lower than 0.1, 0.05 and 0.01.

Table C4: Excluding respondents whose recommendations did not vary

	Competences			Invitation		
	(1)	(2)	(3)	(1)	(2)	(3)
Female CV	-0.154	-0.446	-0.136	-0.000888	-0.0601**	0.00231
	(0.749)	(1.243)	(1.071)	(0.0201)	(0.0289)	(0.0292)
HC	0.278	-0.325	0.676	-0.0533***	-0.0809***	-0.0511**
	(1.071)	(1.318)	(0.931)	(0.0205)	(0.0290)	(0.0257)
Female CV × HC	0.584	1.822	1.291	0.0330	0.122***	0.0457
	(1.069)	(1.697)	(1.423)	(0.0295)	(0.0407)	(0.0401)
Minority		0.427			-0.0217	
		(1.209)			(0.0330)	
Female CV × Minority		0.860			0.140***	
		(2.352)			(0.0523)	
HC × Minority		1.730			0.0773*	
		(1.700)			(0.0436)	
Female CV × HC × Minority		-3.713			-0.210***	
		(3.162)			(0.0743)	
Female CV × Excellent			-0.262			-0.0106
			(1.868)			(0.0419)
HC × Excellent			-1.219			-0.00689
			(1.745)			(0.0403)
Female CV × HC × Excellent			-1.468			-0.0294
			(2.571)			(0.0601)
Resume FE	Yes	Yes	No	Yes	Yes	No
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2316	2316	2316	2316	2439	2316
R-squared	0.296	0.297	0.297	0.388	0.341	0.389

Notes Estimates obtained using linear regression. Column title indicate specifications. Standard errors in parentheses. Columns 1 and 2 cluster standard errors at the individual level, while in Column 3 we use heteroscedasticity consistent standard errors. ***, **. * indicate p-values lower than 0.01, 0.05 and 0.1 .

Table C5: Is gender of respondent a source of heterogeneity

	Competence		Invitation	
	(Men)	(Women)	(Men)	(Women)
Panel 1: all respondents				
Female CV	-0.251	0.861	0.000530	0.00957
	(0.574)	(0.834)	(0.0128)	(0.0177)
HC	-0.509	-0.198	-0.0349*	-0.0388
	(1.005)	(1.376)	(0.0193)	(0.0240)
Female CV × HC	-0.264	1.009	0.0311	0.0148
	(0.811)	(1.175)	(0.0194)	(0.0259)
R-squared	0.172	0.209	0.127	0.144
Panel 2: Restricted sample				
Female CV	-1.400	1.124	-0.0271	0.0101
	(0.885)	(1.293)	(0.0248)	(0.0362)
HC	-0.0766	-0.711	-0.0493*	-0.0573
	(1.393)	(1.849)	(0.0284)	(0.0417)
Female CV × HC	0.352	0.669	0.0639*	0.0281
	(1.316)	(1.998)	(0.0358)	(0.0518)
R-squared	0.253	0.369	0.329	0.380
Observations	1683	756	1683	756

*Notes*Estimates from regression 3 by gender of the respondent. All estimations include resume fixed effects and other respondent characteristics. Standard errors clustered at the individual level in parentheses. ***, **, * indicate p-values lower than 0.1, 0.05 and 0.01.

Table C6: Differences across disciplines

	Humanities	Social sciences	Natural sciences	Engineering and technical	Agricultural sciences	Medical and health sciences
Panel 1: Perceived competence						
Female CV	-1.166	-1.603**	1.749**	-0.727	1.781	3.140**
	(1.467)	(0.813)	(0.854)	(1.033)	(2.477)	(1.461)
HC	-5.913***	1.853	1.245	-3.375	-2.099	-0.0695
	(2.043)	(1.300)	(1.537)	(2.206)	(4.006)	(2.780)
Female CV × HC	1.779	0.915	-1.076	0.988	3.519	-2.716
	(1.965)	(1.076)	(1.304)	(1.676)	(3.143)	(1.985)
R-squared	0.203	0.245	0.197	0.181	0.558	0.212
Panel 2: Invitation						
Female CV	0.0356	-0.00830	0.00961	-0.00812	0.0225	0.0226
	(0.0225)	(0.0216)	(0.0199)	(0.0244)	(0.0608)	(0.0280)
HC	-0.0252	-0.0135	-0.0118	-0.0774*	-0.0714	-0.0956*
	(0.0366)	(0.0266)	(0.0277)	(0.0422)	(0.0595)	(0.0547)
Female CV × HC	-0.0235	0.0254	0.00378	0.0794**	0.107	0.0299
	(0.0390)	(0.0290)	(0.0327)	(0.0349)	(0.0772)	(0.0425)
R-squared	0.176	0.106	0.160	0.126	0.421	0.170
Resume FE	Yes	Yes	Yes	Yes	Yes	Yes
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	801	1636	1427	1094	218	463

Notes Estimates from Equation 3 by disciplines, as defined by OECD. All estimations include resume fixed effects and other respondent characteristics. Standard errors clustered at the individual level in parentheses. ***, **, * indicate p-values lower than 0.1, 0.05 and 0.01.

Table C7: Perceived competences using alternative estimation methods

	FE	FD	Tobit
Female CV	0.448	0.369	0.104
	(0.412)	(1.369)	(0.507)
HC		0.147	-0.625
		(1.427)	(0.886)
Female CV × HC	-0.203	-0.427	0.186
	(0.598)	(1.893)	(0.718)
Resume FE	Yes	No	Yes
Respondent characteristics	No	Yes	Yes
Observations	5639	1026	5639
R-squared	0.709	0.00995	

Notes Columns names indicate estimation procedures. FE stands for inclusion of individual fixed effects, FD stands for first differences between High resumes in set one, and Tobit stands for Tobit model with censoring at values of 0 (8 cases) and 100 (445 cases). In FD column, there is one observation per individual, hence lower N. Standard errors clustered at the individual level in parentheses. ***, **, * indicate p-values lower than 0.1, 0.05 and 0.01.

Table C8: How perceived competences increase probability of invitation

	All	Women	Men
Competences of candidate	0.00862***	0.00866***	0.00861***
	(0.000369)	(0.000480)	(0.000456)
Resume FE	Yes	Yes	Yes
Respondent characteristics	Yes	Yes	Yes
Observations	5639	2570	3069
R-squared	0.266	0.256	0.277

Notes Estimates obtained using linear regression. Column title indicate sample on which regressions were ran. Column (1), *All*, also includes gender of the resume, and an interaction with treatment variable as additional controls. Standard errors clustered at the individual level and recruitment in parentheses. ***, **, * indicate p-values lower than 0.1, 0.05 and 0.01.