GRAPE

Group for Research in APplied Economics

GRAPE Working Paper #94

Gender-neutral hiring of junior scholars

Joanna Tyrowicz, Lucas van der Velde, and Magdalena Smyk

FAME | GRAPE, 2024

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

Gender-neutral hiring of junior scholars

Joanna Tyrowicz
University of Warsaw, University of Regensburg and IZA

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

Magdalena Smyk
FAME|GRAPE, and Warsaw School of Economics

Abstract
We examine the role for external experts in providing unbiased evaluations of candidates in the context of gender. Affirmative action policies can promote the advancement of minority candidates, but the empirical results have been inconclusive on whether they are effective in eliminating bias. Our experimental design was structured to detect bias: we ask senior academics to assess the quality of job applications to junior positions and recommend which candidates should be invited for interviews. We study the role of the strong commitment to ensuring equal opportunity in hiring. Additionally, we vary the gender composition of the applications being evaluated and the quality of candidates. We find no evidence of bias in quality assessments, nor in recommendation to invite candidates for an interview, suggesting alignment between external experts and institutional objectives of unbiased hiring.

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
ull. 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

We study whether external experts can make evaluation of job candidates academia gender-neutral. Helppie-McFall et al. [1] show two important facts: women in academia expect poorer career outcomes than men, and the gender gap in ex ante expectations is remarkably consistent with the actual gender gaps in outcomes. They measure the probability of tenure and publishing in the top journals. However, in order to even enter this path, one has to first land a junior position in academia. Gender inequality prevails in hiring junior scholars [2–4]. Securing female representation in the entry-level echelons of academia is a necessary step to mitigate gender inequality at later stages of the scholarly careers.

Hiring in academia faces a standard challenge, especially for junior positions: talent and ability are not directly observable. The faculty entrusted with making hiring decisions do so under uncertainty that makes them susceptible to the risk of succumbing to gender biases [5, 6]. At an early stage of careers, objectively verifiable achievements of candidates provide an imperfect signal of their potential. The achievements themselves do not need to be directly comparable if women face the greater challenge of providing evidence of potential in the same time frame as men. Research finds that it takes more revisions and more comments for a scientific article to be accepted in a journal [7, 8] and it is more difficult for them to obtain funding for their own research [9, 10]. As documented in Lundberg and Stearns [11], the recruitment of junior faculty is the area where diversity has lagged in the last decade. What distinguishes academia is the broad range of formalized strategies aimed at reducing gender inequality.[1]

Given the widespread institutional commitment to combating gender inequality in academia, effective policy instruments that help achieve unbiased hiring are a burning need. O'Meara et al. [12] review the existing literature and emphasize that a large number of studies identify existing biases, but few studies provide reliable evaluation of instruments aimed at reducing them. Buckles [13] suggests that the training of the members of the selection committee somewhat increases the share of women hired to junior positions, but these studies do not measure bias *per se*. The evidence is compelling that even in controlled environments, when bias against women is made salient to participants, unequal treatment of women continues to emerge [14]. Rivera [2] shows that women are evaluated on non-merit criteria, such as the presumptions about their private life. A way to address such biases could be to use automated algorithms for recruitment based on objective measurement of achievements. However, this approach leads to biases [15, 16], in part because in many aspects of academic work the playing field is not leveled. *Ceteris paribus*: in addition to publications taking longer and the bar being higher [7, 8], there are fewer citations for scientific articles authored by women [17, 18], and women receive less credit if they coauthor with men [3, 17, 19, 20]. Finally, recommendation letters written for women are less likely to emphasize their academic potential as compared to men [4, 21, 22]. Gender diversity in selection committees is yet another tool considered. Bagues et al. [23] and Baron et al. [24] show that it alone is not capable of reducing the gender gap in recruitment and Deschamps [25] shows that it actually harms the chances of women.

We set up an experiment to study the potential of external experts to provide unbiased assessments of candidates for the assistant professor position.[2] Indeed, inviting external experts is a feasible policy instrument. It is also psychologically plausibly effective on the grounds of decision theory [26, 27]. Although external experts are likely to be driven by similar stereotypes as the rest of academia, they have a lower intrinsic motivation to misjudge the quality of a given application for a number of reasons [e.g., they are not going to be collaborating daily with that candidate, see the chapter by 28, for a full account]. Furthermore, when hiring for their own institution, scientists tend to exhibit an implicit gender quota [29]. This effect is likely to be absent when

[1] In Europe, these strategies include Gender Equality Plans, HR Excellence in Research Award, and country-level initiatives. In the US, Diversity, Equality and Inclusiveness initiatives are no less prevalent.

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

assessing candidates as an external expert.

Our experimental design innovates relative to the existing literature in several ways. First, our focus is on detecting bias. Williams and Ceci [30] show that recruitment committees may exhibit small preference to hire a woman over equally qualified men. However, if biases in assessing qualifications persist, this is not sufficient to ensure equal treatment of candidates. We carefully construct the experiment to be able to detect bias and infer meaningful insights on the prevalence of biases. Second, our experiment is deception-free, but we manage to work with actual job market candidates. We thus collect expert assessment on realistic cases. Third, we study the role of explicit commitment to equal opportunity. Comparing two countries that differ in the institutional setting, Gerxhani et al. [31] show indirectly that gender-based preferential selection may be effective in providing equal opportunities to women. In our study, we explicitly test the role of a strong commitment to ensuring equality as opposed to generic statements about equality. We purposely manipulate the strength of commitment to providing equal opportunities to identify if external expert evaluations are driven by policies at hiring institutions. Finally, women are perceived to be less competent in tasks traditionally associated with men [32–34]. This difference could arise because women are considered less capable of performing male-dominated tasks [33], or because incumbents desire to maintain the "purity" of their discipline [35]. To gauge the role of gender congruence on the bias of assessment by external experts, in our experiment, women are evaluated as minority and majority applicants, and all the applications presented are of high quality.

In our field experiment, senior academics were invited to help us assess the fairness of two previously concluded recruitment processes. Participants were informed that our institution is committed to fairness in hiring junior faculty, but we randomized the strength of this commitment between participants. We provided participants with six biographical profiles obtained from actual recruitment processes in our institutions. The biographical profiles describe actual individuals: based on a large number of applications in those concluded recruitment processes, we were able to construct pairs of expressly equivalent applications by a woman and a man. These anonymized biographical profiles provided information on the university that granted the Ph.D., publication record, previous teaching experience, and recommendation letters. We randomize the gender designation of each biographical profile, as well as the gender composition of the set of candidates. The participants evaluated the quality of these applications (on a scale from 1 to 100) and gave qualitative recommendations on whether a given candidate should have been contacted for the interview phase.

Our results show that the evaluations by external experts do not exhibit gender bias. We ask external experts to provide an assessment of the quality of the applications as well as their opinion on whether a given candidate should be invited to an interview. Both of these proxies show no gender bias. In other words, external experts produced recommendations that align well with policy objectives. In the case of strong hiring commitment, we find no bias for women for interview invitation and some penalty for men. These results survive a wide battery of robustness checks.

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 core (pre-registered) results. Then, Section 5 presents the multiple sensitivity checks for our main results as well as an exploratory analysis on the role of extreme scores. Finally, we discuss the implications of our analysis in Section 6.

2 Insights from existing literature and our hypotheses

The existing literature on gender inequality is embedded in the equity-efficiency trade-off [36].[3] Equity implies comprehensively providing the same opportunities to women, compared to men. Efficiency implies allocating resources (such as jobs) to those scholars who can produce the most valuable academic output. In this context, bias is both inequitable and inefficient [37]. Our review of the literature focuses on assessing junior scholars. We organize this section around three ingredients of our experimental design: bias, the strength of commitment to provide unbiased assessment, and the context of assessing the candidates.

2.1 The gender bias

Analyzing the case of economics, Lundberg and Stearns [11] demonstrates that between 1993 and 2017 the share of women in junior academic jobs has barely changed at all: from 20% to 24%. Economics is no exception. In the late 1990s, the pipe was leaky even among biomedical positions in Sweden Wennerås and Wold [38]. Looking across all fields of academia, Ceci and Williams [39] document the disparity between the share of Ph.D. degrees awarded to women and the share of junior positions in academic institutions in the US. Auriol et al. [40] show that the disparity is more acute and starts at earlier stages of academic career in the US than in Europe. Data tracking entry into academic positions across countries does not exist, but Huang et al. [41] leverage the vast Web of Science database and show that since the first published article, women have lower survival rates in academia across countries, disciplines, and time. They show a remarkable resemblance of the yearly performance of men and women who remain active in science (measured by published scientific articles and citations), accompanied by large gaps in the ability of women to continue their academic careers at par with men.[4]

These gaps tend to be associated with biases, although not exclusively: in some areas, some evidence points to unbiased outcomes.

- Presentations. Minondo [45] argues that seminar invitations in the US are similar for men and women, but evidence from higher-ranked departments shows bias against women [46]. Articles authored by women have a lower acceptance probability at conferences [34]. In this case, the differences across genders are driven by male evaluators and concern mostly scholars at early stages of their careers. Similar evidence was provided by Samahita and Devereux [47] for Ireland. It is all the more striking that presenting at conferences raises chances of getting the paper accepted for publication less for women [48]. Notably, some conferences are able to eliminate bias [49].

- Publishing. Working with data from top economics journals, Card et al. [8] find that male and female referees are equally more harsh toward women-authored articles. Holding women to higher standards exhibits by the length of refereeing and increases the quality of published papers authored by women [50].

- Recognition. Women receive less recognition for their contribution when they coauthor with men [3, 19, 20, 51]. Abstracts purportedly written by women were evaluated as lower quality by junior scholars [52]. In blind reviewing, students evaluated coauthored papers lower when they thought they were written by a woman than when they thought men authored them [53]. Awards from prestigious journals disproportionately frequently omit contributions by women [54]. Card et al. [55] argue that promoting

[3] This paper focuses on gender inequality, but the literature studies the context of race, ethnicity, handicap, sexual orientation and other characteristics, which themselves have no impact on research potential, but constitute foundations of negative stereotyping.

[4] We focus on hiring for junior positions. There is rich literature on career advancement: men and women experienced a lower probability of advancing to a tenured position, conditional on application and on qualifications [e.g., 42–44, to mention a few].

women to professional recognition by nominating committees is an effective way to improve their chances to eventually be recognized.

- Citations. Holding quality constant, studies by women receive fewer citations in medicine [56, 57].[5] Nielsen and Börjeson [59] study the case of management sciences and argue that homophily rather than dismissal stands at the core of differences in citations. Implicit biases against women persist in citation practices in political sciences, despite female-authored papers having on average a higher number of citations [60].[6] In economics, the bias against women persists [17, 18, 62]. In a narrow sample of studies published by top economics journals, citations of articles by women are higher than that of men, which is associated with the fact that women writing is of higher quality and papers authored by women are held to a higher standard [8, 50]. Wu [63] shows that dismissive attitudes toward women among academics are prevalent.

- Recognition for junior scholars. Eberhardt et al. [64] study 12 000 reference letters for junior academics participating in the European job market for economists and find that women are described less favorably then men. Exploring the evidence further Baltrunaite et al. [65] show that, in a nutshell, men are described as brilliant, whereas for women the recommendations emphasize hardworking and diligence [see also 21, 22, 66, for similar evidence from other disciplines]. This effect is likely larger among male reference writers.

- Funding. Witteman et al. [10] find minor (though persistent) gender bias against women in evaluating medical research grant proposals in Canada. Bol et al. [67] show that grant applications by female scientists received (marginally) lower scores by external reviewers than those of male scientists; however, when projects are jointly discussed, there is a redistribution of funds toward women-led projects. Farré and Ortega [68] argue that in a prestigious scholarship in Spain, the evaluators have preference for gender balance, hence they promote men in women-dominated fields, and women in male-dominated fields. Guglielmi [69] demonstrates that while the assessment of projects is unbiased, a joint evaluation of applicants and projects are biased. Note that the context of funding applications is particularly informative for the potential validity of relying on external experts in hiring, because predominantly grant proposals are evaluated by peers without any explicit interest in the outcomes of the assessment.

This mixed evidence has ambiguous implications for gender bias in hiring. Ceci et al. [70] argue that in recent years the chances of women to receive an offer of assistant professorship in math-intensive fields are at least as high as those of men in the US. In a similar spirit, Forman-Rabinovici et al. [71] argue that once gender quota are legally mandated on academic boards, the overall presence of women improves. Unfortunately, inference about junior hiring is not possible from this study, whereas in a causally identified setting, Deschamps [25] shows that women stand a lower chance of being offered a job in French academia in committees chaired by men.[7]

A widely cited correspondence study by Williams and Ceci [30] shows that academics state that if applicants have the same qualifications, then they would prefer to hire a woman over a man. This preference varies across disciplines, and it is the lowest in economics. However, even a full realization of this stated preference is not sufficient to eliminate bias in hiring. Buckles [13] makes the point that if the biases discussed earlier persist, then men and women possessing identical *observable* qualifications, actually differ in their *unobserved* skills. In the case of junior positions, this risk is particularly acute as there are fewer observable accomplishments,

[5] This disparity perpetuates to citations outside academia, as demonstrated in the case of Wikipedia [58].

[6] Indeed, Klinowski [61] shows that women are substantially less likely to criticize the work of other scholars.

[7] Similar findings are revealed in a controlled experiment by Leibbrandt and List [72], but in this study the context was not academic hiring.

increasing the reliance on recommendation letters and professional recognition. The use of external experts appear to be particularly suitable, as the evidence from grant assessment relatively more frequently points to unbiased outcomes.

2.2 The role of commitment to equality in reducing bias

Over the years, academic institutions have devised numerous policies aimed at leveling the playing field. Oppenheimer [73] proposes a topology that identifies the degree of commitment in these policies. On the low end, academic institutions may declare being gender-neutral and encourage women to apply. An example of such statement is frequently encountered in job postings: "*[x] is proud to embrace inclusion and cultural diversity. We encourage women to apply*"[8] or "*[x] fosters an inclusive culture that values diverse backgrounds and perspectives.*". On the high end, institutions may discipline their recruitment processes to prioritize the hiring of women, *ceteris paribus*. Potential examples of these policies as per recruitment announcements include phrases such as "*[i]n the event of equivalent qualifications, female applicants will be given preference*" or "*[a]pplications from women will therefore be given preference if they have the same suitability, ability, and professional performance, unless a competitor has a better qualification*". Quotas are an extreme case of such extreme commitments. For example, a legally mandated commitment exists in Germany: in case of equal qualifications, the position must be offered to women and individuals with disability.

Economic theory is ambiguous about whether the strength of commitment is related to equitable outcomes. Weak commitment is merely "cheap talk" [74]; hence bearing little impact on applicants and selection committees. Strong commitment, by contrast, can actually backfire. Coate and Loury [36] shows that hiring quotas lead to an inefficient equilibrium, in which talented minority representatives insufficiently invest in competence. Fershtman and Pavan [75] argue that search costs increase disproportionately for minority applicants in systems with strong commitment.[9] In other words, while persistent biases are inefficient, policies aimed at overcoming them can also usher inefficiency.

Theory in other social sciences focuses on how bias arises, with less attention devoted to the consequences of policies aimed at reducing bias [12]. Drawing on results from a controlled experiment, Foschi et al. [77] introduce the notion of double standards: women judge the same objective performance by men and women equally, but men do not. Hence, objective recruitment criteria are not going to level the playing field for women. In a broader context, Crandall and Eshleman [78] propose a psychological framework that delineates prejudice itself from how it is expressed. They argue that while prejudice itself is suppressed (e.g., because it can be socially costly even among like-minded members of the majority), it can be costlessly expressed in the form of disproportionately harsh judgment of the representatives of the minority. In this case, weak commitment would be irrelevant and strong commitment would be ineffective as prejudice against minority would be given the false pretense of a merit-based judgment. Leslie et al. [79] argues that strong commitment can backfire because the stigma of incompetence can emerge, veiling all the members of the minority, regardless of their individual potential. In other words, such tools can further delegitimize women in academia rather than reduce biases.

In terms of empirical evidence, the results for hiring for junior positions are mixed.[10] Using observational

[8] This and subsequent citations were taken from the job offers at economists posted at the website https://www.europeanjobmarketofeconomists.org.

[9] Similar mechanism exists if commitment is replaced with an enforced quota. Bijkerk et al. [76] propose that quotas reduce competition between firms to poach high-level minority workers, effectively reducing their bargaining power: the signal of high productivity or potential is simply tainted by the quota.

[10] Observational and experimental studies tend to analyze promotion to tenured positions and full professors; Gerxhani et al. [31] provide a synthetic overview of this literature. In a non-academic context, Petters and Schröder [80] and Neschen and Hügelschäfer [81] show that quotas reduce the perceived performance of women, even those women who are not subject to quota. Although quota are not exactly a strong equivalent to the case of strong commitment, if actually implemented, they may be treated similarly by the representatives of both majority and minority.

data, Ooms et al. [82] show that female Ph.D. graduates have a lower chance of getting an assistant professorship or postdoc position in Europe, but there are many confounding factors which blur causal inference on the pure role of gender biases. Aksnes et al. [83] show that equal achievement among the younger birth cohorts in Norway, but this too is an observational study with multiple confounders.

To our knowledge, the only experimental study that looks at the strength of commitment in hiring junior faculty, albeit indirectly, is an evaluation by external experts of Gerxhani et al. [31]. The authors request the evaluation of applicants for junior academic positions from senior scholars in Germany and Italy. In Germany, the 1989 law explicitly states that when women are a minority, they should be preferred over men in case of equal qualifications. In Italy, by contrast, there is no explicit gender norm for hiring. Although Germany resembles the case of strong commitment, it is important to emphasize that there is no external monitoring, much less enforcement of this preference for hiring women. The authors find that German academics give a higher score to women, *ceteris paribus*. They find no gender differences in Italy.

The experimental evidence of Gerxhani et al. [31] is at odds with compelling qualitative evidence that at least some selection committees find ways to make a biased assessment pass as an objective judgment. Rivera [2] shows that during the recruitment of junior academics, the search committees actively debate the willingness of women in relationship to move to a given university, and that this was virtually never discussed in relation to men. Finding plausible and non-merit reasons to not give someone an offer is a *prima facie* evidence of bias.

The evidence on automated scoring of candidates for assistant professor and postdoc positions based on predetermined criteria (the so-called rubrics) alerts to a similar mechanism. *Once* academic institutions emphasize (even only weak) commitment to diversity, some faculty set the evaluation criteria to reduce the chances of minority applicants to achieve an even score with majority applicants [16]. Analyzing the actual scores, Blair-Loy et al. [15] show that the same research output (number of papers) and impact (H-index), receive lower scores for women than for men, even after they adjust for seniority. They show that this bias is the strongest among individuals with few papers and citations, as is frequently the case for PhD graduates applying for assistant professorships and postdocs. Since they have access to the actual completed recruitment documentation, they also scrutinize the verbal justifications for the scores and find that women receive fewer positive comments, are less likely to be judged as exceptional, and are more likely to receive negative comments *ceteris paribus*.

2.3 The context of assessment: being a minority candidates

To our knowledge, studies do not explore whether biases are comparable between male-dominated and female-dominated fields. Leslie et al. [5] shows that in fields where women are the majority, academics generally rely less on the vague and unobservable category of brilliance. In contrast, in fields dominated by men, this quality is more often considered imperative for a successful academic career.

Indeed, when women are rare, their presence is more frequently considered a product of affirmative action, especially if it was not explicitly stated [84, 85]. They are particularly harshly judged if they succeed at stereotypically male tasks [86]. Systematically, women are perceived as less competent in tasks traditionally associated with men [32, 33]. This concerns both academia as a whole; see, for example, the evidence on conference submissions evaluation [34, 47]; and evaluation of academic quality by non-experts [52, 53]. The diluting of the majority by hiring minority candidates in male-dominated fields can be perceived as reducing their prestige [35].

However, there appears to be some wind of generational change. Belot et al. [87] finds that, within economics, women who specialize in male-dominated fields perform the same as men. This may require a

greater effort [as evidenced by studies on publishing papers, e.g., 8, 50]. Furthermore, these results may be due to survivor bias, i.e. some women leaving academia prematurely [41].

2.4 Hypotheses

Taken together, the existing findings suggest that gender bias in academia is fairly prevalent and that strong commitment to hire women can backfire, especially if women are a minority. In contrast, reliance on external experts nearly uniformly reduces bias in assessing research output (publishing in journals is a notable exception). The bias reemerges when the selection committees engage in deliberation [even when it is absent in scoring the candidates 88] or when they are expected to assess the candidate of a salient gender, not merely the research output of this person [69].

Guided by the literature, in our experiment we elicit assessment of junior job market candidates. Our experiment is similar to the study of Henningsen et al. [89], who study the case of the senior job market and find that strong commitment gives advantage to women. With junior candidates, there is less information to assess the application; thus evaluators can recur to heuristics to fill the gaps [as in 34]. We share some commonalities with the study of Gerxhani et al. [31] who indirectly confirm this evidence for junior job market candidates. Similarly to other studies, we study bias in assessment. We compare averages and conditional averages. In addition, we study the entire distribution of scores. Unlike the earlier studies, we study explicitly the role of the strength of commitment: we compare the cases of weak and strong commitments. Further, we study the interaction between the strength of commitment and the share of women among the assessed applications. Finally, we examine the role of academic excellence.

We preregistered the following hypotheses.

Hypothesis 1 *Applications designated as female receive lower scores than those designated as male under the strong commitment.*

Hypothesis 2 *This gender bias is more severe when woman are minority candidates.*

Hypothesis 3 *This gender bias is lower for applications signaling academic excellence.*

Note that the existing theoretical and empirical literature provides justification for both positive and negative verification of our hypotheses. Studies show evidence of both: strong bias against women and some weak preference toward women. Our design is such that the potential null result has meaningful policy implications.

3 Experimental 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 applications, with three applications in each set. To make the task manageable for experts, the candidates were presented through short biographical profiles. 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.

Using these pairs, we construct biographical profiles. 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.

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, the subjects are asked two questions about each candidate and one summarizing question for each set. First, they assess the competence 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 asked 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. 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. This part of the experiment preparations was performed independently by two authors and a research assistant to eliminate personal biases and inconsistencies. The seven pairs selected were consistently approved by each of the authors without knowing the choices of the others. The seven final pairs emerged as a consensus after each of the authors provided their suggestions.

As a final step, we wrote the biographical profiles based on the content 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 A.2.[11]

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. 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.

Manipulation 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

[11] 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, as some information cannot be anonymized. In the interest of the candidates, we did not ask for external assistance in this task. Balancing between assuring the no-deception design and the privacy protection of the candidates, we decided not to ask an external evaluation of our work.

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 policies to promote gender equality. There were two specific policies, and they were randomized between participants. There was no deception in this description either, as our recruitments were performed at two independent institutions, differing with respect to the strength of commitment to hiring women. Treatment varies between subjects.

In the control treatment, the institution declares that it supports equal opportunity and encourages particularly women to participate in the recruitment process.[12] This statement corresponds to a soft policy, as the statement does not involve any specific commitment on the 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.[13] 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 We aim to 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, men are considered to perform better in male dominated tasks/fields, which would give an edge to male candidates [see meta-analyses by 32, 33]. Moreover, if women *"pollute"* occupations or disciplines, then incumbents will exert greater effort to prevent the entry of women [35]. 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 [90].

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 use exceptional and of high quality biographical profiles.[14] We distinguish three sets: (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.

Each participant was offered to review six different biographical profiles, and it was our responsibility to

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

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

[14]The terms *exceptional* and *high* are relative. *Exceptional* candidates are distinguished by more achievements than the other candidates in the set.

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[15] 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. As is clear from this table, we require seven distinct profiles for each participant so that no biographical profile is assessed twice by the same external expert.[16]

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 the biographical profiles. See Appendix A.2 for detailed profiles.

Operationalizng the hypotheses Recall that our experimental design is such that, under a fair and objective assessment, we expect no differences across genders, conditional on biographical profile. Systematic gender differences in the assessment of biographical profiles have no other basis than bias. We identify the overall effect of the experimental treatment from the comparison of each candidate evaluation between the participants assigned to the NHC and the HC conditions. Take, for example, the combination of sets (1,a). Each of the six candidates in this set received a score from a given participant. A different participant also received a combination of sets (1,a), but that participant was randomized to a different experimental condition. As we average over participants, the treatment effect will be measured as a difference in average scores for EW(1) and all other biographical profiles in this set. In other words, an unbiased evaluation implies that:

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

where the expected $response$ in our experiment refers to the response of the participants to both the scoring question and the invitation question. We denote gender by $G \in \{M, W\}$ and biographical profiles by p. In addition to averages, we provide evidence along the entire distribution of scores.

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 coefficient of interest in our experiment. The null condition of an unbiased evaluation regardless of the treatment can be written down as:

$$\begin{aligned} H_0: \quad & \forall_{p\in\{1,\ldots,7\}}, \ \forall_{G\in\{M,W\}}, \ \forall_{T\in\{NHC,HC\}} \quad E(response_p|G,T) = E(response_p) \\ \text{or} \quad & \forall_{p\in\{1,\ldots,7\}} \quad E(response_p) \perp G,T \end{aligned} \tag{2}$$

where $T \in \{NHC, HC\}$ denotes treatment. To understand whether there is a gender effect of the treatment condition, it must be the case that $E(response_p|G,T) \neq E(response_p|G)$. Our estimated effect comes from

[15] This number comes from computing how many different combinations of exceptional / high-quality male and female we can have within a recruitment ($= 2^3 = 8$) times the number of combinations in the second recruitment ($8 \times 8 = 64$).

[16] Note that biographical profile #3 is only required in the female variant.

the difference between the assessment of biographical profiles for both genders with respect to treatment conditions. The estimation method is similar to a difference in 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 two high quality biographical profiles). 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.[17] The invitations to participate in the survey were sent out by email. We explained that two institutions concluded two recruitment processes. This information was truthful. We further explained that these two institutions sought the ex post advice of external experts on whether they have given each candidate objective assessment. Participation was rewarded with an entry into a lottery, with twenty smart watches as a reward.[18]

The survey was administered anonymously. We created two separate links for men and women among Polish faculty, so that we could adjust 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 eight-day period. 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 61,281 to 6,863 potential respondents. A reminder email was scheduled for April 18th. Due to the anonymous nature of the survey, we cannot tell how many *new* respondents were reached with the reminder email. 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 15 percent.[19] This places our paper at a regular spectrum of response rates in this literature. In a similar correspondence study Powdthavee et al. [91] report response rate of 16% (and a sample size of 271 respondents). Gerxhani et al. [20] reports a response rate of 18.8 percent (and a sample size of 289 respondents), while in Williams and Ceci [30] it was close to 35% (and a sample size for three experiments of jointly 711 respondents). Our final sample is relatively large by the standards of this field: more than a

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

[18] 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 take part without the rewards, which they explicitly stated in the text box intended for the e-mail address.

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

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, because randomization occurred when the respondents clicked on the link in the email.

Note that our pool of participants is unique in some respects. First, our invitation email was widely distributed across many disciplines. Indeed, some of the invited faculty has contacted us to ask if we really seek advice of non-economists. Next, asserting objectivity in hiring has not been an important policy target in many institutions. It is fairly recent and fairly rare that academic institutions in Poland develop gender equality plans and roll-out affirmative action policies. These are conditions for potentially greater bias compared to other pools of participants in the literature. Indeed, our participants may have felt less aware of the contemporaneous profiles of candidates in the market and less concerned about ensuring equal opportunity.

Table A1 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.[20] Around two thirds of 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 A1 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 about seven to eight minutes in both conditions. However, a few outliers raised the average time required to complete the survey by ten minutes. 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 Results

We discuss the results in two substantive parts.[21] First, we present a descriptive analysis that details the responses in our experiment. This section serves to study the potential for bias from several perspectives. Next, we move on to regression analyzes. Regression analysis allows us to quantify the drivers of our results.

4.1 Descriptive analyses

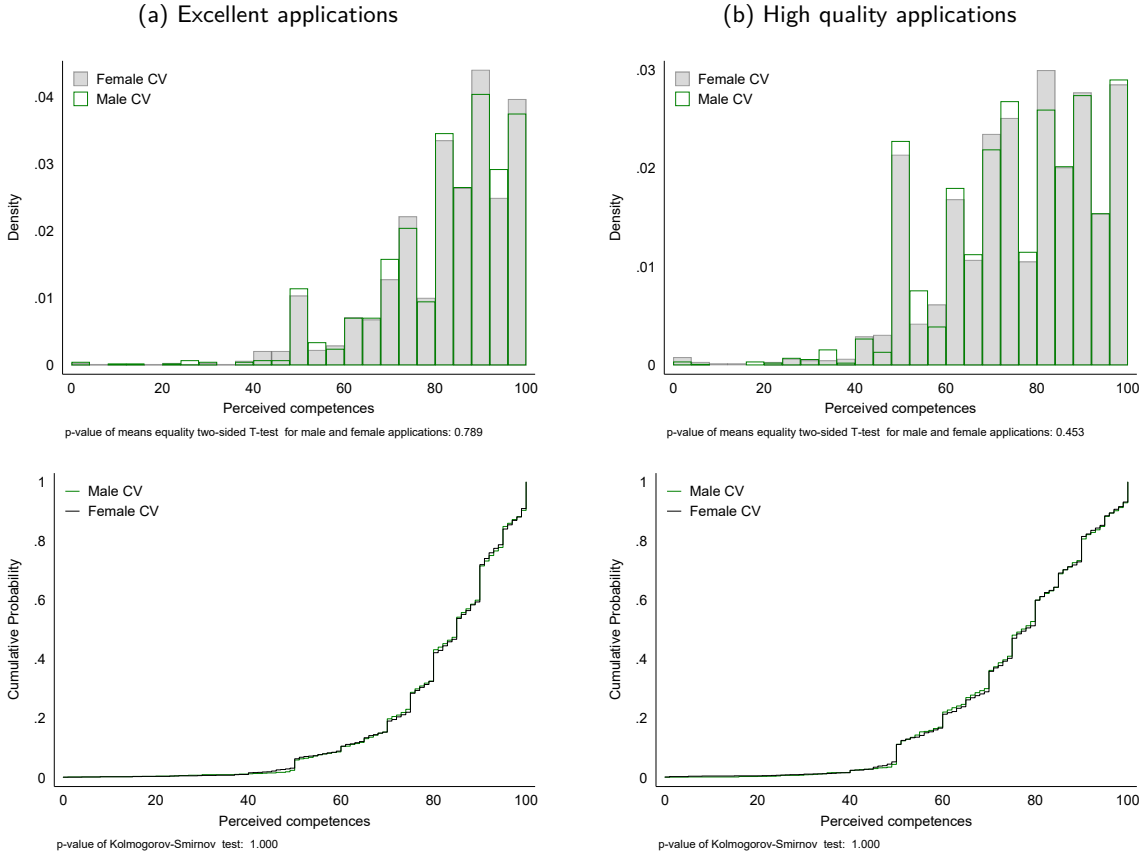
We first portray the histograms and cumulative distribution plots for the assessment of candidate applications in our experiments. In Figure 1 we separately report the scores for the applications designated as excellent (E) and those designated as high-quality (H). Histograms and cumulative distribution plots report the distributions obtained for male and female applications. Since each application exists in both male and female variants, unbiased assessment by external experts should result in statistically indiscernible distributions across genders of applications. Indeed, this is what we find. As expected, the evaluation of H applications is somewhat lower than for E applications. In fact, there is a pronounced spike in assessments at 50 points for the H applications, and the distribution for the E applications is more concentrated around high scores. We find no grounds to reject the null hypothesis that the average EM score by an external expert is equal to the average EW score

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

[21] All the analyses presented in this section were pre-registered.

by the same expert. In fact, the $p - value = 0.79$ for a two-sided t-test. For H applications, the average difference amounts to 0.34 with a $p - value = 0.45$. Kolmogorov-Smirnov tests find no difference in the distribution of the score for the applications of candidates designated as men and women.

Figure 1: Distribution of scores: gender designation and type of profile



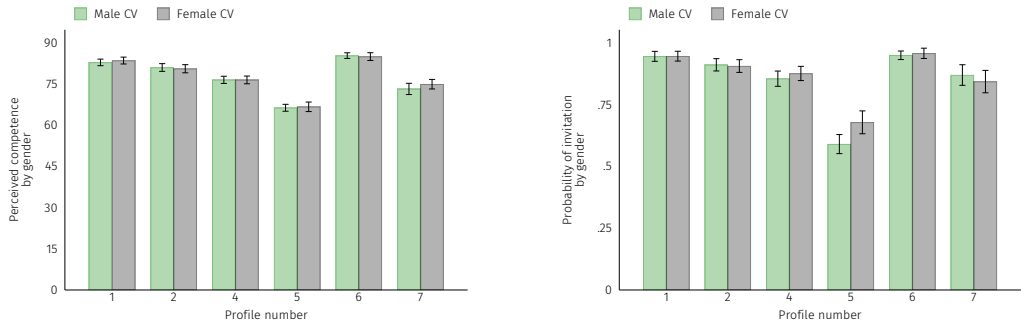
Notes: The figure portrays competence assessment for E and H applications by gender designation.

We do not find differences between the male and female versions of each biographical profile, but we do observe different assessments between biographical profiles. We report these results in Figure 2. In the left panel, we present average competence assessment, 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 the gender groups (as indicated by colors).

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 the case of the invitation score for profile #5, where women have an edge of around five percentage points. Second, the variation between biographical profiles is consistent with our designation of E and H applications. The first three profiles, excellent, scored higher than profiles #4, #5, and #7, which were classified as high. We also observe that profile #6, which was high quality, is evaluated at levels similar to excellent profiles. Overall, these results are consistent with the unbiased evaluations by the external reviewers.

In Figure 3, we report the cumulative distribution function for a measure that directly captures gender bias. We compute the difference in scores assigned when they were designated as men and when they were designated as female candidates. If there is no bias, all scores should be exactly equal. However, one respondent never evaluated the *same* application as two different candidates. Rather, they were provided with two or

Figure 2: Assessed competence and invitation to interview across biographical profiles



(a) Assessed competence

(b) Not inviting is a mistake

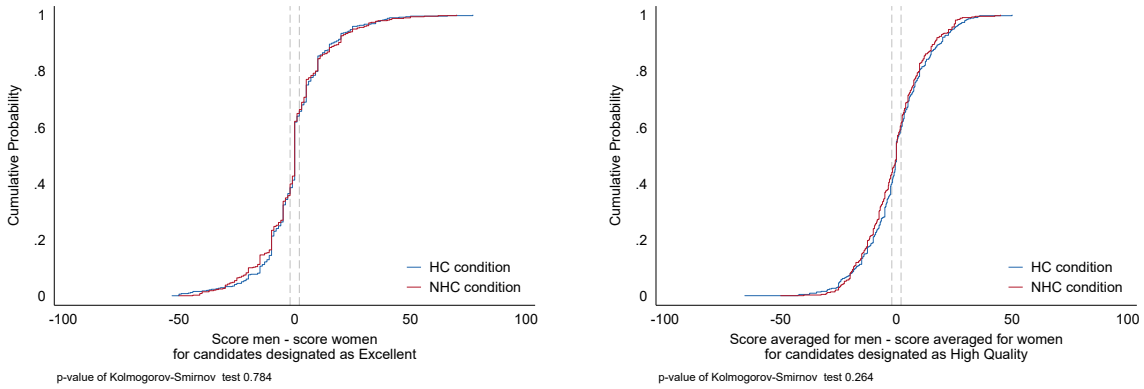
Notes: The figure portrays average competence assessment (left) and the probability of stating that not inviting the candidate would be a mistake (right) for men and women for each profile. Vertical lines represent 95% confidence intervals. We omit biographical profile #3 because it was only used in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

more E applications, at least one for a male candidate and at least one for a female candidate. We compute the average score given to EM applications for each external evaluator and subtract from it the average score given to EW applications by the same evaluator. In the left panel, we show it for the E applications, and in the right panel, we proceed analogously for H applications. We find that 19.51% of the respondents provided an assessment for the HW application that was not different from that of HM (since we compare the averages, we use the absolute difference smaller than 2 points). This proportion increases to 29.92% of respondents for the E applications (with 20% at exactly zero difference). The share of responses with women outscoring men is 41.46% for H applications and 36.10% for E applications. Finally, 33.92% of the respondents judged the EM applications higher than the EW applications, that proportion rises to 39.03% for a comparison between HM and HW. We find no evidence of differences across treatment conditions. Indeed, Kolmogorov-Smirnov test finds no ground to reject the null hypothesis that the distribution under HC is identical to that under NHC.

The external evaluators provided the same unbiased judgment, regardless of the treatment condition. In Figure 4 we report the estimated differences between the average score for the male and female variant of each biographical profile. We compute the difference in means under both treatment conditions. 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, the probability of invitation differs by gender of the application. The probability of invitation is lower for male applicants by around 10 percentage points. This is the profile with the lowest competence assessment; see Figure 4. The only profile for which differences appear to be statistically significant across treatment conditions is the biographical profile #6, where a small and not statistically significant penalty for women under NHC condition is matched with a penalty for men in the HC condition.

These descriptive statistics speak against gender bias among external evaluators. In Appendix C we further explore the magnitude of discrepancies between assessment of profiles when they were designated as female compared to male. Indeed, female designated profiles are evaluated no worse than the male ones, and differences between treatment conditions are minor. In other words, we do not find evidence that external experts exhibit bias against women.

Figure 3: Distribution of differences according to gender designation under HC and NHC treatment conditions

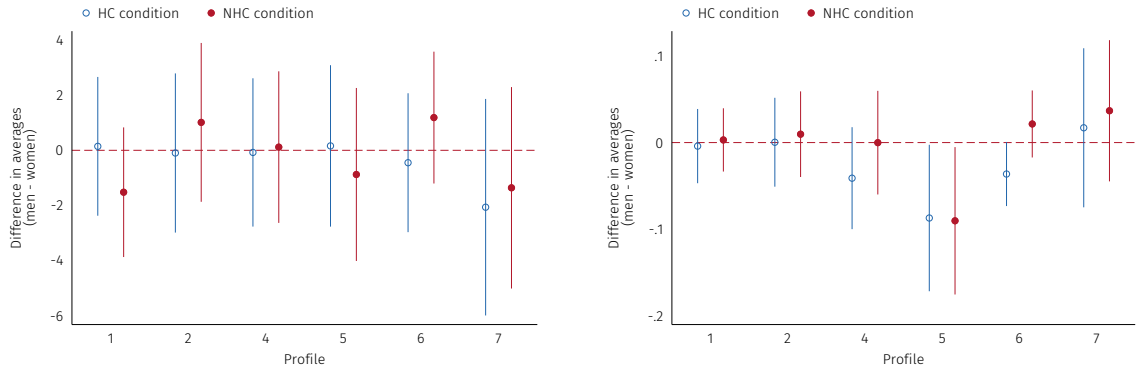


(a) Exceptional applications

(b) High quality applications

Notes: The figure portrays distribution of differences between male and female profile under HC and NHC treatment conditions. HC stands for hiring commitment and NHC stands for no hiring commitment. The left figure presents the differences between Exceptional applications in the first recruitment. Biographical profile #3 is excluded as it only appears in female version. The second figure presents the difference between average scores for high quality applications. Dashed lines indicates values of -2 and 2. Kolmogorov-Smirnov tests show that curves in each subfigure are not statistically different from each other.

Figure 4: Gender gap in assessed competence and invitation to interview across biographical profiles



(a) Competence assessment

(b) Recommendation of interview invitation

Notes: The figure portrays differences in average competence assessment (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 allowing unequal variances. Positive values signify advantage for men.

4.2 Regression analysis

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. Specifically, we estimate:

$$y_{i,p,T} = \beta_0 + \beta_G FemaleCV + \beta_T HC + \beta_{G,T} FemaleCV \times HC + \gamma_p + x_i\beta + e_i \quad (3)$$

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 competence assessment for each candidate, and whether participants considered it would be a mistake not to interview the candidate. The parameters of interest are β_G, and $\beta_{G,T}$. The former captures the effect of female variant of an application relative to male variant under NHC condition. The latter shows the differential effect of treatment: being a female candidate applying to an institution with a HC 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 $\beta_{G,T}$.[22] Finally, $x_i\beta$ identify respondents characteristics (gender, graduation year, previous experience).

In order to test Hypothesis 2, we augment equation (3) to include interactions with an indicator for the minority gender in the current set. This variable is defined separately for each profile in our two sets. If a set contains strictly one profile designated as female, then this profile is given the value of one, and the two other profiles in that set are given the value of zero. Likewise, if a set contains strictly one male profile, then we set the value of the $Minority$ variable to one and zero to the two other female profiles in this set. Thus, the interaction term compares the case of strictly minority female profiles with all other compositions of the set. However, the $Minority$ dummy reports the effect of being a minority candidate for men. The regression is of the following form:

$$\begin{aligned} y_{i,p,T} = \beta_0 \quad &+ \quad \beta_G FemaleCV + \beta_T HC + \beta_M Minority \\ &+ \quad \beta_{G,T} FemaleCV \times HC + \beta_{G,M} FemaleCV \times Minority + \beta_{T,M} HC \times Minority \\ &+ \quad \beta_{G,T,M} FemaleCV \times HC \times Minority + \gamma_p + x_i\beta + e_i \end{aligned} \tag{4}$$

The regression includes additional terms and parameters, which are related to hypothesis two. The parameters $\beta_{G,M}$ and $\beta_{G,T,M}$ 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 $\beta_{G,M} < 0$, i.e. profiles of women are assessed as less competent in male dominated positions, and $\beta_{G,T,M} < 0$, i.e. differences in evaluation are more negative when the institution announces a hiring commitment. To facilitate the interpretation of the three-way interaction models, we will estimate an auxiliary specification: $y_{i,p,T} = \beta_0 + \beta_G FemaleCV + \beta_M Minority + \beta_{G,M} FemaleCV \times Minority + \gamma_p + x_i\beta + e_i$.

Finally, we test Hypothesis 3 by interacting the treatment variable 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.

$$\begin{aligned} y_{i,p,T} = \beta_0 \quad &+ \quad \beta_G FemaleCV + \beta_T HC + \beta_Q Excellent \\ &+ \quad \beta_{G,T} FemaleCV \times HC + \beta_{G,Q} FemaleCV \times Excellent + \beta_{T,Q} HC \times Excellent \\ &+ \quad \beta_{G,T,Q} FemaleCV \times HC \times Excellent + x_i\beta + e_i \end{aligned} \tag{5}$$

To facilitate the interpretation of the three-way interaction models, we will estimate an auxiliary specification: $y_{i,p,T} = \beta_0 + \beta_G FemaleCV + \beta_Q Excellent + \beta_{G,Q} FemaleCV \times Excellent + x_i\beta + e_i$. Unlike previous specifications, these regressions do not include biographical fixed effects γ_p, as doing so would prevent the estimation of the $Excellent$ dummy.

Table 2 presents the results of estimating specifications (3)-(5).[23] The results are consistent with the descriptive statistics. When looking at gender $\hat{\beta}_G$, the coefficients are all very close to zero and are not statistically significant. For example, in the first specification, which corresponds to equation (3), women received 0.154 fewer points in the assessment of competences than men on an average score in excess of 80 points. This result confirms that participants evaluated the applications in a similar way, regardless of gender

[22] We restrict the sample not to include biographical profile # 3, as this profile was only distributed in its female variant.

[23] The full set of coefficients for participant's characteristics is available upon request. Female respondents on average awarded scores higher by 2.3 points (statistically significant) and invitation probabilities of 1.9% higher. Receiving tenure is associated with invitation probabilities higher by 5.2% *ceteris paribus*. The assessment by experts with recruitment experience was not statistically significantly different from individuals without such experience. Respondents in STEM disciplines were considerably less likely to invite candidates to participate in the interview relative to the humanities. We did not identify other systematic drivers of assessing competence, nor recommendation concerning the invitation.

designation.

Table 2: Regression results

	Competences					Invitation				
	Eq (3)	Context: minority		Context: quality		Eq (3)	Context: minority		Context: quality	
	(1a)	(2a)	(3a)	(4a)	(5a)	(1b)	(2b)	(3b)	(4b)	(5b)
Female CV	0.110	0.0315	-0.473	0.0665	0.0981	0.00409	0.0135	-0.00859	0.0343***	0.0215
	(0.476)	(0.607)	(0.868)	(0.498)	(0.705)	(0.0109)	(0.0118)	(0.0165)	(0.0121)	(0.0170)
T = HC	-0.550		-0.987		-0.255	-0.0385***		-0.0441***		-0.0382**
	(0.636)		(0.831)		(0.770)	(0.0131)		(0.0166)		(0.0175)
Female CV × T = HC	0.168		0.977		-0.0702	0.0253		0.0427*		0.0244
	(0.670)		(1.227)		(1.000)	(0.0156)		(0.0236)		(0.0243)
Context		0.0309	-0.626	6.160***	6.348***		0.00415	-0.00438	0.122***	0.118***
		(0.627)	(0.863)	(0.550)	(0.765)		(0.0137)	(0.0187)	(0.0112)	(0.0153)
Female CV × Context		0.579	1.885	0.0603	0.180		0.0134	0.0427	-0.0362**	-0.0275
		(1.189)	(1.679)	(0.682)	(0.974)		(0.0224)	(0.0309)	(0.0153)	(0.0212)
Context × T = HC			1.280		-0.371			0.0161		0.00830
			(1.233)		(1.100)			(0.0264)		(0.0224)
Female CV × Context × T = HC			-2.542		-0.227			-0.0561		-0.0166
			(2.377)		(1.367)			(0.0445)		(0.0306)
Resume FE	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	No
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5639	5639	5639	5639	5639	5639	5639	5639	5639	5639
R-squared	0.171	0.171	0.171	0.0534	0.0536	0.128	0.127	0.129	0.0397	0.0414
H0: $\beta_{G,T} = 0$										
P-value	0.802		0.426		0.944	0.105		0.0702		0.314
Required sample size	241857		23973		3081236	5806		4654		15026
H0: $\beta_T + \beta_{G,T} = 0$										
P-value	0.596		0.991		0.718	0.324		0.932		0.466
Sample needed to reject H0	5000		10000		10000	2500		10000		5000

Notes Estimates obtained using linear regressions. Column titles indicate different specifications. Clustered standard errors in parentheses. The last row indicates that we should have more than n participants ($3n$ observations) to reject the null hypothesis that the sum of coefficients is different from zero. ***, **. * indicate p-values lower than 0.01, 0.05 and 0.1 .

We do not find statistically significant treatment effects in the assessment of the candidates. In all specifications for the competence score, both $\hat{\beta}_T$ and $\hat{\beta}_{G,T}$ are not statistically significant. However, we find significant results for the invitation specifications. For biographical profiles designated as male candidates, under HC the invitations are lower. The additional effect of HC with profiles designated as female candidates is universally of the opposite sign and of similar magnitude. We formally test the relationship between these two estimates: $\beta_T + \beta_{G,T} = 0$. We do not reject this null hypothesis, which corroborates negative treatment effects for profiles designated as male and no treatment effect for profiles designated as female. In case of competence scores, the sample size required to reject the null hypothesis with the actual estimated size effects is ridiculously high. Given the estimated size effects, we would need over 10 000 participants to obtain significant effects.[24] For invitation, the sample sizes are somehow lower, but it would still require more than doubling the number of participants in the column (1b). Rejecting the null hypothesis that the coefficient on the interaction $\hat{\beta}_{G,T}$ also requires a larger sample size, but the difference is somehow smaller. If our study is insufficiently powered to reject this null hypothesis, then it would imply a minor negative treatment effect for women relative to the NHC condition.

As for context, we find a significant coefficient for the two-way interaction between gender and quality of application (Excellent relative to High Quality). We do not find any significance in specifications with three-way interactions. For the minority context, the estimated three-way interactions are large (approximately -2.5 points or -3% of the average score), but these are not precisely estimated, which results in coefficients that are not statistically significant.

Our results on candidate invitations are in line with previous findings on the role of gender [comparing to e.g., 30, 89, note that these studies did not explore the role of the strength of committment]. We do not confirm the preferential assessment of women argued by Williams and Ceci [30] in the United States. In the next section, we explore the sensitivity of our results to various issues related to the data collection process.

Summarizing, we find that external experts provide remarkably unbiased assessment of biographical profiles. We find some minor differences for invitation recommendation, but only in HC treatment. We also did not find evidence of contexts. The respondents were more likely to recommend that candidates with profiles designated as male could have been omitted in the invitations when the institution is committed to hiring women in cases of equal quality. With a sample size larger by 10%, provided that the size effects remained unaffected, we were likely to find a minor negative preference for women as well under HC treatment relative to NHC. Although our results suggest that external experts can give a viable unbiased benchmark assessment and recommendation on invitation of the actual candidates, it is not plausible that any institution can invite 1000+ external experts. In the final exploratory analysis, we study the role of outliers: we quantify the frequency of outliers and their potential impact on the bias of assessment and recommendation.

5 Sensitivity analyses

Although our null result is promising, the estimates presented in Table 2 could be biased towards zero for a number of reasons rendering our conclusions invalid.[25] In this subsection, we explore four sensitivity analyses. The first analysis refers to respondents being inattentive. If that was the case, the participants may have simply not paid attention to the names, in which case the estimated coefficients would underestimate the true effects. Second, inattentive respondents may have failed to read the profiles carefully and thus ignored differences between candidates, for example, in terms of academic achievement. Third, it is possible that the

[24] The required sample sizes would be lower if one is willing to assume lack of correlation between answers in the same set for the same participant.

[25] This section reports exploratory analyses that were not pre-registered.

effects are heterogeneous across respondents, canceling each other out on average. Finally, it is possible that the linear regression models lack the power to detect deviations.

5.1 Were respondents attentive?

We explore whether the scoring by a given participant is consistent.[26] If the participant is attentive, we would expect that higher competence assessment are associated more frequently with an invitation. There can be two potential ways to violate this monotonicity condition: (i) profiles that receive the same competence score can receive different invitation recommendations, and (ii) profiles with lower scores are recommended for invitation, and profiles with higher scores are not. We find thirty violations of the first type and five violations of the second type. Overall, this constitutes less than 1.7% of all observations.

Next, we leverage the evidence from ranking the profiles by the participants: at the end of each recruitment, participants were requested to rank the candidates from the most qualified to the least qualified. To a large extent, these rankings correspond to those implied by the assessed competence. Up to 93% of the participants ranked the profiles in the order implied by their competence assessment.

In addition to inconsistency in assessing biographical profiles, we proxy insufficient attention by the time to complete the survey. We trim the sample to exclude participants whose completion time was among the lowest five percent (they provided all responses to the survey in less than 3 minutes) or among the highest five percent (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 the slowest respondents, we are concerned about simultaneous engagement in other tasks, which could have distracted them.

We reestimate our model on a sample that restricts the participants to consistent scoring on all counts and taking the middle 90% of time to fill in the questionnaire. The results are presented in Table B1 in the Appendix. We find that some of the point estimates are larger in absolute terms 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, the confidence intervals overlap between our main specification and this sensitivity analysis, which suggests that potential differences in the point estimates are not statistically relevant.

Finally, we study the monotonicity between competence assessment and recommendation of invitation to an interview. We find that the correlation is statistically significant and essentially identical for biographical profiles designated as male when compared to those designated as female. The results are reported in Table B6.

5.2 Top coding

As stated in our Hypothesis 3, we expect differences in the evaluation of competences to be lower for excellent profiles. We relied on the pool of actual applicants to our Warsaw-based institutions, but in emails sent to us during the experiment, some of the participants emphasized that all of the proposed candidates were exceptional and – as some phrased it – unheard of in their institution. Accordingly, some respondents may have assessed the applications not vis-a-vis each other, but in a wider context of their experience from recruitment processes in which they were involved. This would imply top-coding in our sample: giving all profiles the same score and recommending universal invitations. We estimate our main regression on the sub-sample of participants whose responses differentiated between the profiles: we keep only participants who did not

[26] In order to avoid making the gender designation of the biographical profiles too salient, we did not include manipulation checks for gender.

recommend inviting all candidates, and whose assessment resulted in at least one point difference between their lowest and their highest competence assessment in each recruitment. These two restrictions reduce the sample by approximately 60% of the participants (the main restriction came from the participants with the universal recommendation to invite the candidates for the interview).

The results are presented in Table B2 and corroborate our main line of interpretation for the assessment of biographical profiles. However, we find new results for the invitations. As previously, men are less likely to be invited for an interview under HC, whereas for women the interaction term is of similar magnitude and opposite sign. But the gender composition of the recruitment set matters: the coefficient $\hat{\beta}_{T,M}$ is positive and statistically significant indicating that the penalty for the male-designated profiles is concentrated in contexts when there were more men than women in the recruitment set. The coefficient $\hat{\beta}_{F,M}$ is also positive: there is a boost in the probability of invitation for female-designated profiles when the recruitment set is majority male. Finally, the three-way interaction represents a reduction of 9 percentage points in the probability that a female-designated profile is recommended for an interview compared to minority women applying to an institution without HC ($0.122 - 0.210 = -0.088$, with a p-value of 0.106). Taken together, the findings provide some tentative support for Hypothesis 2: strong commitment intensifies bias if women are the minority candidate.

5.3 Heterogeneity treatment effects

Gender of the participant Although men and women in our study evaluate the candidates somewhat differently, these differences are not statistically significant. Table B3 reports the estimates. Men appear to perceive profiles of male candidates as more competent, whereas women perceive profiles of female candidates as more competent. However, these differences are minor relative to the heterogeneity among both men and women, resulting in no significance. In terms of treatment effects, the results are consistent with our main specification.

Heterogeneity across disciplines Consistent with Williams and Ceci [30] female-designated profiles are assessed less favorably in economics (this is the most numerous group in the discipline of social sciences). Consistent with our earlier findings, there is stronger bias against female-designated profiles under HC if women constitute a minority. Note that the number of observations varies greatly by discipline, and these comparisons should be taken with a grain of salt.

5.4 Alternative estimation procedures

Participant fixed effects Recall that treatment assignment varies at between-subject level. We cannot adjust for participant fixed effects and still obtain estimates of the treatment effect (HC). However, each participant provided six assessment of competence and six recommendations on invitation for profiles designated to different genders. Hence, we can identify the interaction between treatment and gender. Column 1 of Table B5 contains the results assessed competence: the point estimates remain close to zero, and not statistically significant.

Within subject variation on excellent profiles In the first recruitment, the third profile acted as a signal of whether the composition of applicants was more female- or more male-dominated. We ignore the third profile and focus on comparisons between Anna/Adam and Barbara/Bartosz. The dependent variable is the difference in scores between these profiles, and the independent variables indicate the gender of the first profile (which by construction determines the gender of the second profile). These estimates, as reported in Column 2 of Table B5, show no significant gender biases, nor treatment effects.

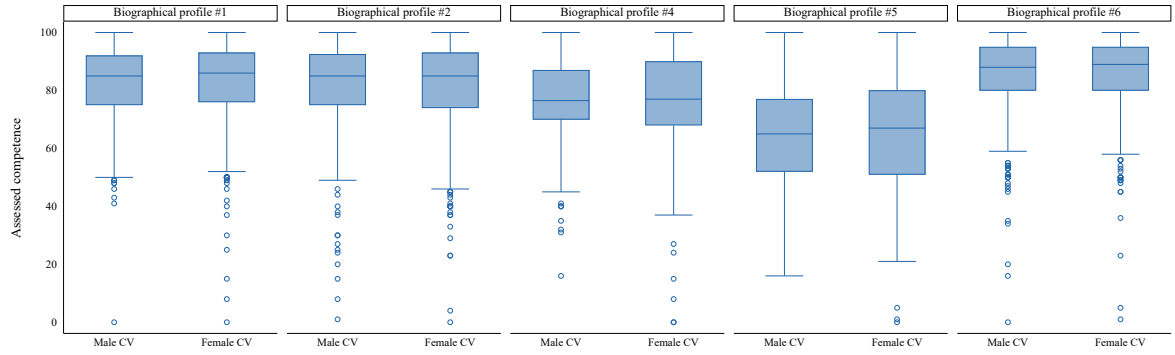
Tobit models Competence could only be assessed on a scale from 1 to 100. The participants may have locked themselves in assigning scores in the first recruitment set, making the scale potentially too short on either end in the second recruitment. This is a different form of censoring than already discussed. We test for this possibility by estimating Tobit models for competence assessment. The point estimates are presented in Column 3 of Table B5. The table corroborates our main findings.

5.5 Can outliers bias the assessment?

Admittedly, most recruitment processes cannot rely on 1000+ external experts. In smaller pools of experts, a biased expert – even if statistically rare – can undermine the objectivity of the whole panel, because the influence of one outlier expert (i.e. the assessments that are unusually high or unusually low) has a stronger bearing on the average. We have in mind the following thought experiment: when relying on fewer external experts, one faces some chance of having an outlier experts. We want to gauge the effect of this event actually happening in one given recruitment.

We operationalize outliers as assessments falling short of the first quartile less than 1.5 of the inter-quartile range. We identify 81 such individuals (if we consider quartiles to be gender specific, the number of outliers grows to 86 individuals; the two groups largely overlap). Consequently, the probability of finding an expert with extreme assessment is below 1%. In Figure 6, we provide box plots for each biographical profile. In our experiment, outliers correspond to an unusually low competence score and tend to be slightly more common among biographical profiles that received on average higher scores, that is profiles #1, #2 and #6. Outliers appear to be more common for the female-designated profiles than for the male-designated ones.

Figure 6: Competence assessment conditional on profile and gender



Note: We report only those biographical profiles for which all participants provided assessment. Hence, profiles #3 and #7 are excluded, see Table 1.

Next, we evaluate the impact of excluding outliers in our main estimates. We consider two cases: (i) exclude the extreme assessments, or (ii) exclude those experts for whom at least one assessment qualifies as extreme. We report the results in Table B7. The first two columns of present estimates from linear probability models, where the dependent variable equals one when an external expert is an outlier. In the first column, outliers are defined for each profile and each gender independently. In the second column, the quartiles are common for both gender designations of profiles. The estimates for competence assessment and invitation recommendation follow in the next columns. Estimates are essentially unaffected by excluding outliers.

6 Discussion and conclusions

Whereas in the case of business organizations, quotas are a common solution, higher education institutions seek alternative policy instruments to level the playing field in recruitment of scholars. We focus on one form of preferential treatment of minority candidates, which requires an unbiased assessment from recruitment committees in order to be effective. Our experiment explores whether external experts can deliver these unbiased assessments.

We focus on assessment by external experts. We study the ability of this instrument to provide an unbiased assessment of candidates. We provide theoretical and empirical evidence for the potential of external experts assessment to be unbiased. We explore their response to implementing a strong commitment to hiring a gender minority candidate. If experts' assessment reproduces prevailing gender stereotypes, strong commitment to hiring women in case of equal qualifications could actually backfire: the experts could reduce the scores assigned to women to diminish their chances of getting the job, thus providing a false legitimization to discriminatory practices. It could be one potential way of providing a false legitimization to discriminatory practices.

We designed a correspondence experiment where external experts were asked to provide an assessment of actual applicants to a junior position. The junior positions are particularly relevant to study for two reasons. First, getting the first job after Ph.D. graduation is the first necessary step to a successful academic career. Second, for junior candidates, there are fewer accomplishments to evaluate than for senior scholars: assessing academic potential is thus more exposed to potentially gender-biased heuristics. We construct biographical profiles based on actual job applications that accurately reflect a pair of candidates of both genders and present them for assessment to external experts randomly designating them as representing a female and a male candidate.

We find no evidence of gender bias. Although proving the null hypothesis is ultimately impossible, we provide rich and robust evidence that an unbiased assessment is an actual feature of the evaluation decisions by the external experts. The distributions of the assessment of the profiles designated as female overlap with those designated as men. We find that a commitment to hiring women does exhibit in somewhat lower probability of invitation recommendation for men, and effectively no absolute effect on likelihood of invitation for women. When applying as a minority candidate, women face somewhat smaller chances of being recommended for invitation, but these effects are small quantitatively and only marginally statistically significant. We find no evidence that particularly negative external evaluations can substantially drive gender differences (even if they are more likely to give unusually low scores to biographical profiles designated as female compared to the male ones).

Our experiment opens several avenues that invite further research. First, in terms of theory, we need to better explore the role of evaluators without "skin in the game." Most existing theory focuses on why insiders may exhibit their biases and implications of these biases for the interactive equilibrium with applicants. Our experiment finds next to no bias with external experts. This finding is in line with the literature on grant applications in several previous studies. However, in some of these earlier studies, the emphasis on the gender of the applicant reintroduced the bias against women [e.g., 69]. In our experiment, the gender of the applicant was salient from the very first sentence. We also made it explicit that we are relying on external experts to help us judge whether our own recruitment process was unbiased. Thus, we prompted a number of triggers that in previous research revealed gender bias (e.g., grant application evaluations), but it did not result in a biased assessment of applicants in our experiment. Indeed, we may need some more theory in psychology, sociology, and economics on why bias does not emerge.

On a related note, a given external expert is unlikely to assess all applicants in a given recruitment process. This implies that when reaching out to obtain the assessment of external experts, the hiring academic

institutions are able to provide gender-balanced subsets of the candidate pool. Presenting no candidates as minority (even if fewer minority individuals applied to a given position) is a cost-less and easy way to eliminate potential bias. We show that even in a generally unbiased assessment, negativity toward a minority candidate arises. The existing literature shows that external experts tend to prefer gender balance in general [67].

The second avenue of further research is related to conveying the information about candidates to experts. We presented external experts with biographical profiles that we edited. This was in some sense similar to relying on predetermined rating criteria [referred to sometimes as rubrics 15, 16], with the main difference that we predetermined which information to report in biographical profiles. In this way, we eliminated the wiggle room for experts to disguise bias as merit-based arguments. Our profiles were stripped of specific names of journals or schools, making it impossible for external experts to construct add hoc arguments for or against candidates of one gender. In evaluations of the implementations of predetermined rating, research has found biasing the criteria and biasing the narratives about the candidates, but we ironed out any detailed information that could be used to specifically raise or reduce the assessment of specific candidate profiles. It appears entirely feasible to compare assessment of raw CVs of candidates and biographical profiles, both on theoretical grounds and in experiments. Given privacy concerns, such experiments should rely on fictitious candidates, but they may deliver useful information on the potential limitations of using external experts in eliciting the unbiased assessment of scholars.

Related, in actual recruitment processes, the assessment typically also comprises actual writing samples from applicants and actual recommendation letters. Existing research demonstrates that scientific articles are kept to a higher standard when authored by women [8, 50] and that recommendation letters differ [4, 64]. Our experiment is not indicative of whether external experts would be immune to gender bias when assessing a specific scientific study. Neither did we present the content of the recommendation letters (we provided information on who authored them). It remains an open research question on how to convey the information on the candidates to external experts to minimize the impact of differences in the source material [see also 92, for an analysis of information transmission and its impact on assessment bias by gender].

The third avenue considers exploring the role of affirmative action instruments, such as a weak or strong commitment to hire a minority candidate in case of equal qualifications. Without the "skin in the game", the participants have proven not to be strongly affected by our experimental manipulation. We find that when women apply as a minority, the scores of external experts are somewhat biased against women under strong commitment. It must be recognized that our inference on the generally negligible role of affirmative action cannot be extended to insider evaluation because they clearly have "skin in the game". Janys [29] shows remarkable preference for exactly two women across departments in all fields in Germany. Such implicit quotas may make it impossible for extremely talented women to be given a job offer in some department and extremely talented men to be denied a job in other departments, depending on the gender composition of the current faculty.

Note that external experts who agreed to participate in our experiment may be particularly sensitive to gender equality. This is likely a common feature across similar correspondence experiments, but we should be cautious about extrapolating our inference to all external experts. We did find cases of extreme assessments. They are always on the negative side (below the range of assessments by other experts) and slightly more likely when assessing the profiles designated as female. Admittedly, they were very few. However, if the reliance on external experts during recruitment or as a means of evaluating its fairness becomes institutionalized in the academic profession, it would potentially be relevant to think about developing "calibration tools ", and more research is needed to identify their validity. For example, one could consider blending in an additional CV between those of actual candidates to obtain an expert-specific scale or measure. Developing such validated tools could be a new avenue for further research.

References

[1] Brooke Helppie-McFall, Eric Parolin, and Basit Zafar. Career Expectations and Outcomes: Evidence (on Gender Gaps) from the Economics Job Market. June 2024. doi: $10.3386/w32446$.

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

[3] Heather Sarsons, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):101–147, 2021.

[4] Audinga Baltrunaite, Alessandra Casarico, and Lucia Rizzica. Women in economics: The role of gendered references at entry in the profession. Temi di Discussione (Working Paper) 1438, Bank of Italy, 2024.

[5] Sarah-Jane Leslie, Andrei Cimpian, Meredith Meyer, and Edward Freeland. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, 347(6219):262–265, 2015.

[6] Asia A. Eaton, Jessica F. Saunders, Ryan K. Jacobson, and Keon West. How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: Professors' Biased Evaluations of Physics and Biology Post-Doctoral Candidates. *Sex Roles*, 82(34):127–141, 2020.

[7] Pleun Van Arensbergen, Inge Van der Weijden, and Peter Van den Besselaar. Gender differences in scientific productivity: a persisting phenomenon? *Scientometrics*, 93(3):857–868, 2012.

[8] David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Are referees and editors in economics gender neutral? *Quarterly Journal of Economics*, 135(1):269–327, 2020.

[9] Anne E Lincoln, Stephanie H Pincus, and Phoebe S Leboy. Scholars' awards go mainly to men. *Nature*, 469(7331):472–472, 2011.

[10] Holly O Witteman, Michael Hendricks, Sharon Straus, and Cara Tannenbaum. 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, 2019.

[11] Shelly Lundberg and Jenna Stearns. Women in economics: stalled progress. *Journal of Economic Perspectives*, 33(1):3–22, 2019.

[12] KerryAnn O'Meara, Dawn Culpepper, and Lindsey L Templeton. Nudging toward diversity: Applying behavioral design to faculty hiring. *Review of Educational Research*, 90(3):311–348, 2020.

[13] Kasey Buckles. Fixing the leaky pipeline: Strategies for making economics work for women at every stage. *Journal of Economic Perspectives*, 33(1):43–60, 2019.

[14] Pascaline Dupas, Alicia Sasser Modestino, Muriel Niederle, Justin Wolfers, and The Seminar Dynamics Collective. Gender and the Dynamics of Economics Seminars. NBER Working Paper 28494, National Bureau of Economic Research, 2021.

[15] Mary Blair-Loy, Olga V Mayorova, Pamela C Cosman, and Stephanie I Fraley. Can rubrics combat gender bias in faculty hiring? *Science*, 377(6601):35–37, 2022.

[16] Dawn Culpepper, Damani White-Lewis, KerryAnn O'Meara, Lindsey Templeton, and Julia Anderson. Do rubrics live up to their promise? Examining how rubrics mitigate bias in faculty hiring. *The Journal of Higher Education*, 94(7):823–850, 2023.

[17] Marc J Lerchenmueller and Olav Sorenson. The gender gap in early career transitions in the life sciences. *Research Policy*, 47(6):1007–1017, 2018.

[18] Sifan Zhou, Sen Chai, and Richard B Freeman. Gender homophily: In-group citation preferences and the gender disadvantage. *Research Policy*, 53(1):104895, 2024.

[19] Heather Sarsons. Recognition for Group Work: Gender Differences in Academia. *American Economic Review*, 107(5):141–145, May 2017. ISSN 0002-8282. doi: 10.1257/aer.p20171126.

[20] Klarita Gerxhani, Nevena Kulic, and Fabienne Liechti. Double standards? Co-authorship and gender bias in early-stage academic evaluations. *European Sociological Review*, 39(2):194–209, 2023.

[21] Toni Schmader, Jessica Whitehead, and Vicki H Wysocki. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles*, 57(7):509–514, 2007.

[22] Juan M. Madera, Michelle R. Hebl, and Randi C. Martin. Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591–1599, November 2009. ISSN 0021-9010. doi: 10.1037/a0016539.

[23] Manuel Bagues, Mauro Sylos-Labini, and Natalia Zinovyeva. Does the gender composition of scientific committees matter? *American Economic Review*, 107(4):1207–1238, 2017.

[24] Justus Baron, Bernhard Ganglmair, Nicola Persico, Timothy Simcoe, and Emanuele Tarantino. Representation is not sufficient for selecting gender diversity. *Research Policy*, 53(6):104994, 2024.

[25] Pierre Deschamps. Gender quotas in hiring committees: A boon or a bane for women? *Management Science*, forthcoming, 2024.

[26] Vincent P. Crawford and Joel Sobel. Strategic Information Transmission. *Econometrica*, 50(6):1431, November 1982. ISSN 0012-9682. doi: 10.2307/1913390.

[27] Paul Milgrom and John Roberts. Relying on the Information of Interested Parties. *The RAND Journal of Economics*, 17(1):18–32, 1986. ISSN 0741-6261. doi: 10.2307/2555625.

[28] Jack B Soll, Katherine L Milkman, and John W Payne. A user's guide to debiasing. In *The Wiley Blackwell handbook of judgment and decision making*, volume 2, pages 924–951. Wiley Online Library, 2015.

[29] Lena Janys. Testing the presence of implicit hiring quotas with application to german universities. *Review of Economics and Statistics*, 106:627637, 2024.

[30] Wendy M. Williams and Stephen J. Ceci. 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, 2015.

[31] Klarita Gerxhani, Nevena Kulic, Alessandra Ruscon, and Heike Solga. Gender bias in evaluating assistant professorship applicants? Evidence from harmonized survey experiments in Germany and Italy. *Social Forces*, forthcoming, 2024.

[32] Heather K. Davison and Michael J. Burke. Sex Discrimination in Simulated Employment Contexts: A Meta-analytic Investigation. *Journal of Vocational Behavior*, 56(2):225–248, April 2000. ISSN 0001-8791. doi: 10.1006/jvbe.1999.1711.

[33] Amanda J. Koch, Susan D. DMello, and Paul R. Sackett. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, 100(1): 128–161, January 2015. ISSN 0021-9010. doi: 10.1037/a0036734.

[34] Laura Hospido and Carlos Sanz. Gender gaps in the evaluation of research: Evidence from submissions to economics conferences. *Oxford Bulletin of Economics and Statistics*, 83(3):590–618, 2020.

[35] Claudia Goldin. 9. a pollution theory of discrimination: Male and female diverences in occupations and earnings. In Leah Platt Boustan, Carola Frydman, and Robert A. Margo, editors, *Human Capital in History: The American Record*, pages 313–354. University of Chicago Press, Chicago, 2014.

[36] Stephen Coate and Glenn C. Loury. Will Affirmative-Action policies eliminate negative stereotypes? *The American Economic Review*, 83(5):1220–1240, 1993.

[37] Harry J. Holzer and David Neumark. Affirmative action: What do we know? *Journal of Policy Analysis and Management*, 25(2):463–490, March 2006. ISSN 1520-6688. doi: 10.1002/pam.20181.

[38] Christine Wennerås and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387(6631):341–343, May 1997. ISSN 1476-4687. doi: 10.1038/387341a0.

[39] Stephen J Ceci and Wendy M Williams. Understanding current causes of women's underrepresentation in science. *Proceedings of the National academy of sciences*, 108(8):3157–3162, 2011.

[40] Emmanuelle Auriol, Guido Friebel, Alisa Weinberger, and Sascha Wilhelm. Underrepresentation of women in the economics profession more pronounced in the united states compared to heterogeneous europe. *Proceedings of the National Academy of Sciences*, 119(16):e2118853119, 2022.

[41] Junming Huang, Alexander J Gates, Roberta Sinatra, and Albert-László Barabási. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9):4609–4616, 2020.

[42] Giulio Marini and Viviana Meschitti. The trench warfare of gender discrimination: evidence from academic promotions to full professor in Italy. *Scientometrics*, 115(2):989–1006, March 2018. ISSN 1588-2861. doi: 10.1007/s11192-018-2696-8.

[43] Clément Bosquet, PierrePhilippe Combes, and Cecilia GarcíaPeñalosa. Gender and Promotions: Evidence from Academic Economists in France. *The Scandinavian Journal of Economics*, 121(3):1020–1053, March 2019. ISSN 1467-9442. doi: 10.1111/sjoe.12300.

[44] Marianna Filandri and Silvia Pasqua. Being good isnt good enough: gender discrimination in Italian academia. *Studies in Higher Education*, 46(8):1533–1551, November 2019. ISSN 1470-174X. doi: 10.1080/03075079.2019.1693990.

[45] Asier Minondo. Who presents and where? an analysis of research seminars in us economics departments. arxiv preprint arxiv:2001.10561, Mimeo, 2020.

[46] Jennifer L Doleac, Erin Hengel, and Elizabeth Pancotti. Diversity in economics seminars: who gives invited talks? *American Economic Review (Papers and Proceedings)*, 111:55–59, 2021.

[47] Margaret Samahita and Kevin Devereux. Are economics conferences gender-neutral? evidence from ireland. *Oxford Bulletin of Economics and Statistics*, 86(1):101–118, 2024.

[48] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Conference presentations and academic publishing. *Economic Modelling*, 95:228–254, 2021.

[49] Anusha Chari and Paul Goldsmith-Pinkham. Gender representation in economics across topics and time: Evidence from the nber summer institute. NBER Working Papers 23953, National Bureau of Economic Research, 2017.

[50] Erin Hengel. Publishing While Female: are Women Held to Higher Standards? Evidence from Peer Review. *The Economic Journal*, 132(648):2951–2991, 2022.

[51] Madeline E. Heilman and Julie J. Chen. Same Behavior, Different Consequences: Reactions to Mens and Womens Altruistic Citizenship Behavior. *Journal of Applied Psychology*, 90(3):431–441, 2005. ISSN 0021-9010. doi: 10.1037/0021-9010.90.3.431.

[52] Silvia Knobloch-Westerwick, Carroll J Glynn, and Michael Huge. The matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science communication*, 35(5):603–625, 2013.

[53] Michał Krawczyk and Magdalena Smyk. Author s gender affects rating of academic articles: Evidence from an incentivized, deception-free laboratory experiment. *European Economic Review*, 90:326–335, 2016.

[54] Lokman I Meho. The gender gap in highly prestigious international research awards, 2001–2020. *Quantitative Science Studies*, 2(3):976–989, 2021.

[55] David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Gender differences in peer recognition by economists. *Econometrica*, 90(5):1937–1971, 2022.

[56] Paula Chatterjee and Rachel M Werner. Gender disparity in citations in high-impact journal articles. *JAMA Network Open*, 4(7):e2114509–e2114509, 2021.

[57] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, 23(8): 918–926, 2020.

[58] Xiang Zheng, Jiajing Chen, Erjia Yan, and Chaoqun Ni. Gender and country biases in wikipedia citations to scholarly publications. *Journal of the Association for Information Science and Technology*, 74(2): 219–233, 2023.

[59] Mathias Wullum Nielsen and Love Börjeson. Gender diversity in the management field: Does it matter for research outcomes? *Research Policy*, 48(7):1617–1632, 2019.

[60] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, 26(3):312–327, 2018.

[61] David Klinowski. Voicing disagreement in science: Missing women. *Review of Economics and Statistics*, forthcoming, 2024.

[62] Abdelghani Maddi and Yves Gingras. Gender diversity in research teams and citation impact in economics and management. *Journal of Economic Surveys*, 35(5):1381–1404, 2021.

[63] Alice H Wu. Gender bias among professionals: an identity-based interpretation. *Review of Economics and Statistics*, 102(5):867–880, 2020.

[64] Markus Eberhardt, Giovanni Facchini, and Valeria Rueda. Gender differences in reference letters: Evidence from the economics job market. *The Economic Journal*, 133(655):2676–2708, 2023.

[65] Audinga Baltrunaite, Piera Bello, Alessandra Casarico, and Paola Profeta. Gender quotas and the quality of politicians. *Journal of Public Economics*, 118:62–74, October 2014. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2014.06.008.

[66] Juan M. Madera, Michelle R. Hebl, Heather Dial, Randi Martin, and Virgina Valian. Raising Doubt in Letters of Recommendation for Academia: Gender Differences and Their Impact. *Journal of Business and Psychology*, 34(3):287–303, April 2018. ISSN 1573-353X. doi: 10.1007/s10869-018-9541-1.

[67] Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. Gender-equal funding rates conceal unequal evaluations. *Research Policy*, 51(1):104399, January 2022. ISSN 0048-7333. doi: 10.1016/j.respol.2021.104399.

[68] Lídia Farré and Francesc Ortega. Selecting talent: Gender differences in success in competitive selection processes. *Journal of Human Resources*, 58(6):1881–1913, 2023.

[69] Giorgia Guglielmi. Gender bias goes away when grant reviewers focus on the science. *Nature*, 554(7690):14–16, 2018.

[70] Stephen J Ceci, Donna K Ginther, Shulamit Kahn, and Wendy M Williams. Women in academic science: A changing landscape. *Psychological science in the public interest*, 15(3):75–141, 2014.

[71] Aliza Forman-Rabinovici, Hadas Mandel, and Anne Bauer. Legislating gender equality in academia: direct and indirect effects of state-mandated gender quota policies in European academia. *Studies in Higher Education*, forthcoming, 2024.

[72] Andreas Leibbrandt and John List. Do equal employment opportunity statements backfire? evidence from a natural field experiment on job-entry decisions. NBER Working Paper 25035, National Bureau of Economic Research, 2018.

[73] David Benjamin Oppenheimer. Distinguishing Five Models of Affirmative Action. *Berkeley Women's LJ*, 4:42, 1988.

[74] Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic perspectives*, 10(3):103–118, 1996.

[75] Daniel Fershtman and Alessandro Pavan. Soft Affirmative Action and Minority Recruitment. *American Economic Review: Insights*, 3(1):1–18, March 2021. ISSN 2640-2068. doi: 10.1257/aeri.20200196.

[76] Suzanne H. Bijkerk, Silvia Dominguez-Martinez, Jurjen Kamphorst, and Otto H. Swank. Labor market quotas when promotions are signals. *Journal of Labor Economics*, 39(2):437–460, 2021.

[77] Martha Foschi, Larissa Lai, and Kirsten Sigerson. Gender and double standards in the assessment of job applicants. *Social Psychology Quarterly*, pages 326–339, 1994.

[78] Christian S Crandall and Amy Eshleman. A justification-suppression model of the expression and experience of prejudice. *Psychological bulletin*, 129(3):414, 2003.

[79] Lisa M. Leslie, David M. Mayer, and David A. Kravitz. The stigma of affirmative action: A stereotyping-based theory and meta-analytic test of the consequences for performance. *Academy of Management Journal*, 57(4):964–989, 2014. ISSN 1948-0989.

[80] Lea M. Petters and Marina Schröder. Negative side effects of affirmative action: How quotas lead to distortions in performance evaluation. *European Economic Review*, 130:103500, November 2020. ISSN 0014-2921. doi: 10.1016/j.euroecorev.2020.103500.

[81] Albena Neschen and Sabine Hügelschäfer. Gender bias in performance evaluations: The impact of gender quotas. *Journal of Economic Psychology*, 85:102383, 2021.

[82] Ward Ooms, Claudia Werker, and Christian Hopp. Moving up the ladder: heterogeneity influencing academic careers through research orientation, gender, and mentors. *Studies in Higher Education*, 44(7): 1268–1289, 2019.

[83] Dag W Aksnes, Shulamit Kahn, Rune Borgan Reiling, and Marte ES Ulvestad. Longitudinal evidence on norwegian phds suggests slower progression for women academics but not a leaky pipeline. preprint DOI: 10.31235/osf.io/pvx8q, SocArXiv, 2022.

[84] Madeline E. Heilman and Steven L. Blader. Assuming preferential selection when the admissions policy is unknown: The effects of gender rarity. *Journal of Applied Psychology*, 86(2):188–193, 2001.

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

[86] Madeline E. Heilman and Tyler G. Okimoto. Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology*, 92(1):81–92, 2007. ISSN 0021-9010. doi: 10.1037/0021-9010.92.1.81.

[87] Michele Belot, Madina Kurmangaliyeva, and Johanna Reuter. Gender Diversity and Diversity of Ideas. IZA Discussion Paper 16631, IZA, 2023.

[88] Friederike Mengel. Gender bias in opinion aggregation. *International Economic Review*, 62(3):1055–1080, 2021.

[89] Levke Henningsen, Lisa K. Horvath, and Klaus Jonas. Affirmative Action policies in academic job advertisements: Do they facilitate or hinder gender discrimination in hiring processes for professorships? *Sex Roles*, 86(1-2):34–48, 2021.

[90] Peter Kuhn and Kailing Shen. Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics*, 128(1):287–336, November 2012. ISSN 1531-4650. doi: 10.1093/qje/qjs046.

[91] Nattavudh Powdthavee, Yohanes E Riyanto, and Jack L Knetsch. Lower-rated publications do lower academics judgments of publication lists: Evidence from a survey experiment of economists. *Journal of Economic Psychology*, 66:33–44, 2018.

[92] Abdelrahman Amer, Ashley Craig, and Clémentine Van Effenterre. Decoding gender bias: The role of personal interaction. Discussion Paper 17077, IZA, 2024.

A Additional experimental materials

Table A1: 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 Sample characteristics across treatment conditions. Each participant evaluated six profiles.

A.1 Instructions to participants

The participants were presented with the following instructions.

The first screen

Dear Professor X,

Thank you for your help in this study.

Upon request of two anonymous research institutions, we conduct an evaluation of their recruitment processes. We study completed recruitments. We aim to verify whether the candidates who participated in these recruitments were objectively assessed.

We will ask you a few questions. The duration of the survey will not exceed 10 minutes. Thank you for your time.

As a token of our gratitude, we will award 20 participants with a smartwatch (Amazfit GTS 3 or Amazfit GTR 3). If you want to enter this lottery, we will ask about your email address at the end of the survey. Your responses will remain anonymous.

All questions about this study should be addressed to msmykgrape.org.pl.

The second screen

On the next screen we will show you biographical profiles of actual candidates in two recent recruitment processes. Both calls for applications were open field in economics. The openings were for an assistant professor position.

We relied on actual applications to construct the biographical profiles, which convey the key facts about each of the candidates, but preserve anonymity of those applicants. The names on biographical profiles are fictitious but names reflect the gender of the applicants.

We ask you to assess each application and to rank the applicants from the best to the worst.

[No Hiring Commitment Treatment] *Our institution values equality, it encourages especially women to apply.*

[Hiring Commitment Treatment] *Our institution values equality. In the case of scores being equal among the top two candidates, we are committed to hiring a woman.*

Naturally, the text following the squared brackets was randomized between participants. The text in the squared brackets was not displayed to the participants.

A.2 Biographical profiles

Below, we list all seven biographical profiles presented in narrative form. The first three profiles correspond to Excellent candidates, while the remaining four correspond to candidates who were of high quality. The biographical profiles are presented in female version. For the male version, all the words that denote gender were changed to their correct form. Note that the Polish language includes gender distinctions in pronouns, declination of nouns and adjectives, and conjugation of verbs.

Recall that the participants were informed that, in order to ensure anonymity to the candidates, the profiles were real, but the actual names were fictitious.

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 her work as a graduate student. She has taught tutorials with her supervisor during her PhD studies.

She provided three references, from Ph.D. 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 interests concern the political economy and inequality.

In addition to the job market paper, she has two revise & resubmit decisions at the top field journals (one coauthored with supervisors and one single-authored) and two more co-authored articles 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 her *alma mater* and previously visited institutions, including a Noble Prize winner and a foreign coauthor.

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

Her research interests concern asset pricing, both on the 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 complete co-authored articles. These papers were presented in high-field and general-interest conferences. In addition, Natalia has worked on two additional studies (one single-authored and one coauthored).

During the Ph.D. 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.

In addition to a single-authored job market paper, she has developed (co-authored) working papers that 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 the labor market.

In addition to the job market paper, she has developed one more single-authored study. The job market paper was presented at 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 the award for Best Paper 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 Ph.D. 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.

B Additional tables and figures

Table B1: 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. *, **, and *** indicate p-values lower than 0.1, 0.05 and 0.01.

Table B2: Excluding respondents whose recommendations did not vary

	Competences			Invitation		
	(1)	(2)	(3)	(1)	(2)	(3)
Female CV	-0.580	-1.111	-0.656	-0.0158	-0.0601**	-0.0198
	(0.763)	(1.226)	(0.975)	(0.0213)	(0.0299)	(0.0308)
HC	-0.416	-0.730	-0.0861	-0.0546**	-0.0809***	-0.0593**
	(0.896)	(1.150)	(1.005)	(0.0220)	(0.0277)	(0.0258)
Female CV × HC	0.542	1.262	1.121	0.0539*	0.122***	0.0764*
	(1.084)	(1.804)	(1.390)	(0.0297)	(0.0411)	(0.0423)
Context		0.0323			-0.0217	
		(1.204)			(0.0332)	
Female CV × Context		1.699	0.0437		0.140**	0.00964
		(2.277)	(1.574)		(0.0552)	(0.0427)
Context × HC		0.910	-0.999		0.0773*	0.0142
		(1.817)	(1.679)		(0.0449)	(0.0409)
Female CV × Context × HC		-2.174	-1.225		-0.210***	-0.0576
		(3.413)	(2.214)		(0.0752)	(0.0603)
Resume FE	Yes	Yes	Yes	Yes	Yes	Yes
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2439	2439	2439	2439	2439	2439
R-squared	0.267	0.267	0.267	0.338	0.341	0.338

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. *, **, and *** indicate p-values lower than 0.1, 0.05 and 0.01.

Table B3: Is gender of respondent a source of heterogeneity

	Competences		Invitation	
	(Men)	(Women)	(Men)	(Women)
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)
Resume FE	Yes	Yes	Yes	Yes
Respondent characteristics	Yes	Yes	Yes	Yes
Observations	3768	1871	3768	1871
R-squared	0.172	0.209	0.127	0.144

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

Table B4: Differences across disciplines

	Humanities	Social Sciences	Natural Sciences	Engineering and technical	Agricultural sciences	Medical and health sciences
Panel 1: Perceived competences						
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 and recruitment level in parentheses. *, **, and *** indicate p-values lower than 0.1, 0.05 and 0.01.

Table B5: Competence assessment 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. *, **, and *** indicate p-values lower than 0.1, 0.05 and 0.01.

Table B6: How competence assessment improves the probability of invitation recommendation

	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. *, **, and *** indicate p-values lower than 0.1, 0.05 and 0.01.

Table B7: On the role of outliers

	Probability of outliers		Competence w/o outliers			Invitation w/o outliers		
	Gender-specific	Joint distribution	Baseline	Drop observations	Drop participants	Baseline	Drop observations	Drop participants
Female CV	0.0133**	0.0135**	0.110	0.699	0.433	0.00409	0.00967	0.0144
	(0.00587)	(0.00559)	(0.476)	(0.439)	(0.439)	(0.0109)	(0.0105)	(0.0108)
HC	0.0136**	0.0130**	-0.550	0.233	-0.0674	-0.0385***	-0.0277**	-0.0204
	(0.00558)	(0.00531)	(0.636)	(0.563)	(0.556)	(0.0131)	(0.0128)	(0.0131)
Female CV × HC	-0.00956	-0.0150*	0.168	-0.198	0.00187	0.0253	0.0210	0.0199
	(0.00822)	(0.00782)	(0.670)	(0.604)	(0.598)	(0.0156)	(0.0152)	(0.0155)
Resume FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Respondent characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	5639	5639	5639	5499	5107	5639	5499	5107
R-squared	0.0310	0.0338	0.171	0.229	0.232	0.128	0.144	0.150

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. *, **, and *** indicate p-values lower than 0.1, 0.05 and 0.01.

C 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 is 1 or -1.[27] 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 C8. 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.[28]

Table C8: 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 C8 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 competence assessment, we can also study how higher scores from external experts translate onto the recommendation whether a candidate should be invited for an interview. For this, we reestimate Equation 3 but including competence assessment as an additional covariate. We present this estimates in Table B6.

[27] In fact, we would not be able to reject the null hypothesis for any value in the confidence interval.
[28] We restrict comparisons to cases when the resumes are different.

The coefficient on competence assessment indicates that increasing 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}) We estimate this coefficients using a linear probability model, which contains clustered standard errors at the level of recruitment and candidate. 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 B6 also includes separate regressions for subsamples of male and female profiles. The resulting coefficients are virtually identical. This result suggests that there is no gender heterogeneity when mapping the competence assessment to probability of invitation. Having similar coefficients further reinforces the result that evaluators are not gender biased.