






How Tinted Are Your Glasses? Gender Views, Beliefs and Recommendations in Hiring*

Anna Hochleitner ^{1,2}, Fabio Tufano ³, Giovanni Facchini ^{4,5},
Valeria Rueda ⁴, and Markus Eberhardt ⁴

¹Centre for Applied Research at NHH (SNF)

²Centre for Experimental Research on Fairness, Inequality and Rationality (FAIR)

³University of Leicester & London School of Economics and Political Science

⁴University of Nottingham & Centre for Economic Policy Research (CEPR)

⁵CESifo & Institute of Labor Economics (IZA)

This draft: August 1, 2025

Abstract

We study gender gaps at different stages of the hiring process, focusing on recommendations and recruitment. First, we document that women receive fewer ‘ability’ and more ‘grindstone’ recommendation letters in the academic job market. Next, we conduct two experiments — with academic economists and an online college-educated sample — analyzing both recommendation and recruitment stages. While recruiters overall favor women, consistent with efforts to diversify hiring, some groups of recommenders write suboptimal letters for them, undermining the initial advantage. Finally, letter choices correlate with gender views and are driven by strategic but erroneous beliefs about the effectiveness of different letter types.

JEL Codes: A11, C93, D90, J16.

Keywords: Gender, Recruitment, Diversity, Experiments.

*Certified random order of the authors ([Ray and Robson, 2018](#)). We thank participants at the Nottingham CeDEx brown bag seminar, seminars at the Norwegian School of Economics, and the FAIR & Max Planck Workshop in Bergen for constructive comments and suggestions. We have benefited greatly from discussions with and inputs from Malte Baader, Björn Bartling, Marianne Bertrand, Leonardo Bursztyn, Stefano DellaVigna, Nickolas Gagnon, Alex Imas, Lawrence Katz, Patrick Kline, John List, Katrine Løken, Paula Onuchic, Amanda Pallais, Evan Rose, Anya Samek, Sofia Shchukina, and Marie-Claire Villeval. We are grateful to Alex Imas for sharing relevant experimental material from [Bohren et al. \(2022\)](#). We also thank Cristina Griffa and Yuliet Verbel for their research assistance. We gratefully acknowledge funding from the University of Nottingham. The usual disclaimers apply. Corresponding author: fabio.tufano@leicester.ac.uk.

1 Introduction

Despite progress in educational attainment and labor force participation, gender gaps in hiring remain persistent — particularly in highly skilled occupations (e.g., [Bertrand et al., 2010](#); [Goldin et al., 2017](#); [Tesar, 2025](#)). These disparities continue to raise important questions about how hiring decisions are made and which aspects of the recruitment process may contribute to inequitable outcomes.

A key challenge is that recruitment processes are often complex, involving multiple stages and actors (e.g., recommenders, hiring committees, etc.). This complexity can conceal the mechanisms driving gender gaps, making it difficult to determine where disparities emerge. It is thus essential to understand how relevant features of these processes affect hiring outcomes, not least to inform the design of more effective policies.

In this paper, we focus on two stages of the recruitment process: the ‘recommendation’ stage, where candidates receive a reference letter; and the ‘hiring’ stage, where recruiters choose candidates based on recommendations and other characteristics. Recommendation letters are a standard component of selection procedures in high skilled occupations such as academia, law, and medicine, as well as in international organizations (e.g., [Avery et al., 2001](#); [Roth, 1984](#); [Coles et al., 2010](#)). They can also increase employment chances for low-skilled occupations (e.g., [Abel et al., 2020](#); [Heller and Kessler, 2021](#)). We find that gender matters differently at the recommendation and hiring stage. In many cases, recruiters even favor women over men with identical characteristics, in line with efforts to diversify the workforce and close existing gender gaps. However, some groups of recommenders write suboptimal letters for women, undermining these efforts. We identify gender views and strategic beliefs as key drivers of these gendered letter choices.

Existing research has examined how recommendations affect the quality of recruitment ([Pallais and Sands, 2016](#); [Abel et al., 2020](#)), but it has predominantly focused on whether a recommendation was sent or not. Notably, little attention has been paid to the content of recommendations or their impact on diversity. Studies analyzing the content of reference letters have documented gendered stereotyping, which correlates with hiring outcomes (see [Trix and Psenka, 2003](#); [Schmader et al., 2007](#), for pioneering contributions). However, these studies cannot fully rule out the possibility that stereotyping arises from unobserved heterogeneity among candidates. Furthermore, recommenders may emphasize different attributes in letters for men and women due to strategic considerations, anticipating different reactions from recruiters depending on the gender of the candidate. Finally, given the complexity of recruitment processes, these studies cannot quantify the direct causal impact of gendered differences in recommendations on hiring outcomes.

Our research addresses these gaps by combining three complementary studies, pinpointing the factors most likely to undermine diversity in recruitment. First, to provide a benchmark and motivate subsequent experiments, we build on [Eberhardt et al. \(2023\)](#) and present stylized facts from an ‘observational study’ of the junior job market for academic economists, a highly structured and competitive hiring process in which reference letters play a pivotal role. The institutional features

of this market allow us to compile a large, representative sample of reference letters (nearly 9,000 letters for 2,900 candidates) and recruitment outcomes for candidates seeking positions at research-intensive institutions. One limitation of this study is that recruitment outcomes are the result of complex factors (e.g., location preferences, interview performance, or department politics), which cannot be observed.

Second, we conduct an ‘academic survey experiment’ to assess the causal impact of gender differences in reference letters on recruitment outcomes. We sample over 1,000 participants from a target population of academic economists in research-intensive institutions. Participants are presented with a hypothetical candidate who obtained their PhD from a top-20 institution. They are randomly assigned to the roles of either recommender or recruiter. Recommenders choose one letter for the candidate from a set of three: one emphasizing ‘ability’ (e.g., cognitive skills), one emphasizing ‘grindstone’ (e.g., hard work), and a ‘neutral’ one. In addition, we elicit their beliefs about the effectiveness of these letters. Recruiters are shown details of a candidate alongside the three letters and asked to select the letter that would most likely prompt them to interview the candidate. The candidate descriptors are identical, except for gender, which is randomized across participants. Finally, for both recommenders and recruiters we assess how informed their views are about gender differences in recruitment documented in recent research. This study allows us to start exploring the causal effect of gender and reference letters on hiring outcomes and begins to unpack mechanisms explaining them. However, decisions are not incentivized and are based on a hypothetical candidate with fixed characteristics.

We thus conduct a third study, the ‘online experiment’ on Prolific. This study involves over 2,000 participants to validate our findings in a controlled, incentivized setting with a broader, college-educated population. We simulate a complete online labor market, randomly assigning participants to candidate, recommender, or recruiter roles. Candidates’ attributes are assessed through two general knowledge quizzes, allowing us to control for candidate productivity, cognitive skills, and effort levels. Recommenders, who observe only the results of the first quiz, choose one of the three letters (ability, grindstone, or neutral) to recommend the candidate. We also elicit their incentivized beliefs about the letter effectiveness. Recruiters hire one of three candidates (all of the same gender, which is randomized across recruiters) based on the letter type, with their payoffs tied to the unobserved performance of their chosen candidate in the second quiz. For both recommenders and recruiters, we assess to what extent they hold stereotyped gender views on effort and ability.

Each study examines a common research hypothesis through complementary yet distinct empirical designs. In line with the framework proposed by [Levitt and List \(2009\)](#) and adopted by [Alempaki et al. \(2019\)](#), our observational study, academic survey and online experiment act as *conceptual replications* of one another. While each study includes unique design elements, the convergent body of evidence strengthens the robustness and generalizability of the findings.¹

¹As [Maniadis et al. \(2017\)](#) point out, even a small number of well-executed replications, especially across different contexts and populations, serves to increase our confidence in the empirical results, often more than pre-registration/analysis alone (e.g., [Coffman and Niederle, 2015](#)).

In our observational study, we find that women receive disproportionately fewer ‘ability’ letters and more ‘grindstone’ letters than their male peers. Interestingly, these patterns are driven by the letters written by more senior recommenders. While grindstone letters correlate with worse placement outcomes for female candidates, women overall place better compared to their male counterparts.

Our experimental studies allow us to causally identify the effects of gender and letter type on hiring decisions. The results confirm that, at the recruitment stage, women are overall more likely to be hired. Moreover, we clearly observe that the most effective letter for recruitment is the ability one. This is particularly true for women, who benefit more from receiving ability letters and are more negatively affected from receiving grindstone letters.

In contrast to the observational results, when looking at the aggregate data, we do not find a gendered pattern in letter choices in the experimental studies. To understand this difference, we explore heterogeneities within our sample. Specifically, in both studies, we elicit the respondents’ gender views and find that these views correlate strongly with their letter choices. Those with uninformed or stereotyped views are less likely to choose ‘ability’ letters for women, whereas those with more informed or non-stereotyped views do the exact opposite. Strikingly, senior, older, or male participants are more likely to hold uninformed gender views, characteristics that also correlate with the likelihood of being a recommender in our observational study. Moreover, we explore the role of motivated beliefs. We show that gendered letter choices are largely driven by recommenders’ erroneous beliefs about the effectiveness of different letter types for women versus men. Most recommenders select the letter they believe maximizes hiring chances but underestimate the value of an ‘ability’ letter for women. These beliefs, which align with gender views, suggest motivated reasoning (e.g., [Bénabou, 2015](#); [Stötzer and Zimmermann, 2024](#)).

A key advantage of our experiments is that they allow us to apply the classic Kitagawa-Oaxaca-Blinder decompositions ([Kitagawa 1955](#); [Oaxaca 1973](#); [Blinder 1973](#)) to break down gender differences in recruitment outcomes into a ‘direct’ and ‘systemic’ components (see [Bohren et al., 2022](#); [Baron et al., 2024](#)). This decomposition helps evaluate how the different stages of the hiring process interact to shape gender gaps. The ‘direct’ component captures differences in how male and female candidates with identical quality signals are treated – specifically, gender differences in hiring outcomes for men and women with the same reference letters. The ‘systemic’ component, in contrast, reflects recruitment disparities that arise when recruiters fail to account for gendered differences in reference letters for candidates with identical letters. Our findings show that the direct component is *positive* for women: they are *more* likely to be recruited than men with identical qualities. This result aligns with the goals of widespread DEI (Diversity, Equity and Inclusion) initiatives aimed at improving women’s recruitment outcomes. However, when assessing the overall outcomes for women, we find that the positive effect of the direct component is undermined by the systemic one. Recruiters’ failure to recognize that women receive sub-optimal letters attenuates the direct component. In cases where recommenders hold uninformed or stereotyped gender views, the systemic component even reverses the direct one, leading to a net negative hiring gap for women.

Our results have potential policy implications for sectors aiming at recruiting more women. We find

that tackling direct discrimination by recruiters—a frequent target of DEI policies—is unlikely to be sufficient to close gender gaps in recruitment. Women are more likely to be disadvantaged at the recommendation stage, where more subtle gendered patterns may appear. Moreover, a major factor explaining why women receive suboptimal reference letters is recommenders’ erroneous beliefs about the effectiveness of different letter types. Addressing these misconceptions and correcting beliefs could improve diversity in hiring at a low cost.²

Related Literature and Our Contribution. Existing research has extensively documented the presence of discrimination in recruitment, both through field data (e.g., [Goldin and Rouse, 2000](#); [Riach and Rich, 2002](#)) and correspondence studies (e.g., [Bertrand and Mullainathan, 2004](#); [Jacquemet and Yannelis, 2012](#)). Recent studies have also explored the role played by recommendations in driving these differences. Although recommendations can improve efficiency, they can also exacerbate gender gaps as men and women differ in their likelihood of being recommended (on efficiency, see, [Pallais and Sands 2016](#); [Abel et al. 2020](#); on gender gaps, see [Beaman et al. 2018](#)). Our paper contributes to this literature by investigating the actual content of recommendations rather than focusing only on whether a recommendation was sent or not. Contemporaneous research investigates this theoretically and, in an application, relies on Large Language Models to illustrate the impact of gendered recommendations on hiring ([Bohren et al., Forthcoming](#)). In contrast, we combine an observational study using real-world references and recruitment outcomes, along with two experiments that analyze reference messages chosen by a population of academics and college-educated individuals. This setup enables us to explore mechanisms in detail, in particular the role of gender views and strategic considerations.

Exploring such mechanisms links us to a body of work that has investigated behavioral channels underlying discrimination in hiring. Specifically, studies have examined the role of idiosyncratic preferences (e.g., [Oreopoulos, 2011](#); [Kuhn and Shen, 2013](#); [Weber and Zulehner, 2014](#)), evaluator’s gender (e.g., [Bagues and Esteve-Volart, 2010](#); [Deschamps, 2023](#)), incentives (e.g., [Szymanski, 2000](#); [Parsons et al., 2011](#)), uncertainty (e.g., [Hendricks et al., 2003](#)), and beliefs (e.g., [Glover et al., 2017](#); [Barron et al., 2024](#)). Our paper contributes to this literature by showing how gender views and strategic considerations – consistent with ‘motivated’ beliefs (e.g., [Eyting, 2022](#); [Stötzer and Zimmermann, 2024](#)) – drive discrimination. We show that recommenders who believe they are acting in the candidate’s best interest may unintentionally harm their prospects. Additionally, we extend recent research on discrimination in layered systems, which has developed empirical strategies to identify discrimination arising at different stages of a process (e.g., [Kline et al., 2022](#); [Bohren et al., 2022](#); [Baron et al., 2024](#); [Jassal, 2024](#); [Shaffer and Harrington, 2024](#)). In particular, our analysis uncovers the importance of gender views and strategic considerations on the *recommendation* rather than the *recruitment* side in explaining differential treatment.

Finally, our paper offers novel insights into the study of the role of gatekeepers in explaining gender disparities in academia, especially in math-intensive subjects such as economics (e.g., [Ceci and](#)

²For an example of an intervention targeting beliefs in an academic context, see [Boring and Philippe \(2021\)](#).

Williams, 2009; Breda and Ly, 2015; Sarsons, 2017; Lundberg and Stearns, 2019; Hengel, 2022).³ Previous studies have used observational data to document gendered patterns in reference letters (e.g., Correll et al. 2020; Eberhardt et al. 2023; Baltrunaite et al. 2024). Our experimental results provide causal evidence on the role of gendered recommendations in recruitment outcomes. Moreover, by highlighting the importance of recommenders’ beliefs, our work points to the role of diffusing information about the efficacy of different types of recommendations as a policy option to diversify the profession.

The remainder of this paper is structured as follows. Section 2 presents observational results from the academic job market that motivate our experiments. Section 3 outlines the design of our two experimental studies. Section 4 reports descriptive statistics. Sections 5 and 6 provide results separately for recruiters’ hiring decisions and recommenders’ letter choices. Section 7 brings both sides of the market together to evaluate the overall effect on hiring. Section 8 concludes.

2 Observational Study: Insights from the Academic Job Market

We start by setting the stage with our ‘observational study’ that draws insights from patterns observed in the junior job market for academic economists. We take advantage of a unique collection of reference letters assembled by Eberhardt et al. (2023) on the basis of applications for assistant professor positions received by a UK research-intensive institution during the 2017-2021 period.⁴ This collection spans almost 9,000 reference letters, written by over 4,000 recommenders in support of 2,900 candidates.⁵

The reference letters are transformed into quantitative data using the *term frequency inverse document frequency* (tf-idf) for each word, which quantifies the importance of a term in each letter compared to its prevalence in all letters.⁶ The focus is on the final paragraph of each reference letter because it summarizes the candidate’s skills, strengths, and prospects. To capture the ‘sentiment’ of each final paragraph, ‘bags of words’ emphasizing different attributes such as natural talent (‘ability’) or perseverance (‘grindstone’) were created (e.g., Trix and Psenka, 2003; Schmader et al., 2007). This resulting categorization was validated by a questionnaire sent to academic economists based in UK research-intensive universities.

We assign the reference letters’ final paragraphs (henceforth ‘reference letters’ for brevity) to mutually exclusive ‘letter types’. We do so by classifying the letters depending on the most emphasized

³Although women’s representation in economics has stagnated in recent years (e.g., Tesar, 2025), there are occasional signs of change as, for instance, positive gender gaps in economics fellowships and awards (e.g., Card et al., 2022, 2023).

⁴Our observational study received ethical clearance by the Nottingham School of Economics Research Ethics Committee in March 2020. Confidential data was handled with the approval of the University of Nottingham ethics board and in accordance with all University of Nottingham regulations, as well as an agreement signed on April 13, 2020, between the university and the job application platform through which the data were collected.

⁵In order to match the experimental setting, we limit the analysis to candidates who either graduated less than a year before submitting their application or will graduate in the Summer after the job market. They make up the vast majority of candidates in the sample.

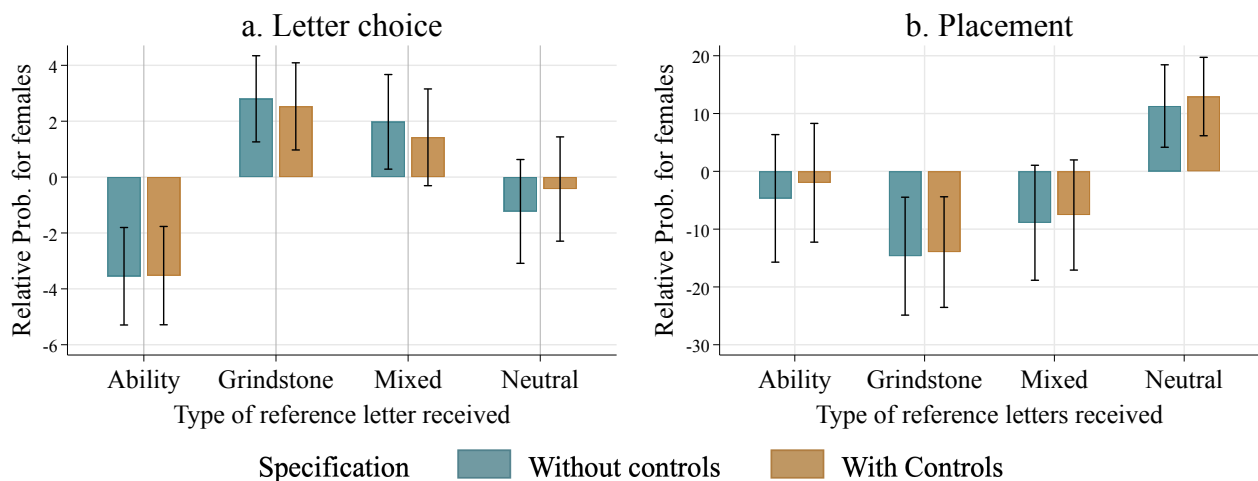
⁶Baltrunaite et al. (2022) use embedding methods to analyze letters, obtaining very similar results.

attributes. This discrete classification differs from the continuous outcome studied in [Eberhardt et al. \(2023\)](#) and informs the design of our experiments, while also enabling closer comparison of results. Letters classified as *ability* (respectively, *grindstone*) type are those containing a frequency of ‘ability’ (‘grindstone’) attributes above the full sample median tf-idf. This naturally leads to two further types: a *neutral* type with letters containing neither attribute and a *mixed* type with letters containing both attributes.

In addition to the reference letter data, we use information on both candidates (e.g., gender; ethnicity; PhD-granting institution; publication record; research field; initial placement outcome) and recommenders (e.g., gender; institutional affiliation and RePEc ranking; seniority; time since their PhD award; number of reference letters in the sample they wrote) – see Appendix [A.1](#) for details.

We begin by examining whether recommenders write letters of different types for men and women using a multinomial logit regression. Our dependent variable equals 0, 1, 2, or 3 if the candidate was given a neutral, grindstone, ability, or mixed (ability *and* grindstone) reference letter, respectively. The primary variable of interest is a gender indicator and our estimates represent the predicted probability that a woman received a specified letter type relative to a man. Figure 1.a reports the estimated marginal effects from models with and without the full set of controls, respectively (see Appendix Table [B.3](#) for more details). Women are 3.55 (3.53 with controls) percentage points less likely to receive an ability letter and 2.30 (2.80 with controls) percentage points more likely to receive a grindstone one. In contrast, the gender of the candidate does not affect the probability of receiving a neutral letter which emphasizes neither attribute. Finally, women have a higher probability of receiving a mixed letter emphasizing both attributes, but this result is not robust to the inclusion of

Figure 1: Letter choice and placement in the job market for academic economists



Note: Panel a presents marginal effects from a multinomial logit regression; the y-axis represents the predicted probability of a female (relative to a male) candidate to receive a given letter type, reported in percentage points. Sample: 8,624 letters for 2,881 candidates (25% ability, 20% grindstone, 25% mixed and 30% neutral letters). Panel b reports marginal effects from a linear probability model; the y-axis shows the predicted probability of a female (relative to a male) candidate to secure a top-200 placement with a given letter type, reported in percentage points. Sample: 1,862 candidates who placed in academia; the outcome mean is 28%. Here, controls for recommenders and letters are averaged at the candidate level. Both panels plot specifications with and without full controls alongside their respective 90% confidence intervals.

control variables.

We explore heterogeneity in the results to ascertain whether these are driven by specific *types* of recommenders. Appendix Table B.4 provides split sample results by academic rank, seniority (PhD cohort before 2004) and gender. We observe that senior recommenders — be it defined by academic rank or PhD cohort — exhibit strikingly more marked gendered patterns in their letters. There are minor differences by recommender gender, though these appear to be driven by the lower seniority of female recommenders. In Section 6, we will examine the role of sample composition in greater detail, drawing comparisons with the heterogeneity observed in the experimental studies.

Next, we explore whether the letter type differentially correlates with top academic placement for male versus female candidates. For this, we use data aggregated at the candidate level: to define the letter types, we sum the tf-idf statistics for all reference letters received by each candidate. This captures the intensity of the vocabulary for each attribute across the set of letters for each candidate. We then apply the approach discussed above to categorize the overall tone of the letters received by each candidate. We end up with a sample of 1,862 candidates, who placed in an academic institution at assistant professor or postdoc level.⁷ We use a linear probability model of placement in a RePEc top-200 institution, regressed on indicators for each letter type, as well as a female interaction term for each of letter type. Figure 1.b presents the results (see Appendix Table B.3, for details). Again, we distinguish models with and without controls. We can see that with a neutral letter (the omitted baseline) women have a 11.30 (12.94 with controls) percentage point higher probability of securing a placement in the top-200 institutions than men. However, while an ability letter has no differential effect across genders, women with a grindstone letter are 14.68 (13.97 with controls) percentage points less likely to secure a top placement. Finally, women with mixed letters emphasizing both attributes are less likely to secure top placement, but this effect is estimated imprecisely and is not statistically significant.⁸

The data in the observational study do not contain any individual information on recruiters. Therefore, we cannot assess whether hiring choices are driven by specific types of recruiters, as we did above for letter choice. In contrast, our experiments capture both sides of the hiring process, allowing us to address this question.

Overall, the patterns in our observational study suggest that women are disproportionately more likely to receive grindstone letters and less likely to receive ability ones. Letter choices correlate with successful academic placement in a gendered way. Despite the differences in the econometric strategies, these results align with those of Eberhardt et al. (2023) — who, by contrast, used a continuous tf-idf attribute measure — thereby reinforcing confidence in the robustness of our findings. While describing salient gendered patterns in the junior job market for academic economists,

⁷Controls for recommenders and letters are averaged at the candidate level. Appendix Table B.7 and Figure B.1 provide additional results excluding postdoc positions or predicting RePEc top 100 placement.

⁸In Appendix Table B.8, we show that the gendered results for ability and grindstone are qualitatively unchanged if we assign the mixed letter to either one or the other category. As Appendix Table B.9 shows, studying the effect of letter type on placement without allowing it to differ by gender leads to the impression that (a) letter types do not matter, and (b) women have an advantage over men (column (3)). Using gender-letter type interaction terms instead reveals the gendered effect of grindstone letters, which eliminates the female advantage discussed above.

these results remain correlational. Despite efforts to control for observable characteristics, it is difficult to assess whether the differences in the type of letter received stem from gender views or from unobserved characteristics. Moreover, the reference letters are only one element of a complex, multi-dimensional recruitment process. They help candidates secure interviews, but many other factors such as communication skills, recruiters’ field preferences, and departmental politics, shape the eventual hiring outcome. For these reasons, we take our empirical exploration to experimental settings to pinpoint causal links and identify the underlying mechanisms explaining the observed patterns.

3 Experimental Studies: Design and Procedures

We conduct two studies to isolate the causal effect of reference letter types on hiring outcomes, the effect of the candidate’s gender on letter types and how these two aspects interact in shaping gender gaps in hiring. Moreover, the studies allow us to shed light on explanatory mechanisms of the observed effects, such as the role of gender views, inherent differences between male and female candidates, and strategic considerations (e.g., whether recommenders choose the letter they believe maximizes their candidate’s success).

The first study relies on a novel survey experiment carried out with academic economists. The second consists of an online experiment featuring a labor market with college-educated participants recruited on Prolific. Both studies received ethical approval from the Nottingham School of Economics Research Ethics Committee in May 2022.

3.1 Academic Survey Experiment

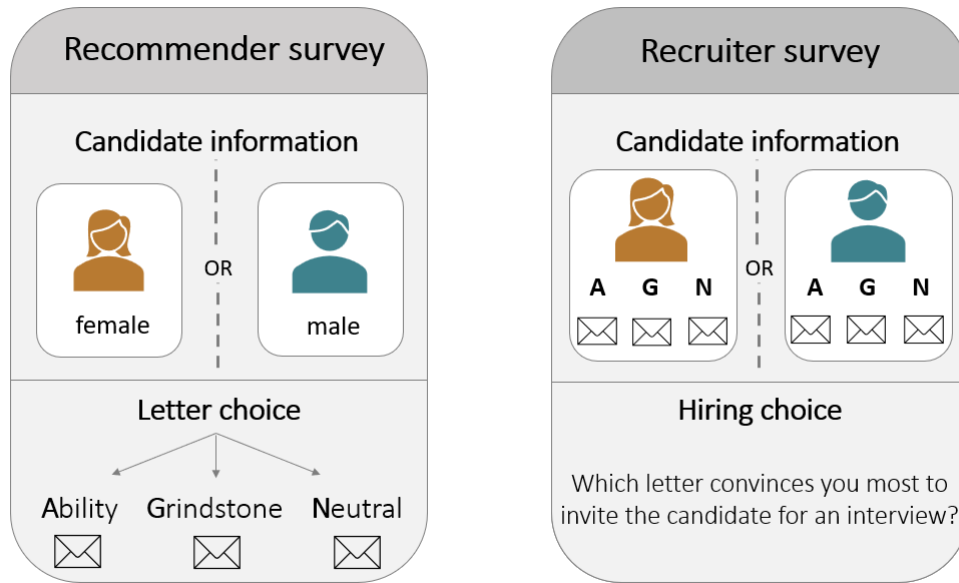
We designed the academic survey experiment (*‘academic survey’* for short) to capture the relevant features of the junior job market for academic economists. In a nutshell, academics are asked to either choose reference letters for PhD candidates or invite them for an interview based on a hypothetical job market application package.

3.1.1 Design

Figure 2 presents a schematic overview of the survey structure. All instructions can be found in Appendix C.1. Participants are randomly assigned to either the *recommender survey* (left panel) or a *recruiter survey* (right panel). All participants see a short biography of one hypothetical candidate including information about their name, PhD institution, job market paper, and research field. As standard in this literature, the gender of the candidate is varied *between* subjects to reduce experimental bias (e.g., [Bertrand and Duflo, 2017](#)), and is signaled to participants through the first name of the candidate (e.g., [Lippens et al., 2023](#)).⁹

⁹We ran a pre-test to select names and chose *Julia* and *Thomas*; both names were respectively rated as unequivocally female and male but not perceived differently with respect to ethnic and socioeconomic background.

Figure 2: Survey structure



Note: We vary the candidate’s gender (as implied by their first name) between subjects. Recruiters see all three letters and decide which one convinces them most to invite the candidate for an interview.

In the **recommender survey**, participants are shown one candidate (either female or male) and asked to select one of three summary recommendation texts for the candidate. The order in which participants view the recommendation texts is randomized between participants. These texts are written to mimic the final paragraphs of reference letters, which usually summarize the overall view of the recommender on the candidate’s strength and potential.¹⁰ Focusing on the summary recommendation allows us to keep the survey manageable (the text is about 140 words).

The summary paragraphs are identical except for three sentences that are either neutral, stress the candidate’s natural ability (e.g., ‘bright and creative’), or emphasize grindstone attributes (e.g., ‘extremely hardworking’). As the effect of mixed letters was very similar to grindstone letters in the observational study, we decided to drop this category in the experimental surveys. This allows us to streamline the design and increase the sample size in each treatment cell. Independently of the survey assignment, all participants had the option to view the full, three-page version of each letter (see Appendix C.2).

Next, we elicit beliefs on which letter is perceived to be the best by most other participants. In particular, we ask which letter maximizes a candidate’s chance of being invited to an interview. The beliefs question is incentivized; if recommenders choose the letter that is preferred by most other participants, they increase their chance of winning a voucher (see Section 3.1.2 for more details).

In the **recruiter survey**, participants receive the profile of one candidate (either female or male) along with all three letters (i.e., grindstone, ability, or neutral). Again, the order in which letters are presented is randomized between participants. Recruiters then choose which one out of the three

¹⁰See Appendix C.1.2 for the exact phrasing of the paragraphs.

letters would most likely lead them to invite the candidate for an interview.¹¹ We chose to ask about individual preferences for invitations to interviews instead of final hiring prospects to abstract from department politics and strategic considerations in the hiring process. Nevertheless, there might still be some strategic concerns that vary across recruiters with different backgrounds. For this reason, we control for recruiter characteristics in all analyses. For simplicity, in the rest of the paper, we still refer to the recruiters’ choice as their ‘hiring decision’.

For all participants (recommenders and recruiters), we collect additional information, including demographics (i.e., gender, academic seniority, year of PhD, the country and the RePEc ranking of their institution), previous experience of writing reference letters and recruiting from the economics job market. In addition, to ascertain individual awareness of gender issues in academia, we ask about their familiarity with research on the topic, measured through self-reported knowledge of five recent research papers.¹² We use this variable as a proxy for how informed are participants’ gender views and test how the latter shape letter choices and hiring decisions.

3.1.2 Procedures

To construct the target population for our survey, we collected the email addresses of academic economists from publicly accessible institutional websites. We chose institutions that regularly engage with the international job market for economists, resulting in a population of 13,185 academics based in 21 countries (see Appendix A.2 for more details).

The survey was programmed in [Qualtrics \(2005\)](#), pre-registered (AsPredicted #101261), and conducted in July 2022.¹³ To incentivize participation, we donated £2 to UNICEF for each completed survey and respondents who finished the study received a lottery ticket to win one of fifty £50 Amazon vouchers. If they answered the beliefs question correctly, they received an additional 5 tickets. Every lottery ticket was equally likely to be drawn. In total, we received 1,020 complete responses (506 in the recommender and 514 in the recruiter survey), which is in line with our pre-specified target sample of 1,000 participants. We address representativeness and self-selection into the survey in Section 4. The median response time was 6.8 minutes.¹⁴

¹¹In addition, we ask recruiters how likely they would invite a candidate with a given letter on a scale from 0 to 10. The median score for each letter was 9 out of 10, indicating that our hypothetical candidates were perceived as very strong. Given the lack of variation in this measure, we focus on the relative hiring choice in our analysis.

¹²We asked participants about their knowledge of the following five papers: [Wu \(2020\)](#); [Dupas et al. \(2021\)](#); [Koffi \(2021\)](#); [Hengel \(2022\)](#); [Eberhardt et al. \(2023\)](#).

¹³Both experimental studies (see 3.2.2 below) were pre-registered following common practice. Given ongoing debates on the scope of pre-registration (e.g., [Imai et al., 2025](#)), we registered design-level commitments without a fully specified pre-analysis plan. The data stem from the first and only study implementations, with all variables reported; experimental data are publicly available at [Hyperlink TBA]. Despite the condensed pre-registration, our reproducible methodology, the studies’ role as conceptual replications, and the convergence of results reinforce the robustness of our inferences.

¹⁴All regressions presented in this paper are robust to dropping the fastest and slowest 5 (10)% of responses. If anything, effects are estimated more precisely when winsorizing on response time. To avoid redundancy and conserve space, we do not report these robustness checks in the appendix, but the results are available upon request.

3.2 Online Experiment

The academic survey allows us to observe experts’ behavior on both sides of the job market in a controlled environment. As references are widely used in labor markets (particularly, for highly skilled professions), a natural question is whether the patterns observed hold for a broader population facing choices that are not hypothetical, but have real consequences.

To answer this question, we designed an online experiment that complements the academic survey in several ways. It provides monetary incentives for all decisions, which thus have real consequences for participants. This setting also strengthens experimental control, since we are able to elicit candidates’ actual productivity, cognitive skills, and self-reported effort. To ascertain individual gender views, we ask participants about their opinions on whether men/women are characterized more by ability, effort, or possess both traits equally. For this experiment, we recruit college-educated individuals from the online platform Prolific.

3.2.1 Design

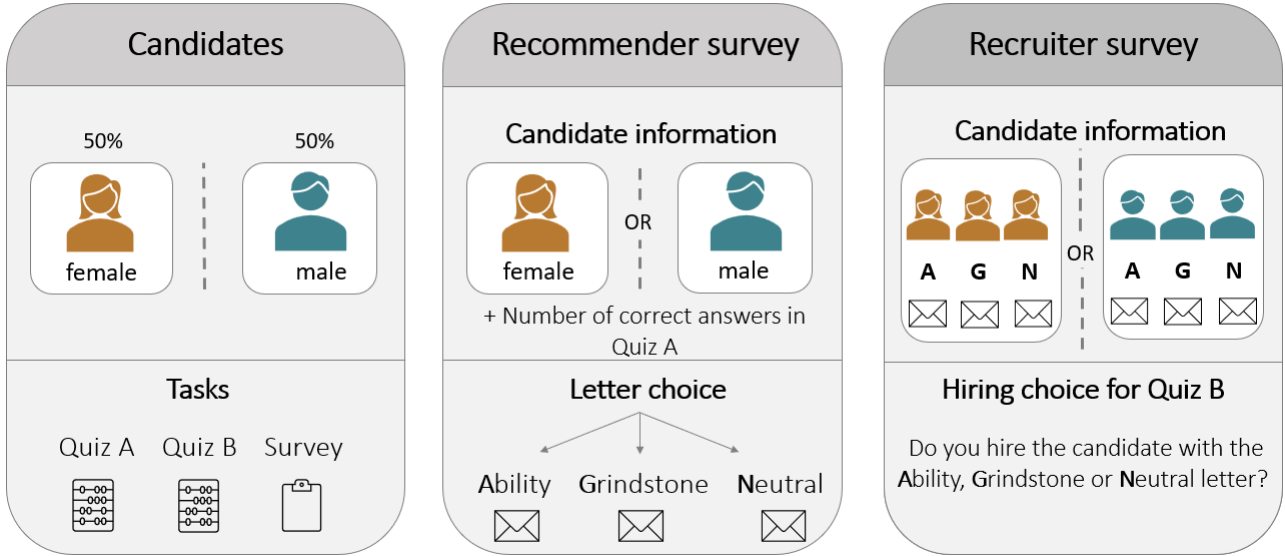
In the experiment, participants are part of an online labor market and are assigned to three different roles: *candidates*, *recommenders*, and *recruiters*. The setup builds on [Bohren et al. \(2022\)](#) and is designed to study the effects of job recommendations. The basic structure is as follows: Candidates complete a range of tasks that generate an imperfect signal of their productivity. Recruiters want to hire candidates for a task, but do not know their productivity. Recommenders observe an informative signal and send a recommendation message – which we call ‘letter’ – to recruiters. We again use a between-subjects design to minimize experimenter demand effects: recommenders and recruiters are shown information for either female or male candidates. We provide a more complete description of participants’ incentives and choices below, while [Figure 3](#) offers an overview of the design of the online experiment. All instructions can be found in [Appendix C.3](#).

Candidates have to solve two quizzes: Quiz A and B. Each quiz consists of ten incentivized questions about finance, mathematics and history. To reduce the number of performance levels to evaluate, we focus on candidates who correctly answered either 5, 6, or 7 Quiz-A questions (13% of all candidates). Given that the average number of correctly solved questions was 3.48, these are high-performing individuals (ranked between the 86th and 99th percentile). We restricted the sample to high-performing candidates to align the quality of candidates with our highly positive reference letters and to reflect the context of recruitment for high-skilled positions.

We also collect proxies for grindstone and ability characteristics. In particular, we survey participants on how much effort they put into the quiz and elicit cognitive skills using a standard Cognitive Reflection Test (‘CRT’, [Thomson and Oppenheimer, 2016](#)). Although our effort measure is self-reported – and thus imperfect – we find that it significantly correlates with time spent on the quizzes (the correlation coefficient is $r = 0.46$, $p < 0.001$).

Recommenders observe the performance of one candidate only in Quiz A, as measured by the number of correctly answered questions, along with their gender. While a high correlation between

Figure 3: Experimental design



Note: We vary candidate's gender between subjects. Recommenders see one of three performance levels. Recruiters see all three letters and have to decide which letter convinces them most to hire the candidate.

quizzes is likely, recommenders are thus not perfectly informed about the potential of a candidate. First, half of all recommenders are assigned a woman, the other half a man. Next, for each assignment, recommenders are equally likely to be assigned a candidate who answered 5, 6, or 7 questions correctly. They are then asked to choose one of three letters (i.e., grindstone, ability, or neutral) to recommend the candidate to recruiters. The order in which letters are presented is randomized between participants. The exact phrasing of the letters can be found in Appendix C.3.2. If the candidate is hired, the recommender receives a bonus (see Section 3.2.2 for details on incentives). Recommenders thus earn a monetary reward for sending the strongest possible letter. To enhance task engagement and experimental control, our design abstracts from reputational concerns that real-world recommenders may face in repeated interactions. This simplification is particularly valuable in our online setting, as it allows us to capture key behavioral drivers (as shown by our convergent body of evidence) and more cleanly identify the role of beliefs and biases in letter choices. In addition to the letter choice, we elicit incentivized beliefs about the most successful letter.

Recruiters are shown three different candidates (as indicated by different candidate IDs). The three candidates have the same gender, but different letters. Recruiters lack any information about the quiz performance and can thus only rely on the letter and/or gender to make a hiring decision. While they do not know that candidates solved either 5, 6, or 7 questions correctly and are high performers, all three letters are overall very positive and strongly recommend the candidate. Recruiters are incentivized to choose the best candidate, as their bonus depends on how many questions the candidate answered correctly in Quiz B.¹⁵

¹⁵We again elicit an absolute hiring choice in addition to the relative choice between candidates. Recruiters are shown one candidate and have to decide whether to hire that candidate. If hired, the candidate's performance defines the recruiter's bonus ranging between £0 and £1. If the candidate is not hired, recruiters receive a fixed payment of £0.40. In line with the academic survey, we focus on the relative hiring choice in our analysis.

To sum up, the experiment follows a 2x3 design for both recommenders (male/female candidate x 5/6/7 correct answers in Quiz A) and recruiters (male/female candidate x ability/grindstone/neutral letter). For both recommenders and recruiters, we collect a measure of their gender views. In line with research in social psychology on gender brilliance bias (e.g., [Del Pinal et al., 2017](#); [Bian et al., 2018](#); [Storage et al., 2020](#)), we ask them if they believe that men/women are successful due to effort, ability or a combination of both. This helps us to understand the role of gender views in driving potential differences in both letter choice and hiring decisions for male and female candidates. Moreover, we collect data on how both recommenders and recruiters perceive the candidate, the skills required to perform well in the quizzes, their demographic background (i.e., age, gender, education), their self-reported risk ([Dohmen et al., 2011](#)), and strategic sophistication, which we measured through a beauty contest game ([Nagel, 1995](#)). Demographics as well as perceptions and preferences are used as additional controls when examining letter choice and hiring decisions.

3.2.2 Procedures

The experiment was programmed in [Qualtrics \(2005\)](#) and pre-registered (AsPredicted #102500). Participants were recruited via Prolific in July 2022. We restricted the study to participants in the US and the UK to minimize language barriers. To enhance comparability with the academic sample and due to the experiment involving hiring decisions, we required participants to have at least an undergraduate degree. We collected data for 100 candidates, as well as 959 recommenders and 959 recruiters. We then match several recommenders to one candidate and we do so in a way that ensures a balance across candidates' gender and performance levels.

All decisions and elicited beliefs were incentivized. Candidates received £0.10 for each correctly solved quiz and CRT question. Recommenders received a bonus of £0.50 if their candidate was hired, plus respectively £0.50 if their beliefs about the recruiters and the candidate were correct. Recruiters received £0.10 for each Quiz B question their hired candidate solved correctly plus £0.50 if their beliefs about the candidate were correct. In addition, participants received a show-up fee, resulting in average earnings of £2.06.¹⁶ As participants took on average 7.2 minutes to complete the experiment, this is equivalent to an hourly wage of £17.20.¹⁷

4 Descriptive Statistics

We first describe and compare the different samples before presenting our results. Figure 4.a illustrates the target population that was invited to participate in the academic survey. It consists of all academic economists (only permanent, full-time, research & teaching faculty) in institutions that regularly engage with the international economic job market. Figure 4.b shows the demographics of individuals who participated in the academic survey.

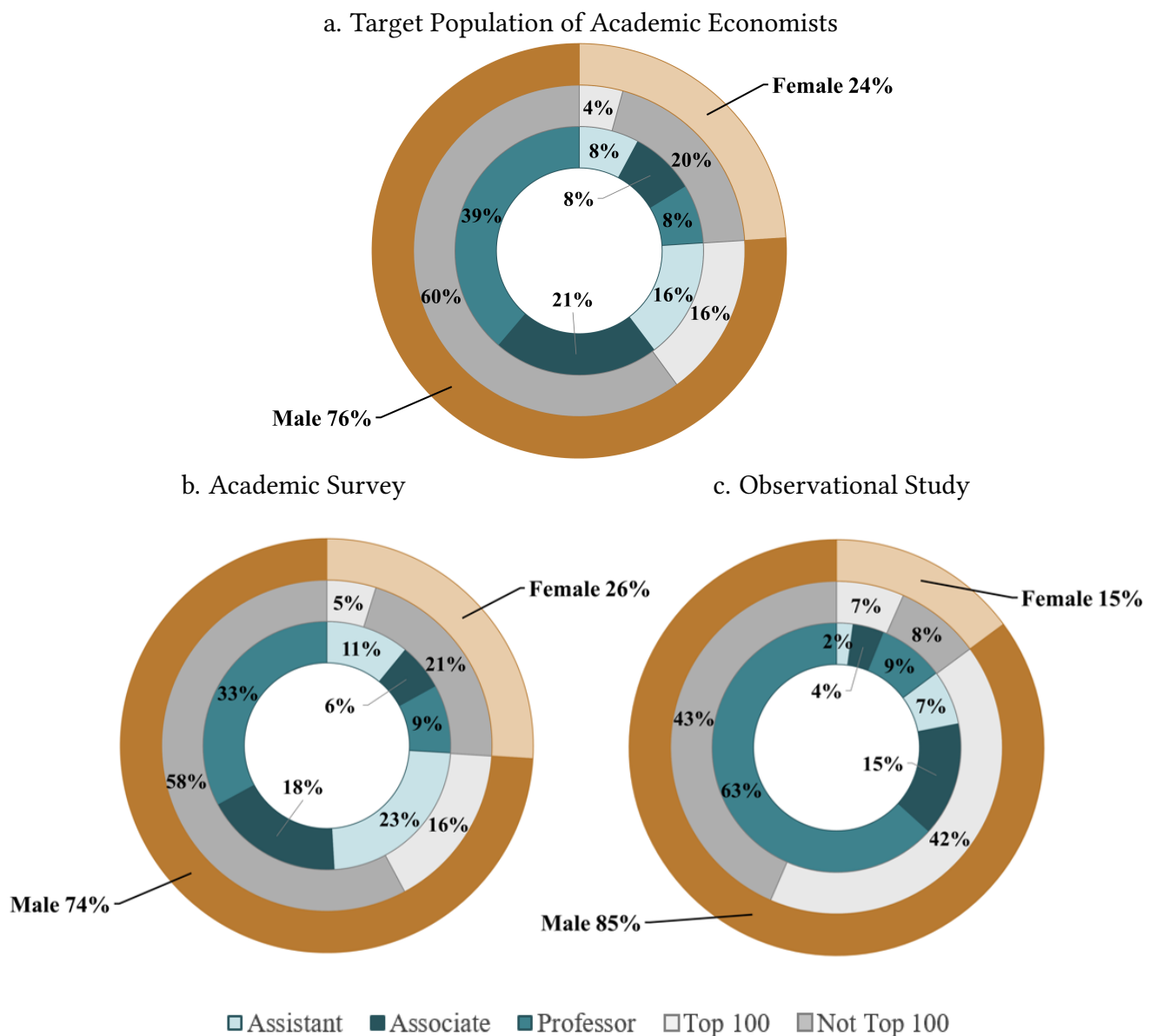
We see that these two groups are highly comparable in their gender, age, seniority, and institutional

¹⁶The average payments by roles were £2.73 for workers, £2.14 for recommenders and £1.91 for recruiters.

¹⁷Again, all regressions are robust to winsorizing on response time and dropping the fastest and slowest 5(10)% of responses.

ranking distributions. In contrast, there are differences with the sample of academic economists who write the reference letters used in our observational study (Section 2). These recommenders, whose baseline characteristics are illustrated in Figure 4.c, are less likely to be women (15% in the observational sample versus 24% in the target population), work at more prestigious institutions (49% in the top-100 versus 20% in the target population), and hold more senior positions (72% versus 47% in the target population). These differences are in part expected, since academic advisors tend to be more senior, and PhD candidates may reach out to academics in more prestigious institutions to write their letters. Interestingly, independently of seniority and rank, female academics are less likely to be recommenders. To account for these differences, in our analysis we re-weight the survey sample to reflect the demographic characteristics of recommenders in the observational sample.

Figure 4: Demographic characteristics across samples



Notes: Panel a shows the demographic composition of academic economists in institutions that regularly engage with the job market and were thus invited to participate in our study. Panel b shows the demographic composition of individuals who decided to participate in the academic survey. Panel c shows the demographic composition of recommenders from the observational study.

Participants in the online experiment are drawn from a different population that does not consist of academic economists but is still college-educated (see Appendix Table B.11 for demographic characteristics). The experimental sample is thus different enough to take a first step to gauge the generalizability of our findings, but similar enough to draw a meaningful comparison to the academic survey. Besides profession, a striking difference with the academic samples is that we have a balanced gender composition. In the main text, we do not weight the experimental sample. In the real world, however, men are often more likely to be in positions of decision power and, if gender biases are not distributed equally within society, our results are a more conservative estimate for any gender differences. We confirm the robustness of all our results to weighted regressions mimicking the gender composition in the observational sample. Indeed, results become more pronounced when using weights.

As mentioned in Section 3, we collect data on gender views in both the academic survey and the online experiment to explore their role in shaping differences in reference letters. To do so, we compare behavior across different sub-samples. In the academic survey, we use participants' knowledge of research on gender differences within academia as a proxy for gender views. Given that this specific sample of academics may have some familiarity with standard measures of gender views, we opted for an alternative measure that is more natural and suffers less from desirability biases. Our working assumption is that academics who know more about gender research are, on average, more interested in the topic and tend to hold more progressive gender views. Our sample of academic economists is roughly split equally into people who know (47.94%) and who do not know this literature (52.06%). We refer to those academic economists who know this gender literature as having 'informed' views and those who do not know the literature as having 'uninformed' views.¹⁸

In the online experiment, as a proxy for gender views, we followed research in social psychology (e.g., Del Pinal et al., 2017; Bian et al., 2018; Storage et al., 2020) and asked participants about their opinions on whether men and women are characterized more by ability, more by effort or possess both traits equally.¹⁹ We define participants as having 'stereotyped' views if they agree that women are more characterized by effort/men are more characterized by ability, and as having 'non-stereotyped' views otherwise (see Appendix Figure B.4). Using this definition, we find that 27.79% of the sample have stereotyped views, while 72.21% have non-stereotyped views.²⁰

¹⁸We find that our measure of gender views correlates in intuitive ways with demographic characteristics such as gender or seniority, with male and older economists more likely to hold uninformed views (see Appendix Figure B.3). However, we are unable to explore whether it also correlates with the respondents' research fields, as this information was not elicited in our survey.

¹⁹Recruiters are directly asked their own opinions. Recommenders are asked what opinion they think most recruiters have. Appendix Figure B.4 shows that answers are very similar.

²⁰Note that the group of non-stereotyped participants includes 80%, who do not differentiate between men and women, and 20% with reversed views, namely participants stating that women are more characterized by ability while men by effort. The behavior of those two sub-groups is qualitatively identical. For this reason and because the second group is very small, we do not distinguish between the two in our analysis.

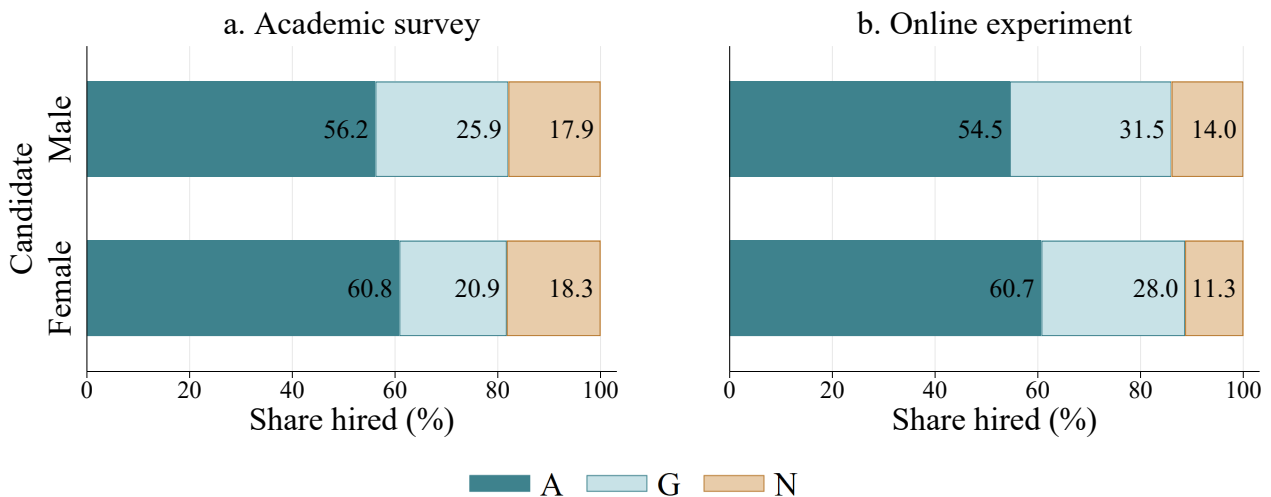
5 Hiring Decisions

In this section, we document that letter type matters for hiring decisions and that the same letter affects recruitment outcomes differently for men and women. Here, we abstract from the fact that men and women may receive different letters. We focus on gender differences in reference letters in Section 6.

To investigate hiring decisions in the recruiter survey, we use a multinomial logit regression model where the dependent variable equals 0 if the candidate with the neutral letter was hired, 1 if the candidate with the grindstone letter was chosen, and 2 if the candidate with the ability letter was appointed. Our explanatory variable of interest is a gender indicator and the corresponding marginal effect estimates the predicted probability of a woman being hired relative to a man with the same letter type. Our analysis controls for order effects (i.e., the order in which reference letters were presented to subjects) and recommender characteristics.

Figure 5 plots the raw data on hiring decisions. It shows that recruiters in both studies have a clear preference for candidates with an ability letter, while grindstone and neutral letters are much less effective. This pattern is particularly pronounced for female candidates. In the academic survey (online experiment), the share of women hired with an ability letter is approximately 40 (33) percentage points higher than the share hired with a grindstone letter. For men, the corresponding difference is only 30 (23) percentage points. This highlights that women benefit disproportionately from an ability letter compared to men.²¹

Figure 5: Hiring choices in the academic survey and online experiment



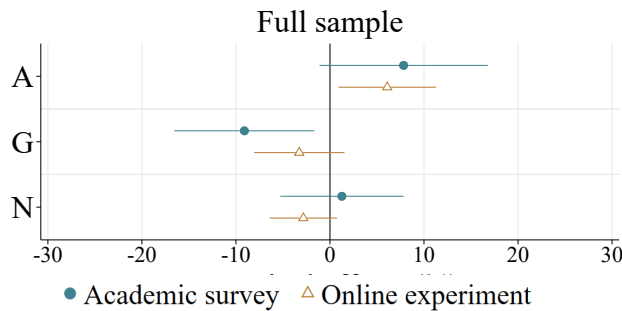
Note: Share of candidates hired with an ability (A), grindstone (G) or neutral (N) letter. In the academic survey, recruiters decide which letter makes them most likely to invite a given candidate to an interview. In the online experiment, recruiters make a hiring choice between three candidates of the same gender but with different letters.

²¹In addition to the choice between candidates, we ask academic recruiters to rate how likely they are to invite a candidate with a given letter to an interview on a scale from 0 to 10. We find that women have about 0.5 points higher rating independently of the letter (Wilcoxon rank-sum tests, $z = -3.84$, $p < 0.001$). This reflects a general propensity to favor women in (at least part of) the recruitment process. Similarly, recruiters in the online experiment are on average 8.56% more likely to hire a woman if asked whether they want to hire a given candidate without comparing them to

Next, we investigate the causal impact of gender on recruitment by letter type using the multinomial regression analysis described above. Figure 6 plots the marginal effects alongside 90% confidence intervals for female relative to male candidates. Compared to their male counterparts, female candidates are more likely to be hired with an ability letter (academic survey: 7.83%; online experiment: 6.09%), and less likely to be hired with a grindstone letter (survey: -9.10%; experiment: -3.27%). These differences are significant for ability letters in the online experiment and for grindstone letters in the academic survey in the full sample. Neutral letters display no statistically significant gendered patterns. Although there are some differences between the two studies, the results thus qualitatively point in the same direction.

Overall, our results show that while all candidates benefit from receiving an ability rather than a grindstone or neutral letter (see Figure 5), this effect is more pronounced for women (see Figure 6). Full regression results are provided in Appendix Tables B.12 and B.13.²²

Figure 6: Differences in hiring between female and male candidates



Note: The figure plots marginal effects from multinomial logit regressions, where the dependent variable is 0 if the candidate with the neutral letter (N) was hired, 1 if the candidate with the grindstone letter (G) was hired, and 2 if the candidate with the ability letter (A) was hired. Regressions control for order effects and include recommender controls. Whiskers represent 90% confidence intervals. Results for the online experiment become even more pronounced when using weighted regressions (see Appendix Table B.14).

6 Letter Choices

The previous section showed that recruiters react differently to the same letter type depending on the candidate's gender. By design, recruiters were equally likely to see each letter type for male and female candidates. However, in practice, recommenders may choose different letters for men and women. We now turn to study whether this is the case. In section 7, we bring both elements — letter choice and recruiter's reaction to letters — together to evaluate the overall impact of gender and letters on hiring outcomes.

To investigate recommender decisions, we again use a multinomial logit regression model where the dependent variable is 0 if the recommender chose a neutral letter for the candidate, 1 if they chose

others (Wilcoxon rank-sum test, $z = -2.72$, $p = 0.007$). On average, recruiters choose to hire in 61% of cases.

²²When looking at recruiter controls, we find that in the academic survey, participants from the Top 100 institutions are more likely to hire a candidate with an ability letter ($p = 0.06$). In the online experiment, female recruiters are more likely to hire candidates with a neutral ($p = 0.07$) or a grindstone letter ($p = 0.01$) compared to male recruiters. Risk preferences and strategic thinking play no role in hiring decisions.

the grindstone letter, and 2 if they chose the ability letter. Our variable of interest is the gender indicator and the corresponding estimated marginal effect captures the predicted probability that a woman receives a given letter type relative to a man.

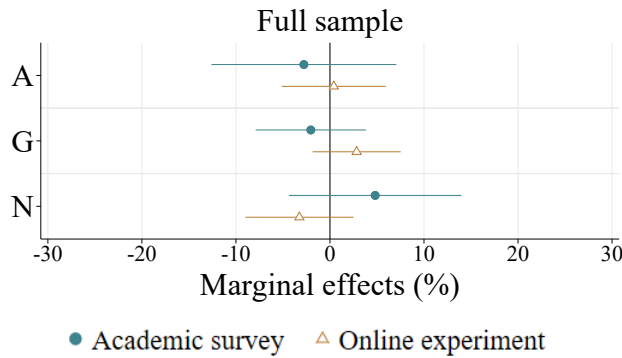
6.1 Aggregate Analysis

When considering the full sample of recommenders, we do not find statistically significant differences in the probability that a woman receives a particular letter type relative to a man (see results in Figure 7). This contrasts with the findings from our observational data. We explore two possible explanations for this difference. This lack of statistical significance could be due to (i) the precision of our estimates, and/or (ii) the aggregate effects masking heterogeneity across groups.

Regarding (i) the precision of our estimates (e.g., [Anderson, 2019](#)), note that the sample size in both the academic survey and the online experiment ($N=505$ and $N=959$, respectively) is substantially smaller compared to the observational study ($N=8,624$). To explore the precision of our estimates in our results further, we conduct a simulation exercise where we vary the sample size of the observational data and record the coefficient distribution (confidence interval) for 100 random sub-samples (see Appendix Figure B.2). The sub-sample with 515 (1,032) observations has roughly the same number of observations as our academic survey (online experiment). We find that the confidence intervals in the simulations match those in the experiments. The results of this precision analysis suggest that the statistically insignificant estimates in the academic survey could be driven by small sample size.

Regarding (ii) heterogeneity, we examine how sample composition affects the results. As a proof of concept, we re-estimate the baseline letter choice regressions from the observational study (Figure 1.a), reweighting the data to match the academic survey sample in terms of gender, seniority,

Figure 7: Difference in letters between female and male candidates



Note: The figure plots marginal effects from multinomial logit regressions, where the dependent variable is 0 if a neutral letter (N) was chosen, 1 if a grindstone letter (G) was chosen, and 2 if an ability letter (A) was chosen. The regressions for the experimental studies (panel a) control for order effects and include recommender controls. The regressions for the observational sample (panel b) show results for the actual full sample and simulations for the average effect sizes when randomly drawing a sub-sample 100 times. The $N=1,032$ sub-sample represents 12% of the original sample size and is similar to the sample size in the online experiment. The $N=515$ sub-sample represents 6% of the original sample size and is similar to the sample size in the academic survey. Whiskers represent 90% confidence intervals.

and institutional rank. Consistent with the aggregate patterns from the academic survey (Figure 7), the effects shown in Appendix Table B.6 are substantially diminished and statistically insignificant. Building on this sample composition analysis, the next section explores heterogeneity in letter choices across subgroups, in particular along gender views.

6.2 Heterogeneity

In this section, we explore whether gendered patterns in letter choices vary across sub-samples in our data, potentially explaining the aggregate results. Specifically, in the spirit of some existing theoretical contributions (e.g., Onuchic, 2024) and empirical studies that investigate mechanisms and sources of discrimination (e.g., Fershtman and Gneezy, 2001; List, 2004; Neumark, 2018), we assess the roles of the recommenders’ gender views. Recommenders’ preconceptions about gender could influence their opinions about candidates, independently of the candidate’s quality. To investigate this possibility, we split the sample depending on gender views and find stark differences.²³

While we pre-registered heterogeneity analyses based on observable demographics such as gender and seniority, we also explore heterogeneity by gender views — a dimension that was not pre-specified but is strongly motivated by the theoretical and empirical literature discussed above. This dimension is closely linked to our research question and plays a central role in our conceptual replication effort, as each study uses a different design to test the same hypothesis and investigate this heterogeneity. Moreover, gender views are correlated with key demographics (e.g., seniority) and allow us to compare results across the academic and online samples using a consistent behavioral measure.

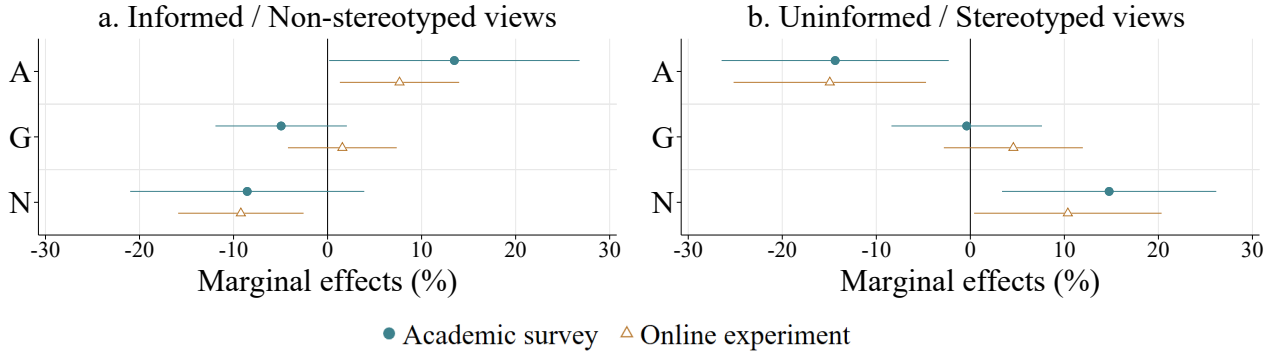
Figure 8 plots the marginal effects for the female dummy obtained from estimating the multinomial logit regression (see Appendix Tables B.15 and B.16). For the online experiment, the regression includes the candidate’s quiz performance as a control, in addition to recommender controls. In both the academic survey and the online experiment, we uncover a similar pattern. Recommenders who have uninformed/stereotyped views are substantially less likely to give women an ability letter (-14.38% and -14.94% in the respective studies) and more likely to give them a neutral letter (14.76% and 10.37%, respectively). These effects are significant at the 10% level and robust to excluding recommender controls. Among recommenders who have informed/non-stereotyped views, we instead see the exact opposite pattern. They are substantially more likely to give ability letters to female candidates (13.48% and 7.65%) and less likely to give them a neutral letter (-8.55% and -9.22%, only significant for the online experiment).²⁴ Overall, this heterogeneity contributes to explaining the absence of statistical significance in the aggregate results.

Interestingly, we observe that participants with uninformed views are more likely to be male ($t =$

²³We also analyzed *hiring* decisions in split samples but found that the same qualitative patterns hold in both groups — see Appendix Tables B.12 and B.13 for details.

²⁴When interacting gender views with the candidate’s gender, we see in both the academic survey and the online experiment that recommenders with informed or non-stereotyped views are significantly more likely to choose an ability letter for a female candidate compared to recommenders with uninformed or stereotyped views ($p = 0.02$ and $p = 0.001$ respectively).

Figure 8: Differences in letters between female and male candidates



Note: The figure plots marginal effects from multinomial logit regressions, where the dependent variable is 0 if a neutral letter (N) was chosen, 1 if a grindstone letter (G) was chosen, and 2 if an ability letter (A) was chosen. Regressions control for order effects and include recommender controls. Whiskers represent 90% confidence intervals. Results for the online experiment become even more pronounced when using weighted regressions (see Appendix Table B.17).

7.02, $p < 0.001$) and already an associate or full professor ($t = -2.31$, $p = 0.02$ and $t = -2.35$, $p = 0.02$, see Appendix Figure B.3). Participants with uninformed views are therefore more similar to recommenders in the observational sample with respect to their demographic characteristics. This can explain why we find the same results for this sub-sample and the observational sample.

We push this comparison one step further by creating a variable measuring the propensity to hold informed views in the *observational data*, using the participant characteristics of the *academic survey*. Specifically, we use the survey data to estimate a linear probability model to predict the likelihood that participants hold informed views ($n=1,018$), conditioning on respondent gender, academic rank, PhD cohort, institution rank and geographical location. Then, using the observational data, we use the marginal effects from this model to predict the propensity of informed views ($n=7,717$).²⁵ In Appendix Table B.5, we present split-sample results for three alternative cutoffs of this propensity measure. The results suggest that gendered patterns in letter choice are driven by the group of recommenders who are more likely to hold *uninformed* views.

Overall, the results from the observational and experimental studies suggest that letter choices vary meaningfully across subgroups: those more likely to hold uninformed/stereotyped views tend to exhibit gendered patterns in letter choices.²⁶ This behavioral pattern is conceptually replicated across our three complementary studies, thus strengthening the robustness of our inference.

²⁵Our sample is reduced by our focus on academic recommenders and the required information on the country of primary academic affiliation of the recommender.

²⁶When looking at how other recommender controls affect letter choice, we find that in the academic survey, participants with more hiring experience are less likely to choose the grindstone letter ($p = 0.006$). In the online experiment, participants with a higher level of education ($p < 0.001$) and men ($p = 0.02$) are more likely to send an ability letter. Risk preferences and strategic thinking play no role for letter decisions.

6.3 Are male and female candidates different?

In the previous subsection, we established that men and women receive different types of letters, although with varying patterns depending on the gender views of recommenders. A possible reason for this pattern is that men and women could differ in their capacity to work hard, as well as in their natural ability.

When analyzing the observational data, we control for a rich set of candidate characteristics that were manually collected from CVs. However, this cannot completely rule out confounds about candidates' quality. In the academic survey, by contrast, candidates have identical observed characteristics by design. While this makes male and female candidates highly comparable, we cannot fully rule out that academic recommenders draw different inferences about a male or female candidate's potential from the observed characteristics.

In the online experiment, we go one step further and provide direct information to recommenders on an objective performance measure that is a signal of what is sought by the recruiter. Nevertheless, we find that gender differences in letters persist in both environments, making it unlikely that actual differences in ability and hard work across genders play an important role. We also do not see any significant differences in quiz performance or cognitive skill between male and female candidates, nor differences in recommenders' or recruiters' expectations about their performance (see Appendix Figures B.5–B.7). The only dimension on which female and male candidates differ is average self-reported effort, with men reporting significantly higher levels (*ibid.*, $z = 2.74$, $p = 0.006$). If anything, this would justify more grindstone letters for male candidates. However, this result should be interpreted carefully as there may be gender differences in confidence or reporting (e.g., Niederle and Vesterlund, 2007; Adamecz-Völgyi and Shure, 2022).²⁷

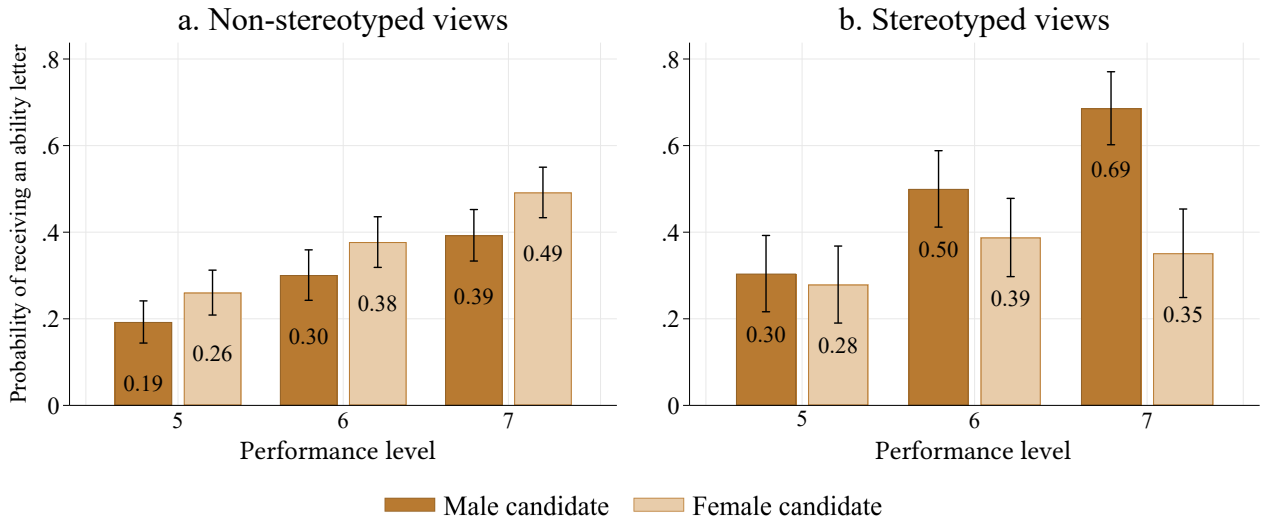
Finally, the online experiment allows us to assess whether recommenders react differently to candidate quality depending on gender views. Figure 9 shows the share of male and female candidates receiving an ability letter by performance level, separately for recommenders who hold (respectively, do not hold) gendered stereotyped views. Among recommenders with non-stereotyped views, there is a clear positive relationship between performance and the likelihood of receiving an ability letter. Among recommenders with stereotyped views, in contrast, this relationship only holds for male candidates.

Taken together, the analysis in this section sheds light on the role of candidate quality in explaining the type of recommendation candidates receive. It is not that the distribution of quality is different between men and women; instead, recommenders who hold stereotyped views react differently to the same quality depending on the gender of the candidate.²⁸

²⁷In fact, we find that the correlation between reported effort and time spent on the quizzes is slightly stronger for women ($r = 0.5$ versus $r = 0.46$).

²⁸In workplaces, these gendered reactions to individuals' signals or behavior continue over time, potentially generating cumulative effects (e.g., Egan et al., 2022; Sarsons, 2022).

Figure 9: Probability of receiving an ability letter by performance level



6.4 Strategic Considerations and Beliefs

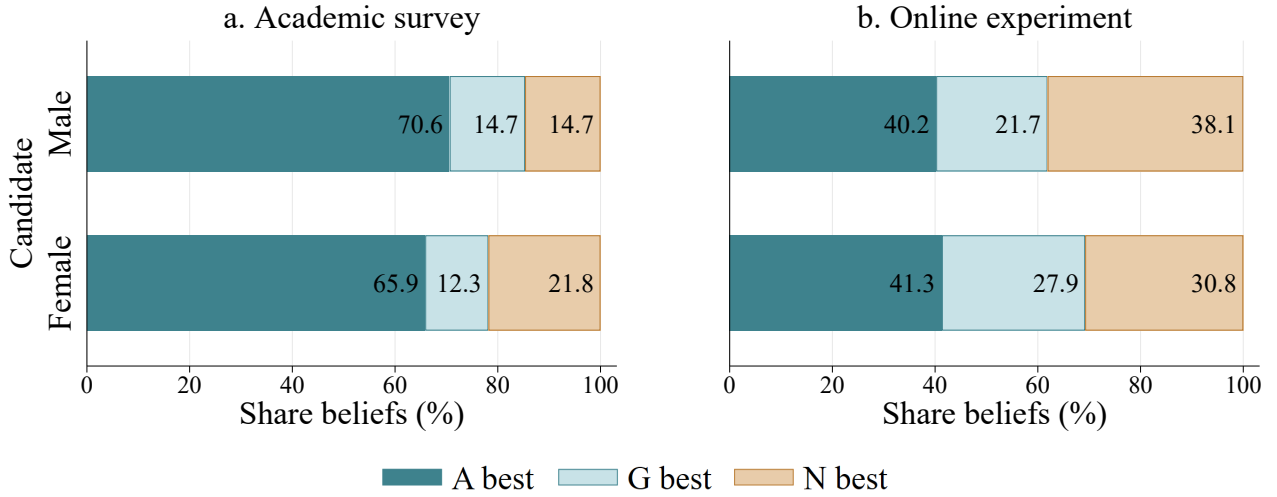
A final reason why men and women receive different letters is that recommenders may believe that it is beneficial for women to be described as hard-working or to understate their abilities. Recommenders may for instance think that a message strongly emphasizing ability will portray women as unlikable (e.g., [Cooper, 2013](#)). To gauge the importance of this mechanism, we conduct two analyses.

Both the academic survey and the online experiment contain a ‘hiring’ component, that allows us to test whether it is optimal for female and male candidates to receive different letters. We showed in Section 5 that the majority of recruiters prefer to hire a candidate with an ability letter (see Figure 5). Moreover, Figure 6 highlights that, if anything, this effect is more pronounced for female candidates. Similarly, receiving a grindstone message is slightly worse for them. These patterns do not support the view that recruiters’ preferences can explain why men and women receive different letters.

Although it is not optimal for female candidates to receive fewer ability letters, recommenders may not know this. To study this possibility, we elicit recommenders’ beliefs about the most effective letter. We do so by asking them to predict which letter maximizes the candidate’s chance of being hired. As we just discussed, the correct answer to this question is the ability letter. However, Figure 10 reveals that many recommenders do not choose the right answer. A non-negligible share thinks that most recruiters will prefer to hire a candidate with a grindstone or a neutral letter. Moreover, in the academic survey, this error is more pronounced for female candidates. Although an ability letter is *more* effective for women than men, Panel a. of Figure 10 reveals that *fewer* recommenders select the correct answer for women compared to men.

Given these inconsistencies between the recommenders’ beliefs and the recruiters’ hiring behavior, we explore whether accounting for beliefs can explain the gender gap in letter choice. To do so, we control for these beliefs in the letter choice regressions. In particular, we add binary control variables

Figure 10: Beliefs about the most effective letter



Note: Share of recommenders believing a given letter maximising the candidate’s probability to be hired. In reality, the best letter for a candidate (independent of gender) is the ability one.

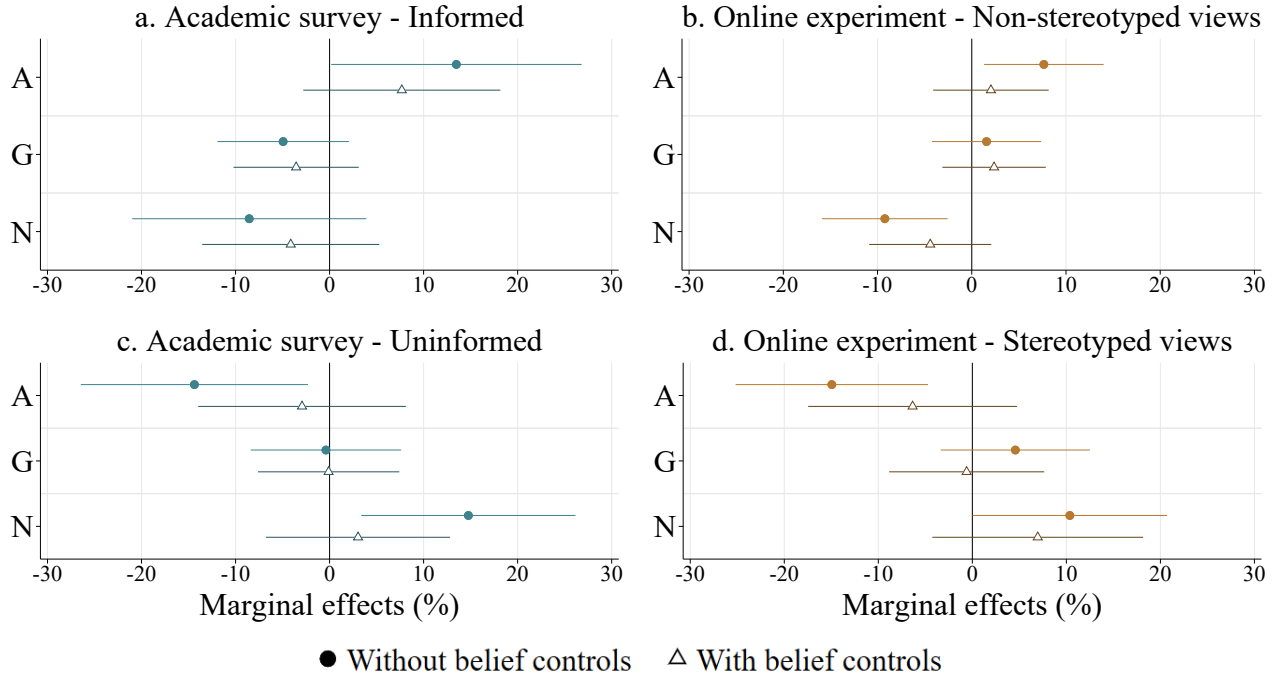
flagging which of the letters was selected as the one maximizing the candidate’s chances of being recruited (see Appendix Tables B.20 and B.21). Figure 11 presents the results from our baseline models alongside those in which we control for recommenders’ beliefs. Across all studies and sub-samples, adding beliefs as a control attenuates the size of the gender effect on letter choice, which is now always statistically insignificant. Moreover, we use nonparametric bootstrapping to test whether the difference between marginal effects in models with and without belief controls is statistically significant. We find that, in several cases, the marginal effects of gender differ significantly across model specifications.²⁹ These results suggest that beliefs are a key driver of letter choices.

A separate question is whether recommenders either act with the candidate’s best interest at heart or hold motivated beliefs (e.g., Bénabou, 2015). Although the majority of recommenders send recruiters what they believe to be the most effective letter, these beliefs also correlate with gender views (see Figure B.8). In particular, recommenders with uninformed/stereotyped views are more likely to hold the erroneous belief that ability letters are *better* for men than for women. This alignment between gender views and beliefs is broadly consistent with motivated reasoning, with recommenders developing a narrative that justifies their choices.

Independently of the source of beliefs, our results suggest that informing recommenders about the effectiveness of ability letters could be a promising approach to decreasing differences in letters written for female and male candidates. Recent evidence suggests this also applies in the case of motivated beliefs (e.g., Zimmermann, 2020).

²⁹ Among recommenders with stereotyped views, the effect is significant for ability letters ($z = -2.41$, $p = 0.016$). For recommenders with non-stereotyped views, significant differences are found for both ability ($z = 3.36$, $p = 0.001$) and neutral letters ($z = -2.88$, $p = 0.004$). Similarly, for recommenders with uninformed views, we observe significant differences for ability and neutral letters ($z = -2.5$, $p = 0.013$ and $z = 2.46$, $p = 0.014$).

Figure 11: Differences in letters between female and male candidates after controlling for beliefs



Note: The figure plots marginal effects from multinomial logit regressions, where the dependent variable is 0 if a neutral letter (N) was chosen, 1 if a grindstone letter (G) was chosen, and 2 if an ability letter (A) was chosen. Regressions control for order effects and include recommender controls. In each panel, the same regression is once run with and without controlling for recommenders' beliefs. Whiskers represent 90% confidence intervals. Results are qualitatively the same when using a weighted online sample (see Appendix Table B.22).

7 Discrimination Decomposition

The discussion of gender representation in the workplace has been taken seriously by many companies and organizations, for instance through DEI policies. In the context of recommendations, we have documented systematic differences in treatment that may undermine these efforts. Our analysis has focused on two instances: the choice of letters on the one hand, and the hiring of candidates conditional on having received the same letter on the other.

An important next step is to understand the interplay of these two instances, and their implications for diversity in hiring. More specifically, despite existing policies to promote the recruitment of women, hiring decisions may still be distorted because reference letters differ systematically for men and women.

To conceptualize differential treatment at different points in the hiring process, we apply the classic Kitagawa-Oaxaca-Blinder decompositions (Kitagawa 1955; Oaxaca 1973; Blinder 1973) to break down 'total discrimination' into a 'direct' and 'systemic' component (see Bohren et al. 2022, whose terminology we draw on in this section, and Baron et al. 2024 for an application). Favoring female over male applications with similar attributes is 'direct discrimination'. In contrast, gender gaps caused by recruitment decisions blindly relying on reference letters without realizing that these differ for men and women as 'systemic discrimination'. Below, we outline the decomposition framework and how we apply it to our data.

7.1 Theoretical Framework

Denote $j \in \mathcal{J}$ the recruiters, and $i \in \mathcal{I}$ the candidates, whose gender $G_i \in \{F, M\}$ is observed by recruiters (with F denoting female and M male). Candidates have ex-ante productivity Y_i . Recruiters receive a signal of this productivity through the reference letters $L_i \in \{\text{Ability}, \text{Grindstone}, \text{Neutral}\}$ and decide on a recruitment outcome $\text{Recruit}_j(G_i, L_i) \in \{0, 1\}$.

7.1.1 Total Discrimination

Total Discrimination (TD) is the expected difference in recruitment outcomes between male and female candidates, conditional on productivity y . This is the empirical hiring gap observed by the researcher and can be written as:

$$\text{TD}(y) = \mathbb{E}[\text{Recruit}_j(G_i, L_i) | G_i = F, Y_i = y] - \mathbb{E}[\text{Recruit}_j(G_i, L_i) | G_i = M, Y_i = y]. \quad (1)$$

7.1.2 Direct Discrimination

Direct discrimination is the difference in recruitment outcomes between male and female candidates with identical letters. In other words, direct discrimination arises when recruiters do not apply the same hiring rule for men and women upon receiving the same letter. Formally, a measure of direct discrimination by recruiter j after receiving a given letter l is as follows:

$$\tau_j(l) = \text{Recruit}_j(F | L_i = l) - \text{Recruit}_j(M | L_i = l). \quad (2)$$

The Average Direct Discrimination (ADD for short) is the average of $\tau(l)$ over the letter distribution in the population.³⁰ Formally,

$$\text{ADD}(g, y) = \mathbb{E}[\tau(L_i) | G_i = g, Y_i = y]. \quad (3)$$

7.1.3 Systemic Discrimination

Finally, Systemic Discrimination (SD) captures the disparities in recruitment outcomes because recruiters ignore that letter types are correlated with candidate gender. A measure of systemic discrimination at productivity y is given by:

$$\text{SD}(g, y) = \mathbb{E}[\text{Recruit}_j(g, L_i) | G_i = F, Y_i = y] - \mathbb{E}[\text{Recruit}_j(g, L_i) | G_i = M, Y_i = y]. \quad (4)$$

When $g = M$, the second term on the right-hand side of the equation gives the expected realized recruitment outcomes for men (as in the second term on the right-hand side of equation 1). The first term gives the expected recruitment outcomes for women if they are assessed using the recruitment

³⁰Alternatively, we can calculate ADD using the letter distribution just for the female, or just the male population. All three specifications produce qualitatively similar results. We refer to the respective results in the appendix when discussing our results.

rules for men (i.e., shutting down any channel of direct discrimination based on gender, but not taking into account that the distribution of letters differs for men and women). The two terms can only differ if the underlying distribution of letter types is not the same across genders. In other words, $SD(M, y)$ captures recruitment gaps that arise only as a result of differences in the distribution of letters.

7.1.4 Decomposition

Combining equations 1 and 4, we can decompose total discrimination as the sum of the average direct discrimination and systemic discrimination. Specifically,

$$\begin{aligned}
TD(y) &= \mathbb{E}[\text{Recruit}_j(F, L_i)|G_i = F, Y_i = y] - \mathbb{E}[\text{Recruit}_j(M, L_i)|G_i = M, Y_i = y] \\
&\quad + \mathbb{E}[\text{Recruit}_j(M, L_i)|G_i = F, Y_i = y] - \mathbb{E}[\text{Recruit}_j(M, L_i)|G_i = F, Y_i = y] \\
&= \underbrace{\mathbb{E}[\text{Recruit}_j(F, L_i) - \text{Recruit}_j(M, L_i)|G_i = F, Y_i = y]}_{\text{Average Direct Discrimination for Women } ADD(F, y)} \\
&\quad + \underbrace{\mathbb{E}[\text{Recruit}_j(M, L_i)|G_i = F, Y_i = y] - \mathbb{E}[\text{Recruit}_j(M, L_i)|G_i = M, Y_i = y]}_{\text{Systemic Discrimination } SD(M, y)} \\
&= ADD(F, y) + SD(M, y).
\end{aligned} \tag{5}$$

7.2 Estimation and Results

We now turn to the estimation of total, direct, and systemic discrimination. In addition, we evaluate heterogeneity with respect to gender views. Given the importance of beliefs as discussed in Section 6, we also analyze a hypothetical scenario that assumes all recommenders to have correct beliefs.

To obtain these estimates, we first need to link recommenders and recruiters. In both the academic survey and the online experiment, recruiters see an equal number of ability, grindstone, and neutral letters. However, to understand the total effect, we need to account for the distribution of letters generated by recommenders. To do so, we run 1,000 simulations per specification, randomly matching one recommender and their letter choice to one recruiter. We determine whether each match between a recruiter and a letter would result in a successful hire, based on the recruiter's decision in the experiment.

7.2.1 Estimation

Total Discrimination. TD is the expected difference in recruitment outcomes between male and female candidates. Since gender is randomized in our two experiments, we can estimate this difference by regressing the hiring outcome on gender:

$$\text{Recruit}_{il} = \alpha + \beta \text{Female}_i + \varepsilon_{il}, \tag{6}$$

where $\text{Recruit}_{i,l}$ is a binary variable which takes the value of 1 if a candidate i with letter l is hired and 0 otherwise; Female_i is a binary variable equal to one if the candidate is a woman. The coefficient $\hat{\beta}$ provides an estimate for TD.

Average Direct Discrimination. ADD captures the average gender difference in recruitment outcomes for candidates with the same letter. To quantify ADD we estimate the following regression equation using Ordinary Least Squares (OLS):

$$\text{Recruit}_{i,d} = \sum_{k=1}^3 (\gamma_k \mathbf{1}\{L_i = l_k\} + \lambda_k \mathbf{1}\{L_i = l_k\} \times \text{Female}_i) + \nu_{i,d}, \quad (7)$$

where $\mathbf{1}\{L_i = l_k\}$ is a binary variable equal to one if the letter for candidate i is of type $l_k \in \{\text{Ability}, \text{Grindstone}, \text{Neutral}\}$.³¹

If the classical OLS assumptions hold, then λ_k gives us the average direct discrimination *for letter type* l_k ($\mathbb{E}[\tau(l_k)]$) (see equation 2, where k was omitted for simplicity of notation). An estimator of the average direct discrimination is then the average of $\hat{\lambda}_k$, weighted by the probability of each letter type:

$$\widehat{\text{ADD}}(g) = \sum_{k=1}^3 \hat{\lambda}_k \hat{p}(l_k | \text{Female}_i \in g). \quad (8)$$

We estimate the probability of each letter type $p(l_k | \text{Female}_i = g)$ using the distribution of letter types in the sample of all candidates.³²

Systemic Discrimination. Systemic Discrimination is the difference between Total and Average Direct Discrimination ($\hat{\beta} - \widehat{\text{ADD}}(g)$).

Before presenting the results of our decomposition exercise, it is worth noting two specific features of our experimental studies.

First, in the academic survey, we ask recruiters which letter they find most convincing, rather than which candidate they would hire. While this does not rule out the possibility that recruiters might hire candidates with weaker letters, given the competitiveness of the academic job market, we regard recruiters' relative preferences as a meaningful proxy for actual hiring decisions. Moreover, this concern does not apply to the online experiment, where recruiters make direct hiring choices among candidates with different letters.

Second, we do not elicit hiring decisions for both male and female candidates from the same recruiter. Although this between-subject design may introduce some noise due to individual variability, cru-

³¹The regression has no constant because the sum of all $\mathbf{1}\{L_i = l_k\}$ equals the constant. The regression does not include the female binary variable because the sum of all $\mathbf{1}\{L_i = l_k\} \times \text{Female}_i$ equals the female binary.

³²We compute ADD using the signal distribution just for the female, or just the male population in Appendix B.8.

cially, it minimizes experimenter demand effects while enabling us to detect average treatment effects, which we decompose in the following section.

7.2.2 Results

Table 1 shows the results of the decomposition exercise. Columns 1 and 4 give results for the full sample in the academic survey and online experiment, respectively, while columns 2-3 and 5-6 contrast results depending on gender views.

The results show that for the full sample in the academic survey, TD is small but *positive*: overall, female candidates have a 1.12 percentage points *higher* chance of being hired compared to their male counterparts. ADD in favor of women is more than twice as large (3.13). However, this is counter-acted by a *negative* SD stemming from women receiving worse letters on average. When focusing on the sub-sample of recommenders who have uninformed views (column 3), the negative SD is exacerbated and becomes even larger than the positive ADD. If recruiters only receive letters from recommenders who have uninformed views, female candidates face an overall 4.65 percentage point lower probability of being hired. By contrast, if recommenders have informed views (column 2), SD becomes positive, resulting in a TD in favor of women of 7.63 percentage points.

The online experiment shows the same qualitative patterns, with positive ADD in favor of women throughout, while SD disadvantages women in the sub-sample of recommenders with stereotyped views and advantages them in the sub-sample with non-stereotyped views.³³

Table 1: Decomposition of total into systemic and direct discrimination - Gender views

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1) Full sample	(2) Informed views Yes	(3) No	(4) Full sample	(5) Stereotyped views No	(6) Yes
TD	1.12	7.63	-4.65	1.34	4.13	-5.21
ADD	3.13	3.40	2.79	0.38	0.10	0.99
SD	-2.01	4.23	-7.44	0.96	4.03	-6.19
Recommenders	506	232	274	959	679	280
Recruiters	514	514	514	959	959	959

Note: The table reports average differences in hiring probability for female versus male candidates across 1,000 simulations per column. It presents results using the distribution of letters for all candidates. Tables B.23 and B.24 show that results are qualitatively and in terms of magnitude identical when using the distribution for only female or male candidates or using weighted regressions for the online sample. TD, ADD and SD are Total, Average Direct and Systemic Discrimination, respectively.

In Table 2, we conduct the decomposition for recommenders with correct and incorrect beliefs, i.e. whether a recommender believes that an ability letter is best or not. Columns 1 and 4 report the

³³Table B.25 in Appendix B shows additional decomposition results when distinguishing between *recruiters* with different gender views. Results are qualitatively similar to using the full sample of recruiters.

full sample baseline results, whereas columns 2-3 and 5-6 contrast findings depending on recommender beliefs. In Section 6.3, we show that the beliefs of recommenders affect the distribution of letters for men and women. Different distributions will in turn affect SD. In line with this argument, in Table 2, we observe that in the academic survey the gap in SD between recommenders holding correct and incorrect beliefs is large (columns 2 and 3: 3.43 vs. -8.12). This gap in turn flips the sign of TD (7.74 vs. -7.66). In the online experiment, a similar pattern emerges even though TD remains positive even under incorrect beliefs.

Table 2: Decomposition of total into systemic and direct discrimination - Beliefs

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Correct beliefs Yes	No	Full sample	Correct beliefs Yes	No
TD	1.12	7.74	-7.66	1.34	2.35	0.36
ADD	3.13	4.31	0.45	0.38	2.11	-0.83
SD	-2.01	3.43	-8.12	0.96	0.24	1.20
Recommenders	506	344	160	959	391	568
Recruiters	514	514	514	959	959	959

Note: The table reports average differences in hiring probability for female versus male candidates across 1,000 simulations per column. Correct beliefs restricts the analysis to recommenders who correctly believe that the ability letter will be most successful. In the academic survey, two recommenders did not answer the beliefs question. The table presents results using the distribution of letters for all candidates. Tables B.26 and B.27 show that results are qualitatively and in terms of magnitude identical when using the distribution for only female or male candidates or using weighted regressions for the online sample. TD, ADD and SD are Total, Average Direct and Systemic Discrimination, respectively.

Overall, these findings highlight the importance of taking systemic elements into account when assessing discrimination patterns. The positive ADD we find in all specifications is indicative of the widespread efforts to improve female recruitment, in line with current DEI policies. However, when looking at final outcomes for female candidates (TD), these efforts may be insufficient if there are systemic elements that negatively affect women’s chances of being hired. In the full sample, the attempt to hire more women is almost counter-balanced by the negative SD. In the sample of recommenders having uninformed/stereotyped views, we see that SD is substantial enough to lead to an overall negative gender hiring gap (TD). Similarly, incorrect beliefs by recommenders undermine the efforts to diversify hiring. The resulting SD can be large enough to drive a negative TD.

8 Concluding Remarks

In this paper, we have studied the gendered impact of recommendation content on recruitment, focusing on reference letters. Using both observational and experimental data, we first show that women have overall higher recruitment prospects compared to men, which is in line with recent DEI efforts. However, this advantage can be undermined by differences in how women and men

are described in recommendation letters. In the observational data, women are less often described in terms of ability and more often characterized with so-called ‘grindstone’ attributes. While the experimental evidence does not show this pattern on aggregate, we do observe it among recommenders who hold stereotypical or uninformed gender views. The demographic profile of these recommenders closely resembles that of the recommenders in the observational sample, indicating a common underlying mechanism. Thus, taken together, our three complementary studies serve as a conceptual replication of each other and offer a convergent body of evidence that strengthens the credibility of our findings beyond what each study could achieve in isolation.

Our experimental analysis suggests that differences in recommendations received by men and women are a more significant barrier to women’s recruitment than outright unequal treatment of candidates with identical characteristics. In other words, to tackle barriers to diversity it is crucial to understand more subtle and indirect forms of discrimination, focusing not only on the recruiter side but also the recommender side of the market. This matters because a significant portion of the efforts to improve diversity in recruitment targets the hiring process, for example through training programs, quotas in interviewing panels, shortlisting targets, or systematic compiling and analyzing of diversity data for Human Resource departments. While recruiters could be trained to take gender imbalances in recommendations into account, a more direct approach would be to target recommenders. In our setting, the weak link in overcoming gender recruitment gaps is represented in this earlier step in the recruitment process, namely by recommenders emphasizing different qualities in their recommendations for male and female candidates. Therefore, correcting recommender beliefs about the effectiveness of different letter types or making them aware of their own — potentially uninformed or stereotyped — gender views could meaningfully improve diversity in the workplace and labor market efficiency.

References

- Abel, Martin, Rulof Burger, and Patrizio Piraino**, “The value of reference letters: Experimental Evidence from South Africa,” *American Economic Journal: Applied Economics*, 2020, 12 (3), 40–71.
- Adamecz-Völgyi, Anna and Nikki Shure**, “The gender gap in top jobs-the role of overconfidence,” *Labour Economics*, 2022, 79, 102283.
- Alempaki, Despoina, Emina Canic, Timothy L Mullett, William J Skylark, Chris Starmer, Neil Stewart, and Fabio Tufano**, “Reexamining how utility and weighting functions get their shapes: A quasi-adversarial collaboration providing a new interpretation,” *Management Science*, 2019, 65 (10), 4841–4862.
- Anderson, Andrew A.**, “Assessing Statistical Results: Magnitude, Precision, and Model Uncertainty,” *The American Statistician*, 2019, 73 (sup1), 118–121.
- Avery, Christopher, Christine Jolls, Richard A Posner, and Alvin E Roth**, “The Market for Federal Judicial Law clerks,” *University of Chicago Law Review*, 2001, 68, 793.
- Bagues, Manuel F and Berta Esteve-Volart**, “Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment,” *The Review of Economic Studies*, 2010, 77 (4), 1301–1328.
- Baltrunaite, Audinga, Alessandra Casarico, and Lucia Rizzica**, “Women in economics: the role of gendered references at entry in the profession,” CEPR Discussion Paper 17474 2022.
- , —, and —, “Women in Economics: The role of gendered references at entry in the profession,” Bank of Italy Temi di Discussione (Working Paper) No 1438 2024.
- Baron, E Jason, Joseph J Doyle Jr, Natalia Emanuel, Peter Hull, and Joseph Ryan**, “Discrimination in Multiphase Systems: Evidence from Child Protection,” *The Quarterly Journal of Economics*, 2024, 139 (3), 1611–1664.
- Barron, Kai, Ruth Dittmann, Stefan Gehrig, and Sebastian Schweighofer-Kodritsch**, “Explicit and implicit belief-based gender discrimination: A hiring experiment,” *Management Science*, 2024, *forthcoming*.
- Beaman, Lori, Niall Keleher, and Jeremy Magruder**, “Do Job Networks Disadvantage Women? Evidence from a Recruitment Experiment in Malawi,” *Journal of Labor Economics*, 2018, 36 (1), 121–157.
- Bénabou, Roland**, “The economics of motivated beliefs,” *Revue d’économie politique*, 2015, 125 (5), 665–685.
- Bertrand, Marianna, Claudia Goldin, and Lawrence F. Katz**, “Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors,” *American Economic Journal: Applied Economics*, 2010, 2 (3), 228–255.

- Bertrand, Marianne and Esther Duflo**, “Field experiments on discrimination,” in Abhijit V Banerjee and Esther Duflo, eds., *Handbook of Economic Field Experiments*, Vol. 1, Elsevier, 2017, chapter 8, pp. 309–393.
- **and Sendhil Mullainathan**, “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination,” *American Economic Review*, 2004, 94 (4), 991–1013.
- Bian, Lin, Sarah-Jane Leslie, Mary C Murphy, and Andrei Cimpian**, “Messages about brilliance undermine women’s interest in educational and professional opportunities,” *Journal of Experimental Social Psychology*, 2018, 76, 404–420.
- Blinder, Alan S**, “Wage Discrimination: Reduced Form and Structural Estimates,” *Journal of Human Resources*, 1973, 8 (4), 436–455.
- Bohren, J. Aislinn, Peter Hull, and Alex Imas**, “Systemic Discrimination: Theory and Measurement,” Working Paper 29820, National Bureau of Economic Research 2022.
- , — , **and** — , “Systemic Discrimination: Theory and Measurement,” *Quarterly Journal of Economics*, Forthcoming.
- Boring, Anne and Arnaud Philippe**, “Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching,” *Journal of Public Economics*, 2021, 193, 104323.
- Breda, Thomas and Son Thierry Ly**, “Professors in core science fields are not always biased against women: Evidence from France,” *American Economic Journal: Applied Economics*, 2015, 7 (4), 53–75.
- Card, David, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri**, “Gender gaps at the academies,” *Proceedings of the National Academy of Sciences*, 2023, 120 (4), e2212421120.
- , — , — , **and Naorre Iriberri**, “Gender Differences in Peer Recognition by Economists,” *Econometrica*, 2022, 90 (5), 1937–1971.
- Ceci, Stephen J and Wendy M Williams**, *The mathematics of sex: How biology and society conspire to limit talented women and girls*, Oxford University Press, 2009.
- Coffman, Lucas C and Muriel Niederle**, “Pre-analysis plans have limited upside, especially where replications are feasible,” *Journal of Economic Perspectives*, 2015, 29 (3), 81–98.
- Coles, Peter, John Cawley, Phillip B. Levine, Muriel Niederle, Alvin E. Roth, and John J. Siegfried**, “The Job Market for New Economists: A Market Design Perspective,” *Journal of Economic Perspectives*, 2010, 24 (4), 187– 205.
- Cooper, Marianne**, “For Women Leaders, Likability and Success Hardly Go Hand-in-Hand,” *Harvard Business Review*, 2013, 30, 7–16.

- Correll, Shelley J., Katherine R. Weisshaar, Alison T. Wynn, and Jennifer D. Wehner**, “Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment,” *American Sociological Review*, 2020, 85 (6), 1022–1050.
- Deschamps, Pierre**, “Gender quotas in hiring committees: A boon or a bane for women?,” *Management Science*, 2023, 70 (11), 7486–7505.
- Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner**, “Individual risk attitudes: Measurement, determinants, and behavioral consequences,” *Journal of the European Economic Association*, 2011, 9 (3), 522–550.
- Dupas, Pascaline, Alice Sasser Modestino, Muriel Niederle, Justin Wolfers, and The Seminar Dynamics Collective**, “Gender and the Dynamics of Economics Seminars,” NBER Working Paper 28494 2021.
- Eberhardt, Markus, Giovanni Facchini, and Valeria Rueda**, “Gender differences in reference letters: Evidence from the economics job market,” *The Economic Journal*, 2023, 133 (655), 2676–2708.
- Egan, Mark, Gregor Matvos, and Amit Seru**, “When Harry fired Sally: The double standard in punishing misconduct,” *Journal of Political Economy*, 2022, 130 (5), 1184–1248.
- Eyting, Markus**, “Why do we discriminate? the role of motivated reasoning,” *SAFE Working Paper*, 2022.
- Fershtman, Chaim and Uri Gneezy**, “Discrimination in a segmented society: An experimental approach,” *The Quarterly Journal of Economics*, 2001, 116 (1), 351–377.
- Glover, Dylan, Amanda Pallais, and William Pariente**, “Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores,” *The Quarterly Journal of Economics*, 2017, 132 (3), 1219–1260.
- Goldin, Claudia and Cecilia Rouse**, “Orchestrating impartiality: The impact of ‘blind’ auditions on female musicians,” *American Economic Review*, 2000, 90 (4), 715–741.
- , **Sari Pekkala Kerr, Claudia Olivetti, and Earling Barth**, “The Expanding Gender Earnings Gap: Evidence from the LEHD-2000 Census,” *American Economic Review Papers and Proceedings*, 2017, 107 (5), 110–114.
- Heller, Sara B and Judd B Kessler**, “The Effects of Letters of Recommendation in the Youth Labor Market,” *National Bureau of Economic Research*, 2021, (29579).
- Hendricks, Wallace, Lawrence DeBrock, and Roger Koenker**, “Uncertainty, hiring, and subsequent performance: The NFL draft,” *Journal of Labor Economics*, 2003, 21 (4), 857–886.
- Hengel, Erin**, “Publishing while female: Are women held to higher standards? Evidence from peer review,” *The Economic Journal*, 2022, 132 (648), 2951–2991.

- Imai, Taisuke, Séverine Toussaert, Aurélien Baillon, Anna Dreber, Seda Ertaç, Magnus Johannesson, Levent Neyse, and Marie Claire Villeval**, “Pre-Registration and Pre-Analysis Plans in Experimental Economics,” Technical Report, I4R Discussion Paper Series 2025.
- Jacquemet, Nicolas and Constantine Yannelis**, “Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market,” *Labour Economics*, 2012, 19 (6), 824–832.
- Jassal, Nirvikar**, “Does victim gender matter for justice delivery? Police and judicial responses to women’s cases in India,” *American Political Science Review*, 2024, 118 (3), 1278–1304.
- Kitagawa, Evelyn M**, “Components of a difference between two rates,” *Journal of the American Statistical Association*, 1955, 50 (272), 1168–1194.
- Kline, Patrick, Evan K Rose, and Christopher R Walters**, “Systemic discrimination among large US employers,” *The Quarterly Journal of Economics*, 2022, 137 (4), 1963–2036.
- Koffi, Marlène**, “Innovative ideas and gender inequality,” Canadian Labour Economics Forum (CLEF), Waterloo, Working Paper Series 35 2021.
- Kuhn, Peter and Kailing Shen**, “Gender discrimination in job ads: Evidence from China,” *The Quarterly Journal of Economics*, 2013, 128 (1), 287–336.
- Levitt, Steven D and John A List**, “Field experiments in economics: The past, the present, and the future,” *European Economic Review*, 2009, 53 (1), 1–18.
- Lippens, Louis, Siel Vermeiren, and Stijn Baert**, “The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments,” *European Economic Review*, 2023, 151, 104315.
- List, John A**, “The nature and extent of discrimination in the marketplace: Evidence from the field,” *The Quarterly Journal of Economics*, 2004, 119 (1), 49–89.
- Lundberg, Shelly and Jenna Stearns**, “Women in Economics: Stalled Progress,” *Journal of Economic Perspectives*, 2019, 33 (1), 3–22.
- Maniadis, Zacharias, Fabio Tufano, and John A. List**, “To Replicate or Not to Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study,” *The Economic Journal*, 2017, 127 (605), F209–F235.
- Nagel, Rosemarie**, “Unraveling in guessing games: An experimental study,” *American Economic Review*, 1995, 85 (5), 1313–1326.
- Neumark, David**, “Experimental research on labor market discrimination,” *Journal of Economic Literature*, 2018, 56 (3), 799–866.

- Niederle, Muriel and Lise Vesterlund**, “Do women shy away from competition? Do men compete too much?,” *The Quarterly Journal of Economics*, 2007, 122 (3), 1067–1101.
- Oaxaca, Ronald**, “Male-female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 1973, 14 (3), 693–709.
- Onuchic, Paula**, “Recent Contributions to the Theories of Discrimination,” London School of Economics, Mimeo 2024.
- Oreopoulos, Philip**, “Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes,” *American Economic Journal: Economic Policy*, 2011, 3 (4), 148–171.
- Pallais, Amanda and Emily G. Sands**, “Why the referential treatment? Evidence from field experiments on referrals,” *Journal of Political Economy*, 2016, 124 (6), 1793–1828.
- Parsons, Christopher A, Johan Sulaeman, Michael C Yates, and Daniel S Hamermesh**, “Strike three: Discrimination, incentives, and evaluation,” *American Economic Review*, 2011, 101 (4), 1410–1435.
- Pinal, Guillermo Del, Alex Madva, and Kevin Reuter**, “Stereotypes, conceptual centrality and gender bias: An empirical investigation,” *Ratio*, 2017, 30 (4), 384–410.
- Qualtrics**, “Qualtrics software, Version [January 2021] of Qualtrics. Copyright ©[2021],” Qualtrics, Provo, Utah, USA 2005. Available at: <https://www.qualtrics.com>.
- Ray, Debraj and Arthur Robson**, “Certified random: A new order for coauthorship,” *American Economic Review*, 2018, 108 (2), 489–520.
- Riach, Peter A and Judith Rich**, “Field experiments of discrimination in the market place,” *The Economic Journal*, 2002, 112 (483), F480–F518.
- Roth, Alvin E.**, “The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory,” *Journal of Political Economy*, 1984, 92 (6), 991–1016.
- Sarsons, Heather**, “Recognition for Group Work: Gender Differences in Academia,” *American Economic Review*, 2017, 107 (5), 141–45.
- , “Interpreting Signals in the Labor Market: Evidence from Medical Referrals,” University of British Columbia, Mimeo 2022.
- Schmader, Toni, Jessica Whitehead, and Vicki H. Wysocki**, “A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants,” *Sex Roles*, 2007, 57 (7-8), 509–514.
- Shaffer, Hannah and Emma Harrington**, “Discrimination in Sequential Systems: Prosecutors’ Response to Police Information,” Unpublished mimeo 2024.

- Storage, Daniel, Tessa ES Charlesworth, Mahzarin R Banaji, and Andrei Cimpian**, “Adults and children implicitly associate brilliance with men more than women,” *Journal of Experimental Social Psychology*, 2020, *90*, 104020.
- Stötzer, Lasse S and Florian Zimmermann**, “A note on motivated cognition and discriminatory beliefs,” *Games and Economic Behavior*, 2024, *147*, 554–562.
- Szymanski, Stefan**, “A market test for discrimination in the English professional soccer leagues,” *Journal of Political Economy*, 2000, *108* (3), 590–603.
- Tesar, Linda**, “Report of the Committee on the Status of Women in the Economics Profession,” *AEA Papers and Proceedings*, May 2025, *115*, 918–43.
- Thomson, Keela S and Daniel M Oppenheimer**, “Investigating an alternate form of the cognitive reflection test,” *Judgment and Decision Making*, 2016, *11* (1), 99–113.
- Trix, Frances and Carolyn Psenka**, “Exploring the color of glass: letters of recommendation for female and male faculty,” *Discourse & Society*, 2003, *14*, 191–220.
- Weber, Andrea and Christine Zulehner**, “Competition and gender prejudice: Are discriminatory employers doomed to fail?,” *Journal of the European Economic Association*, 2014, *12* (2), 492–521.
- Wu, Alice H**, “Gender bias among professionals: an identity-based interpretation,” *Review of Economics and Statistics*, 2020, *102* (5), 867–880.
- Zimmermann, Florian**, “The dynamics of motivated beliefs,” *American Economic Review*, 2020, *110* (2), 337–363.

Appendix

Contents

A	Additional information	38
A.1	Controls in the observational data analysis	38
A.2	Sample construction academic survey	39
B	Additional analysis	39
B.1	Descriptives for Observational Study	39
B.2	Results for Observational Study	42
B.2.1	Letter Choice and Placement	42
B.2.2	Robustness — Letter Choice	43
B.2.3	Robustness — Placement	46
B.3	Observational Study: Simulation Results	48
B.4	Experimental Studies: Descriptive statistics	50
B.5	Experimental Studies: Hiring decision	52
B.6	Experimental Studies: Differences in reference letters	54
B.7	Experimental Studies: Mechanisms	55
B.8	Experimental Studies: Discrimination decomposition	61
C	Experimental Studies: Instructions	66
C.1	Academic survey	66
C.1.1	Consent	66
C.1.2	Recommender survey	67
C.1.3	Recruiter Survey	71
C.1.4	Questionnaire (common to recommender and recruiter surveys)	75
C.2	Full reference letter	77
C.3	Online experiment	78
C.3.1	Workers	78
C.3.2	Recommenders	86
C.3.3	Recruiters	93
C.3.4	Questionnaire for recommenders and recruiters	99

A Additional information

A.1 Controls in the observational data analysis

In this section, we provide details on the additional controls included in the multinomial logit (Letter choice) and least squares (Placement) regressions for the observational data. All regressions include dummies for the year/cohort of the job market (2017-2021).

Candidates In addition to the candidate gender dummy, we include indicator variables for candidate self-reported racial and ethnic minority status (Asian, Black, American Indian, Hispanic, as well as a dummy if the ethnicity information is withheld). We also include indicators variables for PhD institution RePEc ranking (in ranges: top-25, 26-50, 51-100, etc. with institutions ranked above 500 or not ranked in RePEc as the omitted category), the years since they completed their PhD (0 or 1), a dummy for the broad research field and information on their job market paper (JMP) and publication status: the counts of total published articles, of top-5, top general interest and top field journals (see details in [Eberhardt et al., 2023](#)); a dummy for a single-authored JMP, the number of individuals and organisations/universities mentioned in the JMP acknowledgements, respectively, and a dummy if the JMP is missing any acknowledgments.

Recommenders We include a dummy for female recommenders as well as the RePEc rank of the writer’s institution (following the same strategy laid out above). We also add dummies for academic rank (assistant, associate, full professor, or other) and the count of years since the letter writer completed their PhD. Finally, we add a count of how many letters the recommender has written for candidates in our full sample (2017-2021).

Letters We add the total number of words in the letter and dummies for specific ‘signals’ added in the final paragraph of the letter: one for the case that a candidate is recommended positively (“I recommend them to all economics department on the market, including the top-10”), one for a ‘negative’ signal (“I recommend them to all economics department on the market, with the exception of the top-10”), and one which makes a comparison (“This candidate can be compared to X who placed at Y in the market two years ago.”).

Placement analysis The above description refers to the letter choice analysis, which is carried out at the letter level. In the placement analysis, we conduct least squares regressions at the candidate level, hence information at the writer and letter level are averaged by candidate: the average number of words across all letters received by a candidate, the share of letters containing positive/negative signals or comparisons, the proportion of female recommenders and of recommenders at top-25, 26-50, etc institutions, and the average number of years since PhD for all recommenders. We include the share of recommenders who are full professors and the average of letters they have written in our sample. Although our placement analysis excludes teaching jobs (or ‘adjunct’ positions or posts like ‘assistant professor of instruction’) our sample includes both assistant professor and post-doc positions, so we include a dummy for the latter.

A.2 Sample construction academic survey

Invitations to the academic survey were sent out in two waves to the email addresses of academics that we collected from publicly available websites.

For the first wave, we used email addresses of academics in the RePEc top 5% institutions (universities and colleges) based in the following countries: Australia, Belgium, Canada, Denmark, France, Germany, Hong Kong, Ireland, Israel, Italy, Netherlands, New Zealand, Norway, PR China, Singapore, South Korea, Spain, Sweden, Switzerland, the United States, and the United Kingdom. We excluded international organisations, central banks, and research-only or policy institutions.

Our goal was to reach an overall sample size of one thousand survey participants. As this was not reached in our first wave, we invited more academics to participate in a second wave. For this we extended our study population to the top 8% institutions in the US, UK, Ireland, Canada, and Australia, because institutions in these countries are the most likely to engage regularly with the economics job market.

B Additional analysis

B.1 Descriptives for Observational Study

Table [B.1](#) provides sample statistics for our letter analysis in the observational study. The unit of observation here is the reference letter ($n=8,624$). Table [B.2](#) provides sample statistics for our placement analysis in the observational study. The unit of observation here is the individual candidate ($n=1,862$), letter writer and letter-specific characteristics are averaged or constructed as described in the main text.

Table B.1: Summary Statistics — Letter Analysis

	Mean	SD	N
<i>Candidate:</i>			
Female	0.30	0.46	8624
Ethnic White	0.42	0.49	8624
Ethnic Asian	0.33	0.47	8624
Ethnic Black	0.02	0.13	8624
Ethnic American Indian	0.00	0.06	8624
Ethnic Hispanic	0.10	0.30	8624
dto withheld	0.13	0.33	8624
Years since PhD	0.10	0.30	8624
Publication count	0.68	1.38	8624
Top 5 count	0.01	0.09	8624
Top Field Journal count	0.02	0.16	8624
Top General Interest Journal count	0.01	0.09	8624
Single-Authored JMP	0.80	0.40	8624
Persons acknowledged in JMP	20.87	18.63	8624
Organisations acknowledged in JMP	8.40	11.06	8624
JMP acknowledgements missing	0.05	0.22	8624
PhD Institution top-25 (RePEc)	0.21	0.40	8624
PhD Institution top-26 to 50 (RePEc)	0.15	0.35	8624
PhD Institution top-51 to 100 (RePEc)	0.15	0.36	8624
PhD Institution top-101 to 200 (RePEc)	0.19	0.40	8624
PhD Institution top-201 to 500 (RePEc)	0.21	0.41	8624
PhD Institution outside top-500 or not listed (RePEc)	0.09	0.28	8624
<i>Letter writer:</i>			
Female	0.15	0.36	8624
Institution top-25 (RePEc)	0.20	0.40	8624
Institution top-26 to 50 (RePEc)	0.12	0.33	8624
Institution top-51 to 100 (RePEc)	0.16	0.37	8624
Institution top-101 to 200 (RePEc)	0.17	0.37	8624
Institution top-201 to 500 (RePEc)	0.18	0.38	8624
Institution outside top-500 or not listed (RePEc)	0.17	0.38	8624
Assistant Professor	0.09	0.29	8624
Associate Professor	0.18	0.39	8624
Full Professor/Chair	0.70	0.46	8624
Other position	0.02	0.15	8624
Years since PhD	19.91	11.10	8624
Total number of letters in our sample	4.10	3.36	8624
<i>Letter:</i>			
Total word count	1121.79	550.02	8624
Positive signal ("All departments hiring...")	0.26	0.44	8624
Negative signal ("All departments outside the top-20...")	0.13	0.34	8624
Comparison with past JM candidates	0.06	0.24	8624

Table B.2: Summary Statistics – Placement Analysis

	Mean	SD	N
<i>Candidate:</i>			
Top-200 Academic Placement (RePEc)	0.28	0.45	1862
Top-100 Academic Placement (RePEc)	0.17	0.38	1862
Female	0.32	0.47	1862
Ethnic White	0.41	0.49	1862
Ethnic Asian	0.36	0.48	1862
Ethnic Black	0.01	0.12	1862
Ethnic American Indian	0.00	0.06	1862
Ethnic Hispanic	0.09	0.29	1862
dto withheld	0.13	0.33	1862
PhD Institution top-25 (RePEc)	0.21	0.41	1862
PhD Institution top-26 to 50 (RePEc)	0.14	0.35	1862
PhD Institution top-51 to 100 (RePEc)	0.15	0.36	1862
PhD Institution top-101 to 200 (RePEc)	0.20	0.40	1862
PhD Institution top-201 to 500 (RePEc)	0.20	0.40	1862
PhD Institution outside top-500 or not listed (RePEc)	0.09	0.29	1862
Years since PhD	0.10	0.29	1862
Publication count	0.72	1.41	1862
Top 5 count	0.01	0.10	1862
Top Field Journal count	0.03	0.17	1862
Top General Interest Journal count	0.01	0.11	1862
Single-Authored JMP	0.79	0.41	1862
Persons acknowledged in JMP	21.74	18.90	1862
Organisations acknowledged in JMP	8.61	11.27	1862
JMP acknowledgements missing	0.05	0.22	1862
<i>Average Letter writer:</i>			
Female	0.15	0.22	1862
Institution top-25 (RePEc)	0.20	0.34	1862
Institution top-26 to 50 (RePEc)	0.12	0.28	1862
Institution top-51 to 100 (RePEc)	0.16	0.31	1862
Institution top-101 to 200 (RePEc)	0.17	0.31	1862
Institution outside top-200 or not listed (RePEc)	0.20	0.40	1862
Writer is a professor/chair	0.71	0.28	1862
Years since PhD	20.06	6.87	1862
<i>Average Letter:</i>			
Total word count	1127.55	366.83	1862
Positive signal ("All departments hiring...")	0.27	0.27	1862
Negative signal ("All departments outside the top-20...")	0.13	0.20	1862
Comparison with past JM candidates	0.07	0.14	1862
<i>Candidate-level letter type:</i>			
Ability candidate	0.24	0.43	1862
Grindstone candidate	0.21	0.41	1862
Mixed Ability-Grindstone candidate	0.27	0.45	1862
Neutral candidate	0.28	0.45	1862

B.2 Results for Observational Study

B.2.1 Letter Choice and Placement

Table B.3: Letter choice (multinomial logit) and placement determinants (OLS)

Dependent Variable	(1)	(2)	(3)	(4)
Estimator	Letter type (1-4)		Top-200 Placement (dummy)	
Unit of Analysis	MN Logit		OLS	
Sample	Reference Letter		Candidate	
Controls	All candidates		AP & Postdoc Placements	
	None	Full	None	Full
<i>Relative Propensity of a Female Candidate...</i>				
	<i>... to receive reference type</i>		<i>... to secure placement with...</i>	
Ability letter	-3.55*** (1.06)	-3.53*** (1.07)	-4.68 (6.71)	-1.98 (6.24)
Grindstone letter	2.80*** (0.94)	2.53*** (0.95)	-14.68** (6.20)	-13.97** (5.82)
Letter with mixed attributes	1.98* (1.03)	1.42 (1.05)	-8.90 (6.05)	-7.56 (5.79)
Baseline letter (neither attribute)	-1.23 (1.13)	-0.43 (1.13)	11.30*** (4.34)	12.94*** (4.12)
Letters	8624	8624		
Candidates	2881	2881	1862	1862
Writers	4257	4257		
Pseudo R ²	0.00	0.01	0.01	0.14

Note: In models (1) and (3) we only include year dummies, in (2) and (4) we further include full controls for candidate (including about their JMP), recommender (years since PhD, rank, gender, institutional rank), and letter characteristics (including predictions/signals about their placement prospects and other sentiments expressed in the letters). Results from the multinomial logit models in (1) and (2) are predicted probabilities (standard errors are clustered at the candidate level), the letter types are ability, grindstone, both, and neither (without any ordering imposed). Letter types (dichotomous) are determined by adopting the median td-idf value by candidate for ability and grindstone as cut-off — see main text for details. We exclude candidates who have graduated more than one year ago at the time of application. Results from the linear probability model (OLS) in (3) and (4) are marginal effects. Here, all the letter and writer characteristics are averages across the set of letters of each candidate, with writer academic rank replaced by share of recommenders who are full professors.

B.2.2 Robustness — Letter Choice

Table B.4: Letter choice (multinomial logit) — Heterogeneity by writer characteristics

Writer Characteristics	(1) All	(2) Professor Yes	(3) No	(4) PhD Cohort Early	(5) Late	(6) Gender M	(7) F	(8) Female Writer Prof	(9) Early
<i>Relative Propensity of a Female Candidate to receive reference type...</i>									
Ability letter	-3.53*** (1.07)	-4.92*** (1.28)	-0.13 (1.89)	-4.97*** (1.39)	-1.24 (1.61)	-4.02*** (1.18)	-1.19 (2.38)	-4.50 (3.25)	-5.17 (3.64)
Grindstone letter	2.53*** (0.95)	3.21*** (1.08)	1.00 (1.84)	4.44*** (1.19)	-0.35 (1.57)	2.37** (1.02)	3.14 (2.47)	5.77* (3.05)	9.24*** (3.23)
Letter w/ mixed attributes	1.42 (1.05)	1.84 (1.24)	0.02 (1.93)	0.34 (1.35)	2.83* (1.65)	1.31 (1.15)	1.47 (2.60)	-0.00 (3.57)	-0.94 (3.82)
Baseline letter	-0.43 (1.13)	-0.13 (1.37)	-0.90 (1.96)	0.19 (1.53)	-1.23 (1.68)	0.35 (1.24)	-3.42 (2.47)	-1.26 (3.38)	-3.14 (3.65)
Letters	8624	6064	2560	5060	3564	7340	1284	729	633
Candidates	2881	2726	1792	2624	2241	2833	1056	664	593
Writers	4257	2678	1579	2258	1999	3527	730	361	321
Pseudo R ²	0.01	0.01	0.02	0.01	0.02	0.01	0.05	0.07	0.09
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: We conduct heterogeneity analysis by splitting the sample by writer characteristics. All models include year dummies and full controls for candidate (including about their JMP), recommender (years since PhD, rank, gender, institutional rank), and letter characteristics (including predictions/signals about their placement prospects and other sentiments expressed in the letters). Results from the multinomial logit models are predicted probabilities (standard errors are clustered at the candidate level), the letter types are ability, grindstone, both, and neither (without any ordering imposed). Letter types (dichotomous) are determined by adopting the median td-idf value by candidate for ability and grindstone as cut-off. We exclude candidates who have graduated more than one year ago at the time of application as well as recommenders who are not in academia. Columns (2) and (3) split the sample by letter writer academic rank, (4) and (5) by the year they were awarded their PhD (before 2004 is labelled as ‘early’). Columns (6) and (7) are for male and female letter writers, respectively. In (8) and (9) we only look at female letter writers who are Professors or were awarded their PhD before 2004. The model in column (4) excludes the dummy for Assistant Professor rank, which is not identified.

Table B.5: Letter choice (multinomial logit) – Heterogeneity by Predicted Gender Awareness

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Propensity Cutoff (%ile)	N/A	60th		70th		80th	
Gender awareness	All	Low	High	Low	High	Low	High
Female candidate (Ability)	-3.50*** (1.13)	-4.03*** (1.39)	-2.63 (1.88)	-4.08*** (1.32)	-2.13 (2.10)	-4.45*** (1.25)	0.27 (2.43)
Female candidate (Grindstone)	2.26** (1.00)	3.52*** (1.23)	0.07 (1.73)	2.88** (1.15)	1.01 (1.99)	2.36** (1.09)	1.64 (2.47)
Female candidate (Mixed)	1.81 (1.12)	-0.21 (1.40)	5.38*** (1.87)	1.24 (1.30)	3.41 (2.10)	1.39 (1.22)	3.41 (2.63)
Female candidate (Baseline)	-0.57 (1.22)	0.72 (1.56)	-2.82 (1.90)	-0.04 (1.46)	-2.29 (2.08)	0.70 (1.34)	-5.31** (2.50)
Letters	7717	4911	2806	5560	2157	6387	1330
Candidates	2853	2452	1786	2590	1487	2751	1073
Writers	4023	2464	1563	2785	1241	3256	768
Pseudo R ²	0.01	0.01	0.02	0.01	0.03	0.01	0.05
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: We present split sample results using different cutoffs of the predicted propensity of gender awareness (based on the linear probability model in the academic survey data): the 60th, 70th and 80th percentiles, in each case splitting the sample into low and high ‘gender awareness’. All models include full controls for candidate (including about their JMP), recommender (years since PhD, rank, institutional rank), and letter characteristics (including prediction-s/signals about their placement prospects and other sentiments expressed in the letters). Results from the multinomial logit models are predicted probabilities (standard errors are clustered at the candidate level), the letter types are ability, grindstone, both, and neither (without any ordering imposed). Letter types (dichotomous) are determined by adopting the median td-idf value by candidate for ability and grindstone as cut-off. We exclude candidates who have graduated more than one year ago at the time of application as well as letter writers from non-academic institutions and those from countries which were not represented among participants in the academic survey.

Table B.6: Letter choice (multinomial logit) — Results using Academic Survey Weights

Survey Weights	(1) No	(2) No	(3) Yes	(4) Yes
Female candidate (Ability)	-3.50*** (1.07)	-3.47*** (1.08)	-1.79 (1.61)	-1.67 (1.59)
Female candidate (Grindstone)	2.78*** (0.95)	2.50*** (0.96)	0.79 (1.50)	0.57 (1.54)
Female candidate (Mixed)	2.17** (1.04)	1.56 (1.07)	1.35 (1.62)	0.06 (1.62)
Female candidate (Baseline)	-1.45 (1.15)	-0.59 (1.15)	-0.35 (1.68)	1.03 (1.68)
Letters	8427	8427	8427	8427
Candidates	2873	2873	2873	2873
Writers	4102	4102	4102	4102
Pseudo R ²	0.00	0.01	0.00	0.02
Controls	No	Yes	No	Yes

Note: We present weighted regressions for letter choice in our observational study, mimicking the participant distribution of our academic survey in the observational study. In models (1) and (3) we only include year dummies, in all others we further include full controls for candidate (including about their JMP), recommender (years since PhD, rank, gender, institutional rank), and letter characteristics (including predictions/signals about their placement prospects and other sentiments expressed in the letters). Results from the multinomial logit models are predicted probabilities (standard errors are clustered at the candidate level), the letter types are ability, grindstone, both, and neither (without any ordering imposed). Letter types (dichotomous) are determined by adopting the median td-idf value by candidate for ability and grindstone as cut-off. We exclude candidates who have graduated more than one year ago at the time of application as well as recommenders who are not in academia. Weights are constructed based on the proportion of letter writers (Senders in the Survey Experiment) who are (1) male/female, (2) assistant/associate/full professor, (3) from top-100 institutions/not. Weights applied here represent the Field Study Share/Survey Experiment Share.

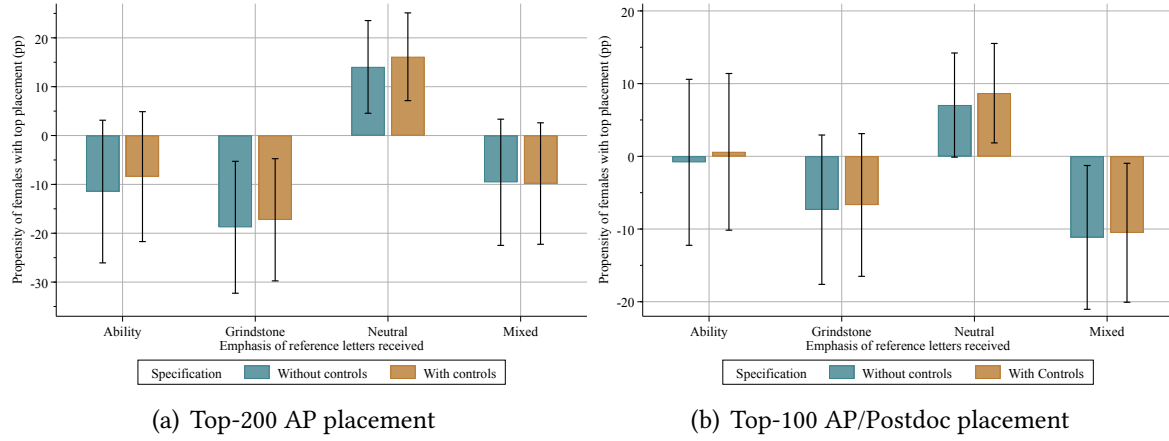
B.2.3 Robustness — Placement

Table B.7: Placement determinants — AP positions and top-100 placement (OLS)

	(1)	(2)	(3)	(4)	(5)	(6)
Placement position	APs & Postdocs		APs only		APs & Postdocs	
Placement rank	Top-200	Top-200	Top-200	Top-200	Top-100	Top-100
Controls	No	Yes	No	Yes	No	Yes
Ability letter	-4.68 (6.71)	-1.98 (6.24)	-11.47 (7.45)	-8.41 (6.79)	-0.82 (5.82)	0.62 (5.50)
Grindstone letter	-14.68** (6.20)	-13.97** (5.82)	-18.78*** (6.89)	-17.25*** (6.38)	-7.34 (5.24)	-6.69 (5.01)
Letter with mixed attributes	-8.90 (6.05)	-7.56 (5.79)	-9.57 (6.59)	-9.83 (6.34)	-11.15** (5.04)	-10.52** (4.88)
Baseline letter (neither attribute)	11.30*** (4.34)	12.94*** (4.12)	14.05*** (4.84)	16.13*** (4.58)	7.05* (3.65)	8.69** (3.49)
Candidates	1862	1862	1370	1370	1862	1862
Female Candidates	595 (32%)	595 (32%)	439 (32%)	439 (32%)	595 (32%)	595 (32%)
Uncond. top placement probability	0.26	0.26	0.23	0.23	0.17	0.17
Uncond. top placement prob. Fem	0.30	0.30	0.26	0.26	0.18	0.18
Female Share Ability type	30%	30%	31%	31%	30%	30%
Female Share Grindstone type	35%	35%	35%	35%	35%	35%
Female Share Mixed type	34%	34%	36%	36%	34%	34%
Adjusted R2	0.01	0.14	0.01	0.17	0.00	0.13

Note: The analysis mirrors that in columns (3) and (4) of Table B.3, which are replicated in (1) and (2) here. In columns (3) and (4), we limit the sample to candidates who secured Assistant Professor positions, and in columns (5) and (6) the dependent variable is now a dummy for a RePEC top-100, rather than a top-200, placement. In models (1), (3) and (5) we only include year dummies, in (2), (4) and (6) we further include full controls for candidate (including about their JMP), recommender (years since PhD, rank, gender, institutional rank), and letter characteristics (including predictions/signals about their placement prospects and other sentiments expressed in the letters) — importantly, all the letter and writer characteristics are averages across the set of letters of each candidate (writer academic rank is the share of recommenders who are full professors). Results presented are marginal effects. The unconditional probability of a top-200 (top-100) placement is 0.26 (0.17), in the reduced sample (only AP placements) in (3) and (4) it is 0.23.

Figure B.1: Robustness — Placement in the Observational Sample



Note: We plot marginal effects from the linear probability models presented in columns (3) and (4) as well as (5) and (6) of Table B.7 in panels (a) and (b), respectively. See Table notes for all details. Compared with the benchmark results for top-200 placement presented in the main text, panel (a) limits the sample to candidates who secure an assistant professor position (no postdocs), panel (b) uses the AP and postdoc sample but studies top-100 placement. The implications of combining the ‘Mixed’ letter type with either ‘ability’ or ‘grindstone’ are analysed in Table B.8 below.

Table B.8: Robustness — Placement in the Observational Sample — Combined Letter Types

Combination	(1) None	(2) None	(3) Ability+Mixed	(4) Ability+Mixed	(5) Grindstone+Mixed	(6) Grindstone+Mixed
Ability \times Female candidate	-4.68 (6.71)	-1.98 (6.24)	-7.48 (5.42)	-5.49 (5.14)	-4.68 (6.70)	-1.96 (6.24)
Grindstone \times Female candidate	-14.68** (6.20)	-13.97** (5.82)	-14.68** (6.20)	-14.00** (5.81)	-11.54** (5.31)	-10.41** (5.04)
Mixed \times Female candidate	-8.90 (6.05)	-7.56 (5.79)				
Female candidate	11.30*** (4.34)	12.94*** (4.12)	11.30*** (4.33)	12.96*** (4.12)	11.30*** (4.33)	12.92*** (4.12)
Candidates	1862	1862	1862	1862	1862	1862
Adjusted R2	0.01	0.14	0.01	0.14	0.01	0.14

Note: We experiment with the categorization of ‘mixed’ letters, which contain both ‘ability’ and ‘grindstone’ elements. In models (1) and (2) we keep ‘mixed’ letters separate, in (3) and (4) we assign them as ‘ability’ and in (5) and (6) as ‘grindstone’ letters. In models (1), (3), and (5) we only include year dummies, in all others we further include full controls for candidate (including about their JMP), the recommender team (share of professors, share female, proportion from different institutional ranks) and the letters (average word length, average ‘signals’) — the latter two are averaged by candidate. Results from the linear probability models are predicted probabilities (standard errors are clustered at the candidate level) of placing in a top-200 institution relative to male candidates.

Table B.9: Robustness — Placement in the Observational Sample — No Gender Interaction

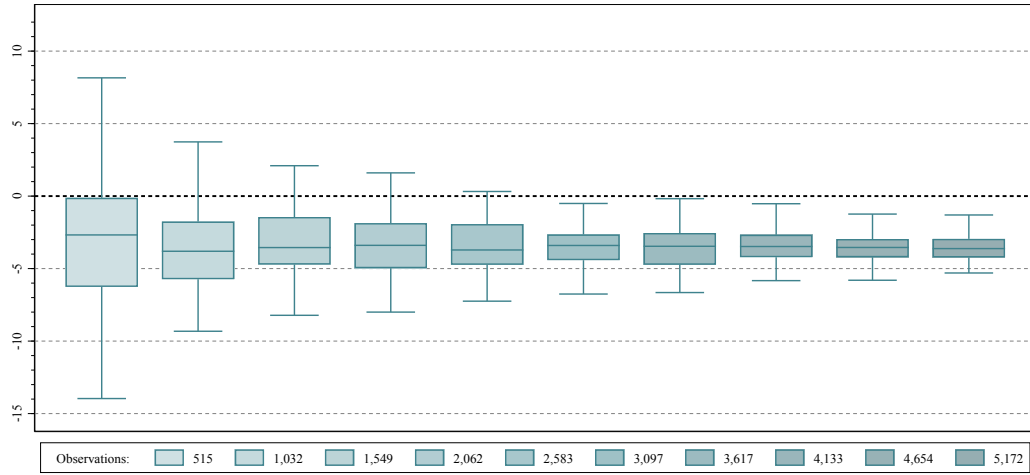
	(1)	(2)	(3)	(4)
Controls	Year dummies only		Full Controls	
Female Interaction	No	Yes	No	Yes
Ability	4.80 (2.95)	6.43* (3.42)	2.34 (2.76)	3.19 (3.16)
Ability \times Female candidate		-4.68 (6.71)		-1.98 (6.24)
Grindstone	-3.39 (2.91)	1.64 (3.56)	-2.61 (2.71)	2.21 (3.26)
Grindstone \times Female candidate		-14.68** (6.20)		-13.97** (5.82)
Mixed	0.20 (2.79)	3.09 (3.32)	-1.64 (2.66)	0.84 (3.13)
Mixed \times Female candidate		-8.90 (6.05)		-7.56 (5.79)
Baseline	23.35*** (2.96)	21.32*** (3.12)	-2.94 (6.94)	-4.50 (7.04)
Female candidate	4.48** (2.25)	11.30*** (4.34)	7.23*** (2.14)	12.94*** (4.12)
Candidates	1862	1862	1862	1862
Adjusted R2	0.00	0.01	0.14	0.14

Note: The table presents results for our full placement model (with female interaction terms) as well as for specifications without female interaction. See notes to Appendix Table B.8 for all other details.

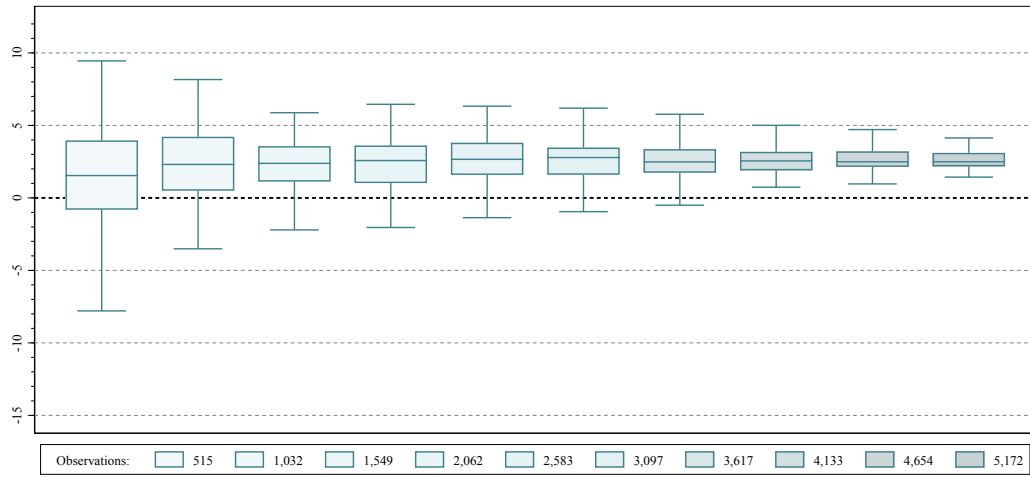
B.3 Observational Study: Simulation Results

In Figure B.2 we present the distribution of estimates from a simulation exercise performed on the observational dataset. Our full sample comprises 8,624 letters, from which we repeatedly (100 times) draw random sub-samples of varying sizes, ranging from 6% (around 500 letters) to 60% (around 5,100 letters) of the full sample. The aim of this exercise is to understand whether the result patterns for restricted samples match those of our experimental studies of similar sample size (just over 500 letter writers in the academic survey and just under 1,000 recommenders in the online experiment). The results presented in the three plots are distributions (box plots) of estimates for the female indicator in reduced samples, adopting the specification with additional controls. This analysis corresponds to the full sample observational study results presented in Figure 1 of the main text. While medians are relatively stable across simulated sub-samples, it can be seen for all three outcomes (ability, grindstone, or neutral letter) that the inter-quartile ranges (box) and the 95% confidence intervals (whiskers) become smaller as the sample size increases. The propensity for female candidates (relative to males) to receive an ability letters is statistically significantly lower once the sample size reaches 3,000 letters, while the propensity to receive a grindstone letter is statistically significantly higher in samples over 4,100 letters.

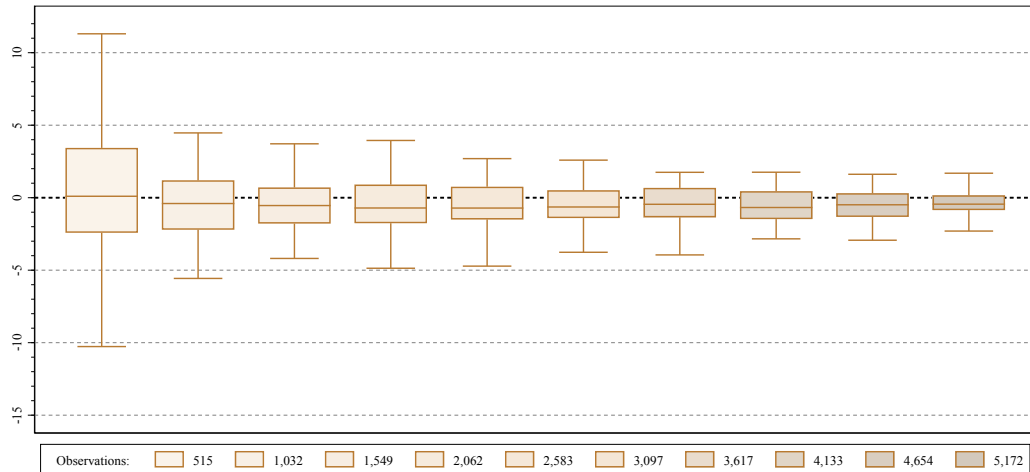
Figure B.2: Simulations — observational study — Multinomial logits (box plots for female indicator)



(a) Propensity of an Ability Letter



(b) Propensity of a Grindstone Letter



(c) Propensity of a Baseline Letter

Note: The figures are coefficient box plots (marginal effects) for marginal effects from multinomial logit regressions in 100 random subsamples of the reference letter data sample (8,624 letters) with the sample size indicated (from 500 to 5,100 letters, i.e. 6 to 60% of the full sample in steps of 6%). We follow the strategy laid out in Section 2 of the main text to categorize letters as emphasizing ‘ability’, ‘grindstone’, or neither (baseline) — we omit the analysis for letters with both attributes. The results presented are estimates for reduced samples corresponding to the full sample results presented in Figure 1 of the main text. We adopt the specification with additional controls.

B.4 Experimental Studies: Descriptive statistics

Tables B.10 and B.11 give an overview of the demographic characteristics in the academic survey and the online experiment. In addition to the characteristics presented in Figure 4 in the main text, the tables show that a large majority of participants in the academic survey has experience with hiring on the junior academic job market (87%) as well as with letter writing (77%). In the online experiment, we required participants to have at least an undergraduate degree, which is also the most common educational degree in the sample (67%).

Table B.10: Summary Statistics — Academic Survey

	Mean	SD	N
Female	0.26	0.44	1,018
<i>Academic rank:</i>			
Assistant Professor	0.34	0.47	1,019
Associate Professor	0.24	0.43	1,019
Full Professor	0.41	0.49	1,019
RePEc rank	291.89	206.62	1,019
Year of PhD	2,006.02	11.34	972
RePEc Top100	0.21	0.41	1,020
<i>Hiring experience:</i>			
Never	0.13	0.33	1,019
1-5	0.50	0.50	1,019
6-10	0.19	0.39	1,019
11-20	0.12	0.32	1,019
>20	0.07	0.25	1,019
<i>Letter writing experience:</i>			
None	0.23	0.42	1,017
1-5	0.30	0.46	1,017
6-10	0.18	0.38	1,017
11-20	0.12	0.32	1,017
_cons	0.17	0.37	1,017
Literature known	0.48	0.50	1,020

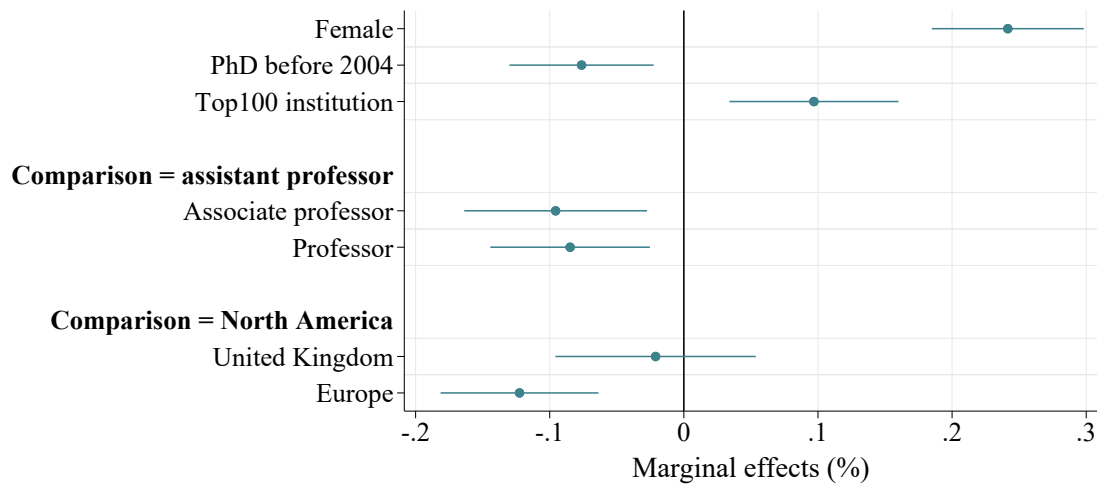
In the academic survey, we distinguish between participants with informed and uninformed views based on their knowledge of recent literature on gender differences in hiring. Figure B.3 shows which demographic characteristics in the academic survey correlate with participants having uninformed views.

In the online experiment, we distinguish between participants with stereotyped and non-stereotyped views. To do so, we ask recruiters whether male/female workers are characterized more by ability, effort, or equally possess both traits. Recommenders are asked to indicate what opinion they think most recruiters have. Figure B.4 shows how participants answered when asked about male or female candidates. We see a similar pattern for both recommenders and recruiters. Male candidates are much more likely to be perceived as mainly characterized by ability, while female candidates are more likely to be mainly seen as characterized by effort. Participants who express these opinions are defined as having stereotyped views in our study.

Table B.11: Summary Statistics — online experiment

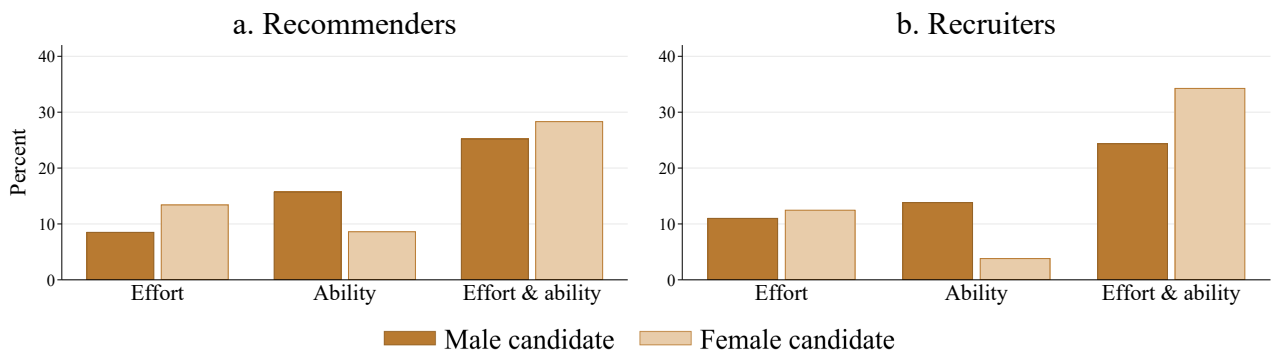
	Mean	SD	N
Female	0.50	0.50	1,913
Age	39.20	12.88	1,918
<i>Education:</i>			
High school	0.01	0.08	1,918
Undergraduate	0.67	0.47	1,918
Masters	0.27	0.45	1,918
PhD	0.04	0.21	1,918
<i>Nationality:</i>			
UK	0.89	0.31	1,711
US	0.07	0.25	1,711
Others	0.04	0.19	1,711
Stereotyped views	0.28	0.45	1,918

Figure B.3: Correlates of gender views in the academic survey



Note: The figure shows marginal effects from regressing gender views on demographic characteristics in the academic survey. Each estimate comes from a separate regression using only one explanatory variable. The dependent variable is 0 if the participant is classified as having uninformed views and 1 if they are classified as having informed views. Whiskers represent 90% confidence intervals.

Figure B.4: Perception of men and women in the online experiment



Note: We asked participants whether they thought that women/men tend to be characterized more by high effort, high ability or equally by high effort and high ability.

B.5 Experimental Studies: Hiring decision

Columns 1 and 2 in Table B.12 present results from a multinomial logit regression of hiring decisions on candidate gender and other control variables in the academic survey. Table B.13 does the same for the online experiment. These results are presented in Figure 6 of the main text.

In addition to looking at the full sample, we explore whether hiring decisions depend on the recruiters' gender views. Columns 3 and 4 show results for the sub-sample of informed/non-stereotyped recruiters, while columns 5 and 6 show results for uninformed/stereotyped recruiters. We see that the same qualitative pattern holds in both groups, but some differences emerge. In the online experiment, the effects are larger in magnitude for participants with stereotyped views. In the academic survey, recruiters who have informed gender views place a larger penalty on grindstone letters for women relative to men, while those who have uninformed views place a larger premium on ability letters for women relative to men.

Table B.12: Hiring choice in the academic survey - Weighted multinomial logits

	Full sample		Informed		Uninformed	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	6.21 (5.54)	7.83 (5.44)	3.16 (8.38)	5.10 (8.27)	9.99 (6.94)	11.54* (6.18)
Grindstone - Female candidate	-7.61* (4.61)	-9.10** (4.53)	-11.43 (6.99)	-13.02* (6.95)	-4.53 (5.82)	-6.28 (5.02)
Neutral - Female candidate	1.39 (3.98)	1.27 (3.98)	8.27 (5.43)	7.92 (5.59)	-5.46 (5.70)	-5.26 (5.14)
Observations	513	513	257	257	256	256
Referrer controls	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.10	0.12	0.07	0.11	0.16	0.21

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

Note: The dependent variable is 0 if a candidate with a neutral letter was chosen, 1 if a candidate with a grindstone letter was chosen, and 2 if a candidate with an ability letter was chosen. Recruiter controls include gender, academic position, institutional rank, letter writing experience, and hiring experience. All regressions control for order effects.

In the main text, we do not weight the experimental sample. As men are often more likely to be in positions of power, our results are thus a more conservative estimate for any gender differences. As a robustness check, we re-weight the experimental sample using the gender distribution in the observational sample as the underlying distribution. Our findings on the hiring side are presented in Table B.14. Results indeed become stronger/more differentiated after re-weighting. We find that women are now significantly less likely to be hired with a grindstone or neutral letter compared to men in the full sample.

Table B.13: Hiring choice in the online experiment - multinomial logits

	Full sample		Non-stereotyped		Stereotyped	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	6.46** (3.16)	6.09* (3.15)	6.54* (3.66)	5.87 (3.64)	7.83 (6.38)	8.88 (6.42)
Grindstone - Female candidate	-3.63 (2.94)	-3.27 (2.92)	-1.92 (3.36)	-1.36 (3.35)	-9.54 (6.04)	-11.19* (5.92)
Neutral - Female candidate	-2.84 (2.15)	-2.83 (2.19)	-4.61* (2.48)	-4.51* (2.51)	1.71 (4.54)	2.32 (4.68)
Observations	959	942	706	697	253	245
Recruiter controls	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.01	0.04	0.01	0.05	0.03	0.09

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

Note: The dependent variable is 0 if a candidate with a neutral letter was chosen, 1 if a candidate with a grindstone letter was chosen, and 2 if a candidate with an ability letter was chosen. Recruiter controls include age, gender, education level, nationality, strategic thinking ability, risk aversion, perceived candidate quality, and perceived required skills to perform well in the quiz. All regressions control for order effects.

Table B.14: Hiring choice in the online experiment - weighted multinomial logits

	Full sample		Non-stereotyped		Stereotyped	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	11.91*** (3.69)	11.79*** (3.71)	11.93*** (4.19)	11.44*** (4.25)	12.72 (7.82)	15.85** (7.86)
Grindstone - Female candidate	-6.34* (3.46)	-6.02* (3.46)	-5.87 (3.90)	-5.18 (3.92)	-7.37 (7.45)	-10.24 (7.41)
Neutral - Female candidate	-5.57** (2.51)	-5.77** (2.51)	-6.07** (2.79)	-6.26** (2.82)	-5.35 (5.60)	-5.62 (5.73)
Observations	959	942	706	697	253	245
Recruiter controls	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.02	0.05	0.02	0.06	0.05	0.10

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

Note: The dependent variable is 0 if a candidate with a neutral letter was chosen, 1 if a candidate with a grindstone letter was chosen, and 2 if a candidate with an ability letter was chosen. Recruiter controls include age, gender, education level, nationality, strategic thinking ability, risk aversion, perceived candidate quality, and perceived required skills to perform well in the quiz. Sampling weights are calculated taking the observational sample as the underlying population. All regressions control for order effects.

B.6 Experimental Studies: Differences in reference letters

Table B.15 presents results from a multinomial logit regression of letter choices on candidate gender and other control variables in the academic survey. Table B.16 does the same for the online experiment. These results are illustrated in Figures 7a and 8 of the main text .

Table B.15: Letter choice in academic survey - Weighted multinomial logits

	Full sample		Informed		Uninformed	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	-4.48 (6.01)	-2.77 (5.98)	9.87 (8.39)	13.48* (8.10)	-15.28** (7.47)	-14.38* (7.34)
Grindstone - Female candidate	-0.83 (3.68)	-2.04 (3.57)	-2.09 (5.53)	-4.94 (4.25)	-1.13 (4.80)	-0.39 (4.87)
Neutral - Female candidate	5.31 (5.62)	4.81 (5.57)	-7.78 (7.82)	-8.55 (7.57)	16.41** (7.04)	14.76** (6.93)
Observations	505	505	231	231	274	274
Referrer controls	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.02	0.04	0.04	0.12	0.07	0.08

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

Note: The dependent variable is 0 if a neutral letter was chosen, 1 if a grindstone letter was chosen, and 2 if an ability letter was chosen. Recommender controls include gender, academic position, institutional rank, letter writing experience, and hiring experience. All regressions control for order effects.

Table B.16: Letter choice in the online experiment - multinomial logits

	Full sample		Non-stereotyped		Stereotyped	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	0.68 (3.06)	0.41 (3.37)	8.25** (3.53)	7.65** (3.86)	-15.62*** (5.51)	-14.94** (6.21)
Grindstone - Female candidate	3.77 (2.64)	2.84 (2.85)	2.01 (3.23)	1.57 (3.52)	6.93 (4.28)	4.57 (4.48)
Neutral - Female candidate	-4.45 (3.15)	-3.25 (3.49)	-10.25*** (3.68)	-9.22** (4.05)	8.70 (5.66)	10.37* (6.04)
Observations	959	769	679	555	280	214
Referrer controls	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.03	0.06	0.03	0.06	0.06	0.14

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

Note: The dependent variable is 0 if a neutral letter was chosen, 1 if a grindstone letter was chosen, and 2 if an ability letter was chosen. Recommender controls include age, gender, education level, nationality, strategic thinking ability, risk aversion, perceived candidate quality, and perceived required skills to perform well in the quiz. All regressions control for order effects and the candidate's performance in Quiz A.

Table B.17 uses the same specification as Table B.16 but employs weights based on the gender distribution in the observational sample. Again, marginal effects are larger and more precisely estimated when using weights. In the sub-sample of participants with stereotyped views, women are now not

only significantly less likely to receive an ability letter but also significantly more likely to receive a grindstone or neutral one.

Table B.17: Letter choice in the online experiment - weighted multinomial logits

	Full sample		Non-stereotyped		Stereotyped	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	1.72 (3.87)	0.22 (4.26)	10.31** (4.40)	8.42* (4.76)	-21.56*** (6.76)	-26.52*** (6.89)
Grindstone - Female candidate	3.24 (3.07)	2.67 (3.25)	1.00 (3.72)	1.76 (3.91)	10.20** (4.78)	6.68 (4.46)
Neutral - Female candidate	-4.96 (3.86)	-2.89 (4.27)	-11.31** (4.44)	-10.18** (4.80)	11.35 (7.02)	19.84*** (7.21)
Observations	954	769	675	555	279	214
Referrer controls	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.02	0.06	0.03	0.07	0.07	0.16

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

Note: The dependent variable is 0 if a neutral letter was chosen, 1 if a grindstone letter was chosen, and 2 if an ability letter was chosen. Recommender controls include age, gender, education level, nationality, strategic thinking ability, risk aversion, perceived candidate quality, and perceived required skills to perform well in the quiz. Sampling weights are calculated by taking the observational sample as the underlying population. All regressions control for order effects and the candidate's performance in Quiz A.

B.7 Experimental Studies: Mechanisms

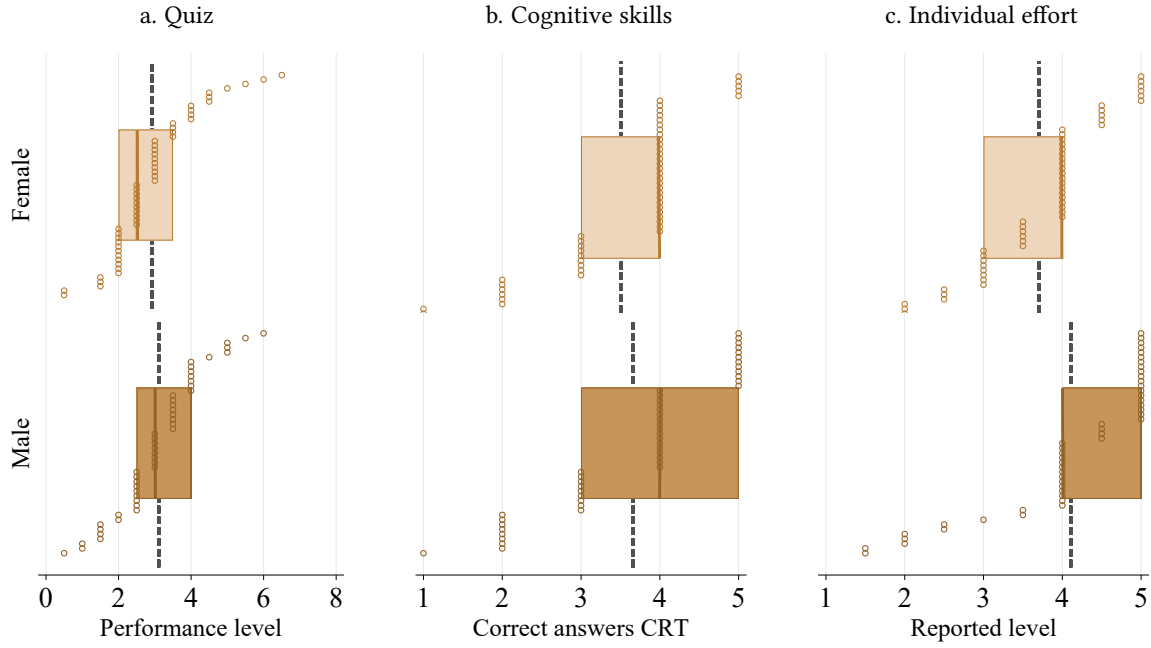
Figure B.5 displays comparisons between female and male candidates in the online experiment ($N=100$). The figure compares the average values and distributions for female and male candidates across three different metrics: quiz performance (panel a), cognitive skills (panel b) and individual effort (panel c). We find no difference in either the average Quiz performance (Wilcoxon rank-sum test, $z = 1.17$, $p = 0.25$) or measured cognitive skills (ibid., $z = 0.74$, $p = 0.46$).³⁴ The only dimension on which female and male candidates differ is average self-reported effort, with men claiming significantly higher levels (ibid., $z = 2.74$, $p = 0.006$). Thus, if anything, men should be more likely to receive a grindstone letter. Finally, if we restrict the analysis to workers with a performance of 5, 6, or 7 correct answers in Quiz A, all differences are insignificant.

Figures B.6 and B.7 show expected candidate qualities for recommenders and recruiters respectively in the online experiment. Recommenders observe a candidate's performance in Quiz A when forming their assessment, whereas recruiters only observe the text of the letter (and candidate gender). We find no significant differences between male and female candidates. If anything, female candidates of the same quality as their male counterparts (same Quiz A performance, same letter) are seen as having slightly higher ability and effort. This stresses that differences in letter or hiring choices are not driven by candidates being different or being perceived as different.

Tables B.18 and B.19 confirm that there are no differences in expected candidate qualities for recom-

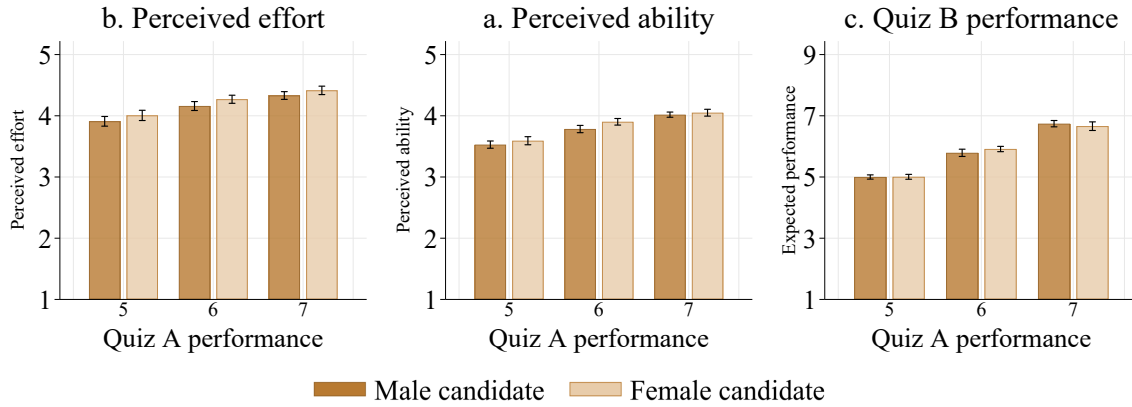
³⁴This is also true when looking at Quiz A and Quiz B performance separately.

Figure B.5: Comparison of female and male candidates



Note: Quiz performance level measures how many questions (out of 10) were solved correctly by the candidate. Cognitive ability is measured by the number of correctly answered questions in a Cognitive Reflection Test (CRT) ranging from 0 to 5. Effort is self-reported on a scale from 1 to 5. Small circles represent individual responses, dashed lines means, thick lines medians and boxes the interquartile range.

Figure B.6: Expected candidate qualities among recommenders in the online experiment



menders and recruiters, respectively — even when splitting the sample by gender views. The only significant difference we uncover is for recommenders with stereotyped views, who expect a worse Quiz B performance from a female than a comparable male candidate. In addition, Table B.19 shows that recruiters expect better performance, higher ability and effort from a candidate with an ability letter compared to one with a neutral one, as well as more effort from a candidate with a grindstone letter. This provides a manipulation check confirming that letters are perceived as intended.

Table B.20 shows results for regressing letter choice on candidate gender in the academic survey while controlling for beliefs about the best letter for a candidate. Table B.21 does the same for the online experiment. We see that any significant gender differences disappear after controlling for beliefs, highlighting the crucial role of strategic considerations. These are the results are represented

Figure B.7: Expected candidate qualities among recruiters in the online experiment

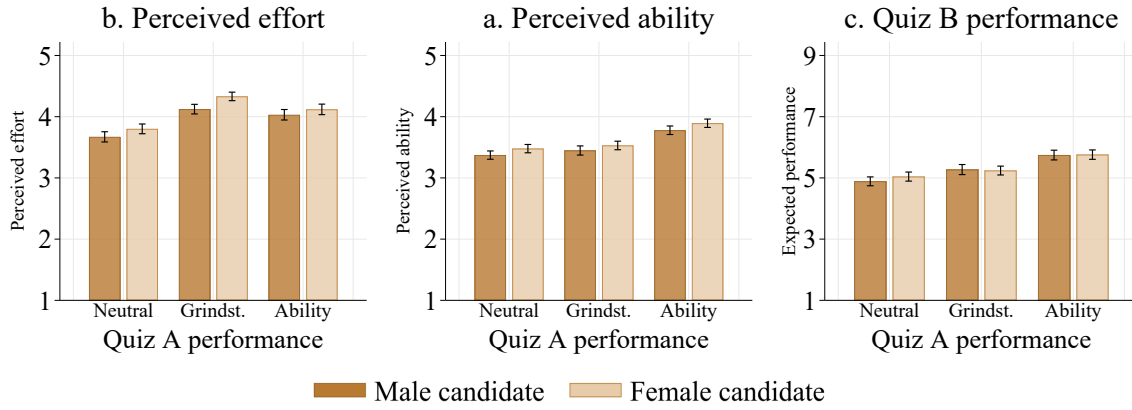


Table B.18: Expected candidate qualities among recommenders by gender views - OLS

	Non-stereotyped			Stereotyped		
	Effort	Ability	Quiz B	Effort	Ability	Quiz B
Female candidate	0.08 (0.10)	0.06 (0.08)	0.15 (0.09)	0.13 (0.17)	0.09 (0.14)	-0.36** (0.17)
<i>Quiz A performance (baseline=5)</i>						
6	0.27*** (0.10)	0.29*** (0.08)	1.03*** (0.11)	0.23 (0.15)	0.17 (0.11)	0.29 (0.23)
7	0.37*** (0.10)	0.49*** (0.07)	1.73*** (0.13)	0.55*** (0.14)	0.49*** (0.09)	1.77*** (0.14)
<i>Interactions</i>						
Female x 6	-0.01 (0.14)	0.07 (0.11)	-0.14 (0.15)	0.06 (0.22)	0.01 (0.18)	0.67** (0.29)
Female x 7	0.03 (0.14)	-0.00 (0.10)	-0.05 (0.19)	-0.09 (0.23)	-0.11 (0.17)	-0.20 (0.32)
Constant	3.94*** (0.07)	3.50*** (0.06)	4.92*** (0.06)	3.83*** (0.12)	3.59*** (0.08)	5.20*** (0.11)
Observations	679	679	679	280	280	280
R ²	0.05	0.12	0.34	0.08	0.10	0.32

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

Note: The dependent variable is either the recommender's perceived effort, ability or expected Quiz B performance of the candidate. Recommenders are informed about the candidate's gender and Quiz A performance when rating the candidate.

in Figure 10 of the main text.

Table B.22 shows results for the online sample when using weights. Results are qualitatively similar to Table B.21. The only difference is that in the sub-sample with gendered views, women are still significantly less likely to receive an ability letter, even after controlling for beliefs. This is again in line with the stronger effects we uncover on hiring and letter choices when re-weighting the online sample.

Table B.19: Expected candidate qualities among recruiters by gender views - OLS

	Non-stereotyped			Stereotyped		
	Effort	Ability	Quiz B	Effort	Ability	Quiz B
Female candidate	0.11 (0.11)	0.11 (0.09)	0.23 (0.19)	0.18 (0.17)	0.09 (0.16)	0.01 (0.33)
<i>Letters (baseline=Neutral)</i>						
Grindstone	0.44*** (0.11)	0.09 (0.09)	0.53*** (0.20)	0.49*** (0.17)	0.05 (0.15)	-0.03 (0.33)
Ability	0.38*** (0.12)	0.39*** (0.09)	0.91*** (0.20)	0.33** (0.16)	0.43*** (0.13)	0.70** (0.30)
<i>Interactions</i>						
Female x Grindstone	0.12 (0.14)	-0.02 (0.13)	-0.15 (0.28)	-0.02 (0.24)	-0.02 (0.22)	-0.19 (0.46)
Female x Ability	-0.10 (0.16)	0.08 (0.13)	-0.13 (0.28)	0.14 (0.25)	-0.25 (0.21)	-0.21 (0.47)
Constant	3.69*** (0.08)	3.37*** (0.06)	4.78*** (0.13)	3.63*** (0.10)	3.39*** (0.10)	5.15*** (0.21)
Observations	706	706	706	253	253	253
R ²	0.07	0.08	0.05	0.09	0.05	0.05

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

Note: The dependent variable is either the recommender's perceived effort, ability or expected Quiz B performance of the candidate. Recruiters are informed about the candidate's gender and letter when rating the candidate.

Table B.20: Letter choices and beliefs in the academic survey - weighted multinomial logits

	Full sample		Informed		Uninformed	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	-8.66 (6.52)	-0.08 (5.34)	5.90 (9.07)	4.65 (6.04)	-17.57** (7.56)	-5.12 (7.13)
Grindstone - Female candidate	0.45 (3.98)	-0.30 (3.48)	-3.06 (4.47)	-3.57 (3.65)	2.40 (4.81)	1.78 (4.58)
Neutral - Female candidate	8.20 (6.23)	0.39 (4.83)	-2.85 (8.53)	-1.07 (5.94)	15.17** (7.35)	3.34 (6.53)
Observations	415	413	182	181	233	232
Referrer controls	Yes	Yes	Yes	Yes	Yes	Yes
Beliefs	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.07	0.33	0.17	0.46	0.11	0.33

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

Note: The dependent variable is 0 if a neutral letter was chosen, 1 if a grindstone letter was chosen, and 2 if an ability letter was chosen. Beliefs control for which letter recommenders believe to be preferred by most recruiters. It is a categorical variable that takes the value 0 if the neutral letter is believed to be preferred by most, 1 if it is the grindstone, and 2 if it is the ability letter. Recommender controls include gender, academic position, institutional rank, letter writing experience, and hiring experience. All regressions control for order effects.

Table B.21: Letter choices and beliefs in the online experiment - multinomial logits

	Full sample		Non-stereotyped		Stereotyped	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	0.41 (3.37)	0.46 (3.17)	7.65** (3.86)	2.04 (3.73)	-14.94** (6.22)	-6.27 (6.91)
Grindstone - Female candidate	2.84 (2.85)	0.55 (2.71)	1.57 (3.52)	2.37 (3.34)	4.57 (4.63)	0.22 (4.75)
Neutral - Female candidate	-3.25 (3.49)	-1.01 (3.40)	-9.22** (4.05)	-4.41 (3.94)	10.37* (6.16)	6.05 (6.80)
Observations	769	769	555	555	214	214
Referrer controls	Yes	Yes	Yes	Yes	Yes	Yes
Beliefs	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.06	0.15	0.06	0.18	0.14	0.16

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

Note: The dependent variable is 0 if a neutral letter was chosen, 1 if a grindstone letter was chosen, and 2 if an ability letter was chosen. Beliefs control for which letter recommenders believe to be preferred by most recruiters. It is a categorical variable that takes the value 0 if the neutral letter is believed to be preferred by most, 1 if it is the grindstone, and 2 if it is the ability letter. Recommender controls include age, gender, education level, nationality, strategic thinking ability, risk aversion, perceived candidate quality, and perceived required skills to perform well in the quiz. All regressions control for order effects and the candidate's performance in Quiz A.

Table B.22: Letter choices and beliefs in the online experiment - weighted multinomial logits

	Full sample		Non-stereotyped		Stereotyped	
	(1)	(2)	(3)	(4)	(5)	(6)
Ability - Female candidate	0.22 (4.26)	-1.07 (3.98)	8.42* (4.76)	3.58 (4.61)	-25.79*** (7.12)	-15.24* (8.24)
Grindstone - Female candidate	2.67 (3.25)	1.15 (3.08)	1.76 (3.91)	1.94 (3.59)	7.95 (5.63)	2.29 (4.66)
Neutral - Female candidate	-2.89 (4.27)	-0.08 (4.04)	-10.18** (4.80)	-5.52 (4.50)	17.84** (7.80)	12.95 (8.04)
Observations	769	769	555	555	214	214
Referrer controls	Yes	Yes	Yes	Yes	Yes	Yes
Beliefs	No	Yes	No	Yes	No	Yes
Pseudo R ²	0.06	0.17	0.07	0.21	0.16	0.21

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

Note: The dependent variable is 0 if a neutral letter was chosen, 1 if a grindstone letter was chosen, and 2 if an ability letter was chosen. Recommender controls include age, gender, education level, nationality, strategic thinking ability, risk aversion, perceived candidate quality, and perceived required skills to perform well in the quiz. Sampling weights are calculated using the observational sample as the underlying population. All regressions control for order effects and the candidate's performance in Quiz A.

Figure B.8 decomposes the beliefs about which letter maximizes a candidate's probability to be hired by gender views. In both the academic survey and the online experiment, we find that participants with uninformed or stereotyped views are much more likely to believe that the ability letter is best for male candidates than for female candidates. In the online experiment, we observe the opposite pattern for people with non-stereotyped views, while participants with informed views in the academic survey have similar beliefs independent of the candidate's gender.

Figure B.8: Beliefs about the most effective letter by gender views



Note: Share of recommenders believing a given letter maximizing the candidate's probability to be hired. In reality, the best letter for a candidate (independent of gender) is the ability one.

B.8 Experimental Studies: Discrimination decomposition

Gender views

Table B.23 shows the same decomposition as Table 1 in the main text. In addition to calculating the average direct discrimination (ADD) based on the distribution of letters in the overall sample, the table also reports results from using the distribution of letters for female and male candidates instead. Results are qualitatively the same, independently of which distribution is used.

Table B.23: Decomposition of total into systemic and direct discrimination - Gender views

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1)	(2)	(3)	(5)	(6)	(7)
	Full sample	Informed views Yes	No	Full sample	Stereotyped views No	Yes
Total discrimination						
TD	1.12	7.63	-4.65	1.34	4.13	-5.21
Direct discrimination						
ADD ($p(s_k)$ sample)	3.13	3.40	2.79	0.38	0.10	0.99
ADD ($p(s_k)$ female)	3.08	3.79	2.32	0.38	0.43	0.18
ADD ($p(s_k)$ male)	3.19	3.00	3.23	0.37	-0.26	1.67
Systemic discrimination						
SD ($p(s_k)$ sample)	-2.01	4.23	-7.44	0.96	4.03	-6.19
SD ($p(s_k)$ female)	-1.96	3.83	-6.97	0.96	3.70	-5.39
SD ($p(s_k)$ male)	-2.07	4.63	-7.88	0.97	4.40	-6.88
Recommenders	506	232	274	959	679	280
Recruiters	514	514	514	959	959	959

Note: The table reports average differences in hiring probability for female versus male candidates across 1,000 simulations per column.

Table B.24 uses weighted regressions for the decomposition in the online experiment. Again, results are qualitatively similar to the unweighted results, albeit more pronounced.

Table B.24: Decomposition of total into systemic and direct discrimination - Gender views, weighted online sample

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1)	(2)	(3)	(5)	(6)	(7)
	Full sample	Informed views Yes	No	Full sample	Stereotyped views No	Yes
Total discrimination						
TD	1.12	7.63	-4.65	2.67	6.10	-6.32
Direct discrimination						
ADD ($p(s_k)$ sample)	3.13	3.40	2.79	1.26	0.93	2.17
ADD ($p(s_k)$ female)	3.08	3.79	2.32	1.39	1.82	0.21
ADD ($p(s_k)$ male)	3.19	3.00	3.23	1.12	-0.04	3.97
Systemic discrimination						
SD ($p(s_k)$ sample)	-2.01	4.23	-7.44	1.41	5.18	-8.50
SD ($p(s_k)$ female)	-1.96	3.83	-6.97	1.28	4.28	-6.53
SD ($p(s_k)$ male)	-2.07	4.63	-7.88	1.54	6.14	-10.29
Recommenders	506	232	274	959	679	280
Recruiters	514	514	514	959	959	959

Note: The table reports average differences in hiring probability for female versus male candidates across 1,000 simulations per column.

Recruiter sub-samples

Table B.25 shows decomposition results when distinguishing between *recruiters* with different gender views. As can be seen there are no systematic differences between recruiters depending on their gender views. This underscores that it is the behavior of recommenders and not that of recruiters driving our results.

Table B.25: Decomposition of total into systemic and direct discrimination - recruiter sub-samples

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1)	(2)	(3)	(4)	(5)	(6)
	Full sample	Informed views Yes	No	Full sample	Stereotyped views No	Yes
Total discrimination	1.12	2.05	1.11	1.34	2.85	0.93
Direct discrimination						
ADD ($p(s_k)$ sample)	3.13	3.74	3.63	0.38	1.91	-0.08
ADD ($p(s_k)$ female)	3.08	3.40	3.85	0.38	1.67	0.01
ADD ($p(s_k)$ male)	3.19	4.05	3.41	0.37	2.16	-0.17
Systemic discrimination						
SD ($p(s_k)$ sample)	-2.01	-1.69	-2.51	0.96	0.94	1.02
SD ($p(s_k)$ female)	-1.96	-1.36	-2.74	0.96	1.19	0.93
SD ($p(s_k)$ male)	-2.07	-2.01	-2.29	0.97	0.69	1.11
Recommenders	506	506	506	959	959	959
Recruiters	514	257	257	959	706	253

Note: The table reports average estimates across 1,000 simulations per column.

Beliefs

Table B.26 shows decomposition results for the sub-samples of participants with correct or incorrect beliefs. This is analogous to Table 2 in the main text. However, here we additionally report results when calculating ADD based on the distribution of letters among male and female candidates. The findings are qualitatively very similar to using the distribution in the overall sample. Table B.27 reports the same analysis with weighted regressions for the online sample.

Table B.26: Decomposition of total into systemic and direct discrimination - Beliefs

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1)	(2)	(3)	(5)	(6)	(7)
	Full sample	Correct beliefs Yes	No	Full sample	Correct beliefs Yes	No
Total discrimination						
TD	1.12	7.74	-7.66	1.34	2.35	0.36
Direct discrimination						
ADD ($p(s_k)$ sample)	3.13	4.31	0.45	0.38	2.11	-0.83
ADD ($p(s_k)$ female)	3.08	4.62	0.27	0.38	2.13	-0.87
ADD ($p(s_k)$ male)	3.19	4.04	0.67	0.37	2.09	-0.80
Systemic discrimination						
SD ($p(s_k)$ sample)	-2.01	3.43	-8.12	0.96	0.24	1.20
SD ($p(s_k)$ female)	-1.96	3.12	-7.94	0.96	0.22	1.23
SD ($p(s_k)$ male)	-2.07	3.69	-8.33	0.97	0.27	1.16
Recommenders	506	344	160	959	391	568
Recruiters	514	514	514	959	959	959

Note: The table reports average differences in hiring probability for female versus male candidates across 1,000 simulations per column. Correct beliefs restricts the analysis to recommenders who correctly believe that the ability letter will be most successful. In the academic sample, two recommenders did not answer the beliefs question. The table presents results using the distribution of letters for all candidates.

Table B.27: Decomposition of total into systemic and direct discrimination - Beliefs, weighted online sample

	<i>Academic survey</i>			<i>Online experiment</i>		
	(1)	(2)	(3)	(5)	(6)	(7)
	Full sample	Correct beliefs Yes	No	Full sample	Correct beliefs Yes	No
Total discrimination						
TD	1.12	7.74	-7.66	2.67	6.45	-1.20
Direct discrimination						
ADD ($p(s_k)$ sample)	3.13	4.31	0.45	1.26	4.46	-1.38
ADD ($p(s_k)$ female)	3.08	4.62	0.27	1.39	4.88	-1.67
ADD ($p(s_k)$ male)	3.19	4.04	0.67	1.12	3.99	-1.09
Systemic discrimination						
SD ($p(s_k)$ sample)	-2.01	3.43	-8.12	1.41	1.99	0.18
SD ($p(s_k)$ female)	-1.96	3.12	-7.94	1.28	1.57	0.48
SD ($p(s_k)$ male)	-2.07	3.69	-8.33	1.54	2.45	-0.10
Recommenders	506	344	160	959	391	568
Recruiters	514	514	514	959	959	959

Note: The table reports average differences in hiring probability for female versus male candidates across 1,000 simulations per column. Correct beliefs restricts the analysis to recommenders who correctly believe that the ability letter will be most successful. In the academic sample, two recommenders did not answer the beliefs question. The table presents results using the distribution of letters for all candidates.

C Experimental Studies: Instructions

Below we provide the full instructions for the academic survey and the online experiment. We added notes (in italics) to provide more details on the survey flow and to explain which questions were randomized.

C.1 Academic survey

Thanks a lot for participating in this study!

C.1.1 Consent

Study information

The aim of this study is to develop a better understanding of the decision making of academics in the recruitment process for junior faculty members. In the following, we will show you a short biography of a **hypothetical candidate along with the final paragraph of three different reference letters**. We will then ask you some questions about how you perceive the final paragraphs.

The study should take **5-7 minutes**. Your participation is crucial for our research, and we are very grateful for your support. As a thank-you token, we donate **£2 to UNICEF** for every completed survey. In addition, for completing this survey you can win one of fifty **£50-Amazon vouchers**. To be considered for the vouchers, you may leave your email address when prompted at the end of the survey. By completing this survey you receive one ticket to win a voucher. Throughout the study you will have the chance to receive more tickets depending on your responses. Winners will be chosen at random after all responses have been collected and will be contacted by the end of July. Similarly, if you would like to be informed about the **results of our study** once it is completed, you may indicate this at the end. Your email address will only be used for these purposes.

Contact information

If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact us at cedex@nottingham.ac.uk.

Consent

Please note that you are free to withdraw at any point while completing the questionnaire. All data collected will be kept confidential and used for research purposes only.

By clicking the **‘I consent’ button**, you agree to the following:

1. I have read and understood the information about the study.

2. I confirm that I have been given enough information about this study, and I voluntarily agree to take part.

If you do not agree, please click the '**I don't consent**' button. Thank you for your time.

- I consent
- I do not consent

C.1.2 Recommender survey

Note: The following questions were only presented to half of the participants who were randomly assigned to act as recommenders.

Application - NAME Miller

Note: Half of the participants are randomized to see Julia instead of NAME, the other half to see Thomas

Imagine you agreed to write a reference letter for NAME Miller, a PhD student who goes on the economics job market. In the box below we provide a very **short description** of this hypothetical student:

NAME Miller is a fifth-year PhD student in Economics with a focus on applied micro. NAME's PhD institution is ranked among the top 20 institutions globally. NAME's job market paper exploits a large and quasi-randomized 4 year-long income tax holiday to identify intertemporal labor responses to taxation among high-wage earners. It is likely to be published in a top general interest journal or at least in a top field journal.

We will now show you the final paragraph of three reference letters. This final paragraph can be seen as a summary and a conclusion of the letter. **Your task is to choose your preferred paragraph for this student's letter.** By preferred we mean, which paragraph is most in line with how you would end a reference letter for this student. You will be given the option to read the full letter, although this is not necessary for the study.

Note: There was a button participants could click to read the full letter. The full letter is provided in Section [C.2](#) below. It is identical across the three letter types save for the final paragraph.

Reference letter - NAME Miller

1) Imagine you were asked to write a reference letter for NAME. Which of the three final paragraphs would you choose?

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

Note: The order of the three final paragraphs is randomized across participants. Passages were highlighted in bold as indicated below.

Letter 1 - final paragraph

NAME is a remarkable job market applicant who **has the talent to come up with interesting, innovative, and policy-relevant ideas**. This is well reflected in HIS/HER completed projects. Furthermore, NAME has a very promising and unusually developed research agenda. NAME has shown that HE/SHE **has the technical skills and versatility to push HIS/HER projects until completion**. I expect NAME to continue generating good research ideas and have a productive career. As for HIS/HER classroom performance, HE/SHE has already proven to be a great teacher. NAME received excellent student evaluations for the courses HE/SHE TA'd for me. Moreover, NAME is also a pleasant and collegial person who is a pleasure to interact with. In sum, NAME is an **extremely bright and creative deep thinker**. All departments should consider HIS/HER application seriously and carefully. I recommend NAME enthusiastically.

Note: This is the ability letter.

Letter 2 - final paragraph

NAME is a remarkable job market applicant who **has always been keen to put in the effort needed to come up with interesting, innovative, and policy-relevant ideas**. This is well reflected in HIS/HER completed projects. Furthermore, NAME has a very promising and unusually developed research agenda. NAME has shown that SHE/HE **has the drive and stamina to push HIS/HER projects until completion**. I expect NAME to continue generating good research ideas and have a productive career. As for HIS/HER classroom performance, HE/SHE has already proven to be a great teacher. NAME received excellent student evaluations for the courses HE/SHE TA'd for me. Moreover, NAME is also a pleasant and collegial person who is a pleasure to interact with. In sum, NAME is an **extremely hard-working and dedicated researcher**. All departments should consider HIS/HER application seriously and carefully. I recommend NAME enthusiastically.

Note: This is the grindstone letter.

Letter 3 - final paragraph

NAME is a remarkable job market applicant who **has come up with interesting, innovative and policy relevant ideas**. This is well reflected in HIS/HER completed projects. Furthermore, NAME has a very promising and unusually developed research agenda. NAME has shown that HE/SHE **can push HIS/HER projects until completion**. I expect NAME to continue generating good research ideas and have a productive career. As for HIS/HER classroom performance, HE/SHE has already proven to be a great teacher. NAME received excellent student evaluations for the courses HE/SHE TA'd for me. Moreover, NAME is also a pleasant and collegial person who is a pleasure to interact with. In sum, NAME is a **great candidate**. All departments should consider HIS/HER application seriously and carefully. I recommend NAME enthusiastically.

Note: This is the neutral letter.

Hiring prospects for NAME Miller - Own evaluation

2) Now imagine that you yourself received an application for a candidate with NAME's profile, how likely would you be to invite NAME for an interview if the application was supported by...

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

3) Among the three final paragraphs, which one convinces you most to invite NAME to an interview?

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

Hiring prospects for NAME Miller - Views of other participants

4) We show the selected final paragraphs together with NAME's profile to **other participants** in the study who are all academic economists. Among all participants, what do you think was the most frequently chosen answer to the question: "How likely would you be to invite NAME for an interview?" after participants read...

For each answer that you guess correctly you will receive 5 additional tickets to win one of the £50 Amazon vouchers.

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

5) What do you think most **other participants** answered to the following question: "Among the three final paragraphs, which one convinces you most to invite NAME to an interview?"

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

C.1.3 Recruiter Survey

Application - NAME Miller

Note: Half of the participants are randomized to see Julia instead of NAME, the other half to see Thomas

Imagine you are recruiting on the economics job market. In the box below we provide a very **short description** of this hypothetical student:

NAME Miller is a fifth-year PhD student in Economics with a focus on applied micro. NAME's PhD institution is ranked among the top 20 institutions globally. NAME's job market paper exploits a large and quasi-randomized 4 year-long income tax holiday to identify intertemporal labor responses to taxation among high-wage earners. It is likely to be published in a top general interest journal or at least in a top field journal.

In addition, you received a reference letter for this candidate. We will now show you the final paragraph of the letter. This final paragraph can be seen as a summary and a conclusion of the letter. **Your task is to decide whether you want to invite NAME for an interview.** Although this is not necessary for the study, you may click on the link below the paragraph to read the full letter.

Note: There was a button participants could click to read the full letter. The full letter is available in Section [C.2](#) below. It is identical across letter types besides the final paragraph.

Letter 1 - NAME Miller

Note: The order of the three final paragraphs is randomized across participants.

Letter 1 - final paragraph

NAME is a remarkable job market applicant who **has the talent to come up with interesting, innovative, and policy-relevant ideas**. This is well reflected in HIS/HER completed projects. Furthermore, NAME has a very promising and unusually developed research agenda. NAME has shown that HE/SHE **has the technical skills and versatility to push HIS/HER projects until completion**. I expect NAME to continue generating good research ideas and have a productive career. As for HIS/HER classroom performance, HE/SHE has already proven to be a great teacher. NAME received excellent student evaluations for the courses HE/SHE TA'd for me. Moreover, NAME is also a pleasant and collegial person who is a pleasure to interact with. In sum, NAME is an **extremely bright and creative deep thinker**. All departments should consider HIS/HER application seriously and carefully. I recommend NAME enthusiastically.

Note: This is the ability letter.

1) How likely would you be to invite NAME for an interview?

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

2) We ask the same question to **other participants** in this study who are all academic economists. Out of all participants, what do you think was the most frequent answer to the question: “How likely would you be to invite NAME for an interview?”

If you guess correctly you will receive 5 additional tickets to win one of the £50 Amazon vouchers.

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

Letter 2 - NAME Miller

Imagine the same candidate received a different recommendation letter with the following final paragraph. Although this is not necessary for the study, you again have the option to read the full letter.

As a reminder we also show you again NAME’s profile:

NAME Miller is a fifth-year PhD student in Economics with a focus on applied micro. NAME’s PhD institution is ranked among the top 20 institutions globally. NAME’s job market paper exploits a large and quasi-randomized 4 year-long income tax holiday to identify intertemporal labor responses to taxation among high-wage earners. It is likely to be published in a top general interest journal or at least in a top field journal.

Letter 2 - final paragraph

NAME is a remarkable job market applicant who **has always been keen to put in the effort needed to come up with interesting, innovative, and policy-relevant ideas**. This is well reflected in HIS/HER completed projects. Furthermore, NAME has a very promising and unusually developed research agenda. NAME has shown that SHE/HE **has the drive and stamina to push HIS/HER projects until completion**. I expect NAME to continue generating good research ideas and have a productive career. As for HIS/HER classroom performance, HE/SHE has already proven to be a great teacher. NAME received excellent student evaluations for the courses HE/SHE TA’d for me. Moreover, NAME is also a pleasant and collegial person who is a pleasure to interact with. In sum, NAME is an **extremely hard-working and dedicated researcher**. All departments should consider HIS/HER application seriously and carefully. I recommend NAME enthusiastically.

Note: This is the grindstone letter.

3) How likely would you be to invite NAME for an interview?

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

4) We ask the same question to **other participants** in this study who are all academic economists. Out of all participants, what do you think was the most frequent answer to the question: “How likely would you be to invite NAME for an interview?”

If you guess correctly you will receive 5 additional tickets to win one of the £50 Amazon vouchers.

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

Letter 3 - NAME Miller

Finally, imagine the final paragraph of NAME’s reference letter reads as shown below.

As a reminder we also show you again NAME’s profile:

NAME Miller is a fifth-year PhD student in Economics with a focus on applied micro. NAME’s PhD institution is ranked among the top 20 institutions globally. NAME’s job market paper exploits a large and quasi-randomized 4 year-long income tax holiday to identify intertemporal labor responses to taxation among high-wage earners. It is likely to be published in a top general interest journal or at least in a top field journal.

Letter 3 - final paragraph

NAME is a remarkable job market applicant who **has come up with interesting, innovative and policy relevant ideas**. This is well reflected in HIS/HER completed projects. Furthermore, NAME has a very promising and unusually developed research agenda. NAME has shown that HE/SHE **can push HIS/HER projects until completion**. I expect NAME to continue generating good research ideas and have a productive career. As for HIS/HER classroom performance, HE/SHE has already proven to be a great teacher. NAME received excellent student evaluations for the courses HE/SHE TA’d for me. Moreover, NAME is also a pleasant and collegial person who is a pleasure to interact with. In sum, NAME is a **great candidate**. All departments should consider HIS/HER application seriously and carefully. I recommend NAME enthusiastically.

Note: This is the neutral letter.

5) How likely would you be to invite NAME for an interview?

Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).

6) We ask the same question to **other participants** in this study who are all academic economists. Out of all participants, what do you think was the most frequent answer to the question: “How likely would you be to invite NAME for an interview?”

If you guess correctly you will receive 5 additional tickets to win one of the £50 Amazon vouchers. *Note: Answers for each letter are presented on a Likert scale from 0 (definitely not) to 10 (definitely invite).*

7) Among the three final paragraphs, which one convinces you most to invite NAME to an interview?

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

8) We ask the same question to **other participants** in this study who are all academic economists. What do you think most other participants answered to the following question: “Among the three final paragraphs, which one convinces you most to invite NAME to an interview?”

If you guess correctly you will again receive 5 additional tickets to win one of the £50 Amazon vouchers. You can have another look at the final paragraphs by clicking on the corresponding button.

- Letter 1 - final paragraph
- Letter 2 - final paragraph
- Letter 3 - final paragraph

C.1.4 Questionnaire (common to recommender and recruiter surveys)

Thank you very much! You completed the main part of the survey. Before you go, we would like to ask you a few last questions about yourself and your hiring experience.

1) When considering whether to invite a job market candidate for an interview, how important are the following characteristics to you?

Please rank them from most to least important by **dragging the options** into your preferred order.

Most important

- Intellectual rigour
- Technical skills
- Hard-working
- Teaching experience
- Quality of the job market paper
- Rank of the PhD institution
- Social skills

Least important

2) Your hiring experience

a. How often have you been part of a junior hiring committee?

- Never
- 1-5
- 6-10
- 11-20
- >20

b. How many reference letters have you already written for PhD students (approximately)?

- Never
- 1-5

- 6-10
- 11-20
- >20

3) About this survey

a. How familiar are you with the literature on gender bias?

Note: Answers on a Likert scale from 0 (very unfamiliar) to 5 (very familiar).

b. Have you read any of the following papers?

- Dupas et al. (2021). “Gender and the Dynamics of Economics Seminars”
- Koffi (2021). “Innovative ideas and gender inequality”
- Hengel (2022). “Are Women Held to Higher Standards? Evidence from Peer Review”
- Wu (2020). “Gender Bias among Professionals: An Identity-Based Interpretation”
- Eberhardt et al. (2022). “Gender Differences in Reference Letters: Evidence from the Economics Job Market”

Note: Order of papers is randomized. Answer options are "Yes" and "No".

4) Lottery

a. Would you like to participate in the draw for the Amazon vouchers?

- Yes
- No

b. Would you like to be informed of the results of this study?

- Yes
- No

In case you answered yes to either question 4a or 4b, please enter your **email address** here:

Thank you very much again for your participation!

If you have any comments, you can leave them below. Any feedback/ additional thoughts regarding this study are highly appreciated.

Please click finish to submit your answers.

C.2 Full reference letter

In the study participants could click a link to read the full reference letter of the candidate. The letter was identical in all treatments except for the last paragraph as outlined above. The full letter can be found below:

To whom it may concern

Letter of Recommendation for NAME Miller

This letter is to recommend NAME, a fifth year graduate student at my university, for a position in your department. NAME is an applied-micro researcher, working on taxation in the context of country YYY. I have been one of NAME's advisors over the past four years. Hence, I know NAME well. NAME has produced a very interesting job market paper, convincingly identifying the effects of income taxation on high earning employees using a unique natural experiment carried out in YYY. NAME has been very productive and has two other completed papers on taxation, which NAME will be able to submit very soon.

NAME's job market paper exploits a large and quasi-randomized 4 year-long income tax holiday to identify intertemporal labor responses to taxation among high-wage earners. Eligibility for the tax holiday was based on whether past (and thus pre-determined) wage earnings were below a fixed threshold, creating a discontinuity that affected workers active in the same labor market with sharply different marginal tax rates. Using rich population-wide administrative data and a regression discontinuity design, NAME estimates a precise and large elasticity of earnings with respect to the marginal tax rate.

Interestingly, NAME finds that behavioural responses are larger for more flexible outcomes (overtime hours) and more elastic subgroups of the population. Another important result uncovered by NAME is that a significant number of self-employed workers who regularly worked for the same firms switched their status to permanent salaried workers for their previous clients. This result highlights the importance of tax incentives to determine the nature of work and is suggestive of co-operation between workers and employers in determining the type of contract they choose.

To my knowledge, this is one of the best identified papers on the short-term effects of taxes on labor supply. The additional results indicating that employer and employee cooperation is crucial to

deliver tax avoidance are also in line with a set of very recent studies that have focused on payroll taxes. As a result, I am highly confident that this paper will end up published in a top general interest/top field journal. Importantly, I believe that it will become a reference for future studies on the real effects of income taxation on high wage income earners. Because of the simplicity and clarity of the context, this paper could also become a widely used teaching reference, both at the advanced undergraduate and PhD levels.

NAME has 2 additional completed drafts, again focusing on country YYY.

In the paper “Evidence from a change in the payment system of means tested transfers” NAME and a fellow PhD student show that the way in which child benefits are disbursed affects employer’s behavior and female labor supply. To tackle this question, they take advantage of a change in the payment system that was gradually rolled out between 1995 and 2001 in YYY. Under the old system, employers were intermediaries that paid child benefits to their employees together with their salary and were allowed to deduct this transfer from employer social security contributions. The new system centralized instead the payments in a government agency that started disbursing the allowances directly to eligible workers. Using detailed employer-employee data, NAME and her co-author show that employers reduce wages after the reform, disproportionately affecting female labor supply. The effect is mostly driven by big firms and more generally by firms that enjoy a significant degree of monopsonistic power in the labor market. This paper contributes to the public economics literature showing that employee vs. employer administration, even when economically neutral, can end up having a large impact in practice.

In NAME’s last paper “Taxpayers’ responses to tax changes”, NAME estimates the response of self-employed and firms to a new turnover tax, using rich administrative data. In particular, NAME exploits several revenue-dependent discontinuities that provide incentives to underreport taxable income combined with a bunching design to estimate revenue elasticities. Consistent with the literature, this paper uncovers tax avoidance as a response to such opportunities.

This constitutes the main body of the letter. It is followed by the last paragraph. The last paragraph is either grindstone, ability or neutral and shown in the instructions of the academic survey.

C.3 Online experiment

C.3.1 Workers

Welcome to this study!

Study information In this study, you will be asked to attempt two quizzes (A and B) and fill out a short survey. Each quiz has 10 questions from different subjects: math, business, and history. In other words, you will answer 20 questions in total. All questions for a quiz will be displayed on a single page and you can answer questions in the order you want to. You will have **2 minutes to work on each quiz**.

How much can I earn? You will be paid £1.80 for your participation. In addition, you can earn an additional bonus of up to £1.50 depending on your answers. For your bonus payment, one of the two quizzes will be randomly selected and you will get **£0.10 for every correct answer in that quiz**. It is in your best interest to treat each quiz as if it determines your bonus payment. You can win another £0.50 in the survey following the quizzes.

How long will it take? The study will take around 7 minutes to complete.

Do I have to participate? No, you are under no obligation to take part. You are free to withdraw at any point before or while completing the study. All data collected will be kept confidential and used for research purposes only.

Consent

Please note that you are free to withdraw at any point while completing the questionnaire. All data collected will be kept confidential and used for research purposes only.

By clicking the **‘I consent’** button, **you agree to the following:**

1. I have read and understood the information about the study.
2. I understand that my responses will remain anonymous.
3. I confirm that I have been given enough information about this study, and I voluntarily agree to take part.

If you do not agree, please click the **‘I don’t consent’** button. Thank you for your time.

- I consent
- I don’t consent

In online studies, there are sometimes bots or participants who do not read the instructions. This means that there are random answers which compromise the results of research studies. To show that you read our questions carefully, please enter ‘Turquoise’ as your answer to the next question. Based on the text above, what is your favourite colour?

- Blue
- Green

- Turquoise
- Purple
- Red

ID Please enter your Prolific ID below:

Recall that you have **2 minutes** to work on each Quiz. Once you click continue the timer for Quiz A will start.

Quiz A

Note: The order of Quiz A and Quiz B was randomized between participants.

Q1) How many original colonies were there in the US?

- Eleven
- Twelve
- Thirteen
- Fourteen

Q2) What kingdom is the oldest monarchy in Europe?

- United Kingdom
- Denmark
- Germany
- Sweden

Q3) What city was the US Declaration of Independence signed?

- Washington DC
- New York City
- Philadelphia
- Boston

Q4) Giovanni wants to buy shirts that cost £19.40 each and sweaters that cost £24.80 each. An 8% sales tax will be applied to the entire purchase. If Giovanni buys 2 shirts, which equation relates the number of sweaters purchased, p , and the total cost in dollars, y ?

- $1.08(38.80 + 24.80p) = y$
- $38.80 + 24.80p = 0.92y$
- $38.80 + 24.80p = 1.08y$
- $0.92(38.80 + 24.80p) = y$

Q5) The equation $y=36+18x$ models the relationship between the height y , in inches, of a typical golden delicious apple tree and the number of years, x , after it was planted. If the equation is graphed in the xy -plane, what is indicated by the y -intercept of the graph?

- The age, in years, of a typical apple tree when it is planted
- The height, in inches, of a typical apple tree when it is planted
- The number of years it takes a typical apple tree to grow
- The number of inches a typical apple tree grows each year

Q6) $-3x-4y = 20$

$x-10y = 16$

If (x,y) is the solution to the system of equations above, what is the value of x ?

- -14
- -12
- -4
- 16

Q7) Which of the following is an example of £-cost averaging?

- Index funds
- 401(k) plans
- Certificates of deposit

Q8) Suppose you have £100 in a savings account earning 2 percent interest a year. After five years,

how much would you have?

- More than £102
- Exactly £102
- Less than £102

Q9) If interest rates rise, what will typically happen to bond prices? Rise, fall, stay the same, or is there no relationship?

- Rise
- Fall
- Stay the same
- No relationship

Q10) Buying a single company's stock usually provides a safer return than a stock mutual fund.

- True
- False

Thank you, you have finished Quiz A!

How much effort did you put into doing Quiz A on a scale from 1 to 5?

Note: Answers for each letter are presented on a Likert scale from 1 (little effort) to 5 (a lot of effort).

Once you click continue, the **2 minute timer** for Quiz B will start.

Quiz B

Q1) Who is the only British prime minister to have been assassinated?

- William Pitt the Elder
- Spencer Perceval
- George Canning

Q2) Which infamous incident of treachery in Scotland is said to have inspired the extremely bloody “Red Wedding” massacre scene in the TV series Game of Thrones?

- The Black Dinner of 1440
- The Glencoe Massacre of 1692
- The murder of Lord Darnley in 1567

Q3) In the immediate aftermath of the Boston Massacre, many of the citizens of the city of Boston:

- decided to enlist in the British army in order to obtain weapons to defend themselves.
- organized the Boston Tea Party.
- circulated propaganda in the form of pamphlets and prints.
- looked to George Washington to organize and lead a new colonial military force.

Q4) A line is graphed in the xy-plane. If the line has a positive slope and a negative y-intercept, which of the following points cannot lie on the line?

- (-3, -3)
- (-3, 3)
- (3, -3)
- (3, 3)

Q5) A parachute design uses 18 separate pieces of rope. Each piece of rope must be at least 270 centimeters and no more than 280 centimeters long. What inequality represents all possible values of the total length of rope x , in centimeters, needed for the parachute?

- $270 \leq x \leq 280$
- $4,860 \leq x \leq 4,870$
- $4,860 \leq x \leq 5,040$
- $5,030 \leq x \leq 5,040$

Q6) $(x^2y^2)^{(1/2)}(x^2y^3)^{(1/3)} = x^{(a/3)}y^{(a/2)}$

If the equation above, where a is a constant, is true for all positive values of x and y , what is the value of a ?

- 2
- 3
- 5
- 6

Q7) If the equation $y=(x-6)(x+12)$ is graphed in the xy -plane, what is the x -coordinate of the parabola's vertex?

- -6
- -3
- 3
- 6

Q8) Which of the following orders of the four investment instruments according to their average volatility (from 1 = low to 4 = high) is correct?

- 1. Savings account 2. Government bonds 3. Stocks 4. Corporate bonds
- 1. Savings account 2. Government bonds 3. Corporate bonds 4. Stocks
- 1. Government bonds 2. Savings account 3. Stocks 4. Corporate bonds
- 1. Government bonds 2. Savings account 3. Corporate bonds 4. Stocks

Q9) Which one of the following statements is NOT a possible advantage of investing in investment funds from the perspective of an investor?

- The possibility to invest diversified
- The possibility to invest in special markets
- The possibility to invest small amounts of money
- The possibility to participate in the choice of individual stocks

Q10) Which one of the following terms is synonymous with keeping a sell option?

- Short Put
- Long Put

- Short Call
- Long Call

Thank you, you have finished Quiz B!

How much effort did you put into doing Quiz B on a scale from 1 to 5?

Note: Answers for each letter are presented on a Likert scale from 1 (little effort) to 5 (a lot of effort).

Thank you for finishing the main part of the experiment!

Before you leave, we will ask you 5 last questions to test your reasoning ability. For each question you answer correctly, you will earn an additional £0.10.

1) If you're running a race and you pass the person in second place, what place are you in?

2) A farmer had 15 sheep and all but 8 died. How many are left?

3) Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?

4) How many cubic feet of sand are there in a hole that is 3' deep x 3' wide x 3' long?

5) An expedition on a mountain climbing trip was traveling with eleven horse packs. Each horse can carry only three packs. How many horses does the expedition need?

Finally, we would like to get some information about yourself.

How old are you?

What is your gender?

- Male
- Female
- Non-binary/ third gender
- Prefer not to say

What is the highest level of education you completed?

- No formal education
- High school or equivalent
- College/ undergraduate degree
- Master's degree/ MBA
- Doctoral Degree (PhD)

Thank you very much again for your participation! Any feedback/ additional thoughts regarding this study is highly appreciated. If you have any comments, you can leave them below:

Please click finish to submit your answers.

C.3.2 Recommenders

Welcome to this study!

Study information In this study we will ask you to evaluate and predict the performance of workers on an online crowdsourcing platform. In addition, you will get the opportunity to recommend a worker to a recruiter.

How much can I earn? You will be paid £1.50 for your participation. In addition, you can earn an additional bonus of up to £1.50 depending on your choices and the choices of other participants.

How long will it take? The study will take around 10 minutes to complete.

Do I have to participate? No, you are under no obligation to take part. You are free to withdraw at any point before or while completing the study. All data collected will be kept confidential and

used for research purposes only.

Consent

Please note that you are free to withdraw at any point while completing the questionnaire. All data collected will be kept confidential and used for research purposes only.

By clicking the '**I consent**' button, **you agree to the following**:

1. I have read and understood the information about the study.
2. I understand that my responses will remain anonymous.
3. I confirm that I have been given enough information about this study, and I voluntarily agree to take part.

If you do not agree, please click the '**I don't consent**' button. Thank you for your time.

- I consent
- I don't consent

In online studies, there are sometimes bots or participants who do not read the instructions. This means that there are random answers which compromise the results of research studies. To show that you read our questions carefully, please enter 'Turquoise' as your answer to the next question. Based on the text above, what is your favourite colour?

- Blue
- Green
- Turquoise
- Purple
- Red

ID Please enter your Prolific ID below:

Instructions

Today you will decide how to recommend a worker to a recruiter.

The workers

In a previous study, we asked people to complete 2 quizzes of similar difficulty: Quiz A and B. Each quiz consisted of 10 questions about **math, finance, and history**. Workers were paid for finishing the quiz and received an additional bonus for getting randomly selected questions correct. You can see a few sample questions here:

Q7) If the equation $y = (x-6)(x+12)$ is graphed in the xy -plane, what is the x -coordinate of the parabola's vertex?

☐ -6

☐ -3

☐ 3

☐ 6

Q8) Who is the only British prime minister to have been assassinated?

☐ William Pitt the Elder

☐ Spencer Perceval

☐ George Canning

Q9) If interest rates rise, what will typically happen to bond prices? Rise, fall, stay the same, or is there no relationship?

☐ Rise

☐ Fall

☐ Stay the same

☐ No relationship

Your task

We will show you **Quiz A** performance and demographic information of one worker. You will then be asked to recommend the worker to a recruiter (see below), using one of three pre-written messages.

We will also ask you a few questions about how you perceive the performance of the worker. **If**

your guess about the worker is correct, you will earn an additional bonus of £0.50.

The recruiters

Recruiters receive your message in addition to demographic information about the worker.

They are informed about the content of the quiz but do not know anything about the worker's performance. The recruiter will then decide whether to hire the worker.

If the worker is hired, the **recruiter receives a bonus (between £0 and £1) depending on how many questions the worker answered correctly in Quiz B.** If they do not hire the worker, they receive a fixed payment. If the worker performed above average, the recruiter earns more by hiring the worker. If the worker performed below average it is better for the recruiter not to hire the worker.

If a recruiter hires the worker you recommended, you receive an additional bonus of £0.50.

We will also ask you a few questions about the recruiter's behaviour. Again, **if you guess correctly, you will earn an additional £0.50.**

On the next screen, you will be asked some comprehension questions to make sure that you understand the instructions.

We are now asking you a few questions about the set-up. If you need to consult the instructions again, you can do so by clicking on the button below.

Note: Participants could only continue once they selected the correct answer.

1) Questions in Quiz A are:

- Harder than in Quiz B
- Easier than in Quiz B
- Comparable to Quiz B

2) To make a hiring decision the recruiter sees:

- Quiz A performance
- Quiz B performance
- Your message

3) Which of the following statements is true:

- Recruiters earn more the better a hired worker performs
- Recruiters earn more the more workers they hire
- Recruiters do not care about the worker's performance

4) You are earning an additional bonus if:

- The worker you recommend is hired
- The worker performs well in Quiz A
- The worker performs well in Quiz B

The worker

Below you see information about a worker:

Note: Participants randomly saw 5, 6 or 7 correctly solved questions instead of x and either male or female instead of y .

Correctly solved questions in Quiz A	$x/10$
Worker's gender	y

Note: On average workers answered 3.5/10 questions correctly in Quiz A.

We first ask you how you perceive the performance of this worker. We will randomly select one out of the questions below to determine your bonus.

If your guess about the worker is correct, you will receive an additional £0.50. It is in your best interest to treat each question as if it determines your bonus payment.

1) How many questions do you think did the worker shown above answer correctly in **Quiz B** (out of 10)?

Note: Possible answers range from 0 to 10.

2) In addition to the quizzes, we elicited the effort workers put into the task on a scale from 1 to 5. **How much effort** do you think did the worker shown above put into solving Quiz A?

Note: Possible answers range from 1 (little effort) to 5 (a lot of effort).

3) Finally, we tested the workers' cognitive ability. Workers could score between 1 and 5 points. How do you think did the worker shown above perform in the **cognitive ability** test?

Note: Possible answers range from 1 (lowest ability) to 5 (highest ability).

About Quiz A

1) How much **effort** do you think is required to perform well (answering at least 5 questions correct) in Quiz A?

Note: Possible answers range from 1 (little effort) to 5 (a lot of effort).

2) How much **cognitive ability** do you think is required to perform well (answering at least 5 questions correct) in Quiz A?

Note: Possible answers range from 1 (little cognitive ability) to 5 (a lot of cognitive ability).

3) Recall that your worker was **y** and solved **x** questions correctly in Quiz A. How much do you think was this performance due to effort and how much due to cognitive ability (in %)?

Note that the two values you select should add up to 100%.

_____ Performance due to **effort**

_____ Performance due to **cognitive ability**

Recommendation decision

Here is a reminder of the worker.

Correctly solved questions in Quiz A	x/10
Worker's gender	y

Note: On average workers answered 3.5/10 questions correctly in Quiz A.

Which message do you want to send to the recruiter to recommend your worker?

Recall that recruiters will only see your recommendation message and the worker's gender to make a hiring decision.

Note: The order of the different messages is randomized.

- Having observed the worker's performance in Quiz A, I believe that HE/SHE will **put in the effort** to perform well in Quiz B. I think HE/SHE is **very thorough**.

- Having observed the worker's performance in Quiz A, I believe that HE/SHE has the **talent** to perform well in Quiz B. I think HE/SHE is **very bright**.
- Having observed the worker's performance in Quiz A, I believe that HE/SHE **can perform** well in Quiz B. I think HE/SHE is a **good worker**.

About the recruiters

We now ask you some questions about the recruiters. To determine your bonus, we will randomly select one of the questions below and from the next page.

If your guess about the recruiter is correct, you will receive an additional £0.50. It is in your best interest to treat each question as if it determines your bonus payment.

Note: The order of messages is randomized.

We will ask many recruiters to make a hiring decision based on the three messages that you have seen before.

1) Out of 100 recruiters how many will on average hire a **y worker** after reading the following message?

*"Having observed the worker's performance in Quiz A, I believe that HE/SHE will **put in the effort** to perform well in Quiz B. I think HE/SHE is **very thorough**."*

Note: Answer possibilities go from 0 to 100 in steps of 10.

2) Out of 100 recruiters how many will on average hire a **y worker** after reading the following message?

*"Having observed the worker's performance in Quiz A, I believe that HE/SHE has the **talent** to perform well in Quiz B. I think HE/SHE is **very bright**. "*

Note: Answer possibilities go from 0 to 100 in steps of 10.

3) Out of 100 recruiters how many will on average hire a **y worker** after reading the following message?

*"Having observed the worker's performance in Quiz A, I believe that HE/SHE **can perform** well in Quiz B. I think HE/SHE is a **good worker**."*

Note: Answer possibilities go from 0 to 100 in steps of 10.

4) We will also show recruiters three y candidates with different messages, and ask them to hire one of them. **Which candidate do you think will get hired?**

The candidate with the message...

Note: The order of messages is randomized.

- Having observed the worker's performance in Quiz A, I believe that HE/SHE will **put in the effort** to perform well in Quiz B. I think HE/SHE is **very thorough**.
- Having observed the worker's performance in Quiz A, I believe that HE/SHE has the **talent** to perform well in Quiz B. I think HE/SHE is **very bright**.
- Having observed the worker's performance in Quiz A, I believe that HE/SHE **can perform** well in Quiz B. I think HE/SHE is a **good worker**.

5) Finally, we will ask recruiters whether y workers are characterized more by talent or effort. **What do you think most of them answered?**

- y workers are characterized more by **talent**
- y workers are characterized more by **effort**
- y workers are equally characterized by **effort and talent**

C.3.3 Recruiters

Welcome to this study!

Study information In this study we will ask you to predict the performance of workers on an online crowdsourcing platform. In addition, you will get the opportunity to hire one worker and receive an additional bonus based on their performance.

How much can I earn? You will be paid £1.00 for your participation. In addition, you can earn an additional bonus depending on your responses and the responses of other participants.

How long will it take? The study will take around 10 minutes to complete.

Do I have to participate? No, you are under no obligation to take part. You are free to withdraw at any point before or while completing the study. All data collected will be kept confidential and

used for research purposes only.

Consent

Please note that you are free to withdraw at any point while completing the questionnaire. All data collected will be kept confidential and used for research purposes only.

By clicking the **'I consent'** button, **you agree to the following:**

1. I have read and understood the information about the study.
2. I understand that my responses will remain anonymous.
3. I confirm that I have been given enough information about this study, and I voluntarily agree to take part.

If you do not agree, please click the **'I don't consent'** button. Thank you for your time.

- I consent
- I don't consent

In online studies, there are sometimes bots or participants who do not read the instructions. This means that there are random answers which compromise the results of research studies. To show that you read our questions carefully, please enter 'Turquoise' as your answer to the next question. Based on the text above, what is your favourite colour?

- Blue
- Green
- Turquoise
- Purple
- Red

ID Please enter your Prolific ID below:

Instructions

Today you will decide whether to hire workers.

The workers

In a previous study, we asked people to complete 2 quizzes of similar difficulty: Quiz A and B. Each quiz consisted of 10 questions about **math, finance, and history**. Workers were paid for finishing the quiz and received an additional bonus for getting randomly selected questions correct. You can see a few sample questions here:

Q7) If the equation $y = (x-6)(x+12)$ is graphed in the xy -plane, what is the x -coordinate of the parabola's vertex?

☐ -6

☐ -3

☐ 3

☐ 6

Q8) Who is the only British prime minister to have been assassinated?

☐ William Pitt the Elder

☐ Spencer Perceval

☐ George Canning

Q9) If interest rates rise, what will typically happen to bond prices? Rise, fall, stay the same, or is there no relationship?

☐ Rise

☐ Fall

☐ Stay the same

☐ No relationship

Recommenders

We hired other Prolific workers to act as recommenders and have a closer look at the workers. Each recommender saw demographic information and **performance in Quiz A** for one worker. They were then asked to recommend the worker using one of three pre-defined messages. If you decide to hire the worker that was recommended by the recommender, the recommender will receive a bonus of £0.50.

Your task

We will show you **demographic information and the chosen recommendation message for one worker**. You will then have to choose whether to hire the worker or not.

If you hire the worker, you will receive a bonus of **£0.10 for each question the worker answered correctly in Quiz B**. You thus earn more the more questions were answered correctly by a hired worker. If you decide not to hire the worker, you will receive a fixed bonus of £0.40.

We will then ask you to make a second hiring decision between three workers. You will again receive **£0.10 for each question the hired worker answered correctly in Quiz B**. Finally, we will ask you a few questions about how you perceive the performance of the workers. **If your guess about the worker is correct you will receive an additional bonus of £0.50.**

On the next screen, you will be asked some comprehension questions to make sure that you understand the instructions.

We are now asking you a few questions about the set-up. If you need to consult the instructions again, you can do so by clicking on the button below.

Note: Participants could only continue once they selected the correct answer.

1) Questions in Quiz A are:

- Harder than in Quiz B
- Easier than in Quiz B
- Comparable to Quiz B

2) The recommender knows:

- The worker's performance in Quiz A
- The worker's performance in Quiz B
- The worker's performance in both quizzes

3) If you decide to hire a worker and that worker answered 5 out of 10 questions in Quiz B correctly, what is your additional bonus?

- £0.50
- £0.80
- £1.00

4) What is your additional bonus if you decide not to hire a worker?

- £1.00
- £0.00
- £0.80

Hiring decision 1

Below you see information about a worker:

Note: Participants randomly saw one of the three messages instead of z and either male or female instead of y. The three messages that were randomly assigned to z are the same messages as used in the recommender survey.

recommender's message	z
Worker's gender	y

Note: On average workers answered 3.5/10 questions correctly in Quiz A.

Do you want to hire this worker?

Recall that if you hire the worker, their performance in Quiz B will determine your bonus payment (you receive £0.10 per correct answer). If you decide not to hire the worker, you receive a fixed payment of £0.40.

- Yes, I want to hire the worker.
- No, I don't want to hire the worker.

About the worker

Here is a reminder of the worker:

recommender's message	z
Worker's gender	y

We now ask you how you perceive the performance of this worker. We will randomly select one of the questions below to determine your bonus.

1) How many questions do you think did the worker shown above answer correctly in **Quiz B** (out of 10)?

Note: Possible answers range from 0 to 10.

2) In addition to the quizzes, we elicited the effort workers put into the task on a scale from 1 to 5. **How much effort** do you think did the worker shown above put into solving Quiz A?

Note: Possible answers range from 1 (little effort) to 5 (a lot of effort).

3) Finally, we tested the workers' cognitive ability. Workers could score between 1 and 5 points. How do you think did the worker shown above perform in the **cognitive ability** test?

Note: Possible answers range from 1 (lowest ability) to 5 (highest ability).

4) Among all workers, do you think that y workers are characterized more by high ability or high effort?

- y workers are characterized more by **talent**
- y workers are characterized more by **effort**
- y workers are equally characterized by **effort and talent**

About the quiz

1) How much **effort** do you think is required to perform well (answering at least 5 questions correct) in the quiz?

Note: Possible answers range from 1 (little effort) to 5 (a lot of effort).

2) How much **cognitive ability** do you think is required to perform well (answering at least 5 questions correct) in the quiz?

Note: Possible answers range from 1 (little cognitive ability) to 5 (a lot of cognitive ability).

Hiring decision 2

Below you see information about three workers:

Note: Participants randomly saw a different messages (z1, z2, z3) for each worker. The gender of the

worker was fixed and either male or female instead of y . The three messages that were randomly assigned to $z1$, $z2$, and $z3$ are the same messages as used in the recommender survey.

	Worker 1	Worker 2	Worker 3
recommender's message	$z1$	$z2$	$z3$
Worker's gender	y	y	y

Which of the three workers do you want to hire?

The performance of the hired worker in Quiz B will determine your bonus payment (you receive £0.10 per correct answer).

- Worker 1
- Worker 2
- Worker 3

C.3.4 Questionnaire for recommenders and recruiters

Thank you for finishing the main part of the experiment. Before you leave, we would like to ask a few questions about yourself and your way of thinking.

1) In general, how willing are you to take risks?

Please use a scale from 0 to 10, where 0 means “completely unwilling to take risks” and a 10 means you are “very willing to take risks”.

Note: Answers are presented on a Likert scale from 0 (completely unwilling to take risks) to 10 (very willing to take risks).

2) We asked all participants in this study to guess a number between 0 and 100. We will then calculate the mean of all choices. The person whose guess is closest to $2/3$ of the mean wins an **additional bonus of £5**. Which number (between 0 and 100) do you want to guess?

3) What do you think are the most important characteristics in a potential employee?

Please rank them by **dragging the options** into your preferred order (from most important at the top to least important at the bottom).

Most important

- Hard working
- Intelligence

- Good social skills
- Prior job experience

Least important

4) How old are you?

5) What is your gender?

- Male
- Female
- Non-binary/ third gender
- Prefer not to say

6) What is the highest level of education you completed?

- No formal education
- High school or equivalent
- College/ undergraduate degree
- Master's degree/ MBA
- Doctoral Degree (PhD)

Thank you very much again for your participation! Any feedback/ additional thoughts regarding this study is highly appreciated. If you have any comments, you can leave them below:

Please click finish to submit your answers.