THE ECONOMIC JOURNAL

*The Economic Journal*, 1–33 https://doi.org/10.1093/ej/uead045 

Advance Access Publication Date: 20 June 2023

# GENDER DIFFERENCES IN REFERENCE LETTERS: EVIDENCE FROM THE ECONOMICS JOB MARKET*

*Markus Eberhardt, Giovanni Facchini and Valeria Rueda*

Academia, and economics in particular, faces increased scrutiny because of gender imbalance. This paper studies the job market for entry-level faculty positions. We employ machine learning methods to analyse gendered patterns in the text of 12,000 reference letters written in support of over 3,700 candidates. Using both supervised and unsupervised techniques, we document widespread differences in the attributes emphasised. Women are systematically more likely to be described using 'grindstone' terms and at times less likely to be praised for their ability. Using information on initial placement, we highlight the implications of these gendered descriptors for the quality of academic placement.

Gender disparities in the workplace have received significant attention in public debate. Academia is facing increased scrutiny due to its low female representation, especially in the field of economics (Valian, 1999; Lundberg, 2020, Part I). Recent empirical work has documented that the economics career pipeline for women is 'leaky', meaning that women tend to drop out of the profession during critical transitions, such as the jump from earning a PhD to an assistant professorship, or from assistant to associate professor (for a broad review, see Lundberg and Stearns, 2019). This paper studies the first step of the academic career of an economist, the junior 'job market'—the stage at which the leak has grown the most in the past decade (Lundberg and Stearns, 2019)—and which so far has not received much systematic attention (Lundberg, 2020).

The academic job market in economics is unique in that it is a highly structured institution. It starts every year in late fall, with universities posting their job advertisements and potential applicants preparing a 'job market package'. The latter consists of one or more academic papers, a CV and a set of recommendation letters written by scholars familiar with the candidate. All the parties involved, i.e., the candidates, the letter writers and the hiring committees, interact via centralised platforms. Typically, the candidate uses the same package for the vast majority of

* Corresponding author: Valeria Rueda, School of Economics, University of Nottingham, Sir Clive Granger Building, NG7 2RD Nottingham, UK. Email: valeria.rueda@nottingham.ac.uk

This paper was received on 31 January 2022 and accepted on 13 June 2023. The Editor was Sule Alan.

The authors were granted an exemption to publish their data because access to the data is restricted. However, the authors provided the Journal with temporary access to the data, which allowed the Journal to run their codes. The codes are available on the Journal repository. The data and codes were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.7963179.

We gratefully acknowledge financial support from STEMM-CHANGE, the School of Economics and the Faculty of Social Sciences (all at Nottingham University). Malena Arcidiácono, Edoardo Cefalà, Cristina Griffa, Yuliet Verbel-Bustamante, Thea Zoellner and Diego Marino-Fages have provided excellent research assistance. The University of Nottingham School of Economics Research Ethics Committee gave clearance to use the data on March 17, 2020 and to conduct a survey of academics on January 18, 2021. Handling of confidential data was done with the approval of the university human-subject ethics board and all university regulations, as well as an agreement—signed in April 13, 2020—between the university and the job application platform through which the data were collected. The views expressed in this paper are those of the authors and do not necessarily represent the views of the University of Nottingham. We thank seminar participants at Bocconi University, Lund, St Andrews, SOFI, Surrey, Warwick and Trinity College Dublin, as well as the Monash-Zürich text-as-data conference for comments and suggestions. We are grateful for the input from two anonymous referees and the handling editor, which substantially improved the paper. The usual disclaimers apply.

jobs, making the marginal cost of an additional application low. Reference letters are not tailored to a particular institution and the same letter is usually used for *all* job applications (for more details, see Coles *et al.*, 2010).

In this paper, we investigate the presence of differences in the language used in reference letters, depending on the gender of the candidate being recommended. We use a unique dataset encompassing all applications for entry-level positions received by a research-intensive university in the UK over the 2017–21 period. Deploying natural language processing tools, we analyse the text of almost 12,000 reference letters written in support of 3,700 candidates. A standard letter covers a lengthy discussion of the candidate's job market paper, and some reference to their additional research and to their teaching and citizenship skills. Importantly, the final section of the letter provides a summary assessment of the candidate's academic abilities and recruitment prospects. Since we are primarily interested in the way candidates are described, we focus much of our attention on this final section.

This corpus is then transformed into a term-frequency-inverse-document-frequency (tf-idf) representation. Borrowing from methods developed in cognitive psychology and linguistics, we quantify whether letters written in support of female candidates emphasise systematically different attributes. We use three complementary approaches. First, we employ an unsupervised method to ascertain the terms in the letters that are the best predictors of a candidate's gender. We adopt a LASSO technique that selects the strongest predictors. Among these, we frequently observe terms related to research interests, but also to personality ('nice', 'pleasant') and 'grindstone' attributes ('determined', 'hardworking', etc.). Second, we rely on a supervised method, building dictionaries of words for common attributes emphasised in reference letters. These dictionaries are informed by existing research on the topic (Trix and Psenka, 2003; Schmader *et al.*, 2007). We validate our dictionaries through an original comprehensive survey of academic economists based in UK research-intensive universities. Corroborating the exploratory results from the LASSO, we observe that descriptions of female candidates tend to emphasise significantly more 'grindstone' attributes. In further specifications, we also uncover a tendency to use fewer terms related to ability and research. Third, we also qualify the strength of support received by a candidate by analysing the type of placement recommendation received (e.g., 'I recommend this person to any institution, including the very best'). We observe that women receive fewer positive signals, but find no difference in negative ones. Women are also more likely to be compared to other candidates.

This paper thus documents differences in the language chosen for female and male candidates. In line with previous research, we observe that women are described with more 'grindstone' attributes and at times fewer 'ability' ones (Bourdieu and Passeron, 1977; Trix and Psenka, 2003; Schmader *et al.*, 2007). A natural question following on from these findings is whether the differences uncovered matter. Diligence and working hard are positive attributes (see Alan *et al.*, 2019 on 'grit'). However, given the overwhelmingly positive tone of recommendation letters in the job market, it may be misleading to interpret our findings as suggesting that women receive 'better' recommendations. The opposite may well be true. In fact, as noted by Valian (1999, p. 170) '[a]lthough working hard is a virtue, labelling a woman a hard worker can be damning with faint praise. If someone is not considered able to begin with, working hard can be seen as confirmation of his or her inability.' More generally, sociologists have pointed out that minorities are more often praised for their diligence than for their innate ability and that the signal of diligence is often interpreted as a lack of innate talent (Bourdieu and Passeron, 1977, p. 201).

To illustrate the importance of language in reference letters, we study the correlation between the attributes emphasised in letters and job market placement. We manually collected information from personal websites, academic departments' placement records and LinkedIn profiles, to establish whether the candidate placed in academia or elsewhere. For academic placements, we also link the hiring institution to its RePEc (Research Papers in Economics) rankings. To the best of our knowledge, we are the first to collect such information.[1] The results indicate that language matters differentially for women and men. In particular, for placement in academia, male candidates tend to benefit from 'standout' terminology, while for females, the patterns vary. Within academia, letters emphasising 'grindstone' are associated with obtaining a job in a less prestigious institution, but only for women.

In studying gendered patterns in the language of reference letters, we address several empirical challenges. More specifically, the attributes emphasised in reference letters may be influenced by many factors, such as the institution the candidate graduated from or their research field. Some of these determinants may differ systematically for male and female candidates. We tackle this problem in a variety of ways. In our baseline specifications, we control for the observable candidate and writer characteristics obtained from the job application platform and from additional information we collected manually. On the writer's side, we control for their gender, the number of letters they provide in our sample and the RePEc ranking of their institution. On the candidate side, we control for ethnicity, years since PhD completion, broad field of specialisation, publication record and the ranking of their PhD-awarding institution. The baseline results are not sensitive to these controls, nor to alternative definitions of the reference letter ends. We also check that the results are not driven by alternative explanations that could correlate with the gender of the letter writer such as the location of the PhD-granting institution, the gender or the cultural background of the letter writer, the academic field of the candidate or the extent of networking conducted before the market, among others.

Still, we may worry that unobservable determinants could affect our findings. Therefore, we run more restrictive models that allow us to account for unobserved, time-invariant institutional and letter writer characteristics. A first set of models, which include fixed effects for the PhD-granting institution, confirm the gendered patterns observed even for candidates of the same cohort at the same institution. In further analysis, we restrict the sample to referees who have written letters for both male and female candidates and employ writer fixed effects. These more demanding specifications confirm that differences in describing male and female candidates are detectable even when we focus on individual writers. Further probing indicates that more experience in writing for female candidates attenuates some of these differences.

This article is related to the literature on gender representation in academia. Several papers have shown that women are under-represented in math-intensive fields (for a detailed review of the literature, see Ceci and Williams, 2009, pp. 3–16; Kahn and Ginther, 2017). Investigations of different aspects of academic life have uncovered significant barriers. For example, Nittrouer *et al.* (2018) and Hospido and Sanz (2021), among others, observed that female academics are less likely to be accepted to present their work at academic conferences. Many researchers have emphasised systematic gender biases in student evaluations of teachers, which are frequently used indicators of performance in promotion and tenure packages (MacNell *et al.*, 2015; Boring, 2017; Fan *et al.*, 2019; Mengel *et al.*, 2019; Boring and Philippe, 2021). These patterns are persistent, despite evidence of a demand for diversity (Funk *et al.*, 2019).

[1] In a related paper, Baltrunaite *et al.* (2022) studied placement in the same year for all candidates, i.e., between one and up to 10 years after their initial placement, and focused on the attainment of the associate professor rank.

While other math-intensive fields have shown some improvement, economics has been in the spotlight for its persistently low representation of women (Bayer and Rouse, 2016; Lundberg and Stearns, 2019). Not only is there low female representation at the earliest stages of the profession, but the career pipeline is also 'leaky'. In trying to understand barriers to women's advancement in economics, researchers have looked at different stages of an academic career. Focusing on the first one, Boustan and Langan (2019) documented the wide variation of gender representation across PhD programs, and found that this representation tends to be a persistent attribute of a department. Turning to the next steps as academic professionals, other limitations to the advancement of women have been observed. In particular, there is evidence that females face barriers to promotion (Ginther and Kahn, 2004; Sarsons, 2017; Bosquet *et al.*, 2019; Deschamps, 2022), higher standards to judge the quality of their research (Card *et al.*, 2020; Dupas *et al.*, 2021; Grossbard *et al.*, 2021; Hengel, 2022) and that their work gets cited less (Koffi, 2021). Taken together, all these factors are likely to hamper the progression of women in their academic careers. We contribute to this burgeoning literature by focusing on a major and to date unexplored stepping stone: the junior job market. At this stage, beyond institutional credentials, little information about the candidate's research or teaching is observed. Therefore, reference letters play a crucial role in supporting the applicant.

The professional culture in economics may also be problematic for women's advancement. Wu (2018) reported evidence of gender biases in posts about women in a well-known and widely used anonymous forum in the profession. Similarly, Dupas *et al.* (2021) studied the seminar culture and presented evidence that female speakers face more hostile audiences. By analysing recommendation letters, we are investigating a different aspect of the professional culture, namely mentorship. As opposed to these previous studies, our focus is on a setting in which economists fulfil a supportive and nurturing role.

Existing literature has uncovered gendered patterns in academic reference letters in other disciplines. For example, Trix and Psenka (2003) showed that letters written in support of female applicants to medical faculty positions are shorter and emphasise more 'grindstone' and 'teaching characteristics'. Looking at job applicants in chemistry and biochemistry, Schmader *et al.* (2007) observed similar patterns. Madera *et al.* (2009) documented that letters for female applicants in psychology emphasise their 'communal' attributes ('nice', 'collegial', etc.). This line of research has also uncovered systematic differences in the presence of doubt raisers in geosciences (Dutt *et al.*, 2016) and psychology (Hebl *et al.*, 2018; Madera *et al.*, 2019).

We contribute to this literature in three main ways. First, we validate the 'sentiment' classification previously used by carrying out an original survey of academic economists.

Second, by focusing on economics, we can leverage a substantially larger sample of letters that are broadly representative of a highly structured and globalised academic job market. This allows us to rely on unsupervised techniques to describe gendered patterns in the language used when writing references. More specifically, fitting a LASSO, we show that many of the words that best predict letters written for women relate to 'grindstone' or 'teaching and citizenship' traits, whereas many 'ability' terms are more predictive of letters written for men. In other words, gendered differences in language used are already salient when describing the data with an unsupervised approach. This suggests in turn that the patterns uncovered in this literature with supervised techniques are unlikely to be driven by biases in the selection of the relevant terms.

Third, we also further advance the literature by analysing gendered differences in the quality of placement recommendations (e.g., whether people are explicitly recommended to a top institution). More importantly, we also analyse the implications of gendered language on the

initial placement *outcomes* of candidates. Ongoing work on reference letters by Baltrunaite *et al.* (2022), which relies on word embedding representation of words rather than tf-idf, confirms these patterns for two Italian institutions and focusing on longer-term career outcomes.

The remainder of the paper is organised as follows. In Section 1 we discuss our sample as well as the general approach of our main textual analysis. Section 2 explains the process of data cleaning and preparation, followed by the exploratory analysis using unsupervised methods in Section 3. Section 4 outlines the supervised approach and presents the baseline results, with extensions and additional robustness checks. Section 5 discusses the analysis of job market placement, followed by concluding remarks.

## 1. Data

We collected and cleaned the text of almost 12,000 reference letters written in support of over 3,700 candidates who applied for entry-level positions between 2017 and 2021 at a research-intensive economics department in the UK.[2] In each year in our sample the department advertised multiple positions open to all fields.

The department is one of the largest in the UK, with over 55 regular faculty members and was ranked in the top five in the most recent public evaluation of scientific research carried out in UK universities (Research Excellence Framework, 2021). It has been consistently ranked in the top-75 worldwide according to the RePEc rankings. The majority of the faculty has an international background, with 53% having earned a PhD outside the UK (half of them in the United States, the other half in other European countries). The department has a large PhD program, with over 50 students in residence in a given year. Of staff, 23% is female.[3]

The applications were provided by the institution, as collected from the leading job application platform in economics. The data are limited to the entire set of applications submitted for positions advertised by that institution. Access to and handling of these confidential data are in accordance with the approval of the university human-subject ethics board and all university regulations, as well as an agreement signed between the university and the application platform.

For each letter, we know a number of characteristics of the candidate and the letter writer. For candidates, we know characteristics they entered on the application platform, such as gender, ethnicity and the institution granting their PhD.[4] We also manually collect data from the candidates' CVs: we add information on their publication record at the time of application and their graduation date. The institutional ranking of both letter writers and candidates are taken from RePEc.[5] Information on the main advisor is also collected. Finally, we manually searched online for each individual candidate to establish their first professional placement in the year following their first appearance in our sample. Combining information from personal websites, academic departments' placement records and LinkedIn profiles, we establish whether the candidate placed in academia or elsewhere. For academic placements, we also collect the name of the institution, which we link to RePEc rankings.

[2] All applications were filed exclusively through the leading job application platform in economics, without any additional paperwork required.

[3] A figure slightly below the average for UK research-intensive institutions in the so-called Russell Group. For more details, see De Fraja *et al.* (2019).

[4] If the gender was withheld in the application, it was determined using a manual internet search.

[5] See Online Appendix A for more details on how the ranking is constructed.

For each letter writer, we have information on the institution where they were based at the time the letter was written. Using the R library 'GenderizeR', we also infer their gender from their first names. For this procedure, we adopt a conservative approach and manually search for cases in which the gender probability reported by the algorithm is below 0.75.[6] We also manually collected information on their academic rank, their seniority (year of PhD completion) and their country of birth.[7]

Summary statistics of these characteristics are presented in Tables 1 and 2. The majority of applicants and reference letter writers are based in the top-100 ranked institutions, with slightly more letter writers concentrated at the very top, as also shown in Figure 1. We have 5,655 writers (female share 17.4%) in our sample, and on average each writer has written slightly more than two letters. Overall, approximately 30% of the candidates in our sample are women. This statistic is consistent with the figures reported by Lundberg and Stearns (2019) and has remained stable over time, as shown in Figure 2. Table 3 shows the share of applicants by country. Approximately 50% of the candidates are based at US institutions and 14% in the UK.

Reference letters for the economics job market have a mean length of 1,089 words, which corresponds to around three A4 pages, with an SD of 554 words (around 1.5 pages). A standard letter covers a lengthy discussion of the candidate's job market paper, and some reference to their additional research and to their teaching and citizenship attributes. Importantly, the final section of the letter provides a summary assessment of the candidate's academic abilities and recruitment prospects.

Since we are primarily interested in the way candidates are described, we focus our analysis on this end section. Section 2.1 explains how this section is extracted. A typical example of the information provided is given by the following quotation, in which identifiable and sensitive characteristics have been redacted to protect privacy.

> ... working in this area. In terms of recent students coming out of [Institution X] that I have worked with, [Candidate $\alpha$] would be on a par of with a number of excellent recent placements such as [Candidate $\beta$] who went to [Institution Y], [Candidate $\gamma$], who went to [Institution Z] and [Candidate $\delta$] who went to [Institution W]. These economists are carving out excellent, innovative careers and I can see [Candidate $\alpha$] joining their ranks. What makes [Candidate $\alpha$] stand out from recent cohorts is [Candidate $\alpha$]'s ability to work with governments. [Candidate $\alpha$] has been central to the work that [Institution X] does in [Country A]. Precisely, [Candidate $\alpha$] has done such a good job starting up projects with the government and delivering answers to big, difficult to tackle questions. You can see this hallmark in all [Candidate $\alpha$]'s papers and I have a sense [Candidate $\alpha$] is going to be highly productive in [his/her] career for this reason. I therefore recommend that all top economics departments, business schools and public policy schools interested in hiring someone in [Field $\phi$] take a careful look at this application.

## 2. Methods

### 2.1. *Data Processing*

In this section, we explain the methods employed to transform our collection of letters into data.

Following standard procedure, we pre-process the text. First, we clean all punctuation and clearly separate out the words. Next, we remove all common stop words such as articles or pronouns. Furthermore, we stem the words, i.e., we reduce the words to their common stem (or

[6] The names of only 284 individuals fall below this threshold (5.9% of all letter writers) and their gender has been determined using a manual search.

[7] If the country of birth was unknown, we have attributed it based on the location of the institution that granted their undergraduate degree.

Table 1. *Descriptive Statistics—Candidate Characteristics.*

| | Full sample | | | | | Males | | Females | | Difference (M − F) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | *N* | Mean | SD | Min | Max | *N* | Mean | *N* | Mean | Estimate | *p*-value | Sig. |
| **Characteristics of the candidates** | | | | | | | | | | | | |
| ***Gender*** | | | | | | | | | | | | |
| Female | 3,721 | 0.291 | 0.454 | 0 | 1 | 2,639 | 0 | 1,082 | 1 | −1 | | |
| ***Ethnicity*** | | | | | | | | | | | | |
| Asian | 3,721 | 0.316 | 0.465 | 0 | 1 | 2,639 | 0.286 | 1,082 | 0.389 | −0.103 | 0.000 | *** |
| Black | 3,721 | 0.018 | 0.132 | 0 | 1 | 2,639 | 0.022 | 1,082 | 0.008 | 0.013 | 0.005 | *** |
| American Indian | 3,721 | 0.005 | 0.067 | 0 | 1 | 2,639 | 0.006 | 1,082 | 0.002 | 0.004 | 0.115 | |
| Hispanic | 3,721 | 0.091 | 0.287 | 0 | 1 | 2,639 | 0.100 | 1,082 | 0.068 | 0.031 | 0.003 | *** |
| Hispanic Withheld | 3,721 | 0.133 | 0.340 | 0 | 1 | 2,639 | 0.143 | 1,082 | 0.109 | 0.034 | 0.006 | *** |
| White | 3,721 | 0.435 | 0.496 | 0 | 1 | 2,639 | 0.454 | 1,082 | 0.389 | 0.065 | 0.000 | *** |
| ***PhD location*** | | | | | | | | | | | | |
| US-based Institution | 3,721 | 0.504 | 0.500 | 0 | 1 | 2,639 | 0.501 | 1,082 | 0.510 | −0.009 | 0.610 | |
| ***Research field*** | | | | | | | | | | | | |
| Theory | 3,721 | 0.196 | 0.397 | 0 | 1 | 2,639 | 0.221 | 1,082 | 0.135 | 0.086 | 0.000 | *** |
| Macro | 3,721 | 0.249 | 0.433 | 0 | 1 | 2,639 | 0.244 | 1,082 | 0.262 | −0.018 | 0.238 | |
| Applied | 3,721 | 0.275 | 0.447 | 0 | 1 | 2,639 | 0.254 | 1,082 | 0.325 | −0.071 | 0.000 | *** |
| Residual | 3,721 | 0.259 | 0.438 | 0 | 1 | 2,639 | 0.257 | 1,082 | 0.264 | −0.007 | 0.640 | |
| ***PhD institution*** | | | | | | | | | | | | |
| RePEc rank top 25 | 3,721 | 0.183 | 0.387 | 0 | 1 | 2,639 | 0.189 | 1,082 | 0.166 | 0.023 | 0.098 | * |
| Rank 26−50 | 3,721 | 0.133 | 0.339 | 0 | 1 | 2,639 | 0.135 | 1,082 | 0.127 | 0.009 | 0.480 | |
| Rank 51−100 | 3,721 | 0.147 | 0.354 | 0 | 1 | 2,639 | 0.143 | 1,082 | 0.157 | −0.014 | 0.265 | |
| Rank 101−200 | 3,721 | 0.192 | 0.394 | 0 | 1 | 2,639 | 0.199 | 1,082 | 0.175 | 0.025 | 0.083 | * |
| Rank 201−500 | 3,721 | 0.222 | 0.416 | 0 | 1 | 2,639 | 0.208 | 1,082 | 0.256 | −0.048 | 0.002 | *** |
| Rank 500+ | 3,721 | 0.123 | 0.329 | 0 | 1 | 2,639 | 0.125 | 1,082 | 0.119 | 0.005 | 0.646 | |
| Years since PhD | 3,721 | 1.073 | 2.201 | 0 | 22 | 2,639 | 1.144 | 1,082 | 0.899 | 0.244 | 0.002 | *** |
| ***Publications (counts)*** | | | | | | | | | | | | |
| Total | 3,721 | 1.156 | 2.057 | 0 | 18 | 2,639 | 1.223 | 1,082 | 0.991 | 0.232 | 0.002 | *** |
| Top five | 3,721 | 0.013 | 0.121 | 0 | 3 | 2,639 | 0.016 | 1,082 | 0.007 | 0.008 | 0.062 | * |
| Top general interest | 3,721 | 0.020 | 0.148 | 0 | 2 | 2,639 | 0.020 | 1,082 | 0.018 | 0.002 | 0.765 | |
| Top field | 3,721 | 0.048 | 0.236 | 0 | 2 | 2,639 | 0.055 | 1,082 | 0.033 | 0.021 | 0.012 | ** |

*Notes:* Top-field journals are JIE, JET, JoE, JME, JPubE, JLE, JDE, JEH, JFE, JF, Rand. Top general-interest journals are JEEA, REStat, EJ, IER and all the AEJs. See the main text for additional information.

Table 2. *Descriptive Statistics—Letter and Letter-Writer Characteristics.*

| | Full sample | | | | | Males | | Females | | Difference (M − F) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | *N* | Mean | SD | Min | Max | *N* | Mean | *N* | Mean | Estimate | *p*-value | Sig. |
| *Panel A. Characteristics of the letters* | | | | | | | | | | | | |
| Total word count | 11,846 | 1,089 | 554 | 197 | 5,000 | 8,486 | 1,092 | 3,360 | 1,081 | 11.820 | 0.295 | |
| **Word bags** | | | | | | | | | | | | |
| Ability | 11,846 | 0.257 | 0.203 | 0 | 1.396 | 8,486 | 0.258 | 3,360 | 0.255 | 0.003 | 0.513 | |
| Grindstone | 11,846 | 0.075 | 0.101 | 0 | 0.787 | 8,486 | 0.073 | 3,360 | 0.080 | −0.007 | 0.001 | *** |
| Recruitment | 11,846 | 0.337 | 0.269 | 0 | 2.004 | 8,486 | 0.340 | 3,360 | 0.330 | 0.009 | 0.087 | * |
| Research | 11,846 | 0.914 | 0.484 | 0 | 2.996 | 8,486 | 0.921 | 3,360 | 0.896 | 0.025 | 0.011 | ** |
| Standout | 11,846 | 0.349 | 0.216 | 0 | 1.578 | 8,486 | 0.350 | 3,360 | 0.347 | 0.003 | 0.567 | |
| Teaching and citizenship | 11,846 | 0.368 | 0.333 | 0 | 1.949 | 8,486 | 0.365 | 3,360 | 0.376 | −0.012 | 0.086 | * |
| **Letter writers** | | | | | | | | | | | | |
| Female | 11,846 | 0.148 | 0.355 | 0 | 1 | 8,486 | 0.132 | 3,360 | 0.188 | −0.056 | 0.000 | *** |
| RePEc rank top 25 | 11,846 | 0.188 | 0.391 | 0 | 1 | 8,486 | 0.189 | 3,360 | 0.183 | 0.006 | 0.439 | |
| Rank 26−50 | 11,846 | 0.114 | 0.318 | 0 | 1 | 8,486 | 0.116 | 3,360 | 0.109 | 0.007 | 0.262 | |
| Rank 51−100 | 11,846 | 0.149 | 0.356 | 0 | 1 | 8,486 | 0.146 | 3,360 | 0.157 | −0.010 | 0.161 | |
| Rank 101−200 | 11,846 | 0.161 | 0.368 | 0 | 1 | 8,486 | 0.168 | 3,360 | 0.143 | 0.026 | 0.001 | *** |
| Rank 201−500 | 11,846 | 0.186 | 0.389 | 0 | 1 | 8,486 | 0.179 | 3,360 | 0.205 | −0.026 | 0.001 | *** |
| Rank 500+ | 11,846 | 0.202 | 0.401 | 0 | 1 | 8,486 | 0.201 | 3,360 | 0.204 | −0.003 | 0.746 | |

| | Full sample | | | | | Male writer | | Female writer | | Difference (M − F) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | *N* | Mean | SD | Min | Max | *N* | Mean | *N* | Mean | Estimate | *p*-value | Sig. |
| *Panel B. Characteristics of the letterwriter* | | | | | | | | | | | | |
| Female writer | 5,655 | 0.174 | 0.379 | 0 | 1 | 4,670 | 0 | 985 | 1 | −1 | | |
| Letters in the sample | 5,655 | 2.140 | 1.951 | 1 | 23 | 4,670 | 2.210 | 985 | 1.810 | 0.400 | 0.000 | *** |
| RePEc rank top 25 | 5,655 | 0.163 | 0.369 | 0 | 1 | 4,670 | 0.165 | 985 | 0.153 | 0.012 | 0.362 | |
| Rank 26−50 | 5,655 | 0.093 | 0.290 | 0 | 1 | 4,670 | 0.091 | 985 | 0.099 | −0.008 | 0.429 | |
| Rank 51−100 | 5,655 | 0.132 | 0.338 | 0 | 1 | 4,670 | 0.133 | 985 | 0.125 | 0.008 | 0.494 | |
| Rank 101−200 | 5,655 | 0.161 | 0.368 | 0 | 1 | 4,670 | 0.161 | 985 | 0.164 | −0.004 | 0.777 | |
| Rank 201−500 | 5,655 | 0.217 | 0.412 | 0 | 1 | 4,670 | 0.219 | 985 | 0.209 | 0.009 | 0.511 | |
| Rank 500+ | 5,655 | 0.234 | 0.423 | 0 | 1 | 4,670 | 0.231 | 985 | 0.249 | −0.018 | 0.234 | |
| Assistant Professor | 5,655 | 0.152 | 0.359 | 0 | 1 | 4,670 | 0.140 | 985 | 0.209 | −0.069 | 0.000 | *** |
| Associate Professor | 5,655 | 0.203 | 0.402 | 0 | 1 | 4,670 | 0.190 | 985 | 0.264 | −0.074 | 0.000 | *** |
| Full Professor/Chair | 5,655 | 0.622 | 0.485 | 0 | 1 | 4,670 | 0.649 | 985 | 0.498 | 0.15 | 0.000 | *** |
| PhD year | 5,098 | 2000 | 11.417 | 1953 | 2021 | 4,213 | 1999 | 885 | 2003 | −3.516 | 0.000 | *** |
| Prior to 2000 | 5,098 | 0.42 | 0.494 | 0 | 1 | 4,213 | 0.441 | 885 | 0.318 | 0.124 | 0.000 | *** |
| After (incl) 2000 | 5,098 | 0.58 | 0.494 | 0 | 1 | 4,213 | 0.559 | 885 | 0.682 | −0.124 | 0.000 | *** |

*Notes:* Institutional rankings are based on RePEc. See the main text for additional information.

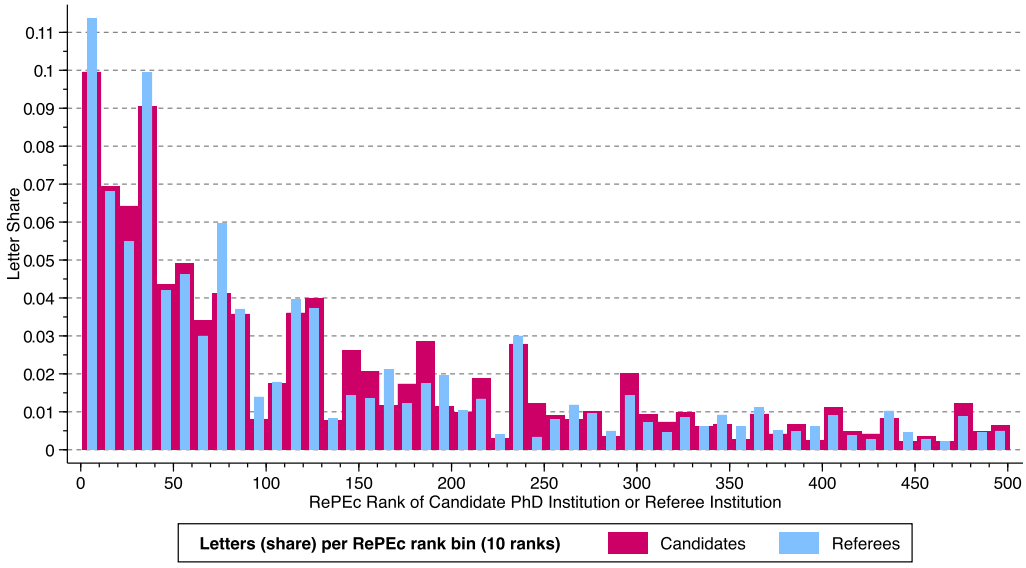


Fig. 1. *RePEc Ranks of Candidate and Letter-Writer Institutions.*
*Notes:* The figure presents the frequency distributions of candidate and letter-writer institutional ranks (in bins of 10 institutions).

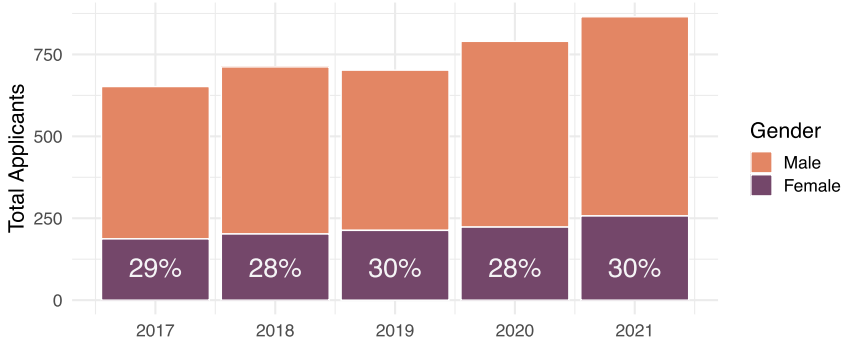


Fig. 2. *Gender Distribution of Applicants in the Sample.*
*Notes:* The figure shows the total number of applicants per year and the share of female applicants each year.

root). For instance, the words 'published', 'publishing' or 'publishes' will all be collapsed to the stem 'publish'. Following these steps, we have converted each reference letter into a collection of (stemmed) words.

We then need to establish a measure of the importance of each word per letter. We compute the tf-idf of each word using Python's Sklearn library.

We now define a few concepts to explain how we transform our collection of letters into data. Each letter is a *document*. Denote each document $d \in \{1, \ldots, D\}$. The corpus $D$ is the set of documents. Each document $d$ contains $N_d$ words $w_i(d)$, $i \in \{1, \ldots, N_d\}$. Words are drawn from a set of terms $t \in \{1, \ldots, T\}$. The set of terms is the entire vocabulary present in the corpus.

Table 3. *Descriptive Statistics: Candidate Country.*

| ISO3 | Candidates | Percent | Cum | ISO3 | Candidates | Percent | Cum |
|---|---|---|---|---|---|---|---|
| USA | 1,874 | 50.4 | 50 | CHN | 8 | 0.2 | 98 |
| GBR | 503 | 13.5 | 64 | IRL | 8 | 0.2 | 98 |
| CAN | 190 | 5.1 | 69 | HUN | 7 | 0.2 | 99 |
| FRA | 190 | 5.1 | 74 | GRC | 6 | 0.2 | 99 |
| DEU | 164 | 4.4 | 79 | IND | 6 | 0.2 | 99 |
| ESP | 157 | 4.2 | 83 | BRA | 5 | 0.1 | 99 |
| ITA | 132 | 3.5 | 86 | RUS | 5 | 0.1 | 99 |
| NLD | 106 | 2.8 | 89 | PSE | 4 | 0.1 | 99 |
| SWE | 62 | 1.7 | 91 | TUR | 4 | 0.1 | 99 |
| AUS | 49 | 1.3 | 92 | IRN | 3 | 0.1 | 100 |
| CHE | 49 | 1.3 | 93 | ISR | 3 | 0.1 | 100 |
| BEL | 41 | 1.1 | 95 | JPN | 3 | 0.1 | 100 |
| HKG | 26 | 0.7 | 95 | MEX | 3 | 0.1 | 100 |
| DNK | 18 | 0.5 | 96 | CHL | 2 | 0.1 | 100 |
| NOR | 16 | 0.4 | 96 | BGR | 1 | 0.0 | 100 |
| SGP | 15 | 0.4 | 97 | CYP | 1 | 0.0 | 100 |
| n/a | 14 | 0.4 | 97 | GEO | 1 | 0.0 | 100 |
| PRT | 12 | 0.3 | 97 | KOR | 1 | 0.0 | 100 |
| AUT | 11 | 0.3 | 98 | MYS | 1 | 0.0 | 100 |
| CZE | 10 | 0.3 | 98 | NZL | 1 | 0.0 | 100 |
| FIN | 9 | 0.2 | 98 | | | | |
| | | | | **Total** | **3,721** | | |

*Notes:* Three-digit ISO code for geographic location of the applicant (PhD institution, not nationality), in order of magnitude. Cum denotes the cumulative sum (rounded).

We represent the corpus of letters with a matrix of dimension $D \times T$. Each row of this matrix represents a document, and each column represents a term. For each document, each cell refers to the tf-idf of the term. The tf-idf is a common measure used to quantify the importance of a term in each document, compared to its prevalence in the corpus. The tf-idf is the product of the term frequency and the inverse-document frequency. The term frequency $\text{tf}(t, d)$ is the number of times term $t$ appears in document $d$:

$$\text{tf}(t, d) = \sum_{i}^{N_d} \mathbf{1}\{w_i = t\}.$$

The inverse-document frequency is the logarithmically scaled inverse fraction of the document frequency of $t$, $\text{df}(t)$, which is the number of documents that contain the term $t$:

$$\text{idf}(t) = \log \frac{1 + D}{1 + \text{df}(t)} \quad \text{with} \quad \text{df}(t) = \sum_{d} \mathbf{1}\{\text{tf}(t, d) > 0\}.$$

The tf-idf is then[8]

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) = \log \frac{1 + D}{1 + \text{df}(t)} \sum_{i}^{N_d} \mathbf{1}\{w_i = t\}.$$

This approach is considered standard for text vectorisation in natural language processing, and researchers have shown that this simple representation is sufficient to infer interesting properties from texts (Grimmer and Stewart, 2013). This approach has many advantages. First, it is easy

[8] By default, Python's Sklearn uses an L-2 normalisation, which means that it normalises the final tf-idf with the vector's Euclidean norm. This is aimed at correcting for long versus short documents. Following standard procedure, we also drop terms that are either too common (i.e., that appear in more than 70% of documents) or too rare (less than 1% of documents).

to implement. Second, the tf-idf for each word has the simple interpretation of capturing the importance of each word in the document, relative to its frequency in the corpus. We can also measure the importance of specific attributes in each letter by summing the tf-idf for the groups of words in the attribute category for each letter.

This approach has two main shortcomings. First, the vector space grows linearly with the vocabulary, which can cause significant computational challenges. In our case, our sample size is not large enough for this to become an issue. The second shortcoming is that the relationships *between* words are not taken into account. More recent deep-learning techniques use word embedding representations, resulting in a vector-space of low dimension. With word embeddings, terms represented with vectors that are close in space are semantically similar. Recent literature in law and economics has pioneered the implementation of word embeddings, for instance, to compare the similarity of different semantic fields inside a given corpus (Ash *et al.*, 2022; 2023, among others). Many of these papers are interested in exploring whether different semantic fields are correlated in different corpora (e.g., whether 'female' words tend to be associated with 'career' words or 'family' words). Unfortunately, word embeddings may perform disappointingly compared to traditional methods in smaller samples (Shao *et al.*, 2018; Ash *et al.*, 2023), and our sample is much smaller than those used in the new economics literature applying word embeddings.[9]

### 2.2. *Separating Ends*

In most of our analysis, we concentrate on the end of the letter. The rationale behind this choice is that reference letters in economics follow a fairly rigid structure, and the end of the letter is where the referees summarise their opinion about the candidate, including their job market prospects.

We use a two-step procedure to separate the letter ends. First, we create a dictionary of commonly used closing phrases (e.g., 'Yours sincerely'). These phrases flag the end of the letter, and permit cleaning out long signatures (with multiple affiliations, addresses, etc.). We then take the 200 words *before* the first closing phrase flagged, which roughly corresponds to the length of one large paragraph. With this approach, we cover more than 89% of the letters. For letters without any identifiable closing phrase, we use the last 200 words of the document. We also consider 150 and 250 word cuts for the letter ends in the robustness section.

### 2.3. *Language Categorisation*

Reference letters for the economics job market tend to have an overwhelmingly positive tone. Therefore, a standard computational text analysis that aims at weighting positive terms against negative ones is not appropriate in this context. We build instead on the categorisation proposed by Schmader *et al.* (2007) in their analysis of a smaller sample of applicants in chemistry ($n = 277$) for a large US research university, which in turn builds on earlier qualitative work by Trix and Psenka (2003).

Schmader *et al.* (2007) proposed five language categories that can be used to describe relevant features of an applicant, including *ability traits, grindstone traits, research terms, standout adjectives* and *teaching and citizenship terms*. We add a category that refers to the *recruitment*

[9] For instance, Ash *et al.*'s (2023) analysis of judge-specific corpora falls into the category of a 'small' sample for word embeddings. Their analysis relied on corpora with at least 1.5 million tokens (pre-processed words). For comparison, our main sample of interest, which consists of the universe of end of letters, contains approximately 852,000 tokens.

*prospects* of the candidate. Ability traits involve language aimed at highlighting the applicant's suitability for the advertised position and include words such as *talent*, *intellectual*, *creative*, etc. Grindstone traits refer to language that, in the words of Trix and Psenka (2003, p. 207), resemble 'putting one's shoulder to the grindstone'. Words in this category include *hardworking*, *conscientious*, *diligent*, etc. Research terms are descriptors of the type of research carried out by the candidate and related matters, e.g., *applied economics*, *game theory*, *public economics*, etc. Standout terms highlight especially desirable attributes of the applicant, like *excellent*, *top*, *strongest*, etc. Teaching and citizenship is a broad category that refers to both the candidate's skills in the classroom as well as their behaviour with colleagues. Language in this group includes *good teacher*, *excellent colleague*, *friendly*, etc. The last category, recruitment prospects, has been added to identify words that, in the highly competitive and globalised labour market for fresh economics PhDs, are widely used to describe the expected placement of the candidate. Words in this group include *highly recommended*, *top department*, *tenure track*, etc. Online Appendix Figure B.2 shows word clouds for each of our language categories.

To corroborate our word classification, we carried out a survey of all faculty employed at UK economics departments that were submitted to the 2014 Research Excellence Framework (REF).[10] Each participant was shown a sample of 20 words and asked to classify them in one of the six categories listed above. The survey was run between the end of March and the beginning of April 2021, and a total of 1,205 individuals were contacted. Participants were incentivised with a lottery of Amazon vouchers worth £20 each. A total of 195 took part in the survey, corresponding to 16% of the underlying population.

Figure 3 provides a breakdown of the population and of the survey respondents by level of seniority and gender. As can be seen, about one-third of the population (left panel) are associate professors, with a slightly higher share represented by full professors, and a slightly smaller one by assistant professors. The share of females declines with seniority, representing 32% of staff at the assistant professor level, and only 15% at the most senior level. Turning to our sample (right panel), respondents are slightly more likely to be full professors, and slightly less likely to be assistant professors than in the underlying population. Not surprisingly, females are over-represented among respondents, especially at the intermediate level of seniority.

Figure 4 illustrates the extent to which our own assessment of an expression is shared by the academics who took part in our survey. For all expressions classified into a language category by the authors, we show the distribution of classifications chosen by the plurality of validators.[11] While there is variation across language categories, there is broad consensus between our categorisation and that of the profession.

## 3. Unsupervised Analysis

### 3.1. *Methodology*

As an initial unsupervised analysis, we ask whether specific terms used are more predictive of the gender of the candidate. To this end, we employ a least absolute shrinkage and selection operator

[10] The REF is a periodic, comprehensive assessment of the research carried out by UK universities. For more information, see De Fraja *et al.* (2019).

[11] See Online Appendix A for more details on how the figure is constructed.

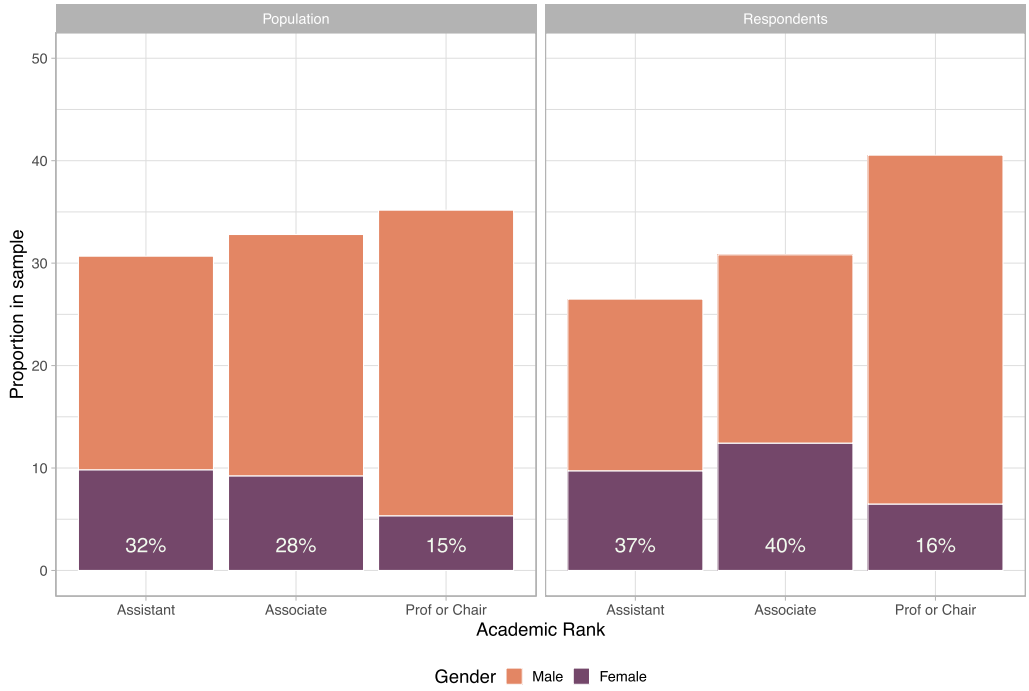


Fig. 3. *Population of Academics Surveyed Compared to Respondents.*
*Notes:* The figure compares the representation of women per academic rank between the total populations surveyed and the respondents of the validation exercise. The percentages at the top of each bar are the share of women inside the category. The category 'others' is accounted for in the calculations, but excluded from the graphs because of its low representation ($< 2\%$ of the sample in both the population and the validation sample).

(LASSO) to select the relevant set of terms. The LASSO estimator $\hat{\beta}$ solves the problem

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2D} \sum_{d=1}^{D} (y_d - \boldsymbol{x}_d' \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\}, \tag{1}$$

where $d$ is the letter. The gender of the candidate is the binary variable $y_d$. Vector $x_d$ is the collection of tfidf$(t, d)$ for the corpus. The second, penalty, term in (1) contains the 'tuning' parameters $\lambda$ and $\omega$ that are selected to reduce the number of non-zero, but small coefficients. Here $p$ is the total number of terms.

We implement different LASSO estimators that vary in their treatment of the penalty function: for a 75% training sample, we consider a cross-validation (CV) LASSO, an adaptive LASSO as well as an elastic net (enet) LASSO. These approaches differ in the way the optimal tuning parameters $(\lambda, \omega)$ are estimated or, in the case of the enet, by the specific form the penalty function takes. Since female candidates make up only 30% of our sample, we also experiment with 'oversampling' females in the training sample. The final set of selected terms is not sensitive to the choice of LASSO method nor to the oversampling choice.

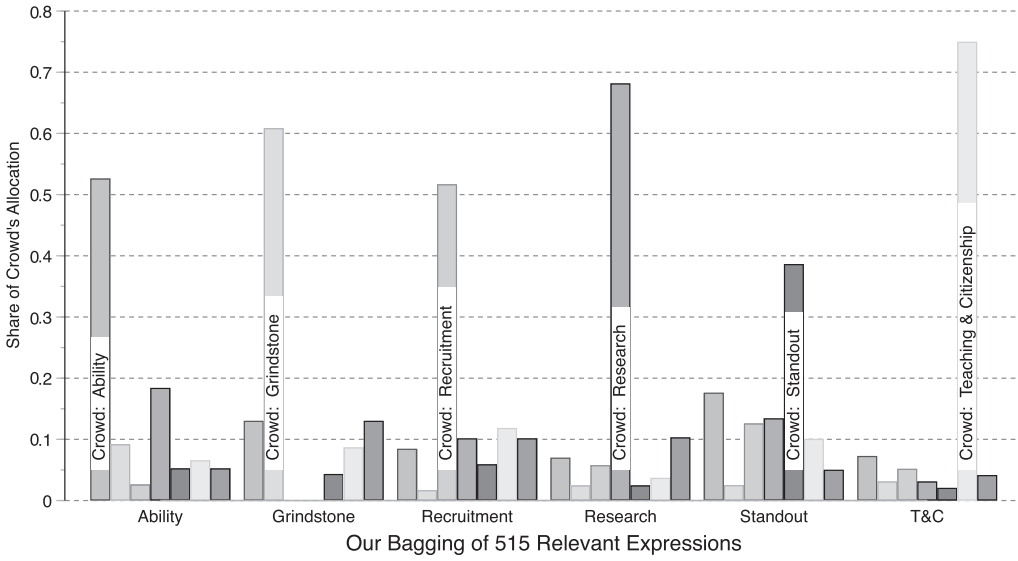


Fig. 4. *Correspondence between Authors' Sentiment Categories and the 'Wisdom of the Crowd'.* *Notes:* This figure shows the correspondence between the authors' chosen classification for each expression and the classification chosen by validators. For any word validated, it is attributed to the category that was chosen by the plurality of validators who were shown that word. See more details in Online Appendix A.

We only present results from the adaptive LASSO because across all specifications, it has higher predictive power than the enet and the CV.[12]

### 3.2. *LASSO Results*

A visualisation of the results is presented in Figure 5. The figure records the 289 predictors selected by the LASSO. We present the standardised beta coefficients of the linear probability model of candidate gender on tf-idf. Each line groups up to six predictors with similar coefficient magnitudes. The bars represent the range of the coefficient of the predictors listed in the line. Positive predictors are associated with female candidates, whereas negative ones are associated with males.

First, the figure reflects that women select across different research fields. Research on 'women', 'health' or 'environment(al)' tends to be disproportionately carried out by female candidates, whereas 'theory', 'history' or 'finance' appear to be associated with male candidates. This 'self-selection' mechanism is one that we also consider carefully in the remainder of the paper.

Second, qualitatively, it appears that certain personality traits are gender specific. While women are disproportionately more likely to be described as 'driven', 'determined' or 'hardworking', men are disproportionately seen as 'thinkers' or 'creative'. This is a pattern that will be confirmed

[12] We compare the areas under the receiving operator curve (AUROC). The ROC is a measure of predictive fit employed in the binary-dependent variable literature, quantifying the correctly predicted 0s and correctly predicted 1s.


Fig. 5. *LASSO Visualisation.*

*Notes:* This figure shows the terms selected in the LASSO exercise. In each line, the vertical bars illustrate the range of the standardised beta coefficient for all the words listed. The beta coefficient is the change in propensity that the candidate is female associated with a one SD increase in the tf-idf of the term. This LASSO exercise is conducted with stemmed words. In this figure, we have attributed to each stem its most frequent corresponding word. Out of 1,425, we select 289 stems by the adaptive LASSO. Here $N = 11,846$, AUROC $= 0.739$.

in the next section. Also aligning with gender stereotyping, descriptives related to 'communal attributes' ('nice', 'pleasant person') are associated with women.

Finally, it is worth noting that traits such as youth ('young researcher') and shyness are reserved to women. This finding conforms to the stereotyping of women as naïve or child-like that has been documented in sociology (see, for instance, Goffman, 1979, pp. 5, 50–1, and Gornick, 1979, pp. vii–ix), and for which there is suggestive evidence that it may harm women's credibility in the workplace (for a review, see MacArthur *et al.*, 2020).

This exploratory analysis shows that even using an unsupervised method such as the LASSO, a portrait of women as 'determined' and 'hardworking' is drawn. This observation is consistent with the previous findings highlighting that female candidates are mostly praised on their 'grindstone' attributes (Trix and Psenka, 2003; Valian, 2005).

## 4. Supervised Analysis with Dictionaries

In our supervised analysis, we employ the dictionaries related to 'ability', 'grindstone', 'research', 'recruitment', 'standout' and 'teaching and citizenship' discussed in Section 2.3—we refer to these as 'sentiments' for ease of discussion.

### 4.1. *Specification and Implementation*

We run regressions defined in the following equation using ordinary least squares:

$$\text{Sentiment}_{diwt} = \alpha + \beta\,\text{Female}_i + \mathbf{X}_i'\gamma + \mathbf{W}_w'\lambda + \nu_t + \varepsilon_{diwt}. \tag{2}$$

$\text{Sentiment}_{diwt}$ is the importance of each sentiment in letter $d$, written for candidate $i$ by letter writer $w$ in year $t$. For each sentiment ('ability', 'grindstone', etc.), $\text{Sentiment}_{diwt}$ is the sum of $\text{tfidf}(t, d)$ of all the terms in letter $d$ associated with that sentiment in our dictionaries. $\text{Female}_i$ is an indicator equal to 1 if the candidate is female and $\beta$ is our coefficient of interest. Vector $\mathbf{X}_i$ is a vector of candidate-level controls and $\mathbf{W}_w$ is a vector of letter-writer controls; both are described in more detail below. We further include recruitment cohort fixed effects $\nu_t$.

It is possible that attributes of candidates or letter writers that influence how a recommendation is written differ systematically between men and women. For instance, publication records may vary by gender, which in turn might affect the recommendation's strength (Hengel, 2022). Similarly, female candidates may not be represented in highly ranked institutions in the same way as males, etc. The variables included in the regression aim at accounting for these differences.

First, with regards to candidate attributes, all specifications include controls for their ethnicity, race, and the year they entered the job market. We sequentially add indicator variables accounting for the RePEc ranking band of the candidate's PhD-awarding institution.[13] Finally, we control for the years since PhD completion, for the broad field of specialisation[14] and for the publication record. For the latter, we include the total number of publications and the number of articles published in top-field, top-five and top general-interest journals.[15]

[13] In particular, we distinguish top 25, top 26–50, top 51–100, top 101–200, top 201–500 and an indicator for institutions not included in our top-5% RePEc ranking in January 2021 (12% of the sample).

[14] Section 4.3 describes in greater detail how we define fields and the robustness of our results to alternative definitions.

[15] We define the following journals as top field: JDE, JEH, JET, JF, JFE, JIE, JME, JoE, JPubE and RAND. Top general-interest journals are the four AEJs, EJ, IER, JEEA and REStat.

Table 4. *Sentiments—End of Letters.*

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Ability | −0.0147 | −0.0124 | −0.0142 | −0.0180 | −0.0182 | −0.0230 | −0.0231 |
| | (0.70) | (0.59) | (0.68) | (0.86) | (0.87) | (1.09) | (1.10) |
| Grindstone | 0.0637 | 0.0615 | 0.0636 | 0.0580 | 0.0575 | 0.0512 | 0.0512 |
| | (3.02)*** | (2.91)*** | (3.01)*** | (2.73)*** | (2.70)*** | (2.41)** | (2.41)** |
| Recruitment | −0.0236 | −0.0237 | −0.0227 | −0.0317 | −0.0286 | −0.0247 | −0.0249 |
| | (1.11) | (1.13) | (1.08) | (1.51) | (1.35) | (1.17) | (1.20) |
| Research | −0.0548 | −0.0520 | −0.0520 | −0.0613 | −0.0607 | −0.0561 | −0.0560 |
| | (2.66)*** | (2.52)** | (2.52)** | (2.96)*** | (2.92)*** | (2.70)*** | (2.71)*** |
| Standout | −0.0035 | −0.0013 | −0.0036 | −0.0146 | −0.0119 | −0.0135 | −0.0136 |
| | (0.17) | (0.06) | (0.17) | (0.69) | (0.56) | (0.64) | (0.66) |
| Teaching and citizenship | 0.0343 | 0.0250 | 0.0240 | 0.0177 | 0.0148 | 0.0071 | 0.0070 |
| | (1.60) | (1.18) | (1.13) | (0.83) | (0.69) | (0.33) | (0.33) |
| FEs/variables absorbed | 10 | 15 | 15 | 19 | 19 | 25 | 25 |
| Additional variables | 0 | 0 | 1 | 1 | 5 | 6 | 7 |
| Number of letters | 11,846 | 11,846 | 11,846 | 11,846 | 11,846 | 11,846 | 11,846 |
| dto for females | 3,360 | 3,360 | 3,360 | 3,360 | 3,360 | 3,360 | 3,360 |
| Number of candidates | 3,721 | 3,721 | 3,721 | 3,721 | 3,721 | 3,721 | 3,721 |
| dto female | 1,082 | 1,082 | 1,082 | 1,082 | 1,082 | 1,082 | 1,082 |
| Number of writers | 5,655 | 5,655 | 5,655 | 5,655 | 5,655 | 5,655 | 5,655 |
| dto female | 985 | 985 | 985 | 985 | 985 | 985 | 985 |
| Letters by female writers | 1,751 | 1,751 | 1,751 | 1,751 | 1,751 | 1,751 | 1,751 |
| Year FEs | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Ethnicity/race FEs | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Institution rank FEs | No | Yes | Yes | Yes | Yes | yes | Yes |
| Years since PhD | No | No | Yes | Yes | Yes | Yes | Yes |
| Research field FEs | No | No | No | Yes | Yes | Yes | Yes |
| Publications | No | No | No | No | Yes | Yes | Yes |
| Writer characteristics | No | No | No | No | No | Yes | Yes |
| Letter length | No | No | No | No | No | No | Yes |

*Notes:* The table shows results of the OLS regression of each 'sentiment' (e.g., 'ability', 'grindstone', etc.) on a female candidate indicator as well as controls mentioned in the text (letter ends with 200 words). The table reports the estimate for the female indicator. Each row reports a different outcome, whereas each column reports a different specification. Standard errors are clustered at the letter-writer level; we report the absolute $t$-statistics in parentheses. The coefficients are reported in terms of SDs of the dependent variable. Here ** and *** indicate statistical significance at the 5% and 1% levels, respectively.

Next, turning to the letter-writers' characteristics, we control for their gender, the RePEc ranking band of their institution and the number of reference letters they provide in our sample. These controls proxy for the quality and prestige of the letter writer. Finally, we also account for the length of the letter (total word count).

Each empirical model is estimated using four different sets of standard errors: robust, clustered by letter writer, clustered by the letter-writer's institution and clustered by the candidate's PhD-awarding institution.[16]

## 4.2. *Main Results*

### 4.2.1. *Baseline*

Table 4 presents baseline results for the six outcomes using standard errors clustered by letter writer. In Figure 6 we visualise these results along with those from a similar analysis carried out by further splitting the sample by letter-writer institutional ranking. Heterogeneity by institutional quality is a natural concern: familiarity with the job market might vary across institutions and in turn this might lead to different reference writing practices. Similarly, institutional culture, which may vary across the hierarchy of economic departments, can also shape language in references.

[16] Exceptions here are naturally the candidate institution FE and writer FE specifications.


Fig. 6. *Regression Results—All Letter Writers Combined.*
*Notes:* This figure shows the coefficient estimates for the regressions specified in (2). We compare all seven specifications described in Table 4. The symbol's filling permits visualising significance. Using four levels of possible standard error clustering (none, candidate's institution, letter-writer's institution or letter writer), we flag significance at three different levels (10%, 5% and 1%). We thus flag 12 possible significance indicators. For each level of clustering, the symbol in the graph is thus shaded with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Filled symbols are significant at the 1% level across all possible clustering. Open symbols do not reach significance for any type of standard error.

For instance, Lundberg and Stearns (2019) highlighted the hierarchical nature of the economics profession, in which a high fraction of potential letter writers comes from the most prestigious institutions. We address this here by focusing on top-25 or top-100 institutions and then probe this issue further in our analysis using institutional fixed effects below.

The standard errors are computed using the four types of clustering described at the end of Section 4.1. The total numbers of outcomes, specifications and clusterings combine into a total of 504 regressions. To visualise all these results in Figure 6, a darker shading of the marker indicates more specifications yielding statistically significant estimates for $\hat{\beta}$ (see the figure's notes for more details). Filled symbols are significant at the 1% level across all possible standard error clusterings. Open symbols do not reach significance for any type of clustering. The coefficient magnitudes are the estimates from (2) normalised by the SD of the respective dependent variable.

Figure 6 shows that no matter the institutional ranking, and across all specifications, female candidates are significantly more likely to be associated with 'grindstone' terms (from 5% to 12% of an SD). These results confirm our interpretation of the unsupervised analysis (see Section 3). We also observe that fewer terms related to research are used in letters supporting

female candidates. Both of these results echo findings from other disciplines (Trix and Psenka, 2003; Valian, 2005).

Furthermore, in all subsamples, female candidates are on average associated weakly and insignificantly so with more 'teaching and citizenship' terms. We also find no statistically significant differences between female and male candidates for 'standout' terms —in contrast with Trix and Psenka (2003) and Schmader *et al.* (2007), who observed a higher frequency of these adjectives in letters supporting male applicants for academic positions in medicine, and chemistry and biochemistry, respectively.[17]

Finally, we find fewer terms related to 'ability' or 'recruitment' for female candidates, but the estimates are not statistically significant.

The magnitude of the estimates of interest does not differ greatly across specifications, even after controlling for proxies capturing determinants of language that correlate with gender. This stability provides some reassurance that other unobserved confounding determinants of language used in references are unlikely to explain away the results.

#### 4.2.2. *Male and female writers*

Figure 7 compares results by letter-writer gender.[18] The pattern uncovered for 'grindstone' words continues to hold when we separately consider male and female referees. When it comes to research, it appears instead that the negative effect we have documented above is entirely driven by male writers. Female referees are actually *more* likely to use research terms for female candidates than for males.

In comparing male and female writers, we may worry that female referees cluster in departments with specific characteristics or that female writers attract different candidate types. We address these concerns later in this section by adding institution, writer or candidate fixed effects.

#### 4.2.3. *Cultural background*

Gender norms differ across cultures and are highly persistent over time (see, e.g., Alesina *et al.*, 2013). Academic economists come from all over the world, and thus we can explore whether the effects we have uncovered so far are driven by writers born in countries with more traditional gender norms. To carry out this analysis, we start by manually collecting, for each referee in our sample, information on their country of birth.[19]

To measure gender norms, we then follow the literature and use data from the World Value Survey (WVS, Wave 7, 2017–20).[20] In particular, we rely on whether the respondent agrees with the following statements: 'A pre-school child suffers with a working mother', 'University is more important for a boy than for a girl' and 'Men make better business executives than women do'. We consider a writer's country of birth as having more 'traditional' gender norms if the share of individuals agreeing/strongly agreeing with each statement is above the median for our sample.[21]

[17] We also experiment with separating the teaching and citizenship 'sentiments'. The coefficients remain insignificant and small in magnitude. Results are available in Online Appendix Figure E.1.

[18] There are 985 female letter writers (who have written 1,751 letters) in total, of whom only 156 are in the top-25 group (with 314 letters), and 382 in the top-100 group (735).

[19] When this information was unavailable, we use the country of the institution granting their undergraduate degree as a proxy.

[20] Because of a lack of later data, we use WVS results from 2010–14 for India.

[21] We average the shares by country across the three responses, akin to a first principal component, and take the median cutoff for this average response. Regression results for each of the three statements as well as the average are provided in Online Appendix Table C.4.


Fig. 7. *Regression Results, by Gender of Letter Writer.*
*Notes:* This figure shows the coefficient estimates for the regressions specified in (2), estimated separately for male and female letter writers. We compare all seven specifications described in Table 4. The symbol's filling permits visualising significance. Using four levels of possible standard error clustering (none, candidate's institution, letter-writer's institution or letter writer), we flag significance at three different levels (10%, 5% and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Filled symbols are significant at the 1% level across all possible clustering. Open symbols do not reach significance for any level of standard error.

The results are reported in Figure 8. We observe that, for all measures of gender norms, writers from all origins still tend to use more 'grindstone' terms for female candidates. Therefore, our results are not driven uniquely by referees born in countries with 'traditional' gender norms. However, we note that the estimates are qualitatively larger for these writers, although the difference is not significant.[22]

#### 4.2.4. *Specifications with fixed effects*

We have uncovered systematic differences in the attributes highlighted for female and male candidates. Here we explore whether these differences are driven by the sorting of female candidates across institutions and/or letter writers.

Boustan and Langan (2019) documented that female representation is a persistent attribute of economics departments, and that it matters to promote women's careers. Hence, it is important to study whether institutional sorting drives our results. We run regressions including fixed effects

[22] The results for a fully interacted model are presented in Online Appendix Table C.4.

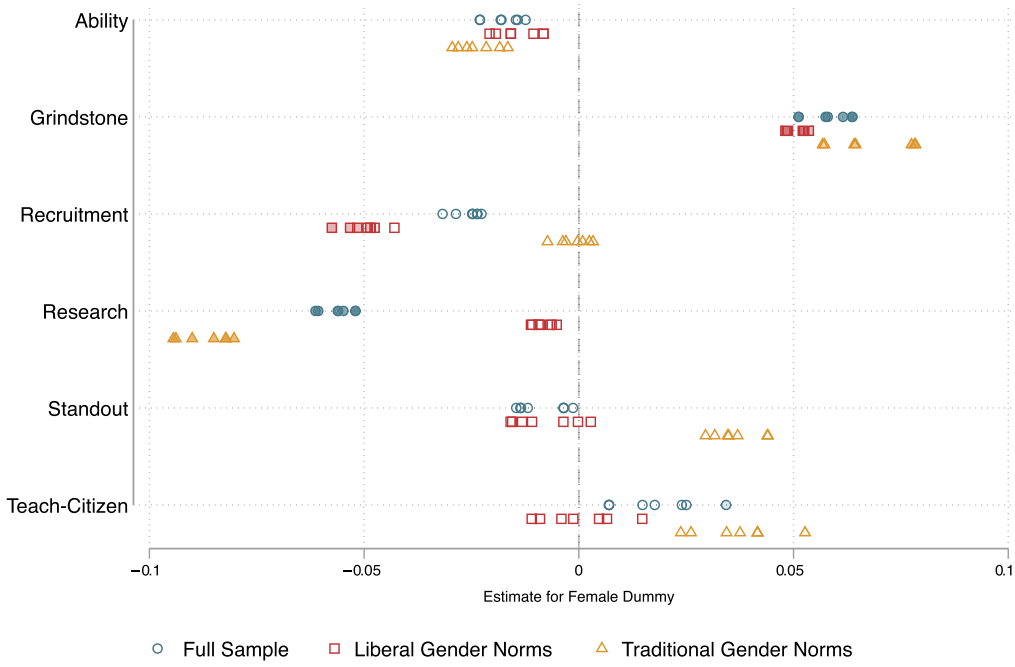


Fig. 8. *Regression Results, by Gender Norms of Writer.*
*Notes:* This figure shows the coefficient estimates for the regressions specified in (2), estimated separately for letter writers from countries with traditional versus liberal gender norms. We compare all seven specifications described in Table 4. The symbol's filling permits visualising significance. Using four levels of possible standard error clustering (none, candidate's institution, letter-writer's institution or letter writer), we flag significance at three different levels (10%, 5% and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shaded with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Filled symbols are significant at the 1% level across all possible clustering. Open symbols do not reach significance for any level of standard error.

for the candidate's institution. The results are reported in Figure 9. They suggest that among students from the same cohort, graduating from the same institution—who, for example, were admitted to PhD programs arguably applying the same entry requirements—women are still significantly more likely to be described with 'grindstone' terms.

We are still concerned that, even within the same graduate program, sorting across letter writers could explain our findings. To address this concern, in Figure 9 we also plot the estimates of a set of specifications including writer fixed effects.[23] Note that these models are identified from referees who have written two or more letters across all five sample years, with at least one for a female and one for a male candidate. This significantly reduces our sample (we can include only 18% of the letter writers).

[23] In our analysis we drop the top-1% most prolific referees ($n = 12$), namely, those with a dozen or more letters in the sample, since fixed effect estimates are sensitive to outliers. Leaving these referees in the sample leads to qualitatively similar results.

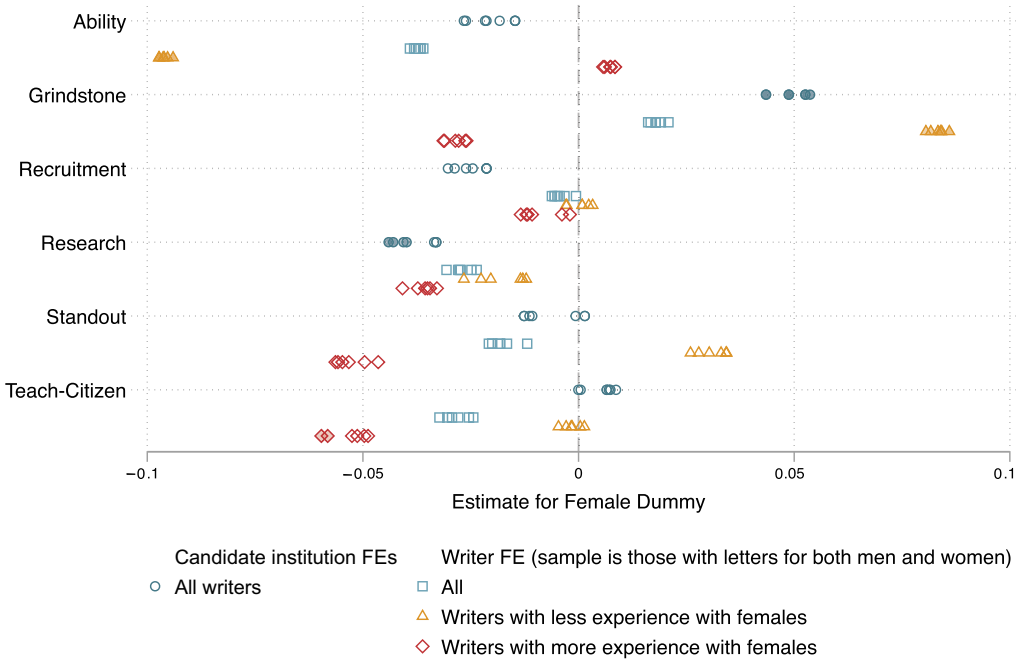


Fig. 9. *Regression Results with Candidate Institution or Writer Fixed Effects.*
*Notes:* This figure shows the coefficient estimates for the regressions specified in (2), estimated separately with candidate institution or letter-writer fixed effects. The symbol's filling permit visualising significance. Using two levels of possible standard error clustering for each fixed effect [none or candidate's institution (respectively letter writer) for candidate's fixed effects (respectively letter-writers' fixed effects)], we flag significance at three different levels (10%, 5% and 1%). We thus flag six possible significance indicators. Then, for each level of clustering, the symbol in the graph is shaded with a 17% ($\approx 100/6$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Filled symbols are significant at the 1% level across all possible clustering. Open symbols do not reach significance for any level of standard error clustering. Additional information on the sample and results for the unclustered, robust standard errors are contained in Online Appendix C.

In the same figure we also separately consider the sample of referees who have less (more) experience with female candidates.[24] This analysis permits unveiling significant heterogeneity. The 'less experienced' group appears to have a significantly higher likelihood of using 'grindstone' terms for women *and* a lower likelihood of using 'ability' ones. These estimates are also larger in magnitude compared to the baseline shown in Figure 6. Experience may matter for two main reasons. On the one hand, referees may vary in their perception of women, and female candidates could sort accordingly to avoid differential treatment. On the other hand, it could be that referees do not differ initially, but that their exposure to female candidates leads them to update preconceptions (a learning effect also observed, for example, by Beaman *et al.*, 2009). Further research is needed to disentangle these two mechanisms.

[24] Less (more) experience is defined to broadly balance the sample of letters across both groups. Writers who have written less (more) than a third of letters for women are considered less experienced. The 'less experienced' group accounts for 42% of referees in the subsample with two or more letters and at least one female candidate.

Furthermore, it is worth pointing out that writers with more experience put less emphasis on 'teaching and citizenship' qualities for women compared to men, although not robustly significantly so.

Finally, fixed effects also allow a more subtle comparison of female and male *writers*. In particular, we can contrast the language chosen by male and female writers for the same candidate by adding *candidate* fixed effects. For job applicants who have both male and female referees, we test whether female writers use different language in general and for female candidates in particular. Our results are presented in Online Appendix Table C.9. They suggest that, for the same candidate, female *writers* use different language. They rely more on 'grindstone' and 'teaching and citizenship' language, and less on 'recruitment'. Importantly, there is no difference in these patterns depending on the gender of the *candidate*, pointing towards the absence of 'same-sex preferences' of letter writers.

### 4.3. *Robustness Checks*

#### 4.3.1. *Alternative letter lengths*

In our baseline analysis, we have defined the end of letter using the last 200 words before the 'polite' end phrase. In Online Appendix Figure E.2 we explore the sensitivity of the baseline results to this choice by experimenting with two alternative cutoffs, using 150 and 250 words. We also study the full reference letters (see Online Appendix Figure E.2).[25] Our findings are unaffected.

#### 4.3.2. *Fields*

We explore heterogeneity of the results according to the candidate's research field to assess possible subcultural differences in the profession. Grouping applicants into meaningful research areas is challenging. On the platform, they typically choose a field, loosely based on JEL codes. Unfortunately, the fields proposed by the platform pool diverse subgroups of the profession, i.e., scholars that are unlikely to publish in the same journals or participate in the same events (conferences, seminars, etc.). For instance, the field 'Development and Growth' includes both macroeconomists working on long-run growth and microeconomists carrying out field experiments in developing countries. Given these shortcomings, we employ an unsupervised data-driven approach to classify candidates into three broad research groups, namely applied economics, theory and macroeconomics, and a residual category. Online Appendix D describes the procedure.[26]

One possible explanation for the association of women to 'grindstone' expressions is that they sort into research fields that require more industriousness than ability. This set of skills is often associated with empirical work. However, the results reported in Online Appendix Figure D.3 show that this association remains strong and significant within applied micro, casting doubt on such a hypothesis. We also uncover a strong negative effect for 'ability' within theory. This finding is worth highlighting as in this field raw talent is arguably valued very highly. This observation sheds new light on earlier findings by Leslie *et al.* (2015), according to whom academic fields that particularly emphasise the role of raw talent are characterised by lower female representation.

[25] Appendix figures are accompanied by corresponding tables providing further details on the specification and sample. For ease of presentation, we do not refer to these tables in the main text.

[26] The baseline analysis reported in Figure 6 employs research field fixed effects using these four clusters. In Online Appendix Figure E.3. we repeat the same exercise using instead the more detailed 145 field definitions from the platform. The findings are robust.

4.3.3. *Main advisors*

So far, we have used all the letters that were submitted for each applicant, i.e., those that were written by the main advisor and those written by other faculty members familiar with the candidate's research. As the main advisor might have better knowledge of the applicant, it is important to investigate whether there are differences in the language he/she used compared to that of the other referees. We collect data on the identity of the letter writers for candidates who were in the job market up to three years after completing their PhD.[27] The results of our analysis are illustrated in Online Appendix Figure E.4, where we report our baseline estimates for the collected sample and those obtained focusing separately on the letters written by the main advisor and the other reviewers.

The findings indicate that the patterns for 'grindstone' terms are generally comparable, but accentuated for letter writers who are *not* the main advisors. Moreover, there is notable divergence in the cases of 'ability' and 'standout'. Compared to main advisors, other referees use significantly fewer 'ability' and 'standout' terms. Overall, this analysis presents suggestive evidence that main advisors are writing more favourable letters for women compared to other referees. Main advisors arguably know the candidates much better and spend more time writing and polishing the letters[28] and through these lengthy processes some of their preconceptions may be toned down.[29]

4.3.4. *Location of PhD-granting institution*

The job market for economists is historically a US institution, and faculty members based there may be better acquainted with the standards of reference writing. We investigate whether our results are driven by letter writers outside the United States, in which case our findings might result from lower levels of experience in the process. Online Appendix Figure E.5 presents the results. Overall, we do not uncover significant differences between the two groups, with the exception of 'research' terms. Referees based outside the United States use significantly fewer research-related words for female candidates compared to their US-based counterparts.

4.3.5. *Candidate's visibility*

So far, our analysis has accounted for the underlying potential of the candidate by controlling for the number and quality of their publications and the ranking of their institution. Additionally, we have shown that our results continue to hold when we compare candidates within the same institution. As a further robustness check, we also account for the circulation of the candidate's job market paper at the time they are on the market. This proxies for the candidate's visibility and/or the extent of networking carried out in this period. We do so by manually collecting information from the job market paper acknowledgements. Using the Stanford Name Entity Recognition tagger, we separate out *people* thanked and *institutions* mentioned. We also compute the length of this note and flag whether the job market paper is single authored. Online Appendix Figure E.3 shows that results with these controls remain unchanged compared to the baseline.

[27] This represents around 50% of the sample of candidates. Candidates who defended earlier were less likely to have a letter from their PhD advisor and were also less likely to report that information on their CV.

[28] Letters from main advisors are on average 33% longer.

[29] The *Oxford English Dictionary* defines a stereotype as a 'widely held but fixed and oversimplified image or idea of a particular type of person or thing.' Describing women as hardworking conforms to the stereotype of women in science (Valian, 1999). Our results suggest that the most informed letter writers—main advisors or writers with greater experience with female candidates—use language that is less in line with these stereotypes. These patterns align with the interpretation that less informed writers are 'stereotyping', in the sense of using an 'oversimplified image' as a shortcut.

#### 4.3.6. *Postdocs*

Female candidates—who may be conscious of potential gender stereotyping—may change their behaviour during their career to make stereotypical traits less salient (e.g., Hengel, 2022, highlighted that women improve their writing throughout their careers, whereas men do not). To assess this possibility, we contrast the estimates for candidates fresh out of PhD programs and those who have been out for 1–3 years ('postdocs'). Results are reported in Online Appendix Figure E.6. Overall, the estimates for the postdoc sample are noisier, as expected due to smaller sample sizes. For 'grindstone' language, the effects remain generally stable. We do however observe an increase in the language related to 'ability' for female postdocs compared to their male counterparts. Further research is needed to establish whether this effect is driven by learning, as found by Hengel (2022), or by differential selection into postdocs.

#### 4.3.7. *Letter-writer seniority/rank*

Results with fixed effects in Figure 9 indicate that writers with less experience with female candidates use more stereotypical language. One alternative explanation for this finding could be that such practice declines with academic experience per se. Using manually collected information on the year the letter writer graduated from their PhD or their academic rank (assistant, associate, full professor), we illustrate that this is not the case. In Online Appendix Figures E.7 and E.8, we observe that the most senior letter writers use gender stereotypical language much more often and that they drive the 'grindstone' results.[30]

### 4.4. *Additional Results*

In this section, we shift our attention away from the analysis of the 'sentiments' expressed in letters ('ability', 'grindstone', etc.) and consider alternative attributes that speak to the quality of the candidate or how the letter is written.

#### 4.4.1. *Placement qualifiers*

Many letter ends carry explicit signals about the candidate, which can be positive or negative, as well as comparisons with placements of recent graduates—see our earlier discussion in Section 1. To analyse potentially gendered patterns in the prevalence of these signals, we compile a dictionary of over 1,000 placement qualifiers. Examples include (for negatives) 'except maybe from those in the top 10/20/30' or 'apart from the very best'; (for positives) 'great hire', 'a star candidate', 'including institutions at the very top'; (for comparatives) 'compared to' or 'on par'. Of the letters in our sample, 24% include at least one positive signal, 13% a negative one and 6% of all letters include a comparative term.

In the baseline specification in (2), we replace the dependent variable with outcomes related to these qualifiers. The first three lines in Figure 10 show that letters written in support of women tend to have significantly fewer positive signals, no significant difference in terms of negative ones and a net negative signal.[31] These results also hold when we consider instead binary variables flagging the presence of either positive or negative signals, or the sign of the net signal.[32] The effects are sizeable. For instance, a letter in support of a female candidate has a 3 percentage

[30] We conduct additional analysis (not reported) splitting the PhD cohorts into five rather than two groups and obtain qualitatively identical findings.

[31] The net signal simply subtracts the count of negative from that of the positive signals.

[32] The dummy for net negative (positive) signal is 1 if the net signal is negative (positive) and 0 otherwise.


Fig. 10. *Regression Results—Placement Signals as Outcomes.*
*Notes:* This figure shows the coefficient estimates for the regressions specified in (2) when outcomes are proxies for academic placement. Rows 1–3 are for counts of positive, negative and net signals; rows 4–7 adopt binary variables for the same outcomes; the final two rows are counts and a dummy for comparative statements in the letter end. The symbol's filling permit visualising significance. Using four levels of possible standard error clustering (none, candidate's institution, letter-writer's institution or letter writer), we flag significance at three different levels (10%, 5% and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shaded with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Filled symbols are significant at the 1% level across all possible clustering. Open symbols do not reach significance for any level of standard error clustering.

point lower probability of containing a positive signal, to be compared with the fact that only 24% of letters contain one.

Finally, we study comparative terms using total counts or an indicator for their presence. This analysis suggests that letters written in support of female candidates have a 1 percentage point higher likelihood of carrying a comparison, a sizeable effect given that only 6% of letters contain one.

Overall, this analysis suggests that women are not shown in a more negative light (in contrast to findings in the literature about 'doubt raisers', see e.g., Trix and Psenka, 2003; Madera *et al.*, 2009). However, they obtain less outright praise, which is consistent with the work of Dutt *et al.* (2016), who found that women in geosciences are less likely to receive 'excellent' letters. The higher prevalence of comparative terms suggests though that the information provided for female candidates might be more 'precise'.

4.4.2. *Letter length and readability*

A standard finding in the literature suggests that letters for female candidates are shorter (Trix and Psenka, 2003). We also investigate proxies for letter-writer effort by looking at letter length and writing clarity in the full reference letter as well as on the discussion of the candidate's job market paper.[33] Our results, reported in Online Appendix Figure F.1, show that female candidates in economics do not receive shorter letters than their male peers. However, the analysis of readability, using the Flesch Reading Ease score suggests that letters for female candidates are harder to read.[34] If we instead use the Dale–Chall Readability score, the results are not statistically significant. Finally, our investigation of the 'research slice' provides no clear evidence of a bias in the discussion of the candidate's JMP.

4.4.3. *Timing of the reference letter*

Work by Baltrunaite *et al.* (2022) revealed that, for their sample of references submitted to two Italian institutions, letter writers are significantly less likely to send references for female candidates by the deadline stated on the application platform.

We test whether this finding also holds in our sample, using two alternative measures of timeliness. First, we construct a binary variable equal to 1 if the application package contains strictly fewer than three letters, the required number in our application process. The results, shown in Online Appendix Figure F.1, confirm that female candidates are significantly more likely to have an incomplete set of references.

Second, we exploit information on the date stated on the reference letter, which is a proxy for the timeliness of the referee given that by default the same letter is automatically submitted by the application platform to all institutions.[35] Our results, shown in the same figure, indicate that letters for female candidates tend to be written earlier (between 0.5 to 1 day earlier, depending on the specification), but this effect is imprecisely estimated and not robust to the inclusion of a full set of controls.

## 5. Placement

So far, our analysis has focused on gendered patterns in the reference letters. We have documented differences in the language chosen for female and male candidates. In line with previous literature, we observe that women are described with more 'grindstone' attributes and at times fewer 'ability' and 'research' ones (Bourdieu and Passeron, 1977; Trix and Psenka, 2003; Schmader *et al.*, 2007). The obvious corollary question is whether being described as a 'grindstone' candidate actually affects job prospects.

To answer this question, we have carried out a systematic collection of the first professional placement of each candidate in the year following their appearance in our sample. Combining information from personal websites, academic departments' placement records and LinkedIn profiles, we establish whether the candidate placed in academia or elsewhere. For academic placements, we also link the name of the institution to its RePEc rankings to proxy for the prestige of the job.

[33] This corresponds to the 'research slice'—see Online Appendix D for more details.

[34] See Hengel (2022, Table 1) for exact definitions. For the Flesch index, a higher score means that the text is easier to read; for the Dale–Chall index, the reverse is the case.

[35] We focus only on recent PhDs for whom letters are presumably written for the first time. We rely on the package `ctparse` to detect and parse dates in the beginning of the letters, which we also double check manually.

It is important to be mindful that reference letters are only one input in the recruitment process. They typically play an important role in enabling candidates to secure interviews at the job market meetings. Many other factors such as presentation skills, research agenda or departmental politics determine whether a candidate actually receives a job offer. Therefore, any result linking letters to actual placement needs to be interpreted with caution.

To study the relation between letter sentiment and placement, we estimate the regression model

$$\begin{aligned}\text{Placement}_{diwt} = {} & \alpha + \beta\,\text{Female}_i \\ & + \sum_{k=1}^{6}(\theta_k\,\text{Sentiment}_{diwt} + \kappa_k\,\text{Sentiment}_{diwt} \times \text{Female}_i) \\ & + \mathbf{X}_i'\gamma + \mathbf{W}_w'\lambda + \nu_t + \varepsilon_{diwt},\end{aligned}$$

where $\text{Placement}_{diwt}$ is the job market outcome of individual $i$, for whom letter $d$ was written by writer $w$ in year $t$. For each sentiment $k$—'ability', 'grindstone', etc.—we are interested in its impact on placement and whether this impact varies with the gender of the candidate. Controls are the same as defined in Section 4.

Our results are reported in Table 5. We consider three measures of placement.[36] The first is a binary variable flagging whether a candidate obtained an academic job and takes into consideration all candidates for whom we found job market outcomes. The second and third measures focus on those who embarked on an academic career, and study the 'quality' of the academic placement. Our first proxy is the RePEc institution score, a continuous variable, which we rescale for ease of interpretation so that a positive coefficient indicates a more prestigious institution. The other is a binary variable indicating whether the candidate placed among the top-200 institutions in RePEc.[37] Odd columns report a parsimonious model, with only the sentiments and cohort fixed effects as controls, whereas even ones report estimates accounting for the full set of controls.

Overall, we observe that women are more likely to place in academia (columns (1) and (2)) and, conditional on embarking on an academic career, they land jobs in more prestigious institutions (columns (4) and (6)). These results are compatible with both positive selection of women in academia and 'positive discrimination'. Evidence of both phenomena has been uncovered by recent literature (for positive selection, see Iaria *et al.*, 2022; for 'positive discrimination', see Card *et al.*, 2022).

We turn now to the analysis of the effect of 'sentiment' on placement. 'Standout' and 'teaching and citizenship' terms are the only ones to significantly affect the probability of placing in academia. Columns (1) and (2) show that a one SD increase in the usage of 'standout' terms is associated with a 2 percentage point higher likelihood of an academic placement. For female candidates, the aggregate effect is effectively nil instead. For 'teaching and citizenship', we find

[36] Models 1 and 2 contain letters for all 2,588 candidates for whom placement information was found; the dependent variable is 1 for an academic placement (Assistant Professor position or postdoc) and 0 otherwise (international organisations, central banks or private sector). Models 3 and 4 contain letters for 957 candidates who placed among the top-500 institutions in RePEc, the only ones for which an RePEc score is computed. The RePEc score is a continuous variable (e.g., third-placed UC Berkeley has a score of 7.12, first-placed Harvard of 1.96). Models 5 and 6 include letters for 1,865 candidates who placed in academia as either AP or postdocs. In this sample we can include all academic institutions (we are not constrained by the availability of an RePEc score). Teaching fellows are included in academic placements in 1 and 2, but results remain identical if we exclude them.

[37] Our results are robust to alternative measures of placement quality; see Online Appendix Table F.3.

Table 5. *Letter Sentiment and Placement.*

| Dependent variable | Academia (dummy) | | Inst. RePEc score | | Top-200 RePEc Inst. | |
|---|---|---|---|---|---|---|
| Sample | All placements | | Academic placements | | AP and postdoc | |
| Controls | Sentiment | All | Sentiment | All | Sentiment | All |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Female candidate | 8.9056 | 8.2124 | 12.3080 | 21.7822 | 7.2032 | 11.4131 |
| | (2.37)** | (2.20)** | (0.84) | (1.52) | (1.56) | (2.55)** |
| Ability | 0.1695 | 0.1138 | 0.7169 | 0.2093 | 0.9848 | 0.8810 |
| | (0.28) | (0.19) | (0.29) | (0.09) | (1.39) | (1.30) |
| Ability × Female candidate | −0.0861 | −0.0697 | 2.5162 | 1.8299 | 0.3787 | 0.1513 |
| | (0.08) | (0.07) | (0.59) | (0.43) | (0.29) | (0.12) |
| Grindstone | −0.4843 | −0.5564 | −2.6959 | 1.2020 | −0.3714 | −0.0088 |
| | (0.82) | (0.95) | (1.12) | (0.51) | (0.51) | (0.01) |
| Grindstone × Female candidate | 0.1153 | 0.1686 | −10.0826 | −10.4156 | −2.4636 | −2.6246 |
| | (0.11) | (0.17) | (2.38)** | (2.49)** | (2.02)** | (2.19)** |
| Recruitment | 0.7008 | 0.8597 | 3.1951 | 1.1869 | 1.8494 | 0.4370 |
| | (1.14) | (1.38) | (1.35) | (0.50) | (2.58)*** | (0.63) |
| Recruitment × Female candidate | 0.4961 | 0.2886 | −3.2295 | −3.2034 | −0.4627 | −0.4311 |
| | (0.48) | (0.28) | (0.80) | (0.80) | (0.37) | (0.35) |
| Research | −0.4928 | −0.2642 | 3.2833 | 5.1540 | 0.6711 | 1.1944 |
| | (0.80) | (0.43) | (1.39) | (2.25)** | (0.95) | (1.76)* |
| Research × Female candidate | −1.7432 | −1.7328 | 0.8732 | −1.1355 | −0.6655 | −1.0630 |
| | (1.63) | (1.63) | (0.21) | (0.28) | (0.51) | (0.83) |
| Standout | 1.9208 | 1.9579 | −0.0326 | −1.5381 | 1.2065 | 0.0060 |
| | (3.28)*** | (3.36)*** | (0.01) | (0.64) | (1.74)* | (0.01) |
| Standout × Female candidate | −1.9979 | −1.9578 | 7.1538 | 6.5266 | 2.5831 | 2.2041 |
| | (1.85)* | (1.84)* | (1.72)* | (1.62) | (1.94)* | (1.71)* |
| Teaching and citizenship | 0.1599 | −0.1094 | 1.5916 | 4.3917 | −1.1806 | 0.5379 |
| | (0.26) | (0.18) | (0.68) | (1.88)* | (1.63) | (0.77) |
| Teaching and citizenship × Female candidate | 1.6494 | 1.8723 | −6.2089 | −8.0217 | −3.4235 | −4.0236 |
| | (1.61) | (1.85)* | (1.55) | (2.04)** | (2.75)*** | (3.30)*** |
| FEs/variables absorbed | 5 | 25 | 5 | 25 | 5 | 25 |
| Additional variables | 0 | 6 | 0 | 6 | 0 | 6 |
| Number of letters | 8,760 | 8,760 | 3,119 | 3,119 | 6,008 | 6,008 |
| dto for females | 2,588 | 2,588 | 991 | 991 | 1,872 | 1,872 |
| Number of candidates | 2,738 | 2,738 | 957 | 957 | 1,865 | 1,865 |
| dto female | 830 | 830 | 313 | 313 | 596 | 596 |
| Number of writers | 4,461 | 4,461 | 2,091 | 2,091 | 3,453 | 3,453 |
| dto female | 774 | 774 | 324 | 324 | 586 | 586 |
| Letters by female writers | 1,339 | 1,339 | 445 | 445 | 910 | 910 |
| Year FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Letter sentiments | Yes | Yes | Yes | Yes | Yes | Yes |
| Ethnicity/race FEs | No | Yes | No | Yes | No | Yes |
| Institution rank FEs | No | Yes | No | Yes | No | Yes |
| Years since PhD | No | Yes | No | Yes | No | Yes |
| Research field FEs | No | Yes | No | Yes | No | Yes |
| Publications | No | Yes | No | Yes | No | Yes |
| Writer chars | No | Yes | No | Yes | No | Yes |
| Letter length | No | Yes | No | Yes | No | Yes |

*Notes:* The table shows OLS regression results of placement outcomes on the letter-specific sum of tf-idf statistics related to the bag of expressions mentioned in the row label and its interaction with a female candidate dummy as well as the additional controls as indicated. We report the absolute $t$-statistics in parentheses. Here $*$, $**$ and $***$ indicate statistical significance at the 10%, 5% and 1% levels, respectively. Results are in percentage points except in (3) and (4) where they are in percent. Furthermore, signs in (3) and (4) are reversed for consistency with the other two placement outcomes. Additional results are presented in Online Appendix Tables F.3 and F.4.

no effect for male candidates, but a positive one for women (amounting to a 1.9 percentage point increase in the likelihood of academic placement).

In columns (3) and (4) three 'sentiments' stand out: 'grindstone', 'standout' and 'teaching and citizenship'. The results for 'grindstone' indicate no statistically significant effect for men, whereas, for women, a one SD increase in this 'sentiment' is associated with a large and significant

(10 points in the rank score) decrease in the ranking of the institutions where they place.[38] Moreover, women benefit more from standout terminology (6.5 to 7.2 increase in the rank score relative to men). Finally, men who receive letters emphasising 'teaching and citizenship' get jobs in higher-ranked institutions (1.6 to 4.4 point increase in rank score), whereas the effect is reversed for women (6.2 to 8.0 point decrease in rank score relative to men).

When considering the likelihood of placing in a top-200 institution (columns (5) and (6)), the 'grindstone' sentiment again plays an important role for women. A one SD increase in 'grindstone' terms is associated with a negligible effect for male candidates, but a large negative effect for women, of the order of a 2.6 percentage point decrease in the probability of obtaining a job in a higher-ranked institution. Letters with more 'standout' or 'research' terms are associated with better placement for both men and women, although significance and magnitude for 'standout' drop for men when adding controls. Finally, 'teaching and citizenship' words also significantly decrease the probability of placing in a top-200 institution, but only so for women.

In this analysis, we do not account for the presence of placement qualifiers, which also exhibit gendered patterns, as shown in Figure 10. The reason is that many placement terms are already included in the 'standout', 'ability', or 'recruitment' sentiments, although without differentiating between positives and negatives. As a robustness check, we present in Online Appendix Table F.4 the results when including the binary variables for placement qualifiers. The results for sentiments commented on above remain unchanged. The analysis also suggests that positive signals improve placement for men and women.

This discussion indicates that the gendered sentiments expressed in job market reference letters are associated with initial placement patterns. Our RePEc score/ranking analysis indicates that 'grindstone' terminology hurts female placement, whether we consider a continuous measure of institutional quality or a binary identifier.[39] Although further research is needed to establish causality, our results indicate that the language in reference letters can play an important role in the first step of the academic career. These results are consistent with findings by Baltrunaite *et al.* (2022) on longer-term career outcomes.

## 6. Concluding Remarks

In this paper, we carried out what is to the best of our knowledge the first systematic analysis of recommendation letters in the junior academic job market in economics. Using both supervised and unsupervised methods, we have documented the presence of important differences in the language used to describe female applicants. Women are more often described with terms praising their 'hard work' or 'dedication' than men. This pattern is robust to alternative specifications and holds across many subsamples of the data. Similarly, we uncover evidence of a lower emphasis on 'ability', especially when comparing individuals within the same institution or for those sharing the same referee.

[38] At the median rank score of 144, a 10 point decline in the score represents a drop of approximately 10 positions in the rank.

[39] Note that more work is needed to establish clear-cut implications in terms of discrimination. On the one hand, one may argue that if employers are seeking to have a balanced workforce in terms of 'grindstone' and other attributes, then penalising 'grindstone' women could just compensate for their greater propensity to exhibit those traits. On the other hand, a fully fledged model of the job market should account for the fact that letter writers could strategically adjust and choose fewer 'grindstone' attributes for their female candidates in order to increase the chances of securing a better placement.

Sociologists characterise these systematic language patterns as possibly resulting from stereotyping, and highlight their potential negative connotations as a strong emphasis on diligence may imply a lack of 'brilliance' (Bourdieu and Passeron, 1977; Valian, 1999). We illustrate the salience of language in reference letters for job market placement by documenting that women receiving letters emphasising that they 'work hard' obtain less prestigious academic positions, while the same is not true for men. On the contrary, those whose letters highlight 'standout' attributes benefit from improved academic placement. Although further evidence is needed, our results thus suggest that, for letter writers who are pushing their candidates towards research-intensive academic employment, using a language emphasising fewer 'grindstone' and more 'standout' attributes increases the chances of achieving the desired objective.

As academics, we know how much time is spent writing and polishing reference letters for job market candidates. This is an occasion where we try our best to promote our students. Therefore, it is unlikely that, on average, we are willingly undermining female students by emphasising less desirable attributes. On a positive note, recent research has shown that unconscious biases can be addressed by providing the actors involved with evidence of the existence of such biases (Boring and Philippe, 2021). By shedding light on these patterns, we hope that this research will be a first step towards increasing awareness of our biases and thereby reducing possible stereotyping in the job markets.

*University of Nottingham & CEPR, UK*
*University of Nottingham, CEPR, UK & IZA, Germany*
*University of Nottingham & CEPR, UK*

Additional Supporting Information may be found in the online version of this article:

**Online Appendix**
**Replication Package**

## References

Alan, S., Boneva, T. and Ertac, S. (2019). 'Ever failed, try again, succeed better: Results from a randomized educational intervention on grit', *The Quarterly Journal of Economics*, vol. 134(3), pp. 1121–62.
Alesina, A., Giuliano, P. and Nunn, N. (2013). 'On the origins of gender roles: Women and the plough', *The Quarterly Journal of Economics*, vol. 128(2), pp. 469–530.
Ash, E., Chen, D.L. and Naidu, S. (2022). 'Ideas have consequences: The impact of law and economics on American justice', Working Paper 29788, National Bureau of Economic Research.
Ash, E., Chen, D.L. and Ornaghi, A. (2023). 'Gender attitudes in the judiciary: Evidence from U.S. circuit courts', *American Economic Journal: Applied Economics*, forthcoming.
Baltrunaite, A., Casarico, A. and Rizzica, L. (2022). 'Women in Economics: The role of gendered references at entry in the profession', Discussion Paper 17474, Centre for Economic Policy Research.
Bayer, A. and Rouse, C.E. (2016). 'Diversity in the economics profession: A new attack on an old problem', *Journal of Economic Perspectives*, vol. 30(4), pp. 221–42.
Beaman, L., Chattopadhyay, R., Duflo, E., Pande, R. and Topalova, P. (2009). 'Powerful women: Does exposure reduce bias?', *The Quarterly Journal of Economics*, vol. 124(4), pp. 1497–540.
Boring, A. (2017). 'Gender biases in student evaluations of teaching', *Journal of Public Economics*, vol. 145, pp. 27–41.
Boring, A. and Philippe, A. (2021). 'Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching', *Journal of Public Economics*, vol. 193, 104323.
Bosquet, C., Combes, P.P. and García-Peñalosa, C. (2019). 'Gender and promotions: Evidence from academic economists in France', *The Scandinavian Journal of Economics*, vol. 121(3), pp. 1020–53.
Bourdieu, P. and Passeron, J.C. (1977). *Reproduction in Education, Society, and Culture*, United Kingdom: Sage Publications.

Boustan, L. and Langan, A. (2019). 'Variation in women's success across PhD programs in economics', *Journal of Economic Perspectives*, vol. 33(1), pp. 23–42.
Card, D., DellaVigna, S., Funk, P. and Iriberri, N. (2020). 'Are referees and editors in economics gender neutral?', *The Quarterly Journal of Economics*, vol. 135(1), pp. 269–327.
Card, D., DellaVigna, S., Funk, P. and Iriberri, N. (2022). 'Gender differences in peer recognition by economists', *Econometrica*, vol. 90(5), pp. 1937–71.
Ceci, S.J. and Williams, W.M. (2009). *The Mathematics of Sex: How Biology and Society Conspire to Limit Talented Women and Girls*, New York: Oxford University Press.
Coles, P., Cawley, J., Levine, P.B., Niederle, M., Roth, A.E. and Siegfried, J.J. (2010). 'The job market for new econonomists: A market design perspective', *Journal of Economic Perspectives*, vol. 24(4), pp. 187–205.
De Fraja, G., Facchini, G. and Gathergood, J. (2019). 'Academic salaries and public evaluation of university research: Evidence from the UK research excellence framework', *Economic Policy*, vol. 34, pp. 523–83.
Deschamps, P. (2022). 'Gender quotas in hiring committees: A boon or a bane for women?', SOFI mimeo.
Dupas, P., Sasser Modestino, A., Niederle, M. and Wolfers, J., The Seminar Dynamics Collective. (2021). 'Gender and the dynamics of economics seminars', Working Paper 28494, National Bureau of Economic Research.
Dutt, K., Pfaff, D.L., Bernstein, A.F., Dillard, J.S. and Block, C.J. (2016). 'Gender differences in recommendation letters for postdoctoral fellowships in geoscience', *Nature Geoscience*, vol. 9(11), pp. 805–8.
Fan, Y., Shepherd, L., Slavich, E., Waters, D., Stone, M., Abel, R. and Johnston, E. (2019). 'Gender and cultural bias in student evaluations: Why representation matters', *PloS One*, vol. 14(2), e0209749.
Funk, P., Iriberri, N. and Savio, G. (2019). 'Does scarcity of female instructors create demand for diversity among students? Evidence from observational and experimental data', Discussion Paper 14190, Centre for Economic Policy Research.
Ginther, D.K. and Kahn, S. (2004). 'Women in economics: Moving up or falling off the academic career ladder?', *Journal of Economic Perspectives*, vol. 18(3), pp. 193–214.
Goffman, E. (1979). *Gender Advertisements*, 1st edn., New York: Harper and Row.
Gornick, V. (1979). Preface. *Gender Advertisements*, by Erving Goffman, New York: Harper and Row.
Grimmer, J. and Stewart, B.M. (2013). 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political Analysis*, vol. 21(3), pp. 267–97.
Grossbard, S., Yilmazer, T. and Zhang, L. (2021). 'The gender gap in citations of articles published in two demographic economics journals', *Review of Economics of the Household*, vol. 19(3), pp. 677–97.
Hebl, M., Nittrouer, C., Corrington, A. and Madera, J. (2018). 'How we describe male and female job applicants differently', *Harvard Business Review*, 27 September.
Hengel, E. (2022). 'Publishing while female: Are women held to higher standards? Evidence from peer review', *Economic Journal*, vol. 132, pp. 2951–-91.
Hospido, L. and Sanz, C. (2021). 'Gender gaps in the evaluation of research: Evidence from submissions to economics conferences', *Oxford Bulletin of Economics and Statistics*, vol. 83(3), pp. 590–618.
Iaria, A., Schwarz, C. and Waldinger, F. (2022). 'Gender gaps in academia: Global evidence over the twentieth century', Discussion Paper 17422, Centre for Economic Policy Research.
Kahn, S. and Ginther, D. (2017). 'Women and STEM', Working Paper 23525, National Bureau of Economic Research.
Koffi, M. (2021). 'Innovative ideas and gender inequality', Working Paper 35, Canadian Labor Economics Forum.
Leslie, S.J., Cimpian, A., Meyer, M. and Freeland, E. (2015). 'Expectations of brilliance underlie gender distributions across academic disciplines', *Science*, vol. 347(6219), pp. 262–5.
Lundberg, S., ed. (2020). *Women in Economics*, London: CEPR Press.
Lundberg, S. and Stearns, J. (2019). 'Women in economics: Stalled progress', *Journal of Economic Perspectives*, vol. 33(1), pp. 3–22.
MacArthur, H.J., Cundiff, J.L. and Mehl, M.R. (2020). 'Estimating the prevalence of gender-biased language in undergraduates' everyday speech', *Sex Roles*, vol. 82(1–2), pp. 81–93.
MacNell, L., Driscoll, A. and Hunt, A.N. (2015). 'What's in a name: Exposing gender bias in student ratings of teaching', *Innovative Higher Education*, vol. 40(4), pp. 291–303.
Madera, J.M., Hebl, M.R., Dial, H., Martin, R. and Valian, V. (2019). 'Raising doubt in letters of recommendation for academia: Gender differences and their impact', *Journal of Business and Psychology*, vol. 34(3), pp. 287–303.
Madera, J.M., Hebl, M.R. and Martin, R.C. (2009). 'Gender and letters of recommendation for academia: Agentic and communal differences', *Journal of Applied Psychology*, vol. 94(6), pp. 1591–9.
Mengel, F., Sauermann, J. and Zölitz, U. (2019). 'Gender bias in teaching evaluations', *Journal of the European Economic Association*, vol. 17(2), pp. 535–66.
Nittrouer, C.L., Hebl, M.R., Ashburn-Nardo, L., Trump-Steele, R.C., Lane, D.M. and Valian, V. (2018). 'Gender disparities in colloquium speakers at top universities', *Proceedings of the National Academy of Sciences*, vol. 151(1), pp. 104–8.
Sarsons, H. (2017). 'Recognition for group work: Gender differences in academia', *American Economic Review*, vol. 107(5), pp. 141–5.
Schmader, T., Whitehead, J. and Wysocki, V.H. (2007). 'A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants', *Sex Roles*, vol. 57(7–8), pp. 509–14.

Shao, Y., Taylor, S., Marshall, N., Morioka, C. and Zeng-Treitler, Q. (2018). 'Clinical text classification with word embedding features vs. bag-of-words features', in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2874–8, Piscataway, NJ: IEEE Press.
Trix, F. and Psenka, C. (2003). 'Exploring the color of glass: Letters of recommendation for female and male faculty', *Discourse and Society*, vol. 14, pp. 191–220.
Valian, V. (1999). *Why So Slow? The Advancement of Women*, Cambridge, MA: MIT Press.
Valian, V. (2005). 'Beyond gender schemas: Improving the advancement of women in academia', *Hypatia*, vol. 20(3), pp. 198–213.
Wu, A.H. (2018). 'Gendered language on the economics job market rumors forum', *American Economic Review Papers & Proceedings*, vol. 108(5), pp. 175–9.