DISCUSSION PAPER SERIES

DP17474

Women in economics: the role of gendered references at entry in the profession

Audinga Baltrunaite, Alessandra Casarico and Lucia Rizzica

ORGANIZATIONAL ECONOMICS

PUBLIC ECONOMICS

ISSN 0265-8003

Women in economics: the role of gendered references at entry in the profession

Audinga Baltrunaite, Alessandra Casarico and Lucia Rizzica

Discussion Paper DP17474
Published 15 July 2022
Submitted 08 July 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Organizational Economics
- Public Economics

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Women in economics: the role of gendered references at entry in the profession

Abstract

We study the presence and the extent of gender differences in reference letters for graduate students in economics and how these may affect the start of young researchers' careers. To these ends, we build a novel rich dataset covering ten cohorts of academic job market applicants to two top institutions hiring on the international market. We collect information from the application packages and conduct text analysis of reference letters using Natural Language Processing (NLP) techniques in order to measure gender differences in the style and content of the letters. We then combine the resulting measures with information on the applicants' subsequent labor market outcomes as extrapolated from the main online repositories. Our results reveal that male and female candidates receive different support from their sponsors and are described in systematically different terms. While female advisors talk more about personal characteristics, only male advisors do so at a different extent for male and female candidates. Such differences in how candidates are talked about affect subsequent career outcomes and explain a non-negligible part (5 to 8% approximately) of the observed gender gaps.

JEL Classification: I23, J16, J44

Keywords: Gender bias, research institutions, professional labor markets

Audinga Baltrunaite - audinga@gmail.com
Bank of Italy and CEPR

Alessandra Casarico - alessandra.casarico@unibocconi.it
Bocconi University, CESIfo and Dondena

Lucia Rizzica - lucia.rizzica@bancaditalia.it
Bank of Italy

Women in economics: the role of gendered references at entry in the profession[*]

Audinga Baltrunaite Alessandra Casarico Lucia Rizzica[†]

July 8, 2022

Abstract

We study the presence and the extent of gender differences in reference letters for graduate students in economics and how these may affect the start of young researchers' careers. To these ends, we build a novel rich dataset covering ten cohorts of academic job market applicants to two top institutions hiring on the international market. We collect information from the application packages and conduct text analysis of reference letters using Natural Language Processing (NLP) techniques in order to measure gender differences in the style and content of the letters. We then combine the resulting measures with information on the applicants' subsequent labor market outcomes as extrapolated from the main online repositories. Our results reveal that male and female candidates receive different support from their sponsors and are described in systematically different terms. While female advisors talk more about personal characteristics, only male advisors do so at a different extent for male and female candidates. Such differences in how candidates are talked about affect subsequent career outcomes and explain a non-negligible part (5 to 8% approximately) of the observed gender gaps.

JEL Classification: I23, J16, J44

Keywords: gender bias; research institutions; professional labor markets

[*] We thank Francesco Billari, Claudia Ciccodicola, Mariateresa Elia, Francesca Felice, Sabina Marchetti, Riccardo Russo, Giovanni Salvucci, Nicola Scalzo and Ivan Triglia for their invaluable help with data acquisition and analysis. In particular, we are indebted to Jonathan Rossi, Julia Milis and Luigi Roncoroni for excellent research assistance and grateful to Luca Bellodi, Michela Carlana, Marta De Philippis, Silvia Giacomelli, Dirk Hovy, Eliana La Ferrara, Francesco Principe, Pilar Sorribas Navarro and seminar participants at Bank of Italy, Bocconi University, University College London, Tinbergen Institute, Free University Berlin, COSME Gender Economics Workshop, and Royal Economic Society conference for insightful comments. *The views expressed in this paper are those of the authors and do not involve the responsibility of the Bank of Italy and of Bocconi University. The usual disclaimers apply.*

[†] Baltrunaite: Bank of Italy and CEPR (email: audinga.baltrunaite@bancaditalia.it); Casarico: Bocconi University, CESIfo and Dondena (email: alessandra.casarico@unibocconi.it); Rizzica: Bank of Italy (email: lucia.rizzica@bancaditalia.it)

1 Introduction

The under-representation of women in academic ranks is a widespread phenomenon and it has been barely improving over time, especially in some fields of study. As discussed by Bayer and Rouse (2016) and Lundberg and Stearns (2019), within the field of economics, women are a minority starting from the undergraduate level, and this gap widens when looking into the higher ranks of academia. Such gender imbalance "likely hampers the discipline, constraining the range of issues addressed and limiting the ability to understand familiar issues from new and innovative perspectives" (Bayer and Rouse (2016), p.221).

The leaky pipeline phenomenon is by no means limited to economics, as testified by the evidence collected in the work of the European Commission (2019), and it is particularly severe in STEM disciplines (Kim and Moser, 2021). Research on the causes of the leaky pipeline in academia, as well as of the low female presence in key and influential institutions has developed a lot in the last decade, prompted by an increased awareness of the costs for society of such under-representation. Lundberg (2020) takes stock of the research on women in economics and provides a comprehensive overview of the available evidence and explanations on why women are still a minority in the field, following all the stages of the career. The research reviewed in the volume highlights several sensitive stages of the career: major choice at the undergraduate level, entry and performance in graduate school, publication records, maternity.

In this paper we focus on the transition from post-graduate program to work, and contribute to the literature by delving into the role of implicit attitudes held by senior academics in feeding the under-representation of women in economics by influencing the early stages of the career. Specifically, we examine whether male and female PhD students receive equal support from their advisors in terms of reference letters, conditional on observed student quality. We focus on the content of the letters that PhD candidates receive from their advisors when applying to the junior economics job market and, by combining modern text analysis tools with insights from the psychology literature, explore whether such letters reveal implicit gender stereotypes in how candidates are talked about and who holds such stereotypes. Finally, we estimate the relative contribution of candidate, advisor, letter characteristics and implicit biases the letters may contain in explaining gender differences in early career outcomes.

In order to conduct the analysis, we collect data from two large institutions recruiting internationally on the academic job market for junior economists. They are both based in Italy. We gather information on ten years of applications, for a total of about 8,000 applications and 25,000 reference letters. We recover detailed information on applicants from their application packages and conduct text analysis on their reference letters employing Natural Language Processing (NLP) tools (specifically, word embeddings). Finally, we map candidates to their current position and publication records using massive web-scraping techniques on several publicly accessible websites.

We find that male and female candidates get to the job market with significantly different support. First, female candidates are more likely to be matched with female advisors, who are generally more junior and less established in the academic community. Second, they receive fewer letters, both because they have fewer sponsors and because there is a higher incidence of advisors not submitting the letter when they are supposed to. Also, letters for female candidates are on average shorter. Third, the content of the letters is consistently different: letters written for female candidates tend to stress more their hardworkingness and diligence, rather than their brilliance and smartness. Fourth, while female advisors generally put more attention to the candidate's personality traits in their letters, they do it to the same extent for male and female candidates, whereas male advisors tend to describe male and female candidates differently. Regressing the candidate position on the academic job ladder and the rank of their current placement institution, and some other early career success proxies - number and quality of publications and citations - on the characteristics of reference letters, we find that how candidates are talked about explains a non negligible part of the lower success of female PhDs in the economics profession.

Our findings are relevant from different standpoints. While outright discrimination is harder to go undetected in the present day compared to the past, implicit bias may be persistent and difficult to capture. Using word embeddings, we aim to provide evidence on the presence of such implicit stereotypes in a natural, rather than experimental (Carlana, 2019), setting. To understand the contribution of biases in how candidates are described to gendered outcomes early in the career, we then incorporate such measures in a standard regression framework. Our analysis eventually advances our knowledge on the roots of the under-representation of women in academic and institutional ranks by opening up the black-box of gendered mentorship configuration and by studying how language used in the reference process can vehicle implicit biases. In particular, we are going to shed light on the (potential) presence of “institutional discrimination” – i.e., when the rules, practices, or non-conscious understandings of appropriate conduct systematically advantage or disadvantage members of particular groups (Haney-Lopez, 2000) – in the academic job market process and, more in general, in all referral based career mechanisms.

The remainder of the paper is structured as follows: in Section 2 we review the literature that is most closely related to our work; in Section 3 we introduce our data collection process and provide some descriptive statistics; Section 4 is dedicated to the text analysis of the reference letters; in Section 5 we estimate gender gaps in early career outcomes and how these are affected by differences in the job market package; in Section 6 we provide some robustness checks. Section 7 concludes.

2 Related literature

This paper stands at the junction between two fields of study: on the one hand, the economics literature that has started digging into the roots of the observed gender imbalance in the profession; on the other hand, a more established literature in applied psychology that aims to pin down the presence and magnitude of stereotypes and implicit (gender) bias in the labor market. Drawing from the latter to qualify the relevant stereotypes, we employ the tools of modern text analysis to quantify whether such stereotypes appear in a large corpus of reference letters and how they affect women’s careers.

Within the field of economics, the literature has extensively documented the gender divide in academia (Bayer and Rouse, 2016; Janys, 2020). In US top departments women represent just 15% of full and 27% of associate professors (Chevalier, 2022). The corresponding figures for Europe are 20% and 32%, respectively (Auriol et al., 2020).

Such imbalances appear very early in the career. According to Lundberg (2020), in the US the share of women among assistant professors in top departments has been stalling over the last decade, while that of women in more senior positions has been rising slowly.

Focusing on the graduate level stage, some factors that are positively correlated with female PhD success include hiring and retaining female faculty, requiring student work-in-progress seminars, a more supportive seminar culture, and general awareness of gender bias issues (Boustan et al., 2020). The gender mix of peers in doctoral programs is also important:

a higher fraction of women in entering PhD cohorts would reduce the gender gap in program attrition, with the effect driven almost entirely by differences in the probability of dropping out in the first year of the program (Bostwick and Weinberg, 2020).

The matching of students and advisors by gender and how such pairing affects job market outcomes are analyzed in Neumark and Gardecki (1998) and Hilmer and Hilmer (2007). The first survey several cohorts of graduate students from institutions granting PhDs in economics from the mid 1970s until the early 1990s and find no link between gendered student-advisor matching and rank of institution of placement. They do find evidence, though, that female students complete their PhD more often and more quickly when paired with a female advisor. Hilmer and Hilmer (2007), instead, focus on 1,900 individuals receiving economics PhDs from the top-30 Economics programs between 1990 and 1994 and examine the differential impact of each of the four possible mentorship configurations (female student–female advisor, female student–male advisor, male student–female advisor, and male student–male advisor) on both initial job placements and early-career research productivity, finding that the female-female pairing is worse than the male-male one, but no worse than the female-male.[1] Finally, more recently, Pezzoni et al. (2016) approach the same question and look at the impact of gender pairing of advisors and their students on research performance of Caltech students during graduate studies. Their evidence suggests that both male and female students publish more when paired with female advisors.

Another important factor affecting PhDs' placement is the field of specialization. Fortin et al. (2021) show that the gap in the likelihood of obtaining an assistant professor position in an institution outside the Top-50 can be fully explained by differences in the field of specialization; in the Top-50 departments, instead, the institution granting the PhD is the most powerful predictor. Similarly, looking directly at earnings, Oaxaca and Sierminska (2021) conclude that 14 percent of either sex academics would have to change specialization in order to achieve complete salary parity across genders.

Some very recent literature, however, has started highlighting the existence of non observable obstacles and implicit discrimination in the field of economics. Paredes et al. (2020),

[1] Focusing on chemistry – Gaulé and Piacentini (2018) find that students working with advisors of the same gender tend to be more productive during the PhD, and that female students working with female advisors are considerably more likely to become faculty themselves. Hence, they argue that the under-representation of women in science and engineering faculty positions may perpetuate itself through the lower availability of same-gender advisors for female students.

for instance, provide evidence that implicit and explicit gender stereotypes are well present in economics from the undergraduate level, with students turning out to be more gender biased than those in other fields and with the gap increasing over the course of studies. Looking directly at faculty members and exploiting the introduction of blind grading of exams in Economics at Stockholm University, Jansson and Tyrefors (2020) show that teachers tend to give higher grades to male students.[2]

Stark gender differences further appear in the process of publication of scientific work. Sarsons (2017) and Sarsons et al. (2021) show that women obtain less recognition for their contribution in coauthored research when collaborating with men, i.e., that coauthored papers affect the probability of being granted tenure less for female economists than for male economists. Similarly, Hengel (2021) employs several NLP tools to show that female authored papers are held to higher writing standards by editors and referees, so that women need to put significantly more effort for publishing their work, and hence are eventually less productive. Koffi (2021a,b) shows that female-authored papers published in the economics top five journals are significantly less likely to be cited than those written by men, even when they are equally closely related to the research considered. However, Card et al. (2020) document no gender disparities along the whole publication process in top economic journals.

Finally, some recent papers have shown how male economists' attitudes expressed in public may be further detrimental to their female colleagues. Dupas et al. (2021) analyze interactions during seminars in economics departments to find that female presenters are systematically asked more and harsher questions by male audience. Sarsons and Xu (2021) survey male and female economists from top departments and show that men are systematically more self-confident than women, providing strong personal judgments even when asked questions on the economy, which are further away from their field of expertise.

Our paper contributes to shedding light on the channels through which gender gaps are generated at the early stages of the economics academic profession by studying the extent and impact of implicit gender stereotypes held by senior faculty. In this respect, we borrow from a literature coming from the fields of psychology and linguistics, which has studied the

[2] Evidence on teachers' gender stereotypes is provided by Carlana (2019), for primary schools, and by Bleemer (2019) at the undergraduate level. The latter, in particular, focuses on the degree of "genderedness" of students' evaluations written by UC Santa Cruz professors and estimates the impact of such trait on the subsequent major choices by student.

presence of gender stereotyping in reference letters.

Trix and Psenka (2003) analyze a corpus of about 300 letters of recommendation for medical faculty at a large American medical school and find that letters written for female applicants differ systematically from those written for male applicants in the extremes of length, in the percentages without basic features, and in the percentages with doubt raisers. Dutt et al. (2016) focus on geoscience and examine the relationship between applicant gender and two outcomes: letter length and letter tone. They show that female applicants are only half as likely to receive excellent letters versus good letters compared to male applicants. In addition, male and female recommenders differ in their likelihood to write stronger letters for male applicants over female applicants.

Some works, then, investigate the content of reference letters in various contexts, using pre-defined semantic classifications. Schmader et al. (2007) employ a text analysis software to examine a corpus of reference letters written for applicants for either a chemistry or biochemistry faculty position at a large U.S. research university. Their findings, though based on a fairly limited sample - 886 letters of recommendation written on behalf of 235 male and 42 female candidates - reveal that recommenders tend to use significantly more standout terms to describe male as compared to female candidates. Letters containing more standout words also include fewer grindstone words.[3] In a similar spirit, Madera et al. (2009) analyze letters written on behalf of applicants for faculty positions in a psychology department, searching for descriptions of candidates that reflect a social role theory of sex differences.[4] The authors find that women are indeed described as more communal and less agentic than men, and that communal characteristics negatively affect the hiring decisions. The latter finding is based on judgments of hireability made by psychologists, rather than on the observation of

[3] Standout words include those referring to the exceptional characteristics of the person or item described. These include, for example, "outstanding", "exceptional", "unique", etc. Grindstone words instead refer to the effort a person exerts in her work. These include for example "hardworking", "tenacious", "work ethic".

[4] According to social role theory (Eagly et al., 2000), behavioral sex differences arise from the division of labor—the differential social roles inhabited by women and men. Historically, men have been more likely to engage in tasks that require speed, strength, and the ability to be away from home for expanded periods of time, whereas women were more likely to stay home and engage in family tasks, such as child rearing. Accordingly, men are perceived and expected to be agentic, and women are perceived and expected to be communal. Agency includes descriptions of aggressiveness, assertiveness, independence, and self-confidence (Eagly and Johannesen-Schmidt, 2001). Agentic behaviors at work include speaking assertively, influencing others, and initiating tasks. Communal behaviors at work include being concerned with the welfare of others (i.e., descriptions of kindness, sympathy, sensitivity, and nurturance), helping others, accepting others' direction, and maintaining relationships (Eagly and Johannesen-Schmidt, 2001).

actual hiring outcomes. Finally, Chapman et al. (2020) carry out a comprehensive study of letters of recommendation for a pool of Radiation Oncology Residency Applicants. Similarly to the previously mentioned studies, they use a dictionary of predetermined themes (LIWC) including standout, grindstone, agentic, communal and also other personality traits. While they do not detect significant differences depending on the gender of the applicant, they document significant linguistic differences related to the gender and other characteristics of the letter-writer, with a general tendency to use a male-biased language.[5]

These two disciplinary strands of the literature have evolved separately: there is no contribution jointly investigating the extent of the bias of sponsors or advisors and the impact this has on real labor market outcomes. A recent partial exception is Eberhardt et al. (2022), which studies the extent of use of gendered language in reference letters for job market candidates applying to a UK institution. However, they provide no analysis of the influence of gendered language on labor market outcomes. Our paper aims to fill this gap by evaluating the extent of bias in reference letters and estimating the contribution that such letters' characteristics give to explaining candidates' subsequent professional outcomes in a set up in which rich student and advisor characteristics, including field of study and proxies for the quality of the candidate, are controlled for. While our analysis is guided by the evidence provided in the psychology literature mentioned above, we advance on these studies taking a massive data analysis approach: we examine around 25,000 letters using modern tools of text analysis and then incorporate them in a novel and rich dataset, which covers graduate students' and advisors' characteristics, so as to estimate comprehensive regression models.

3 Data and descriptive evidence

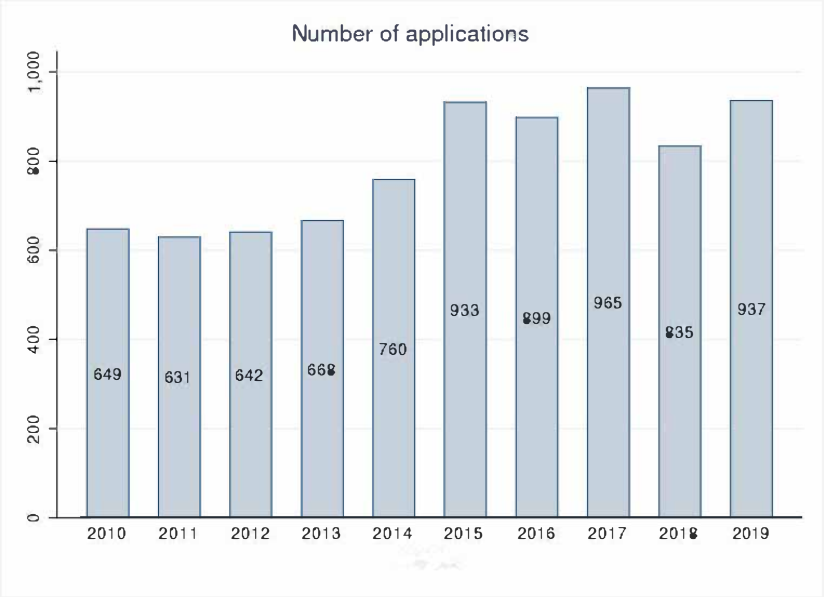
3.1 Data sources

Our work draws from a novel unique dataset that we built for the project. Specifically, we have access to the full package of applications received by two leading institutions recruiting on the international economics job market for positions in Italy. We were granted access to the data under strict confidentiality agreements.

[5] Language is evaluated for gender bias using a publicly available gender bias calculator, available here.

The data cover ten cohorts of applicants for one institution – which features two departments (Economics and Finance) – and five cohorts for the other. Overall, we have data for almost 8,000 applications.[6] Figure 1 shows the distribution of the applications in our sample across years.

Figure 1: Number of applications by year of application.



Source: Number of applications in the sample, by year. Years 2015-2019 include the two institutions, whereas years 2010-2014 only one.

For each candidate we collect information at the time of application available in their CVs and application forms. These allow us to recover the institution in which they obtained their PhD, the main fields of interest[7] and some demographic and career information. We infer the gender of the applicant through gender name dictionaries and, in some residual cases, through manual checking. We then match the institution awarding the PhD with the (yearly) QS world university rankings and with the (2021) Repec ranking of Economics departments to obtain a proxy of PhD quality.

Each candidate's application package also contains the identities of their advisors who are to send their reference letters. Candidates indicate from two to five letter writers, for a

[6] Some applications are repeated across institutions or departments. See Section 3.3 for how we deal with these cases.

[7] This was provided by the candidate in an open-ended question. We thus categorized the answers into JEL codes.

total of 25,778 references. For each reference we can classify the gender of the letter writer and her main affiliation. The actual number of letters in the application package sometimes is lower than the number indicated by the candidate in the application form. This happens when sponsors do not send their reference letter (in time) to the institution.

Finally, we collect information on candidates' labor market outcomes through massive web-scraping of three publicly available websites: Repec, Google Scholar and Linkedin. The first two allow us to collect comprehensive information on candidates who pursued a career in academia and research. Specifically, we retrieve the number and full list of publications, coauthors, the number of citations and the main current affiliation. The Linkedin platform, instead, allows us to obtain information also on those candidates who pursued a non-research career or have not published any work yet. All in all, the combination of these three sources allows us to identify the current placement of 94% of the candidates. As we do for the institution granting the PhD, for those in academia, we further match their affiliation with measures of academic ranking taken from both QS and Repec, to obtain a proxy of their success on the job market.

3.2 Descriptive statistics

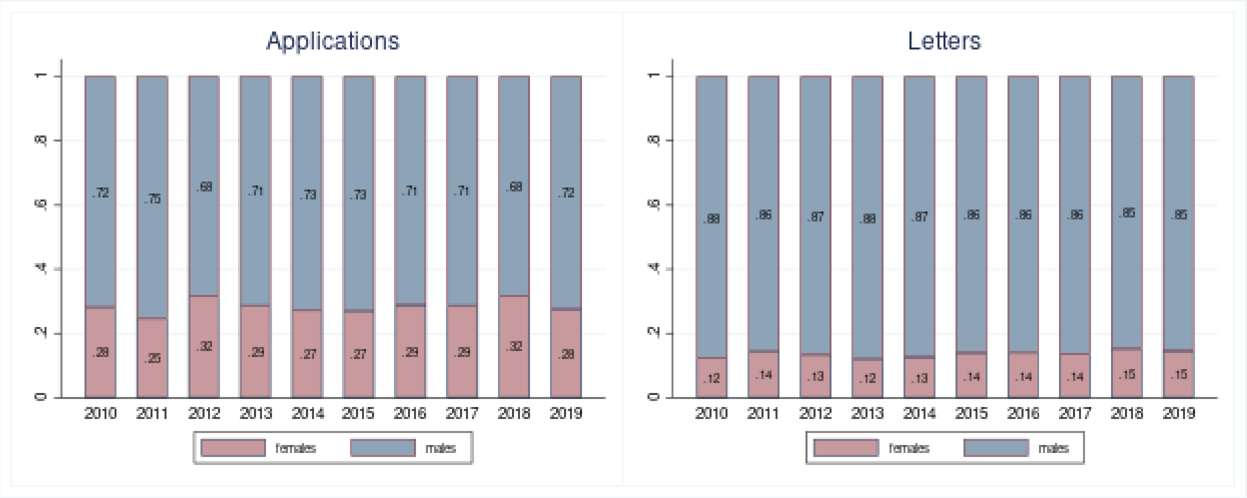
We now present descriptive statistics on the sample of candidates and letter writers.

As Figure 2 shows, less than one third of applications come from female candidates, a share that has remained constant in the ten years considered (left panel). The share of letters written by female sponsors is significantly lower, equal at most to 15% (right panel) and barely rising over the period.

Table 1 summarizes the main characteristics of the job market candidates in our sample and examines the gender differences in their observable characteristics (that may proxy, at least in part, for candidate quality) at the onset of their job market search, in their references and in their labor market outcomes after the job market search. Appendix Table B.1 provides similar descriptive statistics for the sample of European Job market candidates that subscribed to the European Economic Association Candidate Directory for the year 2020/2021. Figures are very similar to those of our sample, thus reassuring us on the external validity of the results of this paper.

Around a half of all candidates apply with a PhD from an institution in the United

Figure 2: Applications and reference letters by gender and year of application.



Notes: Years 2015-2019 include the two institutions, whereas years 2010-2014 only one.

States or Canada, with more males than females (53% vs. 48%) having studied in North America.[8] The opposite is true for European PhDs (43% vs. 48%). With respect to the location of positions advertised by the two institutions, we observe that the pool of applicants is largely international: only 7% of all applicants receive their degree from an institution in Italy, on average. "Domestic" PhD is more common among female candidates: one out of ten women hold or are expected to hold a PhD from an Italian institution. Overall, thus, female candidates tend to come from geographically closer institutions, perhaps signaling their lower willingness to relocate during the job market.[9]

There are significant gender differences in terms of field of specialization of PhD candidates. Female PhD candidates are 10 percentage points more likely to specialize in applied microeconomics research than their male peers, who instead tend to choose topics in macroeconomics, finance, theory or quantitative methods more often. Figure 2 illustrates these differences more in detail, by focusing on 14 categories based on the main JEL codes. Gender differences are mostly driven by macroeconomics or mathematics and quantitative methods

[8] This may be different from the institution of affiliation at the time of the job market application for (a modest fraction of) candidates applying after the conferral of the degree, e.g., the ones applying from a post-doctoral program.

[9] More broadly, this is in line with the literature showing that women have a lower propensity to move away from home for work or study (Rizzica, 2013).

Table 1: Descriptive statistics of job market candidates

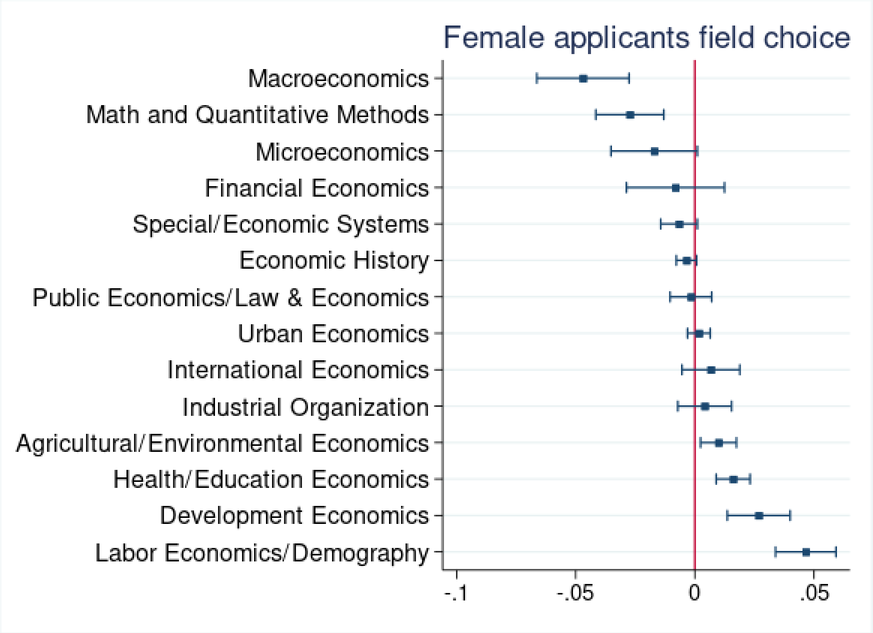
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Male		Female		Difference		
	mean	sd	mean	sd	N	Diff	T-stat
Pre-JM:							
EU PhD	0.43	0.50	0.48	0.50	7077	-0.044***	(-3.329)
Italian PhD	0.07	0.25	0.10	0.30	7077	-0.033***	(-4.454)
Applied micro	0.24	0.43	0.34	0.48	7077	-0.100***	(-8.229)
Macro/International/Finance	0.44	0.50	0.40	0.49	7077	0.048***	(3.750)
Theory/Quantitative	0.24	0.43	0.20	0.40	7077	0.044***	(4.151)
Top-20 QS (general)	0.17	0.38	0.15	0.36	7063	0.020**	(2.143)
Top-20 Repec Econ	0.27	0.44	0.21	0.41	7063	0.052***	(4.725)
Phd ranking Repec Econ	108.03	297.78	112.84	106.91	7063	-4.806	(-0.997)
# Publication pre-JM	0.72	2.17	0.53	1.74	7077	0.186***	(3.784)
References:							
# Letter writers	3.25	0.78	3.21	0.84	7077	0.038*	(1.761)
# Letters uploaded	2.70	1.30	2.63	1.36	7077	0.079**	(2.250)
# Female letter writers	0.39	0.61	0.58	0.74	7077	-0.190***	(-10.320)
Main advisor female	0.11	0.31	0.17	0.37	6913	-0.055***	(-5.833)
Average letter length	1029.74	382.08	992.53	367.73	6028	37.210***	(3.499)
Post-JM:							
Academic placement Linkedin	0.75	0.43	0.75	0.43	6641	0.007	(0.597)
Placement ranking QS (band, general)	5.61	2.30	5.60	2.34	6641	0.014	(0.227)
Placement Top 20 QS (general)	0.07	0.25	0.07	0.26	6641	-0.006	(-0.898)
Placement Top 20 Repec Econ	0.08	0.26	0.08	0.27	6641	-0.006	(-0.807)
Associate professor	0.17	0.37	0.12	0.32	6641	0.050***	(5.498)
Assistant professor	0.46	0.50	0.50	0.50	6641	-0.036***	(-2.667)
PostDoc	0.12	0.32	0.13	0.34	6641	-0.014	(-1.495)
# Publications	2.37	5.49	1.54	3.66	7077	0.838***	(7.477)
Top 5 econ or Top 3 fin publication	0.08	0.28	0.06	0.23	7077	0.030***	(4.642)
# Citations (Repec)	41.07	147.87	26.18	78.62	7077	14.884***	(5.482)
Observations	5041		2036		7077		

Notes: All post-job market variables refer to 2021. * denotes significance at 10%, ** significance at 5% and *** significance at 1%.

(more often chosen by male candidates) and labor economics, demography or development economics (more often chosen by female candidates). Interestingly, there are no pronounced gender differences in financial or international economics, or theory.

In terms of the quality of the PhD granting institution, the pool of male applicants appears to be better selected: they more often come from a top-20 institution according to either QS or Repec rankings. Finally, male applicants' packages are stronger also in terms of

Figure 3: Differences in gender distribution across fields.



Notes. Each point is the difference between female and male candidates in the likelihood of declaring each field as main field (i.e., it is the coefficient associated to a female dummy variable in a linear regression with field dummies as outcomes).

publications they report on their CVs at the time of application: over 70% of male candidates have at least one publication of some kind, while this is the case for only 53% of females.

Significant gender differences are also visible in the application package of candidates. Male candidates have a slightly higher number of academic references, both in terms of designated referees and of actual letters uploaded. Furthermore, there is evidence of assortative matching between students and advisors based on gender: female candidates have a higher number of female letter writers. Moreover, almost 17% of female candidates have a woman as their main (i.e. first) letter writer, while this is the case for 11% of male candidates. Letters are also different in the way they are written. References written for male applicants are longer in terms of the number of words by around half a paragraph.

We then shed light on some job market outcomes. First, there is no evidence of gender differences in terms of obtaining an academic placement, based on information retrieved from Linkedin: about 75% of applicants of both genders ended up in an academic position. Among them, we find that the quality of their job market placement (at the time the online repositories were scraped) does not differ across candidates of different gender: female candidates are more likely than male candidates to end up in a top ranked institution, but this difference is not statistically significant. However, a large gender gap emerges when we consider the position on the academic ladder. Indeed, male scholars are more than 50% more likely to hold an Associate Professor position, while they are less likely to be Assistant professors or Postdocs.[10]

Last, we consider the research output of our pool of candidates. With all the caveats that arise from the literature that we discussed in section 2, these figures suggest that male candidates are more productive during the first years of their academic career: they have almost one publication more than women, are more likely to publish in one of the top-5 journals in economics or top-3 journals in finance, and their research is more often cited.

We next turn to presenting some descriptive statistics regarding the letter writers in our sample. In particular, Table 2 highlights that the pool of advisors is extremely gender-unbalanced, as only 1,449 out of 8,484 referees are women. The fraction of "ghost" referees who happen to be indicated by a candidate (or some candidates), but never upload their letter(s), is larger among women, potentially suggesting their marginal importance in the students' portfolio. Next, although on average female advisors write fewer letters compared to their male counterparts, their letters are longer. In line with assortative matching by gender among advisors and students, illustrated above, female sponsors more often tend to work with at least one female PhD student.

Female letter writers appear to lag behind male advisors in terms of their research output and career achievement. They are generally more junior, both in terms of career length (the average first year of publication is significantly more recent than that of men) and in terms of academic ranks in that they are less likely to hold a full professors status, consistently also with the leaky pipeline phenomenon in economics. Moreover, they have fewer publications in top journals in economics, with nearly half as many citations compared to men.

[10] The measures of the job market placement refer to the placement as indicated in our online sources at the time we scraped the web, i.e., between October 2020 and April 2021. We are currently working on retrieving the placement history for candidates in our sample in order to obtain a more precise indication of their first placement.

Table 2: Descriptive statistics of letter writers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Male letter writer		Female letter writer		Difference		
	mean	sd	mean	sd	Obs	Diff	T-stat
Never uploaded	0.13	0.34	0.16	0.37	8464	-0.029***	(-2.793)
# Letters written	2.66	3.67	1.99	2.83	8464	0.670***	(7.755)
Av. letter length (words)	931.20	467.20	941.48	452.63	7238	-10.281	(-0.715)
At least 1 female advisee	0.45	0.50	0.52	0.50	7238	-0.065***	(-4.130)
Academic affiliation	0.78	0.42	0.74	0.44	8464	0.034***	(2.700)
Full professor	0.24	0.43	0.19	0.39	8464	0.055***	(4.777)
First publication year	1993.99	12.73	1997.97	10.93	6683	-3.979***	(-10.729)
# articles Repec	19.90	30.45	11.33	17.94	8464	8.567***	(14.395)
# Publications GS	70.80	128.88	46.53	74.58	8464	24.268***	(9.741)
# Top 5 publications	2.24	5.04	1.11	2.51	7860	1.124***	(12.129)
At least 1 Top 5 publication	0.42	0.49	0.34	0.47	7860	0.082***	(5.691)
# Citations	1006.27	2646.35	533.30	1472.93	8464	472.970***	(9.468)
Observations	7015		1449		8464		

Notes: * denotes significance at 10%, ** significance at 5% and *** significance at 1%.

3.3 Corpus construction and pre-processing

Starting from our sample of 25,778 (potential) references, we exclude those cases in which the referee did not upload the letter even if she was supposed to, and those cases in which the letter provided was in Italian.[11] This leaves us with 21,533 letters.

As our sample includes two different institutions (one of them having two departments) each year, there are cases in which the same candidate applied to more than one and turned in the same reference letter. In the text analysis, we drop 842 application packages with such duplicate letters, keeping the package with more available letters.

We then anonymize our texts, replacing each reference to the candidate in all letters with the tokens “*candidate_maleID*” for male and “*candidate_femaleID*” for female candidates, *ID* being the individual identifier of each candidate. Moreover, we replace all personal pronouns (eg. him, his, her, etc) in the text with such tokens to identify the majority of instances in which the letter refers to the candidate.

We proceed with a standard pre-processing of the text. We first strip off the header

[11] These are less than 100.

and the footer of each letter, since they typically include emails, addresses and affiliation information of the reference letter writer, with no reference to the candidate herself. Next, we convert the text into lower-case characters, split contractions and remove double spaces, punctuation, numbers and stopwords.[12] We also replaced several bi-grams with a single token to simplify the analysis (e.g,. "job market" was replaced by "job-market", "interest rate' by "interest-rate"). All these steps can help a statistical model to only learn from terms that have a relevant meaning, reducing the dimensionality of our corpus.

After having cleaned the text of each letter, we transform it in a list of "tokens", i.e., words or n-grams. Each occurrence of a given term in the data is a token, while unique tokens are called "types". In other words, tokens may be repeated within each document, whereas types are unique. The full list of types in the corpus of documents is called vocabulary (V). The tokenization allows us to convert texts into lists of units of meaning (words or n-grams) that can be used to map the document to numbers.

In order to reduce the dimensionality of our list of tokens without losing information, we proceed with the *lemmatization* of the text. This process replaces each token with its dictionary base form or lemma. Indeed, all the other forms do not change the core meaning of the lemma but add some information, e.g. temporal, or are required by the syntax. As we are more interested in the meaning rather than in the morphology or syntax, it is useful to replace words with their lemmas and thus further reduce the data dimensionality. All these pre-processing steps allow us to reduce the average length of our documents from 1,029 words in the full letter text, to 988 words in the body text to 536 words after the full pre-processing.

3.4 Corpus description

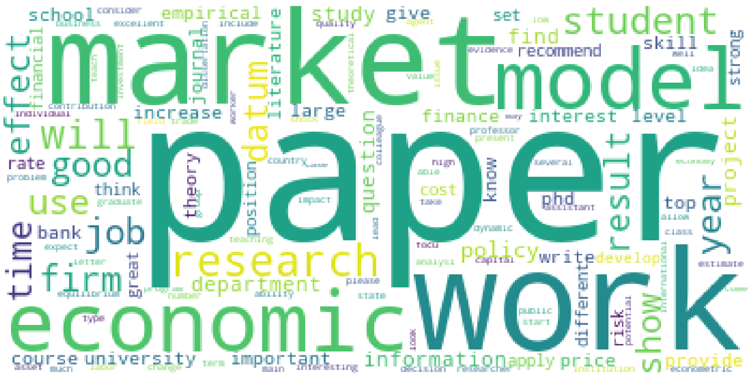
Our final corpus consists of 18,925 documents, which are combinations of 109,744 unique lemmas. The total number of words in our pre-processed documents exceeds 92 millions, it was over 119 millions before pre-processing.

Our corpus can be represented by a term frequency matrix, which has as many rows as documents ($D = 18,925$) and as many columns as lemmatized words ($V = 109,744$) in our vocabulary. Each element in the matrix will be the frequency of the word v in document d. We can use this information to provide a first graphical representation of our data. Figure

[12] These are words which do not carry any information per se but rather have some functional purpose, e.g., "the", "to", "of", etc. For a full list of the stopwords considered see Appendix A.

4 displays the most frequent lemmatized words in our corpus, i.e., in all our letters, put together using a word cloud. The bigger and bolder the word appears, the more often it is mentioned within our corpus and hence the more important it is. Unsurprisingly, in our case the most common words are "market" and "paper", as letters mainly discuss the job market paper of the candidate.

Figure 4: Word cloud for the full corpus of reference letters



Notes: Word cloud based on raw frequencies of all lemmatized words in the corpus.

In order to extrapolate more meaningful information from the letters, we can weigh the raw frequency of each lemma (its *term frequency, tf*) by the (inverse of the) number of documents it occurs in (its *document frequency, df*).[13] This measure is called $tfidf$ and, for every lemmatized word v, it is given by:

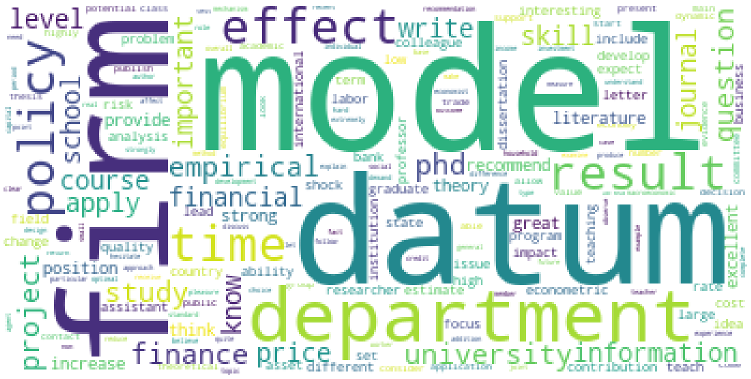
$$tfidf_v = (1 + log(tf_v)) \times (1 + log\frac{N}{df_v}) \qquad (1)$$

Such re-weighting allows us to give low scores to words that occur frequently, but in every document (e.g. function words). Similarly, words that are rare but still appear in most documents in the corpus would also get lower scores. The most prominent example in our setting would be words like "job", "market" and "paper", which were indeed the most frequent ones in Figure 4. Words that are quite frequent but occur only in a few documents get the highest score as these are the words that carry most information about the documents'

[13] This can be easily explained using a term frequency matrix. The tf would be the sum of elements in each column - i.e., the total number of occurrences of each word across all documents - the df would instead be the number of non zero entries in the same column (over the total number of documents N) - i.e., the share of documents in which the frequency of term v is strictly positive.

content. Note also that the use of *logs* dampens the effects of the re-weighting.[14] Figure 5 shows the word cloud that we obtain by reweighting all the words in our corpus by their *tfidf* score. The resulting image is more informative on the content of the letters, highlighting the duality between theory and empirics in the work of the candidates that is described in the letters.

Figure 5: Word cloud with *tfidf* reweighting



Notes: Word cloud based on *tfidf* of words that appear in more than 5% and in less than 75% of documents.

4 Text analysis of reference letters

4.1 Supervised Text Analysis: word embeddings

The description of the corpus provided so far gives little information on what referees say about their students in their letters. Either the use of simple term frequencies or of *Tfidf* are not suitable tools to understand how candidates are described. Indeed, these simple descriptive methods do hint at some information regarding the main content of the letters, but this essentially refers to the main topic of the candidate's research.[15]

[14] Appendix figures B.1.a and B.1.b show the distribution of the two measures in our corpus. Figure B.1.a shows that frequency of words in our corpus declines very quickly: very few words have very high frequency, while most words appear only once. This follows a power-law distribution and is generally referred to as Zipf's law. The use of *tfidf* re-weighting largely reduces the problem.

[15] Moreover, these methods are silent about how words group together into larger constructs. More complex tools that achieve this goal, such as *topic* or *cluster* analysis, are well able to capture the field of research of candidates (the results are available upon request).

Our preferred approach to dig up how candidates are talked about in their reference letters will thus be a *supervised* approach. Unlike unsupervised techniques (such as clustering or topic analysis), supervised methods are algorithms that do require some external classification of the data by the researcher in order to guide the analysis of the text.

In this case, we rely on a model with lists of “target words” that likely capture some meaningful characteristics of the candidates. To do so, we build on the literature in psychology described in Section 2. Following Schmader et al. (2007), we start from two categories utilized to describe job applicants: standout and grindstone terms. They represent, respectively, words referring to the candidates' exceptional character (e.g. outstanding, unique, and exceptional), and words referring to the effort they put in work (e.g. hardworking, conscientious). We then consider two other categories of adjectives that psychologists have identified as often carrying implicit gender stereotypes related to the social role theory. These are agentic and communal adjectives. The first ones refer to personality traits related to self-confidence, assertiveness, tenacity. The latter, in contrast, refer to personality traits that emphasize a person's ability to sympathize with others (e.g. agreeable, caring, warm). We consider lists (of variable length) for each category according to Schmader et al. (2007), Madera et al. (2009) and Chapman et al. (2020). The full lists are reported in Appendix A.[16]

Having defined such lists of target words we aim to understand how these are used in reference to candidates. To do so we transform our target words into mathematical objects (i.e., vectors) that represent their semantic meaning using *word embeddings*.[17] This approach identifies words that are most commonly used together, i.e., in a similar context, to capture their relatedness in semantic terms. This idea of semantic relatedness of context, or distributional semantics, is a concept developed in linguistics, dating back to Firth (1957) who stated that “you shall know a word by the company it keeps”. Mathematically, this translates into representing each target word as a vector in a low dimensional space, where its position and relative proximity to other words capture their semantic similarity in a way that words with similar meanings or semantically related will lie close together. Note that the dimension of such space will be lower than that of the full vocabulary.

[16] We note that some of the words in the original sources never appear in our corpus, thus they are not reported in the lists, for example, “self-starter”, “go-getter”, “endearing”, etc.

[17] Some recent contributions, namely Caliskan et al. (2017), Kozlowski et al. (2019) and Ash et al. (2020) have used word embeddings in a similar spirit to unveil cultural and gender attitudes.

Operationally, our word embedding procedure employs the `word2vec` tool and works as follows. First, we choose the two exogenous parameters to feed into the model. These are the *embedding dimension*, i.e., the dimension of the state space in which we project our text, and the *window size*, i.e., the maximum distance between each word and the target word that defines which tokens are considered. We set the two parameters at 100 and 6, respectively. Moreover, we consider, for computational convenience, only those lemmas that appear at least ten times in our corpus. Second, we estimate our word embedding model which will produce a vector of 100 dimensions for each of the target words initially identified. The algorithm we use is a skipgram model, which computes the probability of observing each context word given the target word we set. The process is iterative: starting from a random embedding, at each iteration the algorithm finds the vectors that minimize a loss function, and then starts again from these vectors. The loss function involves accounting for both the probability of observing each term within the context of the target word and for the probability of *not* observing it. Intuitively, what happens is that at each iteration the word embedding vector becomes more similar to the embeddings of words in its context and less similar to the embeddings of words not in its context. After a predetermined number of iterations (100 in our case), the vectors that minimize the loss function will be the optimal word embedding. These vectors will be our new *vocabulary*.

We measure similarity between the reference to each candidate and the word embedding with each target word by cosine distance, so that word vectors with smaller angles are considered more similar to each other. Cosine similarity can range between -1 and 1. A value of -1 means that the vectors point in diametrically opposite directions, i.e. words have opposite meaning; a value of 1 means that the two vectors point in the same direction, i.e., words are synonyms (or the same word when the vectors exactly overlap); a value of 0 means that the two vectors are orthogonal, i.e. the two words are completely unrelated. We can use cosine similarity to measure the distance between reference to a given candidate in a given text and the predefined categories of (embedded) target words.

4.2 Candidates

In our setting, we compute 42 embeddings, i.e., one for each of the 42 target words. In order to reduce the dimensionality of our vocabulary we further combine those referring to the words in the same category (standout, grindstone, communal and agentic) to obtain four *average vectors*, one for each category. Once we have transformed our lists of target words into just four mathematical objects given by the average embedding vector of the terms in each category, we compute the cosine distance between these vectors and the (embedded) vectors representing tokens for each candidate (i.e., $candidate\_maleID$ and $candidate\_femaleID$) within a specified corpus. Our first exercise considers all the letters written for a given candidate irrespective of the letter writer. This allows us to obtain a measure of how each candidate is described overall.

We calculate cosine similarity for 6,004 candidates and report them by gender in Table 3. Column 6 shows the difference between the cosine similarity between each personality trait average vector and the target token for male candidates and that for female candidates. Positive numbers mean that male candidates are more likely to be described in a given way, negative numbers that female candidates are more likely to be described that way.

Table 3: Cosine similarity between reference to candidate and target average vectors, by candidate's gender

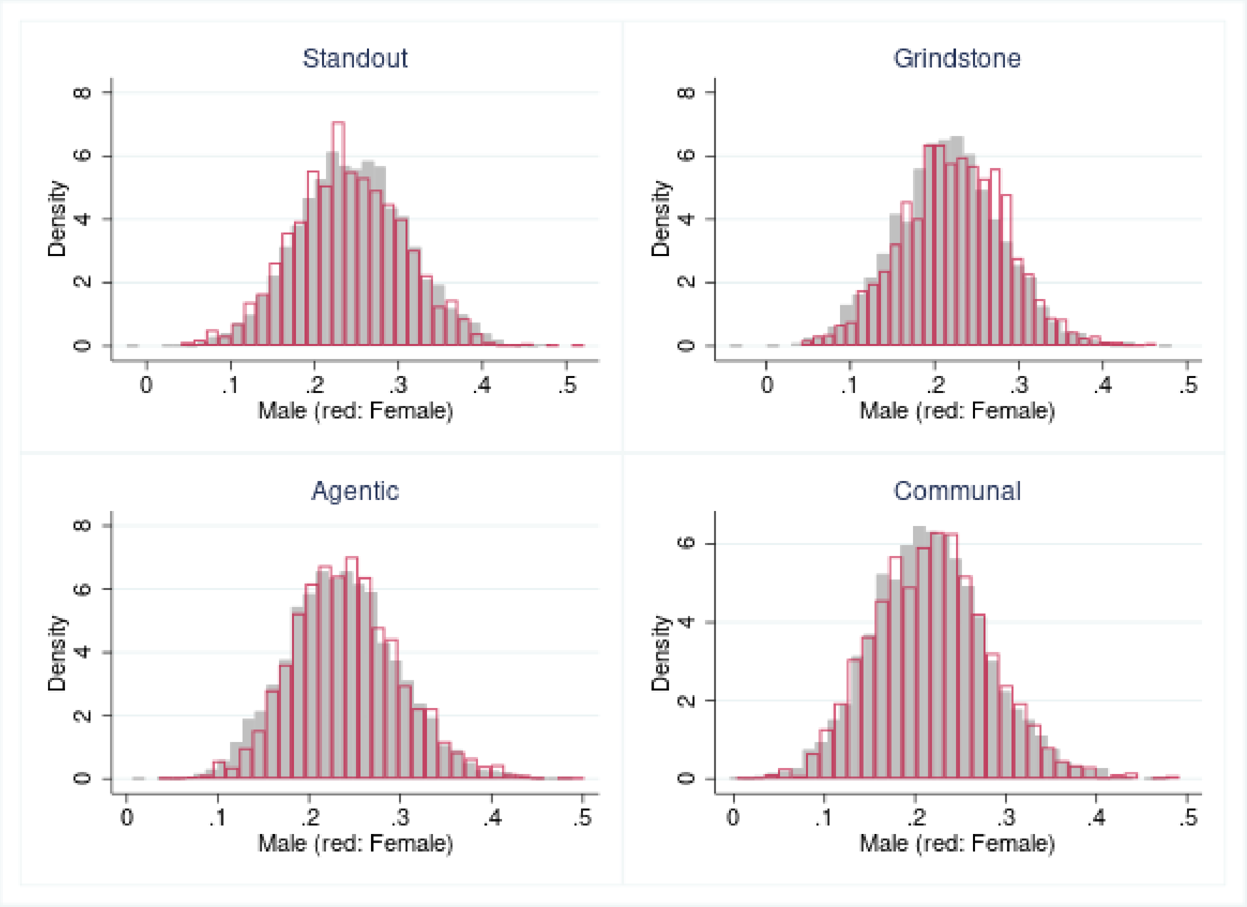
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Male		Female		Difference				
	mean	sd	mean	sd	Obs	Diff	T-stat	Diff (cond)	T-stat
Standout	0.245	0.066	0.240	0.066	6004	0.005***	(2.838)	0.005***	(3.224)
Grindstone	0.216	0.063	0.224	0.064	6004	-0.008***	(-4.538)	-0.006***	(-3.048)
Communal	0.217	0.064	0.219	0.065	6004	-0.002	(-1.109)	-0.002	(0.922)
Agentic	0.236	0.061	0.242	0.061	6004	-0.005***	(-3.066)	-0.002	(0.916)
Observations	4312		1692		6004				

Notes: The conditional differences in column 9 are computed net of candidates' PhD institution ranking band (Repec), year of application, department to which application was sent and field of research. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Our findings reveal that advisors tend to describe male students as outstanding more than they do for female students; also, they tend to stress female candidates' hard-working character more than they do for males. These results are in line with those of Schmader et al. (2007) and persist also controlling for the candidates' PhD institution ranking band (Repec), year of application, department to which application was sent and field of research (Column 8). Considering then the candidate's assertiveness (agentic traits) vis á vis her interpersonal skills (communal traits), the difference in the degree to which both interpersonal skills are stressed is very small and not significant for communal traits. Moreover, the difference in the agentic traits becomes statistically not significant once we condition on the main candidate observables.

Figure 6 further shows the distributions of cosine similarities between the average vector for each category of target words and the embedded vectors of female and male candidates. Indeed, the most significant difference is the one concerning the use of standout and grindstone terms, for which the female candidates' distribution is clearly shifted to the left or right, respectively. Modest shifts may also be noticed for agentic and communal adjectives, yet they are much smaller in magnitude. Taking these pieces of evidence together, we conclude that gender differences are most evident in standout and grindstone categories and, thus, we will focus on them in the regression analysis in Section 5.

Figure 6: Distribution of cosine similarity measures, by candidate's gender



Notes: Each histogram represents the cosine similarity between the average vector of each category of target words and the embedded vector of reference to each candidate.

4.3 Letter writers

We now ask whether female and male letter writers are more prone to describe candidates resorting to words carrying implicit gender bias. To investigate potential differences in letter style for reference letter writers of different genders, we proceed as follows. First, we classify texts depending on who wrote them rather than whom they talk about. Operationally, this entails substituting each reference to candidate with a token identifying the letter writer ($candidate\_refID$), regardless of the candidate gender. Then, we compute the cosine similarity between the average vectors for the target words belonging to the four categories analyzed (i.e., stand-out, grindstone, communal and agentic) and the embeddings corresponding to these new identifiers. Note that these will completely ignore the gender of the candidate who is described in those letters. Our approach thus differs from the candidate-level analysis, because in that case we measure the semantic similarity between the target semantic category and the reference to a single candidate irrespective of who wrote the letter, whereas in this case, between the target semantic category and references to all candidates to whom each referee has written letters for. This provides, for example, the measures of the "average" (across candidates) referee "agenticness" or "grindstoneness", i.e., how often each referee talks about both male and female candidates using agentic or grindstone words in the letters written by him or her. We can thus test for the presence of gender differences in language use between female and male letter writers.

Table 4: Cosine similarity between reference to candidate and target average vectors, by referee's gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Male referee		Female referee		Difference (uncond.)			Difference (cond.)		
	mean	sd	mean	sd	Obs	Diff	T-stat	Obs	Diff	T-stat
Standout	0.237	0.072	0.237	0.074	7097	0.001	(0.390)	6845	0.00006	(0.02)
Grindstone	0.195	0.069	0.210	0.070	7097	-0.016***	(-7.121)	6845	-0.014***	(-6.43)
Communal	0.189	0.074	0.195	0.077	7097	-0.006***	(-2.612)	6845	-0.005**	(-2.15)
Agentic	0.213	0.063	0.225	0.062	7097	-0.012***	(-5.856)	6845	-0.011***	(-5.64)
Observations	5916		1181		7097			6845		

Notes: * $p < 0.1$, ** $p < .05$, *** $p < 0.01$. The conditional difference in column 9 accounts for indicators for those with an academic affiliation, with full professorship and with at least one female advisee and for the letter writer affiliation institution fixed effects.

Table 4 reports the cosine similarity between each personality trait average vectors and references to candidates (of both gender), distinguishing between male and female letter writers. It shows that female letter writers tend to emphasize (all) candidate personality traits more, also accounting for the letter writer affiliation institution fixed effects and several observable referee characteristics. All in all, this indicates that female advisors may provide more information on personal characteristics of the candidates, beside their professional achievements.

Finally, given the evidence of differences in the way in which male and female candidates are talked about provided in Section 4.2, we check whether these patterns are also present in letters written by sponsors of different gender. One could expect that the average effect shown in Table 3 hides significant differences across letter writers' gender, given that for female candidates, sponsor-candidate matching is more often based on gender (see Table 1) and female letter writers stress personality traits more in what they write (see Table 4).

Similarly to the previous exercise, this entails computing the cosine distance between the four average vectors for the target categories (i.e. stand-out, grindstone, communal and agentic) and the vector corresponding to each new identifier which will now capture the referee identity and the candidate's gender ($candidate\_male\_refID$, $candidate\_female\_refID$). Hence, for each referee, we obtain separate measures of cosine similarity to each target semantic category for male candidates and for female candidates (as long as the referee wrote letters for both male and female candidates). In other words, we compute at most two cosine similarity measures for each referee and then compare these measures across the candidate gender in the sample of male and in the sample of female professors.

Interestingly, Table 5 shows that it is only male letter writers that talk about personality traits to a different extent when referring to male and female candidates (panel A). In particular, cosine similarity between standout words and references to male candidates is higher than that for females only when it is a male referee writing a reference letter. Similarly, cosine similarity between grindstone words and references to female candidates is higher than for male candidates only among male referees. Moreover, for those advisors who work with students of both genders, we calculate these differences holding constant the letter writer identity. The pattern we detect is similar: a given male letter writer appears to describe *his students* of different gender differently, in particular, putting more focus on grindstone characteristics when referring to female students (the difference in the standout cosine similarity is not statistically significant in this *within-letter-writer* analysis). Differences in the language use for students of different genders are not detected, instead, among female letter

Table 5: Cosine similarity between reference to candidate and target average vectors, by candidate and referee gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Male candidates		Female candidates		Diff (uncond.)			Diff (cond)		
	mean	sd	mean	sd	Obs	Diff	T-stat	Obs	Diff	T-stat
A. Only male referees										
Standout	0.231	0.072	0.227	0.072	7613	0.004**	(2.359)	3394	0.003	(1.33)
Grindstone	0.193	0.072	0.200	0.070	7613	-0.007***	(-3.876)	3394	-0.013***	(-6.91)
Communal	0.188	0.076	0.185	0.075	7613	0.003*	(1.935)	3394	-0.005***	(-2.86)
Agentic	0.210	0.067	0.213	0.065	7613	-0.002	(-1.566)	3394	-0.0104***	(-6.22)
Observations	4953		2660		7613			3394		
B. Only female referees										
Standout	0.225	0.074	0.226	0.074	1459	-0.001	(-0.216)	556	-0.004	(-0.87)
Grindstone	0.210	0.071	0.215	0.071	1459	-0.005	(-1.309)	556	-0.0002	(-0.05)
Communal	0.196	0.078	0.192	0.079	1459	0.004	(0.879)	556	0.0006	(0.13)
Agentic	0.223	0.065	0.223	0.066	1459	-0.000	(-0.099)	556	-0.005	(-1.27)
Observations	850		609		1459			556		

Notes: * $p < 0.1$, ** $p < .05$, *** $p < 0.01$. The difference in column 9 is from a specification with a letter writer fixed effects and, therefore, only exploits the variation for those who sponsor candidates of both genders.

writers. Overall, this evidence indicates that male referees are more prone to the use of gendered language, whereas female referees are not. In turn, it does not seem likely that the extent of the assortative matching based on gender between students and advisors is what drives gender differences in the emphasis given to different personality traits in references letter for the academic job market.

5 Effects on career outcomes

We now turn to examining labor market outcomes for our sample of job market candidates. In particular, we are interested in understanding the gender differences in early career achievements, both in terms of a raw gap and a conditional one that holds certain candidate, advisor and letter characteristics constant. The comparison between different estimates should illustrate the relevance of these factors in influencing gender gaps (if any are detected). One caveat is that, as explained in Section 3, our career outcomes all refer to 2021 when we retrieved the information from scraping the web. Therefore our measures capture achievements at different stages of the career depending on how long before the candidate was on the market. In order to account for such differences, we will include year of application fixed effects in all our regressions.

5.1 Career achievement

To start analyzing the relation between job market support and labor market outcomes, we point out that the task of ranking different placement along one dimension to proxy for candidate success is a rather difficult one. To start with, there are at least two dimensions of "success", even if we focus on academic placements only:[18] first, the seniority of the position, and, second, the prestige of the institution of affiliation. This is graphically illustrated in Figure 7. The first dimension measures the seniority of the academic titles, at which job market candidates of different cohorts have "arrived" by the current date. We thus build a placement index, that distinguishes across the broad categories of post-doctoral researchers, assistant professors and associate professors, and assigns the highest value of career ladder to the latter (i.e., moving vertically on the y-axis). While one can easily assume that an associate professorship is better than a postdoc position within the same institution, or that an assistant professorship in a Top-5 department is better than an assistant professorship in a lower tier institution, it is very hard to compare placements across different seniority levels and institutions. In fact, the second dimension of career success is defined by the prestige of the placement institution (i.e., moving horizontally on the x-axis).

We start our analysis by combining the two dimensions and estimate the gender difference in terms of the probability of holding an associate professor position in a Top-20 department according to Repec (whether the candidate's position falls in the upper-right box highlighted in the figure). Our pooled linear regression model will have a dummy variable for female candidates and four sets of control variables capturing, respectively, candidates' observable characteristics, letter writers' observable characteristics, some features of the letters and the candidate's characterization in the letters as obtained through word embeddings in Section 4.2.

$$\begin{split} y_i = \alpha + \beta_1 Female_i + \beta_2 Candidate\_X_i + \beta_3 LetterWriter\_X_i + \\ + \beta_4 Letters\_X_i + \beta_5 WE_i + \tau_t + \varepsilon_i \end{split} \tag{2}$$

[18] Since the regression analysis uses all job market candidates in our sample, regardless of the type of their placement, this classification implicitly assumes that non-academic placements are inferior at least in terms of prestige. This may be justified on the grounds that doctoral programs, in fact, are typically meant to train economists for academic careers.

Figure 7: Ranking over career outcomes

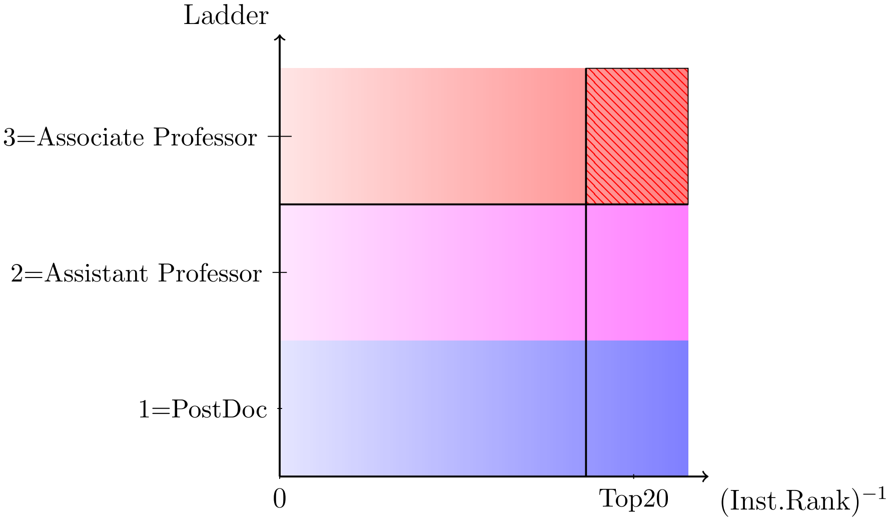


Table 6 shows point estimates from a number of simple OLS regressions. We start with a parsimonious specification, that measures the raw gender gap, and we augment it with batteries of control variables, one at a time, capturing potential determinants of a candidate's success in the job search and subsequent career. Columns 1 and 2 measure the gender gap, respectively, in the full sample of candidates and in the sub-sample of those for whom we are able to observe the full application package and only account for the application vintage (i.e., the year of application). Column 3 accounts for some candidate's characteristics that are meant to control for her strength or quality: indicators for bands of ranking of the institution from which the candidate obtained her PhD, the number of publications the candidate already had when applying to the job market, and fixed effects for each letter of JEL code. Column 4 instead includes controls for the main advisor's gender, academic title (i.e., an indicator for whether the main letter writer is a full professor), the number of top-5 articles they have published as of the beginning of 2021. Column 5 controls for some characteristics of the reference letters in the application package, namely the number of letters and their average (standardised) length. Column 6 examines the role of letter style (i.e., how the candidate's personality is described) by including our measures of distance from several predefined personality traits (Section 4.1). In particular, we control for the average cosine similarity between each reference to the candidate in the letters and the embeddings of standout and grindstone words (Schmader et al., 2007) (we will refer to them, in the interest of brevity, as letter "standout-ness" or letter "grindstone-ness" in the remainder of the paper). Finally, column 7 includes all control variables together. The bottom row indicates, for each column, the fraction of the raw gender gap which is explained by the characteristics included in each column (essentially, comparing column 2 to each of the subsequent columns).

Table 6: Career success: probability of holding an Associate Professorship in a Top 20 Institution

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.00921***	-0.00906***	-0.00683**	-0.00753**	-0.00781**	-0.00829**	-0.00563
	(0.00309)	(0.00343)	(0.00345)	(0.00341)	(0.00343)	(0.00342)	(0.00342)
# Publications Pre-JM			0.00210**				0.00192**
			(0.000982)				(0.000978)
Main lett. writer female				0.00193			0.00136
				(0.00484)			(0.00477)
# Top 5 public. (main lett. writer)				0.00150***			0.00100**
				(0.000378)			(0.000422)
Full professor (main lett. writer)				0.00768*			0.00800**
				(0.00395)			(0.00404)
# Letter writers					0.00992***		0.00780***
					(0.00282)		(0.00283)
Average letter length (std)					0.0118***		0.00939***
					(0.00217)		(0.00235)
Standout cos. sim.						0.0574*	0.0367
						(0.0295)	(0.0302)
Grindstone cos. sim.						-0.0519*	0.00238
						(0.0286)	(0.0317)
Mean dependent variable men	0.017	0.017					
% Raw Gap Explained			24.6	16.9	13.8	8.5	37.9
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.0106	0.0113	0.0264	0.0221	0.0208	0.0123	0.0372
N	6511	5699	5699	5699	5699	5699	5699

Notes: Robust Standard errors in parentheses.* $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Our results show that the coefficient in columns 1 and 2 is negative and significant, hinting that female candidates have lower career success compared to male candidates. Once candidate characteristics (as listed previously) are included in column 3, the point estimate of the female dummy decreases in size, showing that part of the gap is attributable to pre-job market differences between male and female candidates. In particular, the latter explain 25 per cent of the observed gender gap in career success. Similarly, the inclusion of advisor characteristics in Column 4 contributes to the decrease in the raw gender gap, though to a lower extent compared to candidate characteristics, as one may expect. The features of the letter package (number of letters and their length, Column 5) contribute to the reduction in the raw gender gap. Interestingly, Column 6 shows that the emphasis on stand-out words has a positive effect, too, while the opposite holds for grindstone words, whose use in describing candidates penalizes their career success. How candidates are talked about contributes to explaining 8 per cent of the raw gender gap. Finally, column 7 illustrates that more than one third of the raw gender gap can be explained by the control variables considered.

Next, we explore to what extent the two dimensions highlighted in Figure 7 – taken in isolation – are responsible for the observed gender gap. Namely, in Table 7 we replicate a battery of specifications analogous to Table 6, with the dependent variable being an indicator for having the associate (or higher) professor title, while in Table 8 we consider as dependent variable an indicator for holding a placement at a Top-20 institution.

Column 1 and 2 in Table 7 show that female job market candidates appear on lower positions in the seniority ladder compared to their male cohort peers. The raw gender gap in the probability of holding an associate professor position amounts to a point estimate of 4 percentage points in our restricted sample, with a lower value than in the full sample of scholars.

The coefficient decreases when we control for our proxies for the candidate's characteristics and quality (Column 3). For instance, the number of pre-job market publications significantly and positively correlates with a candidate's likelihood of holding an associate professor position. The reduction in the gender gap in the probability of being associate shows that the unfavorable female to male relative position is at least partially due to the selection into and the performance during PhD studies. Column 4 shows that the inclusion of the characteristics of the letter writer has a small influence on the raw gender gap and none of them is statistically significant, whereas Column 5 indicates that the number of letter writers and average letter length are associated with better chances to appear higher on the academic jobs ladder. In column 6 we turn to examining how the content of recom-

Table 7: Career success: probability of holding an Associate Professorship

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.0440***	-0.0409***	-0.0369***	-0.0404***	-0.0401***	-0.0389***	-0.0345***
	(0.00830)	(0.00911)	(0.00915)	(0.00916)	(0.00910)	(0.00913)	(0.00921)
# Publications Pre-JM			0.0235***				0.0231***
			(0.00314)				(0.00313)
Main lett. writer female				-0.00100			-0.000638
				(0.0126)			(0.0126)
# Top 5 public. (main lett. writer)				0.000419			0.000874
				(0.000575)			(0.000621)
Full professor (main lett. writer)				-0.00216			-0.00218
				(0.00917)			(0.00916)
# Letter writers					0.0132*		0.0116
					(0.00714)		(0.00717)
Average letter length (std)					0.00747*		0.0129***
					(0.00417)		(0.00443)
Standout cos. sim.						0.286***	0.223***
						(0.0673)	(0.0684)
Grindstone cos. sim.						-0.0369	0.0164
						(0.0723)	(0.0763)
Mean dependent variable for men	0.159	0.159					
% Raw Gap Explained			9.8	1.2	2.0	4.9	15.6
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.120	0.141	0.168	0.141	0.142	0.143	0.172
N	6913	5699	5699	5699	5699	5699	5699

Notes: Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$

mendation letters and how candidates are talked about relates to the probability of ranking highest in the academic ladder. When accounting for the prevalence of different candidate characterizations based on the semantic classifications described in Section 4.1, we find that a package highly charged with stand-out words correlates positively and significantly with the achievement on the job ladder. On the other hand, grindstone-intense tones are negatively related to the probability of being associate and above (although the coefficient is not statistically significant). The inclusion of our proxies for candidates' characterization in the letters of reference further reduces the observed gender gap, and explains about 5% of it.

The full set of regressors in column (7) accounts for about 15% of the observed gap.

Table 8 reports the estimates of the likelihood of holding a position in a Top-20 Department as of the Repec 2021 ranking.[19] The raw gender gap in placement prestige is not significant in both sub-samples in Columns 1 and 2 of Table 8. Interestingly, the coefficient grows and becomes statistically significant at 10% in specifications with different sets of control variables, suggesting that women with similar job-market "credentials" to their male counterparts may be able to obtain higher ranked positions. We also note that, consistently with the pattern observed in previous tables, the similarity to stand-out words is associated positively and significantly with placement in top departments; conversely, grindstone characterizations are associated with a lower likelihood of working there.

These results suggest that it is the different performance over the career ladder rather than the prestige of the institution of placement, which plays a major role in penalizing female career success. Women are as likely as men to hold a job in a Top-20 institution, but less likely to achieve an associate professor position. This is consistent with diverging career paths, where women, although equally likely to make it to top institutions, appear on less senior positions within these top environments and experience a lower career progression. While this is in line with, for example, the literature on child penalties (Kleven et al., 2019), which highlights how childbirth sets fathers and mothers on different career paths, our results further point out that how candidates are presented on the job market also explains a non negligible part of the gap.

5.2 Research output

In this section we explore whether the differences observed in early career placement further translate into other career success indicators, such as publication records. In particular, in Table 9 we estimate our model on an indicator variable of whether the candidate has any publications in one of the Top-8 journals.[20] The specification in each column replicates the corresponding one in the three previous regression tables.

These results can be interpreted as evidence that the effect of gendered reference letters

[19] The results are qualitatively confirmed using QS ranking instead of Repec.

[20] These include the conventional Top 5 journals in economics – American Economic Review, Econometrica, Review of Economic Studies, Quarterly Journal of Economics, Journal of Political Economy – and the three top journals in finance – Journal of Financial Economics, Review of Financial Studies and Journal of Finance.

Table 8: Career success: placement in top 20 institution (Repec)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.00570	0.0106	0.0169**	0.0158*	0.0151*	0.0144*	0.0213***
	(0.00747)	(0.00831)	(0.00822)	(0.00824)	(0.00820)	(0.00832)	(0.00816)
# Publications Pre-JM			0.00113				0.000379
			(0.00179)				(0.00181)
Main lett. writer female				0.0164			0.0131
				(0.0110)			(0.0107)
# Top 5 public. (main lett. writer)				0.00595***			0.00347***
				(0.000648)			(0.000649)
Full professor (main lett. writer)				0.0147*			0.0164**
				(0.00777)			(0.00771)
# Letter writers					0.0421***		0.0328***
					(0.00608)		(0.00596)
Average letter length (std)					0.0427***		0.0289***
					(0.00403)		(0.00422)
Standout cos. sim.						0.302***	0.208***
						(0.0581)	(0.0572)
Grindstone cos. sim.						-0.250***	-0.0894
						(0.0575)	(0.0603)
Mean dependent variable men	0.078	0.078					
% Raw Gap Explained			-	-	-	-	-
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.00167	0.00216	0.0622	0.0393	0.0342	0.00824	0.0926
N	6511	5699	5699	5699	5699	5699	5699

Notes: Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$

on job placement that we showed in section 5.1 may carry on to early research outcomes - although we acknowledge that other factors such as fertility choices or effort may influence these outcomes to a larger extent than the initial placement - so that also the scientific productivity of female researchers is negatively affected in the medium term.[21] As for the case

[21] In Appendix Tables B.2 and B.3 we estimate our model on two further measures of publication records: the (log 1+) number of publications as retrieved from Repec, and the (log 1+) number of citations to articles as reported in Repec. The results are very much in line with those of Table 9, with a 17 and 29 per cent raw gap in the number of publications and citations, respectively. The gap is mostly driven by candidate's

Table 9: Research productivity: Top 8 publications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.0272***	-0.0276***	-0.0208***	-0.0240***	-0.0243***	-0.0256***	-0.0174**
	(0.00647)	(0.00730)	(0.00720)	(0.00730)	(0.00725)	(0.00734)	(0.00719)
# Publications Pre-JM			0.0285***				0.0282***
			(0.00324)				(0.00324)
Main lett. writer female				0.00812			0.00254
				(0.0103)			(0.0100)
# Top 5 public. (main lett. writer)				0.00392***			0.00275***
				(0.000583)			(0.000617)
Full professor (main lett. writer)				0.00551			0.00704
				(0.00752)			(0.00735)
# Letter writers					0.0213***		0.0147**
					(0.00609)		(0.00579)
Average letter length (std)					0.0313***		0.0280***
					(0.00369)		(0.00385)
Standout cos. sim.						0.190***	0.127**
						(0.0551)	(0.0541)
Grindstone cos. sim.						-0.106*	0.0558
						(0.0578)	(0.0590)
Mean dependent variable men	0.077	0.075					
% Raw gender gap explained			24.6	13.0	12.0	7.2	37.0
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.0283	0.0350	0.101	0.0510	0.0498	0.0371	0.119
N	6913	5699	5699	5699	5699	5699	5699

Notes: Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$

of outcomes related to job market placement, the largest reduction in the raw gender gap is due to the inclusion of candidates' observable characteristics. Interestingly, the sign on the variables measuring the letter emphasis on candidates grindstone or standout characteristics continues to consistently indicate the same pattern, with the former (the latter) being positively (negatively) associated with the research productivity and they appear to explain

characteristics with, however, a significant impact of the letters' characteristics, consistent with that found in the main specification. The results are robust to using a Poisson regression to take into account a potential high incidence of zeros in these dependent variables (available upon request).

around 7% of the gender gap. All together, the control variables used in our regressions account for 37% of the raw gender gap.

6 Robustness checks

6.1 Oaxaca-Blinder Decomposition

In Section 5 we analyzed how the different observable variables – relating to the candidate's and referee's characteristics, the letters in the job market application package and the way in which candidates are characterized in there – can explain gender differences in the observed gender gap in early career outcomes. In order to evaluate the overall contribution of all such variables in explaining the observed gaps and to apportion the part that arises from differences in characteristics rather than unexplained factors, in Table 10 we resort to a pooled Oaxaca-Blinder decomposition where we group together the set of variables as in equation 2.[22] The corresponding model reads:

$$y_m - y_f = (X_m - X_f)\beta' - \delta' \tag{3}$$

The observed gender gap in outcomes $(y_m - y_f)$ is decomposed into a part that is explained by difference in observable characteristics, this corresponds to the first term on the right-hand side of Equation 3 for each set of X, and an unexplained part (δ') which corresponds to (minus) the coefficient of the female dummy (itself negative) in the pooled regressions in Tables 6 and 9.

The table reports the predicted outcome for male and female researchers in the upper panel, together with the difference between the two, i.e., the raw gender gap. This is decomposed into an explained and an unexplained part in the lower panels, the explained part being further split into the contribution of each set of variables.

The results confirm that there are significant gender gaps in both the probability of holding an associate professorship in a Top-20 institution and the probability of having a Top-8 publication. In the first case, the observed characteristics explain about a third of the gap. In particular, candidate's and referee's characteristics account for about 10 and 12% of

[22] Our approach herein follows that in Fortin et al. (2021).

Table 10: Oaxaca-Blinder decomposition

	(1)	(2)	(3)	(4)
	Career success		Top 8 publications	
	Coefficient	std. err.	Coefficient	std. err.
Prediction male	.0209552***	.0022361	.0889376***	.0044438
Prediction female	.0112853***	.0026452	.0595611***	.0059266
Difference	.0096699***	.0034636	.0293765***	.0074076
Explained:				
Candidate chars	.0009868	.0008336	-.0022072	.0018346
Letter writers chars	.0012046**	.0005673	.00338***	.00118
Letters chars	.0005504*	.0003117	.0011433*	.0006399
WEs	.0002832	.0003795	.0009577	.000778
Total	.0030251***	.0010314	.0032739	.0023081
Unexplained:				
Total	.0066448**	.0034291	.0261026***	.0074316

Notes: Robust Standard errors in columns (2) and (4). * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

the gap respectively, whereas letter characteristics and the way candidates are talked about in the letters (i.e., WEs) for about 5 and 3% respectively. Two thirds of the observed gap remain unexplained in this model, consistently with the results in Table 6.

Regarding the research outcomes, the portion of the gap that is explained by the variables included in our regression is slightly above 10%, it was about 15% in Table 9. Of this, referee's and letters' characteristics are the most significant contributors. These results are consistent with those from the pooled linear regressions in Tables 6 and 9.

6.2 Specification checks

In Table 11 we propose a number of different specifications to corroborate our main findings. The table reports only the estimated gender gap and the share of it explained by the model. Full regression results are reported in Appendix Tables B.4 to B.8.

In Panel A we estimate an ordered logistic regression in which the outcome is a discrete variable taking the value 3 if the researcher currently holds an associate professor position, the value 2 if she is an assistant professor one, and the value 1 if she holds a lower ranked academic position (e.g., post-doc, research fellow). The estimated coefficients indicate that

Table 11: Specification checks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Academic ladder							
Female	-0.244***	-0.265***	-0.188***	-0.262***	-0.256***	-0.259***	-0.181**
	(0.0637)	(0.0690)	(0.0708)	(0.0696)	(0.0693)	(0.0691)	(0.0717)
% Raw Gap Explained			29.1	1.1	3.4	2.3	31.7
B. Academic ranking							
Female	-9.764*	-13.46**	-15.39***	-16.69***	-17.35***	-17.29***	-19.92***
	(5.566)	(5.723)	(5.562)	(5.653)	(5.519)	(5.708)	(5.432)
% Raw Gap Explained			14.3	24.0	28.9	28.5	48.0
C. Academic ranking conditional on ladder							
Female	-5.918	-10.47*	-13.00**	-13.22**	-14.48***	-14.32**	-17.64***
	(5.646)	(5.812)	(5.654)	(5.730)	(5.550)	(5.792)	(5.467)
% Raw Gap Explained			24.2	26.3	38.3	36.8	68.5
D. Career success with PhD institution FE							
Female	-0.00767**	-0.00703*	-0.00612	-0.00679*	-0.00664*	-0.00662*	-0.00532
	(0.00331)	(0.00371)	(0.00377)	(0.00375)	(0.00372)	(0.00371)	(0.00378)
% Raw Gap Explained			12.9	3.4	5.5	5.8	24.3
E. Top 8 publications with PhD institution FE							
Female	-0.0210***	-0.0222***	-0.0197**	-0.0223***	-0.0208***	-0.0208***	-0.0170**
	(0.00691)	(0.00774)	(0.00778)	(0.00776)	(0.00769)	(0.00776)	(0.00774)
% Raw Gap Explained			11.3	-5.0	6.3	6.3	23.4
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓

Notes: In panel A, B and C the sample is restricted to candidates who currently hold a position in academia. In panel A the estimated model is an ordered logistic one, in panel B and C a Tobit model with upper censoring at 309, in panels D and E linear models with binary outcomes. Full results are reported in Appendix Tables B.4 to B.8. Robust Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

female researchers are significantly more likely to hold a lower ranked academic position relative to men. The gap is mostly explained by candidate's characteristics and the way candidates are described explain 2.3% of the gap.

In Panel B we only focus on the continuous ranking of the academic institution in which the candidate works. The estimated model is a Tobit in that the Repec ranking variable is censored at $r = 309$. We observe that women generally land in higher ranked institutions and that, accounting for differences in the job market application package, the advantage would be significantly higher. In panel C, we estimate the same model adding fixed effects for the academic position ladder among controls in all specifications. Even holding constant the level of appointment (e.g. among assistant or associate professors only) the ranking of the institutions in which women work is higher and would be significantly higher if they received the same type of support on the market than their male peers.

In the last two panels, we re-estimate the models in Tables 6 and 9 adding fixed effects for the institution granting the PhD. This is meant to better account for unobservable differences in candidates' quality. The estimated gaps, indeed, slightly shrink but remain significant. Similarly, the portion of such gaps explained by differences in job market reference letters is reduced but remains sizable.

7 Conclusions

The goal of this paper is to estimate whether and to what extent female graduate students are subject to differential reference letter writing practices and how the latter affect the start of their careers. In particular, we analyze the extent of gendered student-advisor matching, the presence of explicit differences in the strength of academic support for male and female students, and of implicit gender stereotypes, as captured by the language used, to assess their role in influencing the career path of women.

To these ends, we built a novel dataset containing information on job market candidates applying to two top institutions hiring on the global market for junior economist positions. Our analysis combines standard information on demographic characteristics and labor market outcomes with an innovative set of measures built through the text analysis of the candidates' reference letters.

Our findings reveal significant gender differences in the way male and female job market candidates are presented on the job market. Differences concern not only some observable factors, such as a higher likelihood of being matched with a female advisor, or a lower number of sponsors, but also some more subtle aspects that reveal implicit gender biases on the part of senior academics. In particular, we find that female candidates are consistently described more in terms of being diligent and hardworking rather than outstanding or brilliant. Such differences are driven by letters written by male sponsors, whereas women make no differences based on the gender of the candidate. Linking this information with proxies of career success, we find that such differences explain a non negligible part of the lower success of female PhDs in the economics profession. Indeed, we show that if female candidates were to receive the same recommendation letters as male candidates, the observed gap in their career achievement (both in terms of seniority and of prestige of the placement) would decrease significantly or even be reverted. Similar results hold also when looking at publication records as proxies for early career success. We interpret the latter results as evidence of the (indirect) effect of initial gender differences in job market placement.

The external validity of our results may seem limited by the selected nature of our sample. However, at least for the European market, we showed that the representativeness of the pool of candidates that we analyze is quite large.

In conclusion, the analysis carries important implications for the profession. First, it reveals some implicit gender biases in the way male and female economists are perceived by senior academics. Second, it highlights a potential structural flaw in the academic job market process that, by heavily relying on reference letters, effectively puts female candidates in a weaker position to compete. Our analysis suggests that such mechanism is particularly relevant in contexts, like the economics profession, that are highly male dominated, especially among senior professionals.

Although the referral process is considered an essential ingredient of the labor market, we lack a large-scale assessment of whether it is a gender-neutral process, both on the evaluator (i.e., do female and male evaluators talk about different aspects of candidates in the reference letters/performance appraisals they write?) and on the candidate side (i.e., are female and male candidates described differently?). Our research aims to fill this gap. A higher awareness of such biases can help restructuring the referral process to make it less prone to them, and hence reduce gender gaps at the very first stages of the career and limit their capacity to propagate in the long run.

Note that, while we focus on the academic labor market, the use of references is by no means limited to it. For instance, performance reviews are key tools in organizations to evaluate an employee performance and, while they have the advantage of setting goals and design career trajectories, they could be open to subjective impressions by managers/evaluators and the language used could be indicative of implicit stereotypes on the appropriate characteristics and roles of men and women. Our research thus advances our knowledge on the presence and extent of gendered language in labor market appraisals and their influence on career paths of men and women.

A Words lists

The list of stopwords considered is the following:

[“a”, “about”, “above”, “after”, “again”, “against”, “all”, “also”, “am”, “an”, “and”, “another”, “any”, “are”, “as”, “at”, “back”, “be”, “because”, “been”, “before”, “being”, “below”, “between”, “both”, “but”, “by”, “could”, “did”, “do”, “does”, “doing”, “down”, “during”, “each”, “even”, “ever”, “every”, “few”, “first”, “five”, “for”, “four”, “from”, “further”, “get”, “go”, “goes”, “had”, “has”, “have”, “having”, “he”, “her”, “here”, “hers”, “herself”, “high”, “him”, “himself”, “his”, “how”, “however”, “i”, “if”, “in”, “into”, “is”, “it”, “its”, “itself”, “just”, “least”, “less”, ’like’, ’long’, ’made’, “make”, “many”, “me”, “more”, “most”, “my”, “myself”, “never”, “new”, “no”, “nor”, “not”, “now”, “of”, “off”, “old”, “on”, “once”, “one”, “only”, “or”, “other”, “ought”, “our”, “ours”, “ourselves”, “out”, “over”, “own”, “put”, “said”, “same”, “say”, “says”, “second”, “see”, “seen”, “she”, “should”, “since”, “so”, “some”, “still”, “such”, “take”, “than”, “that”, “the”, “their”, “theirs”, “them”, “themselves”, “then”, “there”, “these”, “they”, “this”, “those”, “three”, “through”, “to”, “too”, “two”, “under”, “until”, “up”, “us”, “very”, “was”, “way”, “we”, “well”, “were”, “what”, “when”, “where”, “whether”, “which”, “while”, “who”, “whom”, “why”, “with”, “would”, “you”, “your”, “yours”, “yourself”, “yourselves”]

To obtain the average vectors that characterize each of the semantic categories described in Section 4.2, we adopt the lists used in the literature (we start from Schmader et al. (2007) for the first two categories and from Chapman et al. (2020) for the last two). Below we report the full lists of words in each category.

- **Standout Adjectives:** [“standout”, “best”, “leader”, “exceptional”, “outstanding”, “star”, “superstar”, “impressive”]
- **Grindstone Adjectives:** [“hardworking”, “tenacious”, “deliberate”, “productive”, “efficient”]
- **Communal Adjectives:** [“likable”, “friendly”, “enthusiastic”, “enthusiasm”, “agreeable”, “caring”, “nice”, “pleasant”, “kind”, “kindness”, “warm”, “warmth”, “cheerful”, “polite”, “smile”, “modest”, “humble”, “genuine”, “collaborative”, “upbeat”]

- **Agentic Adjectives:** [“able”, “competitive”, “proactive”, “accomplished”, “energetic”, “eager”, “ambitious”, “ambition”, “confident”]

B Additional figures and tables

Figure B.1: Document frequency distribution and tfidf distribution of words in our corpus of reference letters.

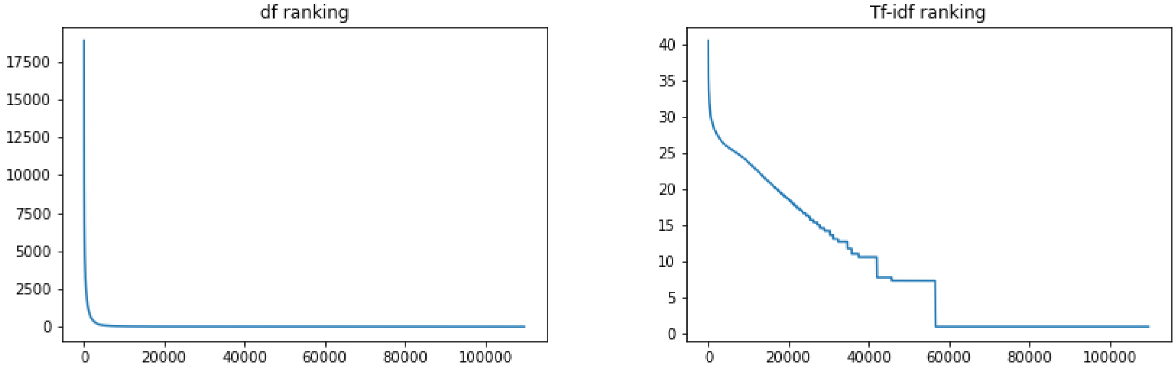


Table B.1: Descriptive statistics of job market candidates on the European Job Market, 2020/2021.

	N	Male	Female	Difference
American/Canadian PhD	787	0.438	0.416	0.022
EU PhD	787	0.436	0.490	-0.054
Italian PhD	787	0.033	0.049	-0.016
Applied micro	787	0.515	0.671	-0.156***
Macro/International/Finance	787	0.210	0.156	0.053*
Theory/Quantitative	787	0.193	0.136	0.057*
Phd Uni Top20 (QS)	787	0.149	0.132	0.017
Phd Uni Top20 Econ	787	0.256	0.198	0.058*
Observations	787			

Notes: Elaborations on data from the European Economic Association job market candidates directory * $p < 0.1$, ** $p < .05$, *** $p < 0.01$

Table B.2: Other research outcomes: (log) number of publications

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.164***	-0.173***	-0.129***	-0.172***	-0.171***	-0.166***	-0.127***
	(0.0212)	(0.0228)	(0.0176)	(0.0229)	(0.0228)	(0.0229)	(0.0176)
# Publications Pre-JM			0.261***				0.260***
			(0.00810)				(0.00805)
Main lett. writer female				0.0519			0.0120
				(0.0327)			(0.0248)
# Top 5 public. (main lett. writer)				0.00360***			0.00201*
				(0.00136)			(0.00118)
Full professor (main lett. writer)				0.0163			0.0359**
				(0.0228)			(0.0175)
# Letter writers					0.0625***		0.0624***
					(0.0188)		(0.0140)
Average letter length (std)					0.0180		0.0458***
					(0.0113)		(0.00882)
Standout cos. sim.						0.558***	0.231*
						(0.175)	(0.133)
Grindstone cos. sim.						-0.378**	0.302**
						(0.178)	(0.144)
Mean dependent variable men	0.638	0.626					
% Raw Gap Explained			25.4	0.6	1.2	4.0	26.6
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.0814	0.0981	0.484	0.0998	0.100	0.1000	0.490
N	6913	5699	5699	5699	5699	5699	5699

Notes: The dependent variable is logarithm of (1 + the number of publications). Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

Table B.3: Other research outcomes: (log) number of citations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.278***	-0.288***	-0.213***	-0.277***	-0.271***	-0.267***	-0.197***
	(0.0488)	(0.0537)	(0.0467)	(0.0537)	(0.0535)	(0.0539)	(0.0465)
# Publications Pre-JM			0.439***				0.436***
			(0.0203)				(0.0201)
Main lett. writer female				0.177**			0.0911
				(0.0745)			(0.0631)
# Top 5 public. (main lett. writer)				0.0219***			0.0142***
				(0.00345)			(0.00325)
Full professor (main lett. writer)				0.0371			0.0802*
				(0.0521)			(0.0450)
# Letter writers					0.213***		0.195***
					(0.0419)		(0.0357)
Average letter length (std)					0.159***		0.190***
					(0.0255)		(0.0236)
Standout cos. sim.						1.541***	1.007***
						(0.394)	(0.344)
Grindstone cos. sim.						-1.530***	0.368
						(0.411)	(0.375)
Mean dependent variable men	1.454	1.426					
% Raw Gap Explained			26.0	3.8	5.9	7.3	31.6
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.0880	0.106	0.347	0.115	0.116	0.109	0.363
N	6913	5699	5699	5699	5699	5699	5699

Notes: Robust Standard errors in parentheses. * $p < 0.1$, **

Table B.4: Robustness checks: ordered logit estimation for academic position (associate, assistant, postdoc)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.244***	-0.265***	-0.188***	-0.262***	-0.256***	-0.259***	-0.181**
	(0.0637)	(0.0690)	(0.0708)	(0.0696)	(0.0693)	(0.0691)	(0.0717)
# Publications Pre-JM			0.144***				0.141***
			(0.0226)				(0.0229)
Main lett. writer female				0.0839			0.103
				(0.0985)			(0.0997)
# Top 5 public. (main lett. writer)				0.00868**			0.0111***
				(0.00342)			(0.00383)
Full professor (main lett. writer)				0.134**			0.0616
				(0.0656)			(0.0673)
# Letter writers					0.221***		0.198***
					(0.0536)		(0.0551)
Average letter length (std)					0.153***		0.179***
					(0.0332)		(0.0362)
Standout cos. sim.						1.308**	0.967*
						(0.510)	(0.523)
Grindstone cos. sim.						-0.153	0.643
						(0.544)	(0.582)
Mean dependent variable men	2.047	2.036					
% Raw Gap Explained			29.1	1.1	3.4	2.3	31.7
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2							
N	4886	4286	4286	4286	4286	4286	4286

Notes: The sample is restricted to candidates who currently hold a position in academia. The estimated model is an order logistic one on a 1-3 variable (3 is associate professor, 2 assistant professor, 1 any other lower ranked academic position). Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

Table B.5: Robustness checks: tobit estimation for academic ranking (Repec 2021 classification)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-9.764*	-13.46**	-15.39***	-16.69***	-17.35***	-17.29***	-19.92***
	(5.566)	(5.723)	(5.562)	(5.653)	(5.519)	(5.708)	(5.432)
# Publications Pre-JM			-0.119				0.371
			(1.291)				(1.249)
Main lett. writer female				-7.954			-4.590
				(8.149)			(7.808)
# Top 5 public. (main lett. writer)				-3.324***			-1.675***
				(0.266)			(0.242)
Full professor (main lett. writer)				-11.14**			-10.64**
				(5.262)			(4.996)
# Letter writers					-33.03***		-25.25***
					(4.074)		(3.957)
Average letter length (std)					-40.64***		-29.91***
					(2.556)		(2.664)
Standout cos. sim.						-246.9***	-199.7***
						(40.86)	(39.05)
Grindstone cos. sim.						304.6***	90.43**
						(43.14)	(42.54)
Mean dependent variable men	176.087	175.597					
% Raw Gap Explained			14.3	24.0	28.9	28.5	48.0
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2							
N	4891	4291	4291	4291	4291	4291	4291

Notes: The sample is restricted to candidates who currently hold a position in academia. The estimated model is Tobit model with upward censoring at $r = 309$. Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

Table B.6: Robustness checks: tobit estimation for academic ranking, conditional on ladder (Repec 2021 classification)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-5.918	-10.47*	-13.00**	-13.22**	-14.48***	-14.32**	-17.64***
	(5.646)	(5.812)	(5.654)	(5.730)	(5.550)	(5.792)	(5.467)
# Publications Pre-JM			-0.675				-0.375
			(1.288)				(1.234)
Main lett. writer female				-10.96			-8.088
				(8.187)			(7.841)
# Top 5 public. (main lett. writer)				-3.337***			-1.755***
				(0.266)			(0.238)
Full professor (main lett. writer)				-13.99***			-13.54***
				(5.302)			(5.005)
# Letter writers					-36.81***		-28.92***
					(4.123)		(3.990)
Average letter length (std)					-43.62***		-34.20***
					(2.567)		(2.674)
Standout cos. sim.						-273.5***	-224.7***
						(41.28)	(39.06)
Grindstone cos. sim.						296.5***	68.68
						(43.43)	(42.69)
Mean dependent variable men	176.087	175.597					
% Raw Gap Explained			24.2	26.3	38.3	36.8	68.5
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2							
N	4729	4155	4155	4155	4155	4155	4155

Notes: The sample is restricted to candidates who currently hold a position in academia. The estimated model is Tobit model with upward censoring at $r = 309$. Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

Table B.7: Robustness checks: career success with PhD institution fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.00767**	-0.00703*	-0.00612	-0.00679*	-0.00664*	-0.00662*	-0.00532
	(0.00331)	(0.00371)	(0.00377)	(0.00375)	(0.00372)	(0.00371)	(0.00378)
# Publications Pre-JM			0.00223**				0.00200*
			(0.00108)				(0.00107)
Main lett. writer female				0.00261			0.00233
				(0.00543)			(0.00533)
# Top 5 public. (main lett. writer)				0.000888*			0.000854*
				(0.000493)			(0.000498)
Full professor (main lett. writer)				0.00979**			0.00964**
				(0.00437)			(0.00439)
# Letter writers					0.00834**		0.00766**
					(0.00324)		(0.00325)
Average letter length (std)					0.0104***		0.0105***
					(0.00256)		(0.00261)
Standout cos. sim.						0.0328	0.0170
						(0.0324)	(0.0331)
Grindstone cos. sim.						-0.0341	0.00484
						(0.0351)	(0.0361)
Mean dependent variable men	0.017	0.017					
% Raw Gap Explained			12.9	3.4	5.5	5.8	24.3
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.0551	0.0544	0.0582	0.0579	0.0596	0.0547	0.0666
N	6382	5573	5573	5573	5573	5573	5573

Notes: The estimated model is equation 2 with the addition of fixed effects for the institution granting the PhD. Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

Table B.8: Robustness checks: Top-8 publications with PhD institution fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.0210***	-0.0222***	-0.0197**	-0.0223***	-0.0208***	-0.0208***	-0.0170**
	(0.00691)	(0.00774)	(0.00778)	(0.00776)	(0.00769)	(0.00776)	(0.00774)
# Publications Pre-JM			0.0290***				0.0284***
			(0.00334)				(0.00333)
Main lett. writer female				0.0188*			0.0103
				(0.0114)			(0.0111)
# Top 5 public. (main lett. writer)				0.00275***			0.00237***
				(0.000708)			(0.000683)
Full professor (main lett. writer)				0.00414			0.00315
				(0.00826)			(0.00800)
# Letter writers					0.0210***		0.0164***
					(0.00649)		(0.00620)
Average letter length (std)					0.0320***		0.0316***
					(0.00433)		(0.00435)
Standout cos. sim.						0.158***	0.119**
						(0.0609)	(0.0600)
Grindstone cos. sim.						-0.0774	-0.00519
						(0.0648)	(0.0650)
Mean dependent variable men	0.077	0.075					
% Raw Gap Explained			11.3	-5.0	6.3	6.3	23.4
Raw	✓	✓					
Candidate chars			✓				✓
Letter writer chars				✓			✓
Letter chars					✓		✓
WEs						✓	✓
R^2	0.0996	0.111	0.154	0.116	0.122	0.112	0.169
N	6775	5573	5573	5573	5573	5573	5573

Notes: The estimated model is equation 2 with the addition of fixed effects for the institution granting the PhD. Robust Standard errors in parentheses. * $p < 0.1$, ** $p < .05$, *** $p < 0.01$.

References

Ash, E., D. L. Chen, and A. Ornaghi (2020). Gender Attitudes in the Judiciary:Evidence from U.S. Circuit Courts. Technical report.

Auriol, E., G. Friebel, and S. Wilhelm (2020). Women in european economics. In S. Lundberg (Ed.), *Women in Economics*, Chapter 7, pp. 26–31. London: CEPR Press.

Bayer, A. and C. E. Rouse (2016). Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives 30*(4), 221–42.

Bleemer, Z. (2019). Gender stereotypes in professor-student interactions.

Bostwick, V. K. and B. A. Weinberg (2020). Peer effects in graduate programmes. In S. Lundberg (Ed.), *Women in Economics*, Chapter 8, pp. 65–71. London: CEPR Press.

Boustan, L., A. Langan, and I. B. Palmer (2020). Variation in women's success across phd programmes in economics. In S. Lundberg (Ed.), *Women in Economics*, Chapter 7, pp. 57–64. London: CEPR Press.

Caliskan, A., J. J. Bryson, and A. Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science 356*(6334), 183–186.

Card, D., S. D. Vigna, P. Funk, and N. Iriberri (2020). Are Referees and Editors in Economics Gender Neutral? *The Quarterly Journal of Economics 135*(1), 269–327.

Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias. *The Quarterly Journal of Economics 134*(3), 1163–1224.

Chapman, B., M. Rooney, E. Ludmir, D. Cruz, A. Salcedo, C. Pinnix, P. Das, R. Jagsi, C. Thomas, and E. Holliday (2020). Linguistic biases in letters of recommendation for radiation oncology residency applicants from 2015 to 2019. *Journal of Cancer Education*.

Chevalier, J. (2022, May). Report: Committee on the status of women in the economics profession (cswep). *AEA Papers and Proceedings 112*, 746–67.

Dupas, P., A. S. Modestino, M. Niederle, J. Wolfers, and T. S. D. Collective (2021). Gender and the Dynamics of Economics Seminars. NBER Working Papers 28494, National Bureau of Economic Research, Inc.

Dutt, K., D. Pfaff, A. Bernstein, J. Dillard, and C. Block (2016). Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience 9*.

Eberhardt, M., G. Facchini, and V. Rueda (2022, January). Gender Differences in Reference Letters: Evidence from the Economics Job Market. IZA Discussion Papers 15055, Institute of Labor Economics (IZA).

European Commission (2019). SHE Figures 2018. Technical report, Directorate-General for Research and Innovation.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society) 1952-59*, 1–32.

Fortin, N., T. Lemieux, and M. Rehavi (2021). Gender differences in fields of specialization and placement outcomes among phds in economics. *AEA Papers and Proceedings 111*, 74–79.

Gaulé, P. and M. Piacentini (2018). An advisor like me? advisor gender and post-graduate careers in science. *Research Policy 47*(4), 805–813.

Haney-Lopez, I. (2000). Institutional racism: Judicial conduct and a new theory of racial discrimination. *The Yale Law Journal 109*, 1717.

Hengel, E. (2021). Publishing while Female. Are women held to higher standards? Evidence from peer review. *Economic Journal Forthcoming*.

Hilmer, C. and M. Hilmer (2007). Women helping women, men helping women? same-gender mentoring, initial job placements, and early career publishing success for economics phds. *American Economic Review 97*(2), 422–426.

Jansson, J. and B. Tyrefors (2020). The Genius is a Male: Stereotypes and Same-Sex Bias in Exam Grading in Economics at Stockholm University. Working Paper Series 1362, Research Institute of Industrial Economics.

Janys, L. (2020). Evidence for a Two-Women Quota in University Departments across Disciplines. IZA Discussion Papers 13372, Institute of Labor Economics (IZA).

Kim, S. D. and P. Moser (2021). Women in science. lessons from the baby boom. Working Paper 29436, National Bureau of Economic Research.

Kleven, H., C. Landais, J. Posch, A. Steinhauer, and J. Zweimuller (2019, May). Child penalties across countries: Evidence and explanations. *AEA Papers and Proceedings 109*, 122–26.

Koffi, M. (2021a). Gendered citations at top economic journals. *AEA Papers and Proceedings 111*, 60–64.

Koffi, M. (2021b). Innovative ideas and gender inequality. Technical report.

Kozlowski, A. C., M. Taddy, and J. A. Evans (2019, sep). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review 84*(5), 905–949.

Lundberg, S. (2020). *Women in Economics.* A VoxEU.org ebook. CEPR Press.

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

Madera, J. M., M. R. Hebl, and R. C. Martin (2009). Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology 94(6)*, 1591–9.

Neumark, D. and R. Gardecki (1998). Women helping women? role model and mentoring effects on female ph.d. students in economics. *Journal of Human Resources 33*(1), 220–246.

Oaxaca, R. L. and E. Sierminska (2021). Field specializations among beginning economists: Are there gender differences? *AEA Papers and Proceedings 111*, 86–91.

Paredes, V., M. D. Paserman, and F. J. Pino (2020). Does Economics Make You Sexist? IZA Discussion Papers 13223, Institute of Labor Economics (IZA).

Pezzoni, M., J. Mairesse, P. Stephan, and J. Lane (2016). Gender and the publication output of graduate students: A case study. *PloS one 11.*

Rizzica, L. (2013). Home or away? Gender differences in the effects of an expansion of tertiary education supply. Questioni di Economia e Finanza (Occasional Papers) 181, Bank of Italy, Economic Research and International Relations Area.

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review 107*(5), 141–45.

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021). Gender differences in recognition for group work. *Journal of Political Economy 129*(1), 101–147.

Sarsons, H. and G. Xu (2021). Confidence men? evidence on confidence and gender among top economists. *AEA Papers and Proceedings 111*, 65–68.

Schmader, T., J. Whitehead, and V. Wysocki (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles 57*, 509–514.

Trix, F. and C. Psenka (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society 14*(2), 191–220.