I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15055

Gender Differences in Reference Letters: Evidence from the Economics Job Market

Markus Eberhardt
Giovanni Facchini
Valeria Rueda

JANUARY 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15055

Gender Differences in Reference Letters: Evidence from the Economics Job Market

Markus Eberhardt
University of Nottingham and CEPR

Giovanni Facchini
University of Nottingham, CEPR and IZA

Valeria Rueda
University of Nottingham and CEPR

JANUARY 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Gender Differences in Reference Letters: Evidence from the Economics Job Market*

Academia, and economics in particular, faces increased scrutiny because of gender imbalance. This paper studies the job market for entry-level faculty positions. We employ machine learning methods to analyze gendered patterns in the text of 9,000 reference letters written in support of 2,800 candidates. Using both supervised and unsupervised techniques, we document widespread differences in the attributes emphasized. Women are systematically more likely to be described using "grindstone" terms and at times less likely to be praised for their ability. Given the time and effort letter writers devote to supporting their students, this gender stereotyping is likely due to unconscious biases.

JEL Classification: J16, A11
Keywords: gender, natural language processing, stereotyping, diversity

Corresponding author:
Valeria Rueda
School of Economics
Sir Clive Granger Building
University of Nottingham
University Park
Nottingham NG7 2RD
United Kingdom
E-mail: valeria.rueda@nottingham.ac.uk

* We gratefully acknowledge financial support from STEMM-CHANGE, University of Nottingham and Econ Job Market for authorizing the use of the data. Malena Arcidiácono, Cristina Griffa, Yuliet Verbel-Bustamante, Thea Zoellner, and Diego Marino-Fages have provided excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily represent the views of the University of Nottingham. We thank seminar participants at Bocconi University and the Monash-Zürich text-as-data conference for their comments.

1 Introduction

Gender disparities in the workplace have received significant attention in public debate. Academia is facing increased scrutiny due to its low female representation (Valian, 1999), especially in the field of economics (Lundberg, 2020, Part I). Recent empirical work has documented that the economics career pipeline for women is "leaky", meaning that women tend to drop out of the profession at critical transitions, such as the jump from earning a Ph.D. to an assistant professorship, or from assistant to associate professor (for a broad review, see Lundberg and Stearns, 2019). This paper studies the first step of the academic career of an economist, the junior "job market" —the stage at which the leak has grown the most in the past decade (Lundberg and Stearns, 2019)— and which so far has not received much systematic attention (Lundberg, 2020).[1]

The academic job market in economics is unique in that it is a highly structured institution. It starts every year in late Fall with universities posting their job advertisements and potential applicants preparing a "job market package". The latter consists of one or more academic papers, a CV and a set of recommendation letters written by scholars familiar with the candidate. All the parties involved, i.e. the candidates, the letter writers and the hiring committees, interact via centralized platforms. Typically, the same package is used for the vast majority of jobs, making the marginal cost of an additional application low. Reference letters are not tailored to a particular institution and the same letter is usually used for *all* job applications (for more details see Coles et al., 2010).

In this paper, we investigate the presence of differences in the language used in reference letters, depending on the gender of the candidate being recommended. We use a unique dataset encompassing all applications for entry level positions received by a research-intensive university in the U.K. over the 2017-2020 period. Deploying Natural Language Processing tools, we analyze the text of over 9,000 reference letters written in support of 2,800 candidates. A standard letter covers a lengthy discussion of the candidate's job market paper, some reference to their additional research, and to their teaching and citizenship skills. Importantly, the final section of the letter provides a summary assessment of the candidate's academic abilities and recruitment prospects. Since we are primarily interested in the way candidates are described, we primarily focus on this final section. This corpus is then transformed into a term-frequency-inverse-document-frequency (tf-idf) representation. Borrowing from methods developed in cognitive psychology and linguistics, we quantify whether letters written in support of female candidates emphasize systematically different attributes.

We use two complementary approaches. First, we employ an unsupervised methodology to ascertain the terms in the letters that are the best predictors of a candidate's gender. We adopt a LASSO technique that selects the strongest predictors. Among these, we frequently observe terms related to research interests, but also to personality and "grindstone" attributes

[1] There is work on this issue in other disciplines, see for example Madera et al. (2009); Dutt et al. (2016); Hebl et al. (2018).

("determined", "diligent", "hardworking", etc.). Second, we rely on a supervised method, building dictionaries of words for common attributes emphasized in reference letters. These dictionaries are informed by existing research on the topic (Trix and Psenka, 2003; Schmader et al., 2007). We validate our dictionaries through an original comprehensive survey of academic economists based in U.K. research-intensive universities. Corroborating the exploratory results from the LASSO, we observe that descriptions of female candidates tend to emphasize significantly more "grindstone" attributes. In further specifications, we also uncover a tendency to use fewer terms related to ability.

Diligence and working hard are positive attributes (see Alan et al., 2019, on 'grit'). However, given the overwhelmingly positive tone of recommendation letters in the job market, it may be misleading to interpret our findings as suggesting that women receive 'better' recommendations. The opposite may well be true. In fact, as noted by Valian (1999, p. 170) "[a]lthough working hard is a virtue, labeling a woman a hard worker can be damning with faint praise. If someone is not considered able to begin with, working hard can be seen as confirmation of his or her inability." More generally, sociologists have pointed out that minorities are more often praised for their diligence than for their innate ability and that the signal of diligence is often interpreted as a lack of innate talent (Bourdieu and Passeron, 1977, p.201).

In additional results, we also find that female candidates are on average weakly and insignificantly associated with more teaching and citizenship terms, but this pattern hides important heterogeneities. In mid-ranking departments, the effect is strongly positive and significant whereas the opposite is true for elite institutions. We observe a similar pattern for the usage of standout terms. All our results are robust to a variety of specifications and across multiple relevant subsamples (e.g. depending on geographical location or institutional ranking).

We expect differences in the language of reference letters to depend on many factors, such as the institution the candidate graduated from or their research field. Some of these determinants may differ systematically for male and female candidates. We tackle this problem in a variety of ways.

In our baseline specifications we control for observable candidate and writer characteristics obtained from the application platform and from additional information we collected manually. On the writer's side, we control for their gender, the number of letters they provide in our sample, and the ranking of their institution. On the candidate's side, we control for their ethnicity, years since Ph.D. completion, broad field of specialization, publication record, and the ranking of their Ph.D.-awarding institution. The baseline results are not very sensitive to these controls, nor to alternative definitions of the reference letter ends.

Still, we may worry that unobservable determinants could affect our findings. We therefore run more restrictive models that allow us to account for unobserved, time-invariant institutional and letter-writer characteristics. A first set of models, which include fixed effects

for the Ph.D.-granting institution, confirm the gendered patterns observed even for candidates of the same cohort at the same institution. In further analysis, we restrict the sample to referees who have written letters for both male and female candidates and employ writer fixed effects. These more demanding specifications confirm that differences in describing male and female candidates are detectable even when we focus on individual writers. Further probing indicates that more experience writing for female candidates attenuates some of these differences.

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

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

The professional culture in Economics may also be problematic for women's advancement. Wu (2018) reports evidence of gender biases in posts about women in a well-known and widely used anonymous forum in the profession. Similarly, Dupas et al. (2021) study the

seminar culture and present evidence that female speakers face more hostile audiences. By analyzing recommendation letters, we are investigating a different aspect of the professional culture, namely mentorship. As opposed to these previous studies, our focus is on a setting in which economists fulfil a supportive and nurturing role.

We also contribute to the literature carrying out text analysis in academic recommendation letters (Trix and Psenka, 2003; Schmader et al., 2007; Dutt et al., 2016; Hebl et al., 2018; Madera et al., 2019). We build on this earlier work to classify the types of attributes usually emphasized in these letters. Additionally, to the best of our knowledge, we are the first to validate our classification by surveying a large sample of academics. Moreover, focusing on economics, we can access a substantially larger sample of letters that are broadly representative of a highly structured and globalized academic job market.

The paper is organized as follows. In Section 2 we discuss our sample as well as the general approach of our main textual analysis. Section 3 explains the process of data cleaning and preparation, followed by the exploratory analysis using unsupervised methods in Section 4. Section 5 outlines the supervised approach and presents the baseline results, with extensions and additional robustness checks provided in Section 6, followed by concluding remarks.

2 Data

We collected and cleaned the text of over 9,000 reference letters written in support of 2,800 candidates who applied for entry-level positions between 2017 and 2020 at a research intensive economics department in the U.K..[2] In each year in our sample the department

The department is one of the largest in the U.K., with over 55 regular faculty members. The majority of the faculty has an international background, with 53% having earned a Ph.D. outside the U.K. (half of them in the U.S., the other half in other European countries). It has been consistently ranked in the top-75 worldwide according to the Research Papers in Economics (RePEc) platform, and in the top-10 in the U.K. according to periodic research assessment exercises carried out since the early 1990s. It has a large Ph.D. program, with over 50 students in residence in a given year. 23% of staff is female.[3]

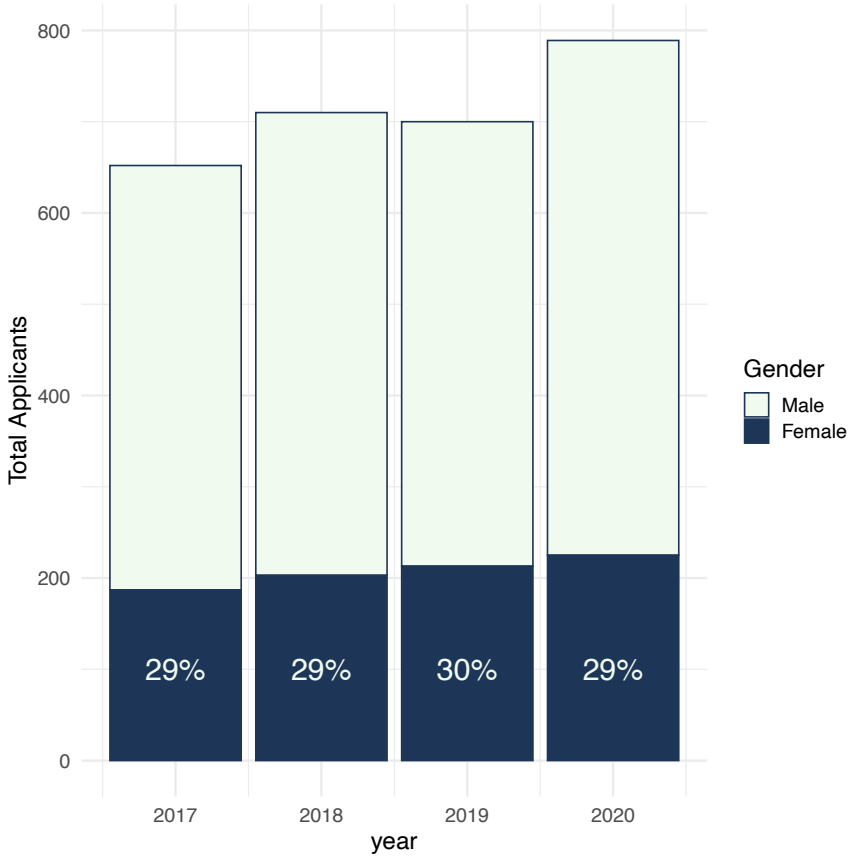
The applications were collected from the platform *EconJobMarket* (EJM). Access to and handling of these confidential data were done in accordance with the data processing agreement signed between the researchers and EJM, which obtained appropriate ethical approval.

For each letter, we know a number of characteristics of the letter writer and the candidate. For the letter writer, we have information on the institution where they were based at the time the letter was written. Using the R library 'GenderizeR', we infer their gender from

[2] All applications were filed exclusively through EJM, without any additional paperwork required.

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

Figure 1: Gender distribution of applicants in the sample



Notes: The figure shows the total number of applicants per year and the share of female applicants each year

their first names.[4] For candidates, we know characteristics they entered on *EconJobMarket*, such as gender, ethnicity, and the institution granting their Ph.D..[5] We also manually collect data from the candidates' CVs: we add information on their publication record at the time of application and their graduation date. The institutional ranking of both letter writers and candidates are taken from RePEc.[6] Information on the main advisor is also collected.

As shown in Figure 1, the number of applications received increased from 652 in 2017 to 789 in 2020.[7] 29-30% of the applicants are female, a figure which is consistent with underlying population data from EJM. Figure 2 shows the share of applicants by country. Approximately 50% of the candidates are based at U.S. institutions and 14% in the U.K. (see also Appendix Table B.1 for a detailed breakdown).

Figure 3 shows that the majority of applicants and reference letter writers are based in the top 100 ranked institutions, with slightly more letter writers concentrated at the very top. In Appendix Figure B.1 we limit ourselves to this group, highlighting that the our sample

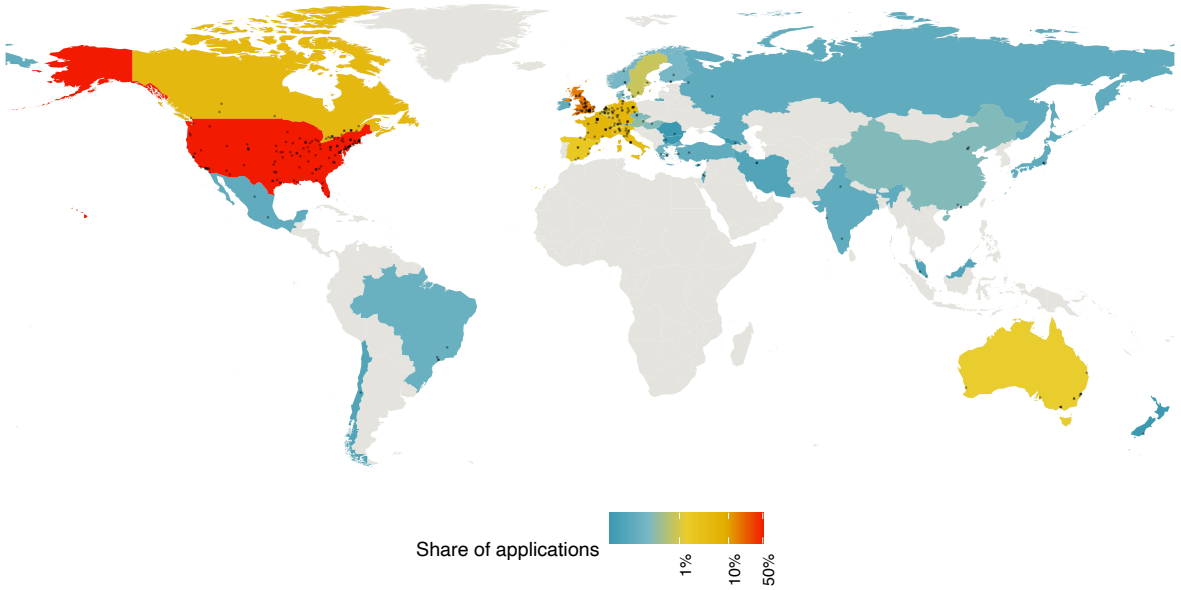
[4]We manually search for any cases of rare first names or where the probability reported by the algorithm is below 0.85.

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

[6]See Appendix A.3 for more details on how the ranking is constructed.

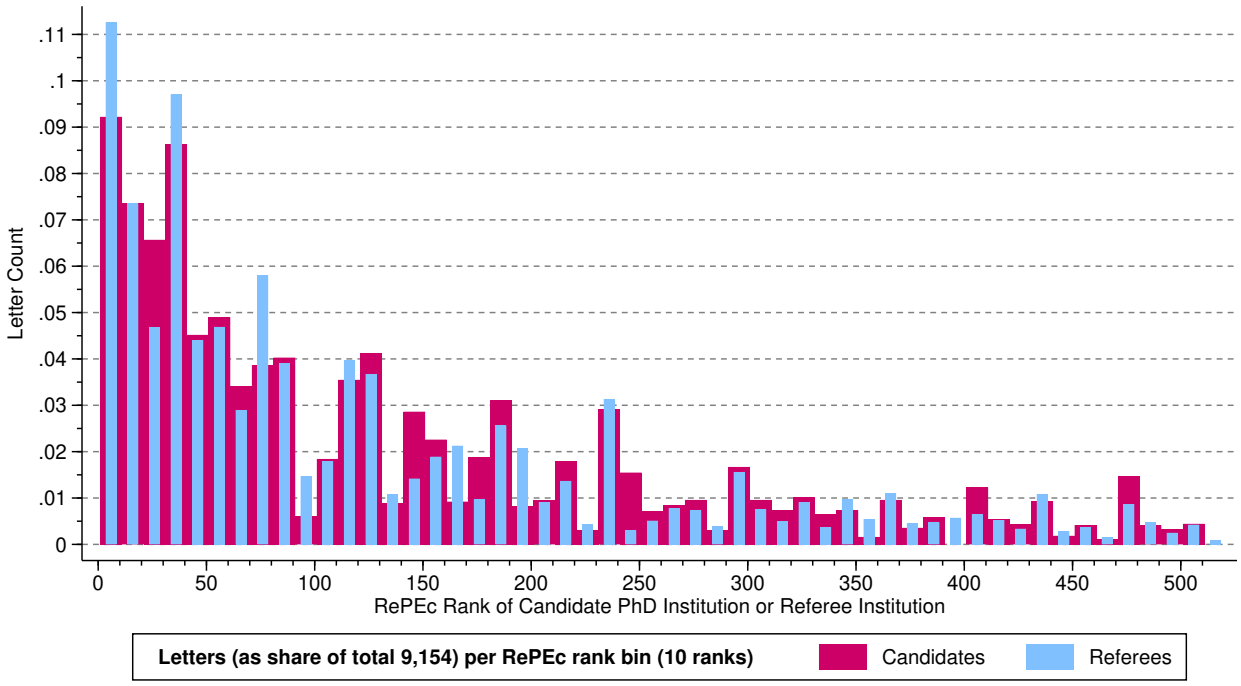
[7]Note that what we label the 2017 cohort refers to those candidates who applied in the Fall of 2016. Hence our analysis is for data which predates the Covid-19 pandemic.

Figure 2: Geographical distribution of applicants in the sample



Notes: The map shows the share of applicants who received their Ph.D.s from each country. The dots indicate the location of the Ph.D.-granting institution.

Figure 3: RePEc Rank of Candidate and Letter Writer Institution



Notes: The figure presents the frequency distribution of candidate and letter writer institution rank (in bins of 10).

contains a considerable number of applications from the very top institutions. We have 4,705 writers (female share 16.5%) in our sample, the median writer has written two letters, a dozen writers have provided twelve or more letters ($n_{max} = 18$). Additional summary statistics are reported in Appendix Table B.2.

Reference letters for the economics job market have a mean (median) length of roughly 7,600 (7,000) characters, which corresponds to around three pages A4, with a standard deviation of 3,600 characters (around 1.5 pages). A standard letter covers a lengthy discussion of the candidate's job market paper, some reference to their additional research, and to their teaching and citizenship attributes. Importantly, the final section of the letter provides a summary assessment of the candidate's academic abilities and recruitment prospects.

Since we are primarily interested in the way candidates are described, we focus our analysis on this end section. Section 3.1 explains how this section is extracted. A typical example of the information provided is given by the following quotation. Identifiable and sensitive characteristics have been redacted to protect privacy.

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

3 Methods

3.1 Data processing

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

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

be collapsed to the stem "publish". Following these steps, we have converted each reference letter into a collection of (stemmed) words.

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

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

In this paper, we use a Bag-of-Words (BOW) vectorization process to represent our corpus as a dataset. Using the vocabulary of applied economics, we can say that the Bag-of-Words is a matrix of dimension $D \times T$. Each row of this matrix represents a document, and each column represents a term. For each document, each cell refers to the term-frequency-inverse-document-frequency (tf-idf) of the term. The tf-idf is a common measure used to quantify the importance of a term in each document, compared to its prevalence in the corpus. The tf-idf is the product of the term frequency and the inverse-document frequency.

The term frequency $\text{tf}(t, d)$ is the number of times term t appears in document d:

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

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

$$\text{idf}(t) = \log \frac{1 + D}{1 + \text{df}(t)} \tag{2}$$

$$\text{with} \quad \text{df}(t) = \sum_{d} \mathbf{1}\{\text{tf}(t, d) > 0\}. \tag{3}$$

The term-frequency-inverse-document-frequency (tf-idf) is then:[8]

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \text{idf}(t) = \log \frac{1 + D}{1 + \text{df}(t)} \sum_{i}^{N_d} \mathbf{1}\{w_i = t\} \tag{4}$$

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

BOW are considered a standard approach for text vectorization in natural language processing, and researchers have shown that this simple representation is sufficient to infer interesting properties from texts (Grimmer and Stewart, 2013). This approach has many advantages. First, it is easy to implement. Second, the tf-idf for each word has the simple interpretation of capturing the importance of each word in the document, relative to its frequency in the corpus. We can also measure the importance of specific "attributes" in each letter by summing the tf-idf for the groups of words in the attribute category for each letter.

This approach has two main shortcomings. First, the vector space grows linearly with the vocabulary, which can cause significant computational challenges. In our case, our sample size is not large enough for this to become an issue. The second shortcoming is that the relationships *between* words are not taken into account. More recent deep-learning techniques use word embedding representations resulting in a vector-space of low dimension. In word embedding representation, terms represented with vectors that are close in space are semantically similar. Recent literature in law and economics has pioneered the implementation of word embeddings, for instance to compare the similarity of different semantic fields inside a given corpus (Ash et al., 2020, 2021, among others). Many of these papers are interested in exploring whether different semantic fields are correlated in different corpora (e.g. whether 'female' words tend to be associated with 'career' words or 'family' words). Unfortunately, word embeddings may perform disappointingly compared to traditional BOW in smaller samples (Shao et al., 2018; Ash et al., 2021), and ours is much smaller than the type of samples used in the new economics literature applying word embeddings.[9]

3.2 Separating ends

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

We use a two-step procedure to separate the letter ends. First, we create a dictionary of commonly used closing phrases (e.g. "Yours sincerely"). These phrases flag the end of the letter, and permit cleaning out long signatures (with multiple affiliations, addresses, etc.). We then take the 200 words *before* the first closing phrase flagged, which roughly corresponds to the length of one large paragraph. With this approach, we cover more than 80% of the letters. For letters without any identifiable closing phrase, we use the last 200 words of the document. We also consider 150 and 250 words cuts for the letter ends in the robustness section.

[9]For instance, Ash et al.'s (2021) analysis of judge-specific corpora falls in the category of a "small" sample for word embeddings. Their analysis relies on corpora with at least 1.5 million tokens (pre-processed words). For comparison, our main sample of interest, which consists of the universe of end of letters, contains approximately 852,000 tokens.

3.3 Language Categorisation

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

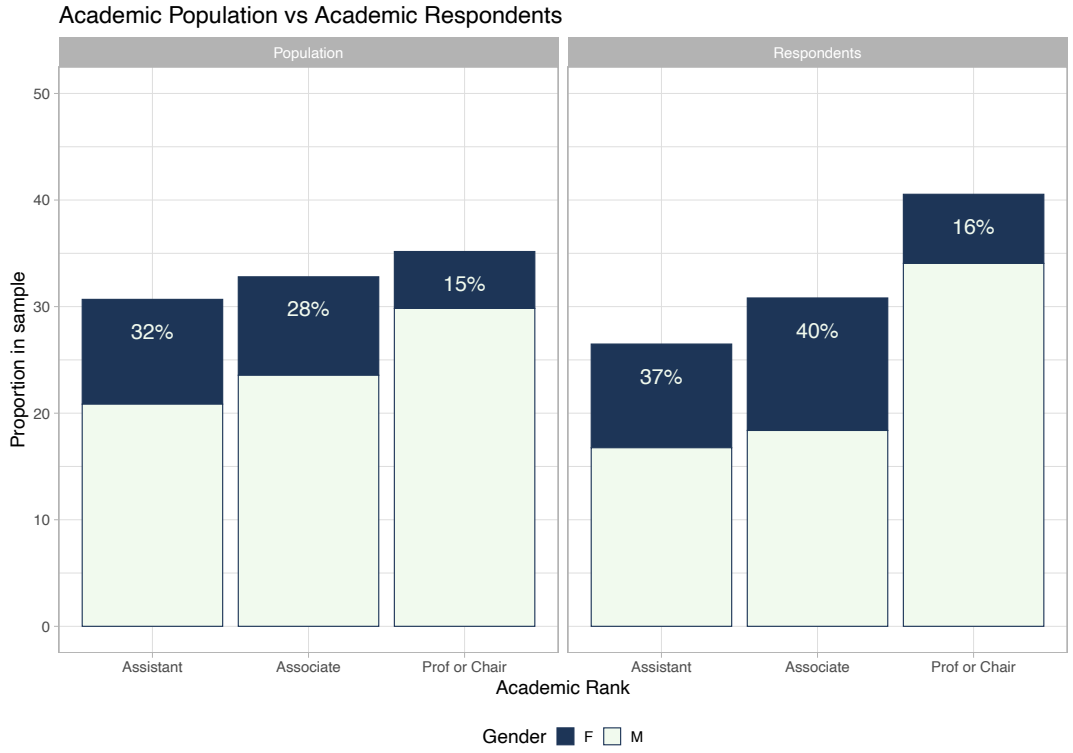
Schmader et al. (2007) propose five language categories that can be used to describe relevant features of an applicant, including *ability traits, grindstone traits, research terms, standout adjectives, and teaching and citizenship terms.* We add a category that refers to the *recruitment prospects* of the candidate. Ability traits involve language aimed at highlighting the applicant's suitability for the advertised position and include words such as *talent*, *brilliant*, *creative*, etc. Grindstone traits refer to language that, in the words of Trix and Psenka (2003, 207), resemble "putting one's shoulder to the grindstone". Words in this category include *hardworking*, *conscientious*, *diligent*, etc. Research terms are descriptors of the type of research carried out by the candidate e.g. *applied economics*, *game theory*, *public economics*, etc. Standout terms highlight especially desirable attributes of the applicant, like *excellent*, *enthusiastic*, *rare* etc. Teaching and citizenship is a broad category that refers both to the candidate's skills in the classroom, as well as their behavior with colleagues. Language in this group includes *good teacher*, *excellent colleague*, *friendly*, etc. The last category, recruitment prospects, has been added to identify words that, in the highly competitive and globalized labor market for fresh economics Ph.D.s, are widely used to describe the expected placement of the candidate. Words in this group include *highly recommended*, *top department*, *tenure track* etc. Appendix Figure B.2 shows word clouds for each of our language categories.

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

Figure 4 provides a breakdown of the population and of the survey respondents by level of seniority and gender. As can be seen, about one third of the population (left panel) are associate professors, with a slightly higher share represented by full professors, and a slightly smaller one by assistant professors. The share of females declines with seniority, representing 32% of staff at the assistant professor level, and only 15% at the most senior level. Turning to our sample (right panel), respondents are slightly more likely to be full professors, and

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

Figure 4: Population of academic surveyed compared to respondents



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

slightly less likely to be assistant professors than in the underlying population. Not surprisingly, females are over-represented among respondents, especially at the intermediate level of seniority.

Figure 5 illustrates the extent to which our own assessment of an expression is shared by the academics who took part in our survey. For all expressions classified into a language category by the authors, we show the distribution of classifications chosen by the plurality of validators.[11] While there is variation across language categories, there is broad consensus between our categorization and that of the profession.

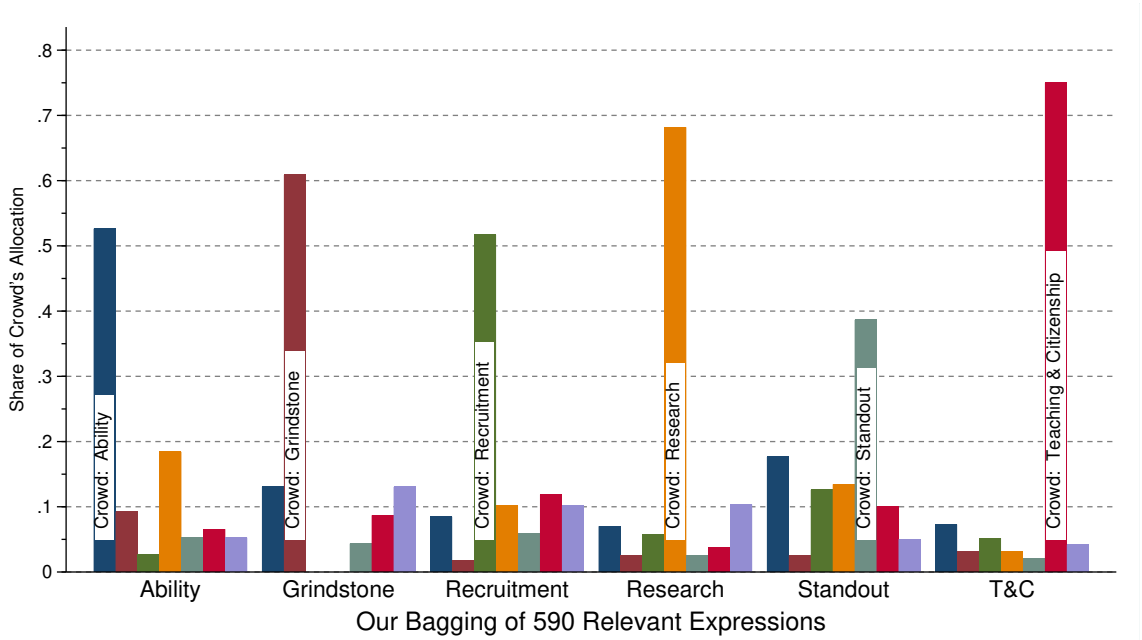
4 Unsupervised Analysis

4.1 Methodology

As an initial unsupervised analysis we ask whether specific terms used are more predictive of the gender of the candidate. To this end, we employ a least absolute shrinkage and selection operator (LASSO) to select the relevant set of terms. The LASSO estimator $\hat{\beta}$ solves the following problem:

[11] See Appendix A.1 for more details on how the figure is constructed.

Figure 5: Correspondence between authors' sentiment categories and the "wisdom of the crowd"



Notes: This figure shows the correspondence between the authors' chosen classification for each expression and the classification chosen by validators. For any word validated, it is attributed to the category that was chosen by the plurality of validators who were shown that word. See more details in Appendix A.1.

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2D} \sum_{d=1}^{D} (y_d - \boldsymbol{x}_d' \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\}, \tag{5}$$

where d is the letter. The gender of the candidate is the binary variable y_d. Vector x_d is the collection of tfidf(t, d) for the corpus. The second, penalty, term in equation (5) contains the 'tuning' parameters λ and ω which are selected to reduce the number of non-zero but small coefficients. p is the total number of terms.

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

In the majority of specifications, the adaptive LASSO has higher predictive power than the enet and the CV.[12] We hence only present results from our preferred model.

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

4.2 LASSO Results

A visualization of the results is presented in Figure 6. The Figure records the 164 predictors selected by the LASSO. We present the standardized beta coefficients of the linear probability model of candidate gender on tf-idf. Each line groups up to 6 predictors with similar coefficient magnitudes. The bars represent the range of the coefficient of the predictors listed in the line. Positive predictors are associated with female candidates, whereas negative ones are associated with males.

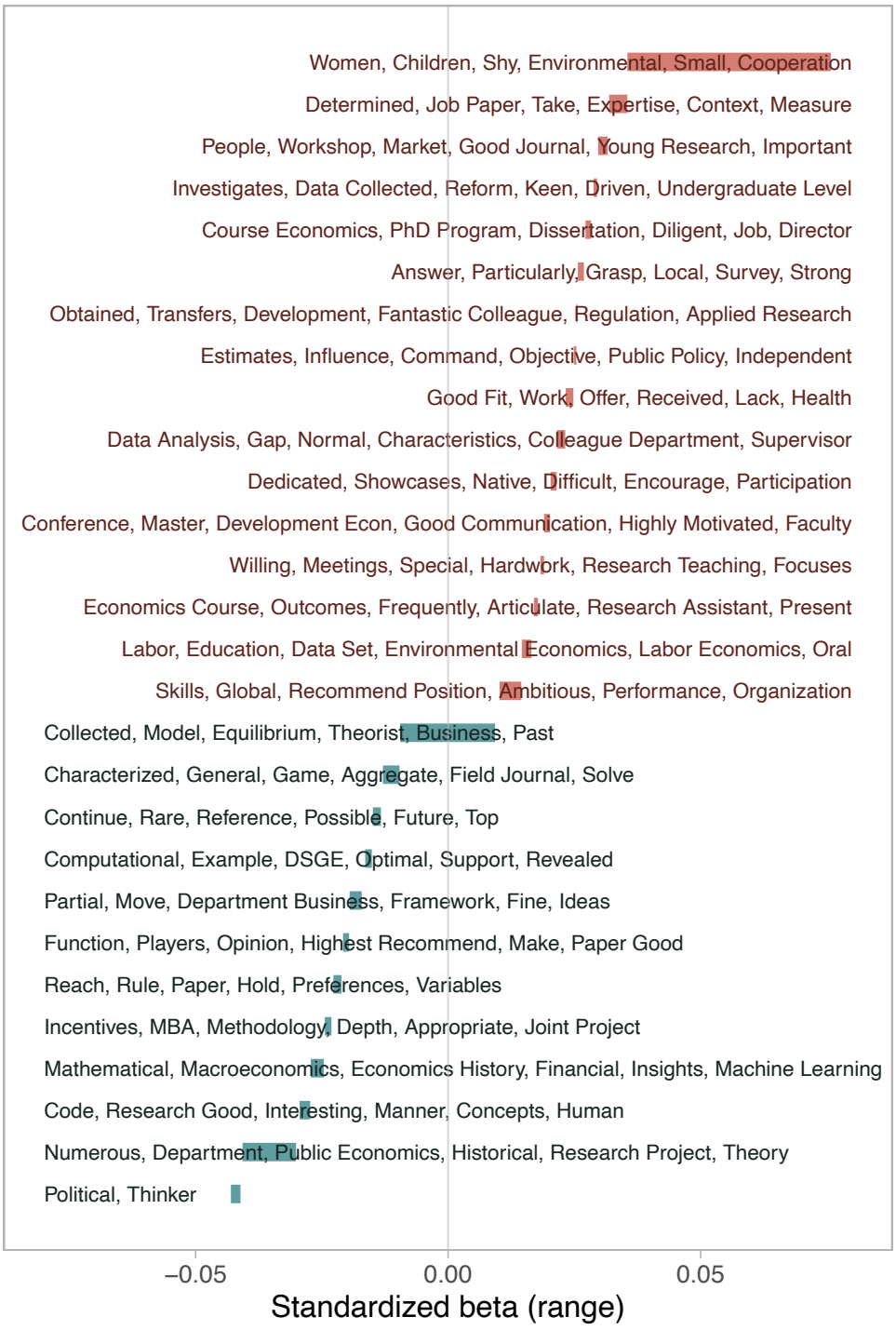
First, the figure reflects that women select across different research fields. Research on "women", "children", or "environmental" tends to be disproportionately carried out by female candidates, whereas "theory", "history", or "political" appear to be associated with male candidates. This "self-selection"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

Second, qualitatively, it appears that certain personality traits are gender specific. While being "determined", "diligent" or "keen" are strong predictors of female candidates, being a "thinker" or having "depth" are attributes more likely to be associated with men. This is a pattern that will be confirmed in the next section.

Finally, it is worth noting that traits such as youth ("young researcher") and shyness are reserved to women. This finding conforms to the stereotyping of women as naïve or child-like that has been documented in sociology (see for instance Goffman (1979, pp. 5 and 50-51) and Gornick (1979)), and for which there is suggestive evidence that it may harm women's credibility in the workplace (for a review see MacArthur et al., 2020).

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

Figure 6: LASSO Visualization



Notes: This figure shows the terms selected in the LASSO exercise. In each line, the vertical bars illustrate the range of the standardized beta coefficient for all the words listed. The beta coefficient is the change in propensity that the candidate is female associated with a one standard deviation increase in the tf-idf of the term. This LASSO exercise is conducted with stemmed words. In this figure, we have attributed to each stem its most frequent corresponding word. 164 stems out of 1408 are selected by the adaptive LASSO. $N = 9,362$, AUROC $= 0.716$.

5 Supervised Analysis with Dictionaries

In our supervised analysis, we employ the word dictionaries related to ability, grindstone, research, recruitment, standout, and teaching & citizenship discussed in Section 3.3 —we refer to these as "sentiments" for ease of discussion.

5.1 Specification and implementation

We run regressions defined in equation (6) using ordinary least squares.

$$\text{Sentiment}_{diwt} = \alpha + \beta\,\text{Female}_i + \mathbf{X}_i'\gamma + \mathbf{W}_w'\lambda + \nu_t + \varepsilon_{diwt} \tag{6}$$

Sentiment_{diwut} is the importance of each sentiment in letter d, written for candidate i by letter writer w in year t. For each sentiment (ability, grindstone, etc.), Sentiment_{diwut} is the sum of $\text{tfidf}(t, d)$ of all the terms in letter d associated with that sentiment in our dictionaries. Female_i is an indicator equal to 1 if the candidate is female, and β is our coefficient of interest. $\mathbf{X}_i$ is a vector of candidate-level controls, $\mathbf{W}_w$ is a vector of letter-writer controls; both are described in more detail below. We further include recruitment cohort fixed effects ν_t.

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

First, with regards to candidate attributes, all specifications include controls for their ethnicity, race, and the year they entered the job market. We sequentially add indicator variables accounting for the RePEc ranking band of the candidate's PhD-awarding institution.[13] Finally, we control for the years since PhD completion, for the broad field of specialization [14] and for the publication record. For the latter, we include the total number of publications and the number of articles published in top-field, top-5, and top general interest journals.[15]

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

To allow for the possibility of heterogeneous effects depending on institutional quality, we

[13] In particular we distinguish: top-25, top-26-50, top-51-100, top-101-200, beyond top-200, and an indicator for institutions not included in the RePEc ranking (11% of the sample).

[14] Section 6 describes in greater details how we define fields and the robustness of our results to alternative definitions.

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

run separate analysis for the full sample of letters, and letters from writers coming from institutions in the top-25 RePEc ranking, ranks 26-100, and top-100. Section 6 also discusses the robustness of our results to including candidate institution or letter-writer fixed effects, to using different cutoffs for the end of letters, and to considering the type of letter writer (main advisors vs other referees) or the location of the Ph.D.-granting institution (U.S. vs non-U.S.).

Each empirical model is estimated using four different sets of standard errors: robust, clustered by letter writer, clustered by letter writer institution, and clustered by candidate Ph.D.-awarding institution.

5.2 Results

Baseline Results Table 1 presents baseline results for the six outcomes using standard errors clustered by letter writer. In Figure 7 we visualize these results along those from a similar analysis carried out by further splitting the sample by the letter-writer institution's ranking, and for the four types of standard errors described above (for a total of 672 regressions). A darker shading of the marker indicates more specifications yielding statistically significant estimates for $\hat{\beta}$ (see the figure's notes for more details). Fully filled symbols are significant at 1% level across all possible standard error clustering. Hollow symbols do not reach significance for any type of clustering. The coefficient magnitudes are the estimates from equation (6) normalized by the standard deviation of the respective dependent variable.

Figure 7 shows that no matter the institutional ranking, and across all specifications, female candidates are significantly more likely to be associated with "grindstone" terms (from 6 to 10% of a standard deviation). These results confirm our interpretation of the unsupervised analysis (see section 4). We also observe that fewer terms related to research are used in letters supporting female candidates. Both of these results echo findings from other disciplines (Trix and Psenka, 2003; Valian, 2005).

Furthermore, in the entire sample, female candidates are on average associated weakly and insignificantly so with more teaching and citizenship terms. However, there are differences across institutions. In mid-ranking departments, letter writers are strongly and significantly more likely to use these terms, whereas the opposite holds for letter writers based in elite departments. We uncover a similar pattern for standout terms —in contrast with Trix and Psenka (2003) and Schmader et al. (2007), who observe a higher frequency of these adjectives in letters supporting male applicants for academic positions in medicine, and chemistry and biochemistry, respectively.

Finally, no significant patterns emerge when we consider ability or recruitment terms. The magnitude of the estimate of interest does not greatly differ across specifications, even after controlling for proxies capturing determinants of language that correlate with gender. This stability provides some reassurance that other unobserved confounding determinants

of language used in references are unlikely to explain away the results.

Table 1: Sentiments — End of Letters (200 words) — 7 Models

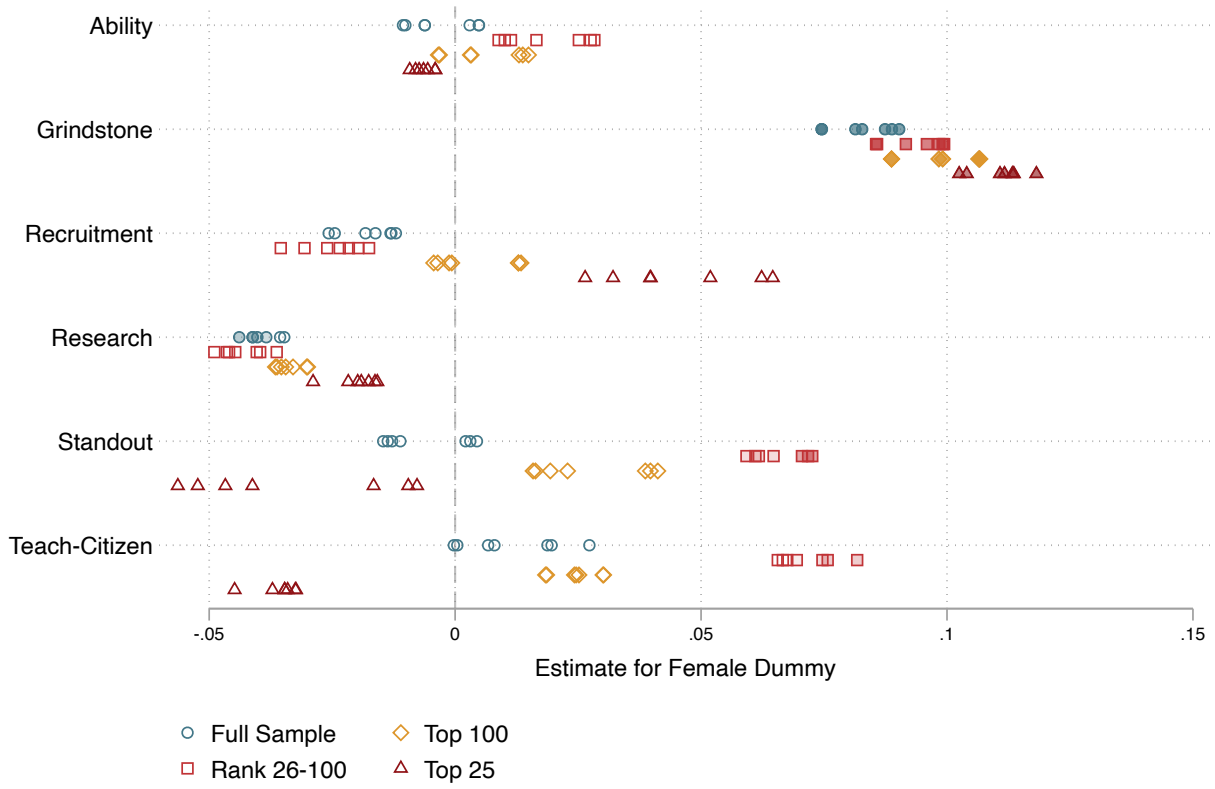
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0049	0.0048	0.0030	-0.0062	-0.0062	-0.0106	-0.0101
	(0.21)	(0.21)	(0.13)	(0.26)	(0.26)	(0.45)	(0.43)
Grindstone	0.0903	0.0874	0.0888	0.0828	0.0814	0.0745	0.0746
	(3.80)***	(3.68)***	(3.74)***	(3.47)***	(3.40)***	(3.12)***	(3.12)***
Recruitment	-0.0131	-0.0130	-0.0120	-0.0257	-0.0245	-0.0182	-0.0162
	(0.55)	(0.54)	(0.51)	(1.08)	(1.02)	(0.76)	(0.69)
Research	-0.0439	-0.0412	-0.0412	-0.0402	-0.0384	-0.0347	-0.0356
	(1.88)*	(1.77)*	(1.77)*	(1.72)*	(1.64)	(1.48)	(1.52)
Standout	0.0031	0.0044	0.0021	-0.0146	-0.0137	-0.0128	-0.0111
	(0.13)	(0.19)	(0.09)	(0.62)	(0.58)	(0.54)	(0.48)
Teaching &	0.0273	0.0196	0.0188	0.0080	0.0067	-0.0002	0.0004
Citizenship	(1.12)	(0.81)	(0.77)	(0.33)	(0.28)	(0.01)	(0.02)
FE/Variables absorbed	9	14	14	17	17	23	23
Additional covariates			1	1	5	6	7
Number of Letters	9154	9154	9154	9154	9154	9154	9154
dto for females	2695	2695	2695	2695	2695	2695	2695
Number of candidates	2791	2791	2791	2791	2791	2791	2791
dto female	817	817	817	817	817	817	817
Number of writers	4705	4705	4705	4705	4705	4705	4705
dto female	778	778	778	778	778	778	778
Letters by fem writers	1315	1315	1315	1315	1315	1315	1315
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively. This is the benchmark analysis for a letter end of 200 words.

Male and Female Writers Figure 8 compares results by letter-writer gender. We consider two samples only: all letter-writers and those based in top-100 institutions (the largest of the samples by institutional rank).[16]

[16]There are only 776 female letter writers (who have written 1,321 letters) in total, of whom 118 are in the top-25 group (with 226 letters), 203 in the top-50 (388), 315 in the top-100 (583), and 197 in the top-25 to 100 group (357).

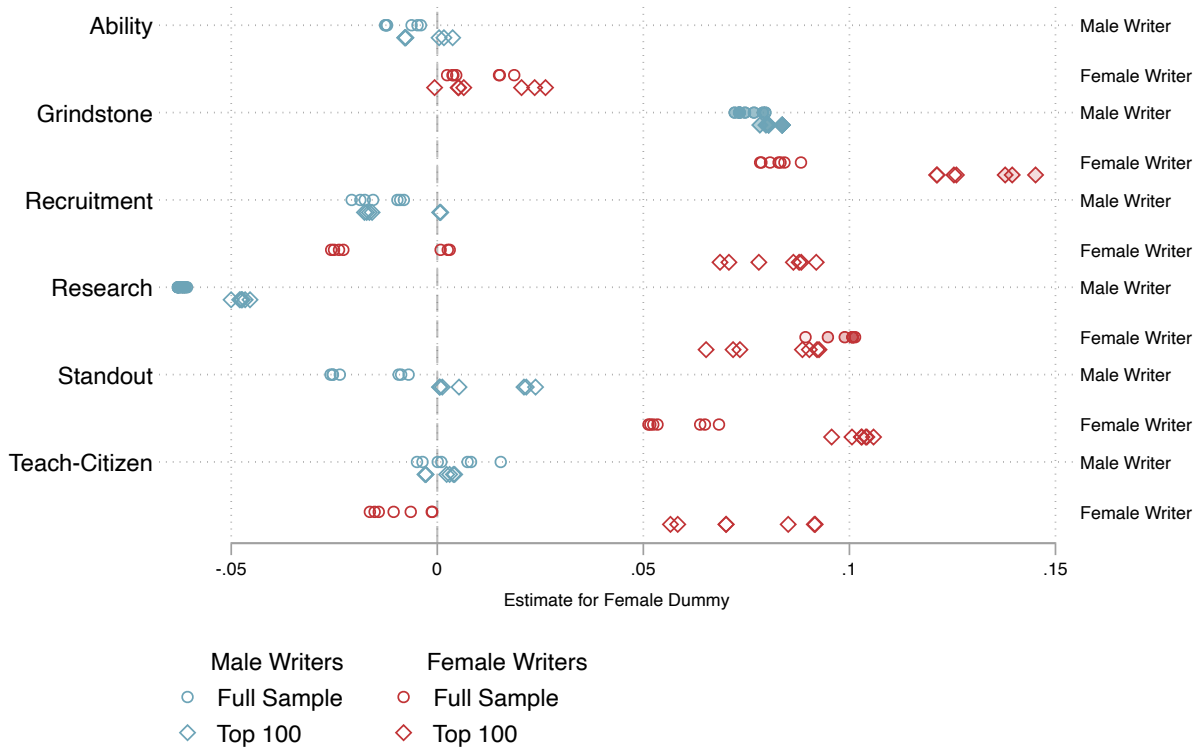
Figure 7: Regression results, all letter writers combined



Notes: This figure shows the coefficient estimates for the regressions specified in 6. We compare different type of specifications, from baseline ones only with candidate controls $\mathbf{X}_i'$ and fixed effects, to the ones with all the controls. The symbol's filling permits visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution,letter-writer's institution, or letter writer), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error. Additional information on the sample and the controls used in each specification are contained in Table 1 and Appendix Table C.1

The pattern uncovered for "grindstone" words continues to hold when we separately consider male and female referees, and independently of the rank of the institution to which they belong. The category where females and males appear to differ most is research, where male writers use significantly fewer research words to describe their female students, whereas the opposite is true for female letter-writers. Some differences appear when we look at "standout" words, which female letter writers are more likely to use to describe female candidates. Across all outcomes, female writers within top-100 institutions also exhibit more outlying patterns, probably reflecting their smaller numbers.

Figure 8: Regression results, by gender of letter writer



Notes: This figure shows the coefficient estimates for the regressions specified in 6, estimated separately for male and female letter-writers. The symbol's filling permit visualizing significance. The symbol's filling permits visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution,letter-writer's institution, or letter writer), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. Additional information on the sample and results for the letter-writer clustering of standard errors are contained in Appendix Section C.2.

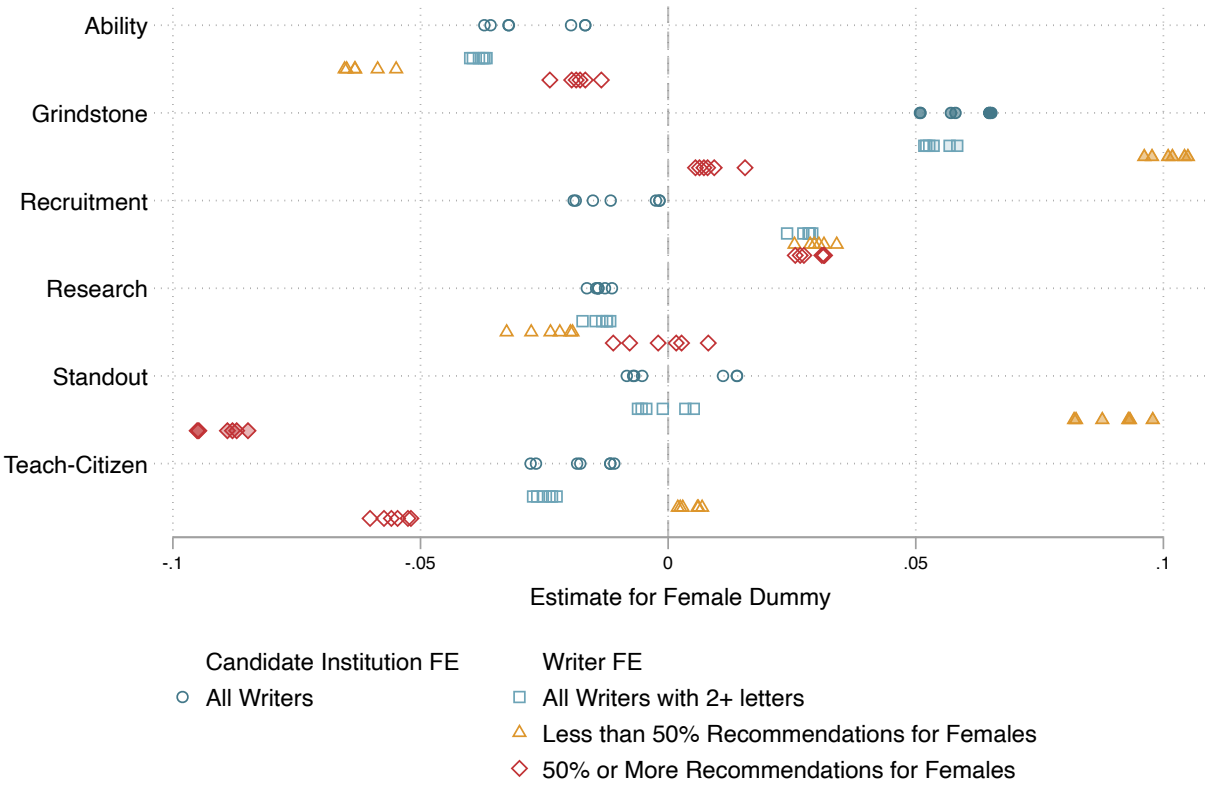
6 Additional Results

6.1 Specifications with Fixed Effects

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

Boustan and Langan (2019) document that female representation is a persistent attribute of economics departments, and that it matters to promote women's careers. Therefore, we need to address whether institutional sorting drives our results. We thus run regressions including fixed effects for the candidates' institution. The results are reported in Figure 9. They suggest that among students from the same cohort, graduating from the same institution —who, for example, were admitted to PhD programs arguably applying the same entry requirements— women are still significantly more likely to be described with "grindstone" terms.

Figure 9: Regression results with candidate institution or writer fixed effects



Notes: This figure shows the coefficient estimates for the regressions specified in 6, estimated separately with candidate institution or letter-writer fixed effects. The symbol's filling permit visualizing significance. Using 2 levels of possible standard error clustering for each fixed effect: none or candidate's institution (resp. letter writer) for candidate's fixed effects (resp. letter-writers' fixed effects). We flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 6 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 17% ($\approx 100/6$) opacity when it reaches significance at each possible level. The darker the symbol the more often they are significant. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. Additional information on the sample and results for the unclustered, robust standard errors are contained in Appendix Section D.1.

We are still concerned that, even within the same graduate program, sorting across letter writers could explain our findings. To address this concern, we run a set of specifications including writer fixed effects.[17] Note that these models are identified from referees who have written two or more letters across all four sample years, with at least one for a female candidate. This significantly reduces our sample (we can include only 22% of the letter-writers). Our results are reported in Figure 9 and they broadly confirm the patterns we have uncovered so far. In the same figure we also separately analyze the sample of referees who have less (more) experience with female candidates (i.e. less (more) than 50% of their references were for women). The "less experienced" group appears to be the one driving the "grindstone" result.[18] Experience may matter for two main reasons. On the one hand, referees may

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

[18] The "less experienced" group accounts for 44% of referees in the subsample with two or more letters and

vary in their perception of women, and female candidates could sort accordingly to avoid stereotyping. On the other hand, it could be that referees do not differ initially, but that their exposure to female candidates leads them to update prior stereotyping (a learning effect also observed for example by Beaman et al., 2009). Further research is needed to disentangle these two mechanisms.

The fixed effects results also uncover a new pattern with regards to "ability". Overall, female candidates are associated with noticeably fewer, albeit insignificantly so, ability terms, with a clearer pattern emerging in the less experienced group.[19]

Finally, we uncover a somewhat puzzling result on the usage of "standout" attributes. While no differential pattern emerges when we look at all referees, our analysis indicates that letter-writers who are less familiar with female candidates tend to use more superlatives to describe them than their more experienced counterparts.

6.2 Heterogeneity by Research Fields

We explore heterogeneity of the results according to the candidate's research field to assess possible sub-cultural differences in the profession.

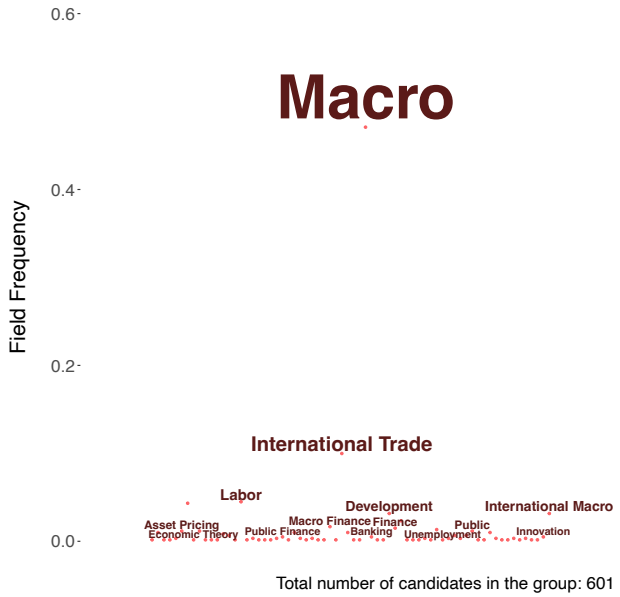
Grouping applicants into meaningful research areas is challenging. On the EJM platform, they typically choose a field, loosely based on a JEL code. Unfortunately, EJM fields do not follow a hierarchical structure. For instance, the EJM field "Development and Growth", which resembles the broad JEL code O: "Economic Development, Innovation, Technological Change, and Growth", is listed alongside "Computational Economics", which resembles the highly specific JEL subcodes C63: "Computational Techniques - Simulation Modeling", C68: "Computable General Equilibrium Models", or D58: "Computable and Other Applied General Equilibrium Models". Moreover, some of the EJM fields pool diverse subgroups of the profession, i.e. scholars that are unlikely to publish in the same journals or participate in the same events (conferences, seminars, etc.). Using the same example, the EJM field "Development and Growth" includes both macroeconomists working on long-run growth and microeconomists carrying out field experiments in developing countries. The lack of a hierarchical structure and the diversity within EJM fields prevents us from meaningfully aggregating this classification into a manageable number of groups.

Given these shortcomings, we employ an unsupervised data-driven approach to classify candidates into three broad research groups. First, from the recommendation letters we extract the text slice that is most likely to discuss the candidates' job market paper. To do so we flag the first instance of the term "job market paper" or "dissertation". We then slice the subsequent 400 words and assemble the research slices from all the recommendation let-

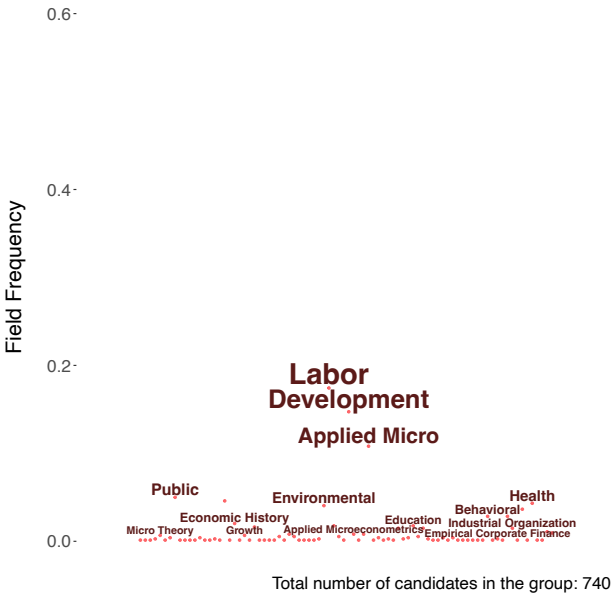
at least one female candidate.

[19]The pattern is significant if we focus on male letter-writers with less experience recommending female candidates. The results are available from the authors upon request.

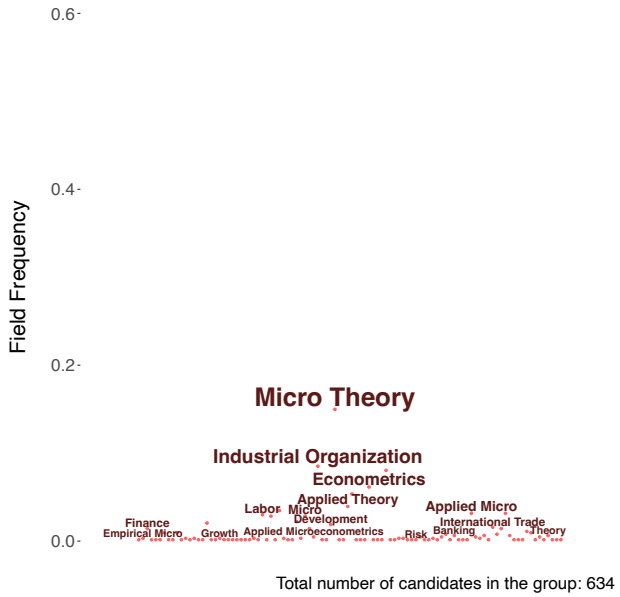
Figure 10: Word clouds for Fields



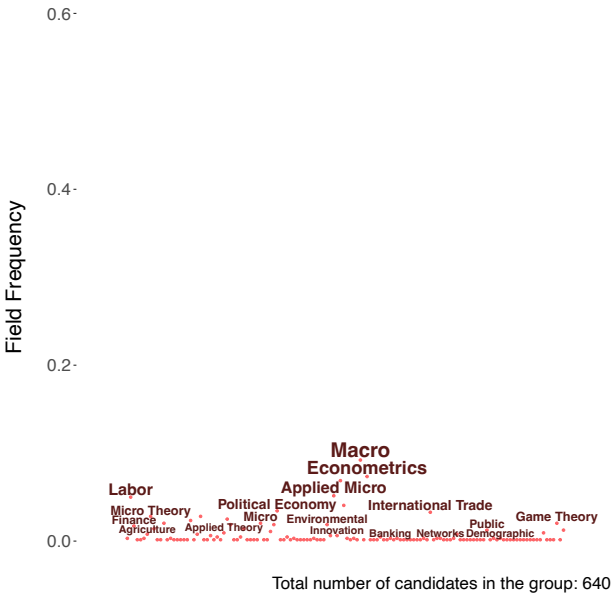
(a) Macro



(b) Applied



(c) Theory



(d) Residual

Notes: The word clouds depict the research fields freely written by candidates for each of the categories. For each category, the y-axis and the font size of the fields reflects its frequency as a primary field in the CVs of candidates that reported them. The fields are randomly distributed across the x-axis.

ters written for the same candidate into a single text.[20] We process these texts as described in section 3.1 and cluster them into four groups using an unsupervised k-means clustering approach.[21]

We assess the credibility of these groupings by highlighting the mapping between them and the self-reported, unstructured primary research field that candidates add to their CV.[22] The

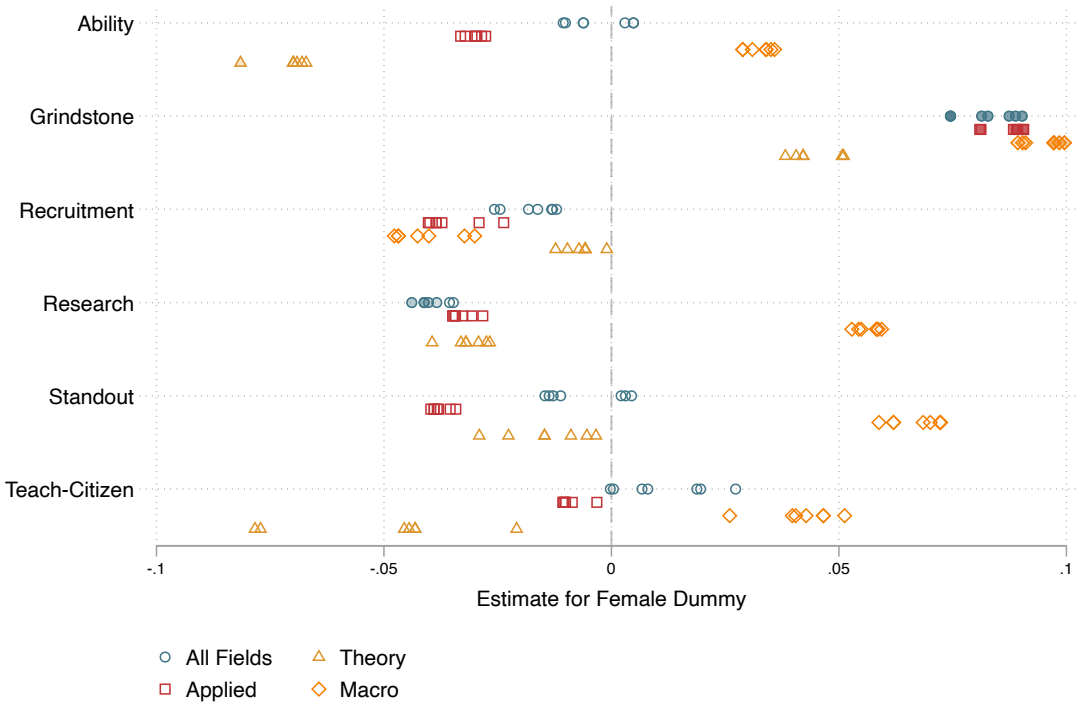
[20] 84% are sliced based on the word "job market paper" and 16% on "dissertation".

[21] For more details about the processing and the choice of cluster numbers, see Appendix section A.2.

[22] 222 distinct fields are reported. While these fields do not necessarily map precisely into an existing JEL

word clouds in Figure 10 show the frequency of the reported main fields for each of the candidates in each broad category. Three clearly identified broad groups emerge: macro, applied, and theory. 47% of candidates report "Macro" as their main field in panel (a). Similarly, applicants listing "Labor", "Development", "Public" or "Applied Micro" make up 45% of those in panel (b); and those indicating "Micro Theory", "Industrial Organization", "Econometrics", "Behavioral", "Applied Theory", "Game Theory" or "Economic Theory" represent 44% of the individuals in panel (c). The clustering procedure also creates a fourth category which we cannot credibly assign to a specific broad area and which as a result has been treated as residual.[23]

Figure 11: Regression results, different candidate research fields



Notes: This figure shows the coefficient estimates for the regressions specified in 6, estimated separately for different (aggregated) research field clusters. We show the three most demanding specifications. The symbol's filling permit visualizing significance. The symbol's filling permits visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution,letter-writer's institution, or letter writer), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. Additional information on the sample and results for the clustered standard errors by letter-writer are contained in Appendix Section D.2.

The results by field are reported in Figure 11. The first important finding is that the association of "grindstone" words with female candidates remains positive across all sub-fields. One possible explanation for the association of women to "grindstone" expressions is that

code, they are typically highly informative when it comes to the actual content of research pursued by the candidates. Moreover, not all candidates report a main field of specialization.

[23] We experiment with alternative definitions of research fields as controls in the baseline regressions in Section 6.3.

they sort into research fields that require more effort, hard work, tenacity and industriousness rather than ability. This set of skills is often associated with empirical work. However, Figure 11 shows that this association remains strong and significant within applied micro, casting doubt on such a hypothesis.

Furthermore, notice also that female candidates are significantly less likely to be praised on their "ability" within the "Theory" field. This finding is worth highlighting as in this field raw talent is arguably valued very highly. This observation sheds new light on earlier findings by Leslie et al. (2015), according to whom academic fields where people particularly emphasize the role of raw talent are characterized by lower female representation.

Finally, in three ("ability", "research", and "standout") of the six outcomes, the coefficients for the "Macro" field are noticeably different in magnitude compared to the other groups. This heterogeneity suggests sub-cultural differences in the profession, which should be further investigated.

6.3 Robustness Checks

Number of words In our baseline analysis we have defined the end of letter using the last 200 words before the 'polite' end phrase. In Appendix Figure E.1 we explore the sensitivity of the baseline results to this choice by experimenting with two alternative cutoffs, using 150 and 250 words. We also study the full reference letters. Our findings are unaffected.

Alternative definitions of research fields Our baseline analysis employs research field fixed effects using the four clusters we obtained in our analysis in Section 6.2. In Figure E.3 we repeat the same exercise using instead the more detailed field definitions from EJM,[24] to analyse the full sample as well as the subsamples by institutional rank. Our findings are robust to this change.

Main Advisors So far we have used all the letters that were submitted for each applicant, i.e. those which were written by the main advisor and those written by other faculty members familiar with the candidate's research. As the main advisor might have better knowledge of the applicant, it is important to investigate whether there are differences in the language he/she used compared to that of the other referees. We collect data on the identity of the letter writers for all candidates who were in the job market up to three years after completing their Ph.D..[25] The results of our analysis are illustrated in Appendix Figure E.4, where we report our baseline estimates for the collected sample and those obtained focusing separately on the letters written by the main advisor and the other reviewers.

[24] There are 25 fields in EJM. We drop those candidates who selected 'Any field'. Results only include specifications (4)-(7) in the full sample since (1)-(3) do not include field FE.

[25] This represents 57% of the sample of candidates. Candidates who defended earlier were less likely to have a letter from their Ph.D. advisor and were also less likely to report that information on their CV.

The findings indicate that the patterns for "grindstone" terms are generally comparable, but accentuated for letter-writers who are *not* the main advisors. Moreover, there is notable divergence in case of "ability", "recruitment" and 'standout", where main advisors appear to use more of these terms for female candidates, compared to other referees. Overall, this analysis presents suggestive evidence that main advisors are writing more favorable letters for women compared to other referees. Main advisors arguably spend more time writing and polishing the letters.[26] an through this lengthy process some of the unconscious stereotyping may be toned down.

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

7 Concluding Remarks

In this paper, we carried out what is to the best of our knowledge the first systematic analysis of recommendation letters in the junior academic job market in economics. Using both supervised and unsupervised methodologies, we have documented the presence of important differences in the language used to describe female applicants. Women are more often described with terms praising their "diligence" or "dedication" than men. This pattern is robust to alternative specifications and holds across many subsamples of the data. Similarly, we uncover evidence of a lower emphasis on ability, especially when comparing individuals within the same institution or for those sharing the same referee. Sociologists have characterised these language patterns as a form of stereotyping, and highlighted the potential negative connotations as a strong emphasis on diligence may imply a lack of "brilliance" (Bourdieu and Passeron, 1977; Valian, 1999).

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

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

References

Sule Alan, Teodora Boneva, and Seda Ertac. Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134 (3):1121–1162, 2019.

Elliott Ash, Daniel L Chen, and Suresh Naidu. Ideas Have Consequences: The Impact of Law and Economics on American Justice. *ETH Mimeo*, 2020.

Elliott Ash, Daniel L Chen, and Arianna Ornaghi. Gender Attitudes in the Judiciary: Evidence from U.S. Circuit Courts. *ETH Mimeo*, 2021.

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

Lori Beaman, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. Powerful Women: Does Exposure Reduce Bias? *The Quarterly Journal of Economics*, 124(4): 1497–1540, 2009. ISSN 0033-5533. doi: 10.1162/QJEC.2009.124.4.1497.

Anne Boring. Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41, 2017. ISSN 00472727. doi: 10.1016/j.jpubeco.2016.11.006.

Anne Boring and Arnaud Philippe. Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching. *Journal of Public Economics*, 193:104323, 2021.

Clément Bosquet, Pierre-Philippe Combes, and Cecilia García-Peñalosa. Gender and promotions: evidence from academic economists in france. *The Scandinavian Journal of Economics*, 121(3):1020–1053, 2019.

Pierre Bourdieu and Jean-Calude Passeron. *Reproduction in Education, Society, and Culture*. Sage, 1977.

Leah Boustan and Andrew Langan. Variation in women's success across PhD programs in economics. *Journal of Economic Perspectives*, 33(1):23–42, 2019. ISSN 08953309. doi: 10.1257/jep.33.1.23.

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

Stephen J Ceci and Wendy M Williams. *The mathematics of sex: How biology and society conspire to limit talented women and girls*. Oxford University Press, 2009.

Peter Coles, John Cawley, Pillip B. Levine, Muriel Niederle, Alvin E. Roth, and John J. Siegfried. The job market for new econonomists: A market design perspective. *Journal of Economic Perspectives*, 24(4):187– 205, 2010.

Gianni De Fraja, Giovanni Facchini, and John Gathergood. Academic salaries and public evaluation of university research: Evidence from the UK Research Excellence Framework. *Economic Policy*, 34:523–583, 2019.

Pascaline Dupas, Alice Sasser Modestino, Muriel Niederle, Justin Wolfers, and The Seminar Dynamics Collective. Gender and the Dynamics of Economics Seminars. *NBER Working Paper*, (28494), 2021.

Kuheli Dutt, Danielle L Pfaff, Ariel F Bernstein, Joseph S Dillard, and Caryn J Block. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience*, 9(11):805–808, 2016.

Y Fan, LJ Shepherd, E Slavich, D Waters, M Stone, R Abel, and EL Johnston. Gender and cultural bias in student evaluations: Why representation matters. *PloS one*, 14(2):e0209749, 2019.

Donna K Ginther and Shulamit Kahn. Women in economics: moving up or falling off the academic career ladder? *Journal of Economic Perspectives*, 18(3):193–214, 2004.

Erving Goffman. *Gender Advertisements*. Harper and Row, New York, 1 edition, 1979.

Vivian Gornick. Introduction to"Gender Advertisements" by Erving Goffman. In *Gender Advertisement*, pages vii–ix. Harper and Row, 1979.

Justin Grimmer and Brandon M Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, 2013.

Shoshana Grossbard, Tansel Yilmazer, and Lingrui Zhang. The gender gap in citations of articles published in two demographic economics journals. *Review of Economics of the Household*, 19(3):677–697, 2021.

Mikki Hebl, Christine Nittrouer, Abigail Corrington, and Juan Madera. How we describe male and female job applicants differently. *Harvard Business Review*, 27, 2018.

Erin Hengel. Publishing while female. *Economic Journal*, forthcoming.

Laura Hospido and Carlos Sanz. Gender gaps in the evaluation of research: evidence from submissions to economics conferences. *Oxford Bulletin of Economics and Statistics*, 83(3): 590–618, 2021.

Shulamit Kahn and Donna Ginther. Women and stem. *NBER WP Series*, 23525, 2017.

Marlène Koffi. Innovative ideas and gender inequality. Technical report, Working Paper Series, 2021.

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

Shelly Lundberg, editor. *Women in Economics*. CEPR, 2020.

Shelly Lundberg and Jenna Stearns. Women in economics: Stalled progress. *Journal of Economic Perspectives*, 33(1):3–22, 2019. doi: 10.1257/jep.33.1.3.

Heather J. MacArthur, Jessica L. Cundiff, and Matthias R. Mehl. Estimating the Prevalence of Gender-Biased Language in Undergraduates' Everyday Speech. *Sex Roles*, 82(1-2):81–93, 2020.

Lillian MacNell, Adam Driscoll, and Andrea N Hunt. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303, 2015.

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

Juan M Madera, Michelle R Hebl, Heather Dial, Randi Martin, and Virgina Valian. Raising doubt in letters of recommendation for academia: Gender differences and their impact. *Journal of Business and Psychology*, 34(3):287–303, 2019.

Friederike Mengel, Jan Sauermann, and Ulf Zölitz. Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2):535–566, 2019. ISSN 15424774. doi: 10.1093/jeea/jvx057.

Christine L. Nittrouer, Michelle R. Hebl, Leslie Ashburn-Nardo, Rachel C.E. Trump-Steele, David M. Lane, and Virginia Valian. Gender disparities in colloquium speakers at top universities. *Proceedings of the National Academy of Sciences*, (1):104–108, 2018. ISSN 0027-8424. doi: 10.1073/PNAS.1708414115.

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

Toni Schmader, Jessica Whitehead, and Vicki H. Wysocki. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, 57(7-8):509–514, 2007. ISSN 03600025. doi: 10.1007/s11199-007-9291-4.

Yijun Shao, Stephanie Taylor, Nell Marshall, Craig Morioka, and Qing Zeng-Treitler. Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2874–2878, 2018. doi: 10.1109/BigData.2018.8622345.

Frances Trix and Carolyn Psenka. Exploring the color of glass: letters of recommendation for female and male faculty. *Discourse & Society*, 14:191–220, 2003.

Virginia Valian. *Why so slow? The advancement of women*. MIT press, 1999.

Virginia Valian. Beyond Gender Schemas: Improving the Advancement of Women in Academia. *Hypatia*, 20(3):198–213, 2005. ISSN 1527-2001.

Alice H Wu. Gendered language on the economics job market rumors forum. In *American Economic Review Papers & Proceedings*, volume 108, pages 175–79, 2018.

A Variable and methods description

A.1 Validation exercise

To construct Figure 5 we assess the correspondence between the validators' chosen categories and ours as follows. Within each of the authors' chosen categories, for each word, we identify the category chosen by a plurality of validators. In the case of ties (e.g. "diligent", which the authors classified as "grindstone", was classified by 28.5% of validators as "ability", and 28.5% as "grindstone"), we attribute that word to both categories ("diligent" is attributed both to "ability" and "grindstone"). For each of our chosen categories, Figure 5 presents the distribution of winning categories. Words for which there are two winning categories count twice in the total, so that the sum of the bars is equal to 1.

Table A.1: Summary Statistics of words in each category

Category	Av. Doc Freq	Av. TF-IDF (x 1000)	N Words	Av. Validators per Word
Ability	8896.56	5.77	57	6.98
Grindstone	8991.60	4.99	20	6.56
Recruitment	9011.69	4.32	118	6.72
Research	9038.12	4.77	210	6.46
Standout	8958.87	5.02	106	6.70
Teach-Citizen	8971.88	5.06	94	6.81

Notes: This table shows summary statistics of words in each category. The First column gives the categories. The second (third) columns give the average TF-IDF (document frequency) of words in each category. The fourth column gives the number of words in each category. The fifth column gives the average number of validators who cross-validated our categorisation for each word.

A.2 Research Fields

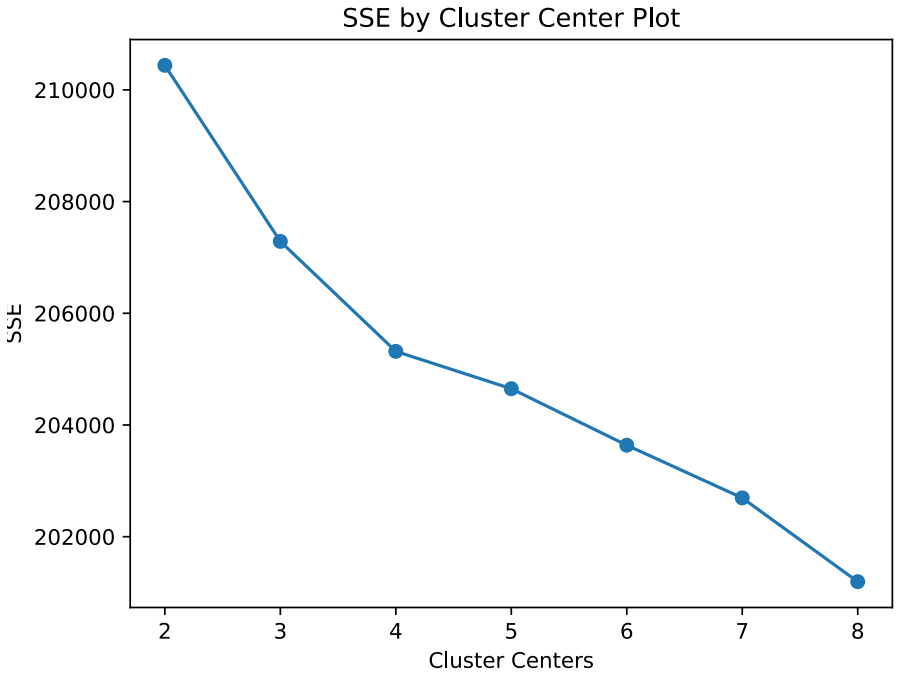
This section describes how we aggregate candidates into three main "research groups". First, we extract from the recommendation letters an excerpt that is highly likely to discuss the candidates' job market paper. To do so, we flag the first instance of the term "job market paper" or "dissertation". We then slice the subsequent 400 words, and assemble the research slices from all the recommendation letters written for the same candidate into a single text.

In the same way as described in section 3.1 we stem the terms and discard stop words. Given that the objective of this procedure is to group texts that use similar terms, we deploy a different approach when transforming the text into a database. Instead of computing the tfidf, which would give more weight to terms that are more frequently used in a document compared to the rest of the corpus, we just use a binary representation in which a term is given a value equal to one if it appears in the text. This approach allows us to more easily identify the research texts that contain broad terms that could characterise a field (e.g. "macro", "Nash equilibrium", "causality"), rather than singling out terms used multiple times to describe the job market paper, but that could be very specific to a particular piece of research (e.g. "assortative matching", "babbling equilibiria"). Finally, following common recommenda-

tions for k-means clustering, we reduce the dimensionality of the problem by carrying out a PCA.

Figure A.1 shows the SSE of the k-means clustering procedure as a function of the number of clusters chosen. We identify a kink at four clusters, hence, using the "elbow method", that is the final number of clusters we select in our analysis.

Figure A.1: Most common words by research cluster



Notes: This figure presents the SSE of the k-means clustering procedure as a function of the number of clusters used to group candidates into research fields.

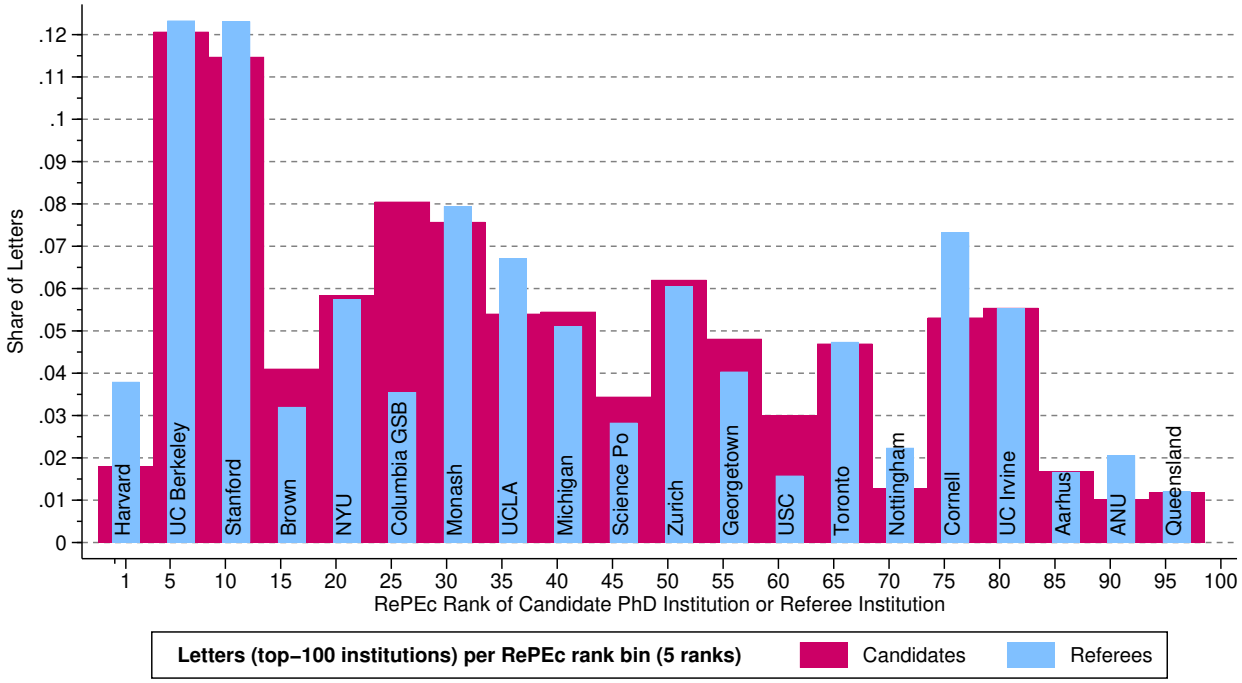
A.3 Institutional Ranking

We used the Research Papers in Economics (RePEc) ranking for the top 5% of economic institutions (version January 2021, see https://ideas.repec.org/top/top.inst.all.html for the current version)[27] as our guide to rank writer and candidate institutions. We drop three research organisations (NBER, IZA, CEPR) but keep international institutions like the IMF as well as the Federal Reserve Banks in the rankings since referees from these institutions are not uncommon. Writer institutional affiliation is collected from their CV via manual internet search and manually matched to the RePEc institutions. We categorise writers into bands on the basis of their institutional ranking: 1-25, 26-50, 51-100, 101-200, 201 and higher. We are missing RePEc-listed affiliation and hence rankings for around 18% of writers, but these only account for 13% of our sample of reference letter. The rank of candidate PhD-institutions was developed in analogy.

[27] The RePEc ranking refers to the top 10% but only the top 5% are ranked, the remainder are unranked within the percentile (all those within the 6th percentile, all those within the 7th percentile, etc).

B Additional Descriptive Statistics

Figure B.1: RePEc Rank of Candidate and Letter Writer Institution, Zooming into Top-100 institutions



Notes: The figure presents the frequency distribution of candidate and letter writer institution rank, zooming in on the top-100 (bin width 5 ranks), highlighting one institutions for each bin.

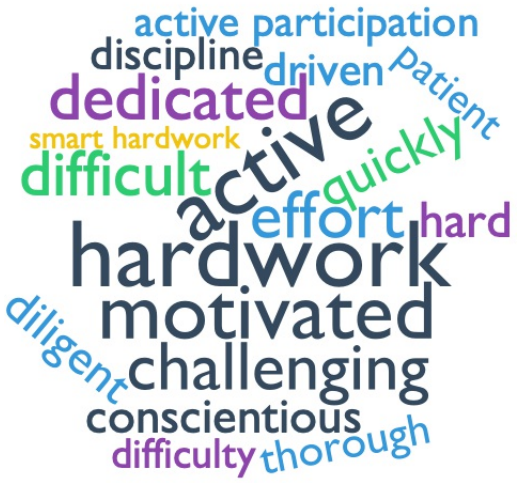
Table B.1: Descriptive Statistics: Candidate Country

ISO3	Candidates	Percent	Cum	ISO3	Candidates	Percent	Cum
USA	1,385	49.6	50	HUN	7	0.3	98
GBR	403	14.4	64	CHN	6	0.2	98
CAN	146	5.2	69	BRA	5	0.2	99
FRA	141	5.1	74	FIN	5	0.2	99
DEU	133	4.8	79	GRC	5	0.2	99
ESP	113	4.1	83	PSE	4	0.1	99
ITA	98	3.5	87	IND	3	0.1	99
NLD	82	2.9	90	IRL	3	0.1	99
SWE	52	1.9	91	ISR	3	0.1	99
AUS	34	1.2	93	MEX	3	0.1	100
CHE	34	1.2	94	TUR	3	0.1	100
BEL	31	1.1	95	CHL	2	0.1	100
HKG	18	0.6	96	JPN	2	0.1	100
SGP	14	0.5	96	RUS	2	0.1	100
NOR	13	0.5	97	CYP	1	0.0	100
DNK	12	0.4	97	GEO	1	0.0	100
AUT	9	0.3	97	IRN	1	0.0	100
PRT	8	0.3	98	MYS	1	0.0	100
CZE	7	0.3	98	NZL	1	0.0	100
				Total	**2,791**		

Notes: Isocode for geographic location of the applicant (PhD institution, not nationality), in order of magnitude. Cum – cumulative sum.

Figure B.2: Word clouds for each sentiment

(a) Ability

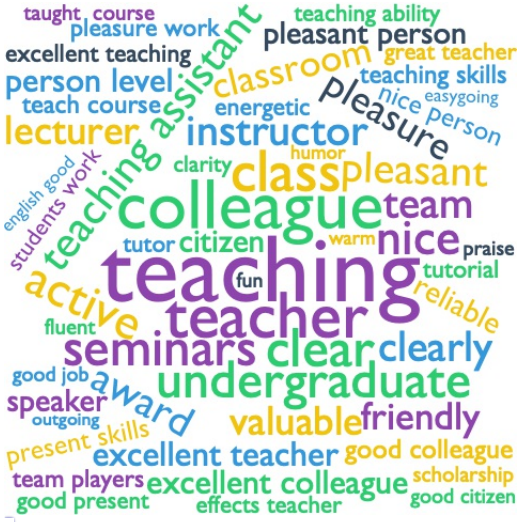


(b) Grindstone

(c) Recruitment

(d) Research

(e) Standout



(f) Teach and Citizen

Notes: The word clouds depict the expressions attributed to each sentiment. The size of the word is illustrative of its document frequency. Within each cloud, larger words are more common in the corpus. The size of the words should not be compared across wordclouds, as the font sizes are adjusted to improve legibility.

Table B.2: Descriptive Statistics — candidate and letter writer characteristics

	N	Mean	SD	Min	Max
Cohort year	9154	2018.566	1.116	2017	2020
Candidate					
Female candidate	9154	0.294	0.456	0	1
Years since graduation	9115†	0.998	2.026	0	16
dto missing	9154	0.004	0.065	0	1
PhD institution top-25	9154	0.185	0.388	0	1
PhD institution top-26 to 50	9154	0.137	0.344	0	1
PhD institution top-51 to 100	9154	0.149	0.356	0	1
PhD institution top-101 to 200	9154	0.198	0.398	0	1
No ranking info	9154	0.110	0.313	0	1
Publication Count	9154	1.174	2.012	0	18
Top-5 count	9154	0.015	0.126	0	2
Top-Field count	9154	0.049	0.232	0	2
Top-General count	9154	0.022	0.160	0	2
Research Field: Theory	9154	0.237	0.425	0	1
Research Field: Applied Micro	9154	0.372	0.483	0	1
Research Field: Macro	9154	0.392	0.488	0	1
Referee					
Referee institution top-25	9154	0.203	0.402	0	1
Referee institution top-26 to 50	9154	0.126	0.331	0	1
Referee institution top-51 to 100	9154	0.165	0.371	0	1
Referee institution top-101 to 200	9154	0.189	0.392	0	1
No ranking info	9154	0.122	0.327	0	1
Referee letter count	9154	3.382	2.783	1	18
Female referee	9154	0.144	0.351	0	1
Letter					
Letter length (ln characters)	9154	8.796	0.492	7.157	10.455
Dependent variable zero (6 bags)	9154	0.886	0.806	0	3

Notes: Notes: Full sample descriptives. Institutional ranking of letter writer and candidate are determined based on RePEc ranking. Top-Field journals are JIE, JET, JoE, JME, JPubE, JLE, JDE, JEH, JFE, JF, Rand. Top-General (interest) are the JEEA, REStat, EJ, IER, and all AEJs. The Research Fields are based on unsupervised cluster analysis of the section of the reference letter discussing the JMP (see data section for all details). We exclude letters for which four or more of the six bags have value 0. † The 39 candidate-letters (0.4% of the sample) for which this information is missing are coded as -999 (not included in these descriptives) and are picked up by the 'missing' dummy.

C Results Tables

C.1 Baseline results — Letter ends — all writers

Table C.1: Sentiments — End of Letters (200 words) — By Writer Institutional Rank

Writer RePEc Rank	(1) Top-25	(2) Top-50	(3) Top-100	(4) 26-100
Female Candidate Coefficient				
Ability	-0.0064	0.0127	-0.0034	0.0101
	(0.12)	(0.31)	(0.10)	(0.23)
Grindstone	0.1025	0.1098	0.0887	0.0858
	(1.95)*	(2.64)***	(2.66)***	(1.98)**
Recruitment	0.0264	-0.0522	-0.0012	-0.0234
	(0.50)	(1.22)	(0.04)	(0.52)
Research	-0.0191	0.0032	-0.0300	-0.0403
	(0.37)	(0.08)	(0.90)	(0.92)
Standout	-0.0564	-0.0393	0.0159	0.0618
	(1.07)	(0.96)	(0.48)	(1.43)
Teaching & Citizenship	-0.0371	-0.0208	0.0184	0.0675
	(0.70)	(0.51)	(0.54)	(1.50)
FE/Variables absorbed	18	19	21	19
Additional covariates	7	7	7	7
Number of Letters	1861	3011	4520	2659
dto for females	550	888	1321	771
Number of candidates	871	1276	1777	1211
dto female	257	362	509	346
Number of writers	818	1304	1998	1180
dto female	118	204	315	197
Letters by fem writers	226	389	583	357
Year FE	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes
Institution Rank (Band) FE	n/a	yes	yes	yes
Years since PhD	yes	yes	yes	yes
Research Field FE	yes	yes	yes	yes
Publications	yes	yes	yes	yes
Writer characteristics	yes	yes	yes	yes
Letter length	yes	yes	yes	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns indicate the different samples related to RePEc ranking of the writer's institution. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Figure 7 in the maintext.

C.2 Baseline results — Letter ends — by letter writer gender

Table C.2: Sentiments — End of Letters (200 words) — Male Writers — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	-0.0040	-0.0048	-0.0062	-0.0125	-0.0122	-0.0127	-0.0123
	(0.16)	(0.19)	(0.24)	(0.48)	(0.47)	(0.49)	(0.48)
Grindstone	0.0796	0.0768	0.0790	0.0747	0.0722	0.0732	0.0734
	(3.06)***	(2.95)***	(3.04)***	(2.86)***	(2.75)***	(2.80)***	(2.81)***
Recruitment	-0.0097	-0.0090	-0.0081	-0.0208	-0.0187	-0.0176	-0.0156
	(0.37)	(0.34)	(0.31)	(0.79)	(0.71)	(0.67)	(0.60)
Research	-0.0629	-0.0616	-0.0612	-0.0625	-0.0607	-0.0620	-0.0629
	(2.48)**	(2.42)**	(2.41)**	(2.45)**	(2.37)**	(2.42)**	(2.47)**
Standout	-0.0088	-0.0070	-0.0094	-0.0259	-0.0254	-0.0253	-0.0236
	(0.34)	(0.27)	(0.36)	(1.00)	(0.98)	(0.98)	(0.92)
Teaching &	0.0154	0.0082	0.0074	-0.0036	-0.0049	0.0001	0.0010
Citizenship	(0.58)	(0.31)	(0.28)	(0.13)	(0.19)	(0.00)	(0.04)
FE/Variables absorbed	9	14	14	17	17	22	22
Additional covariates			1	1	5	6	7
Number of Letters	7839	7839	7839	7839	7839	7839	7839
dto for females	2192	2192	2192	2192	2192	2192	2192
Number of candidates	2775	2775	2775	2775	2775	2775	2775
dto female	806	806	806	806	806	806	806
Number of writers	3927	3927	3927	3927	3927	3927	3927
dto female	0	0	0	0	0	0	0
Letters by fem writers	0	0	0	0	0	0	0
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.
The sample is restricted to male letter writers.

Return to Figure 8 in the maintext.

Table C.3: Sentiments — End of Letters (200 words) — Female Writers — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0150	0.0187	0.0152	0.0037	0.0024	0.0039	0.0046
	(0.26)	(0.33)	(0.27)	(0.06)	(0.04)	(0.07)	(0.08)
Grindstone	0.0882	0.0833	0.0808	0.0828	0.0842	0.0786	0.0783
	(1.51)	(1.44)	(1.39)	(1.40)	(1.42)	(1.35)	(1.34)
Recruitment	0.0026	0.0008	0.0030	-0.0228	-0.0257	-0.0251	-0.0238
	(0.04)	(0.01)	(0.05)	(0.38)	(0.43)	(0.42)	(0.40)
Research	0.0894	0.0989	0.0948	0.1014	0.1007	0.1008	0.1008
	(1.49)	(1.66)*	(1.59)	(1.69)*	(1.68)*	(1.69)*	(1.69)*
Standout	0.0684	0.0649	0.0638	0.0517	0.0512	0.0524	0.0534
	(1.19)	(1.13)	(1.11)	(0.89)	(0.89)	(0.92)	(0.95)
Teaching &	-0.0064	-0.0164	-0.0152	-0.0142	-0.0106	-0.0012	-0.0013
Citizenship	(0.10)	(0.27)	(0.25)	(0.23)	(0.17)	(0.02)	(0.02)
FE/Variables absorbed	9	14	14	17	17	22	22
Additional covariates			1	1	5	6	7
Number of Letters	1315	1315	1315	1315	1315	1315	1315
dto for females	503	503	503	503	503	503	503
Number of candidates	1057	1057	1057	1057	1057	1057	1057
dto female	379	379	379	379	379	379	379
Number of writers	778	778	778	778	778	778	778
dto female	778	778	778	778	778	778	778
Letters by fem writers	1315	1315	1315	1315	1315	1315	1315
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.
The sample is restricted to female letter writers.

Return to Figure 8 in the maintext.

Table C.4: Sentiments — End of Letters (200 words) — Male Writers by Institutional Rank

Male Writer RePEc Rank	(1) Top-25	(2) Top-50	(3) Top-100	(4) 26-100
Female Candidate Coefficient				
Ability	-0.0178	0.0000	-0.0079	0.0106
	(0.31)	(0.00)	(0.22)	(0.22)
Grindstone	0.0618	0.0969	0.0803	0.1033
	(1.10)	(2.15)**	(2.23)**	(2.18)**
Recruitment	-0.0238	-0.0725	-0.0165	-0.0183
	(0.43)	(1.59)	(0.45)	(0.37)
Research	0.0010	0.0067	-0.0473	-0.0803
	(0.02)	(0.15)	(1.33)	(1.70)*
Standout	-0.0612	-0.0592	0.0006	0.0420
	(1.09)	(1.34)	(0.02)	(0.90)
Teaching & Citizenship	-0.0863	-0.0454	0.0039	0.0770
	(1.51)	(1.04)	(0.11)	(1.59)
FE/Variables absorbed	17	18	20	18
Additional covariates	7	7	7	7
Number of Letters	1635	2622	3937	2302
dto for females	463	740	1091	628
Number of candidates	820	1220	1715	1154
dto female	235	344	484	322
Number of writers	700	1100	1683	983
dto female	0	0	0	0
Letters by fem writers	0	0	0	0
Year FE	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes
Institution Rank (Band) FE	n/a	yes	yes	yes
Years since PhD	yes	yes	yes	yes
Research Field FE	yes	yes	yes	yes
Publications	yes	yes	yes	yes
Writer characteristics	yes	yes	yes	yes
Letter length	yes	yes	yes	yes

Notes: In the results presented in this table we exclude letters written by female writers from the sample. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. See Table **??** for all other details. The sample is restricted to male writers.

Return to Figure 8 in the maintext.

Table C.5: Sentiments — End of Letters (200 words) — Female Writers by Institutional Rank

Female Writer RePEc Rank	(1) Top-25	(2) Top-50	(3) Top-100	(4) 26-100
Female Candidate Coefficient				
Ability	0.0252	0.0471	0.0236	0.0108
	(0.20)	(0.45)	(0.28)	(0.09)
Grindstone	0.3238	0.1649	0.1213	-0.0108
	(1.92)*	(1.45)	(1.37)	(0.10)
Recruitment	0.2362	0.0406	0.0686	-0.0075
	(1.37)	(0.32)	(0.71)	(0.07)
Research	-0.0728	0.0155	0.0926	0.2069
	(0.48)	(0.13)	(0.99)	(1.77)*
Standout	-0.0443	0.0470	0.1006	0.1879
	(0.31)	(0.42)	(1.18)	(1.69)*
Teaching & Citizenship	0.2727	0.0804	0.0915	0.0386
	(2.20)**	(0.70)	(0.99)	(0.32)
FE/Variables absorbed	16	17	20	18
Additional covariates	7	7	7	7
Number of Letters	226	388	583	357
dto for females	87	148	230	143
Number of candidates	208	333	495	299
dto female	78	121	182	114
Number of writers	118	204	315	197
dto female	118	204	315	197
Letters by fem writers	226	388	583	357
Year FE	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes
Institution Rank (Band) FE	n/a	yes	yes	yes
Years since PhD	yes	yes	yes	yes
Research Field FE	yes	yes	yes	yes
Publications	yes	yes	yes	yes
Writer characteristics	yes	yes	yes	yes
Letter length	yes	yes	yes	yes

Notes: In the results presented in this table we exclude letters written by male writers from the sample. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. See Table **??** for all other details. The sample is restricted to female letter writers.

Return to Figure 8 in the maintext.

D Additional Results

D.1 Additional results — Letter ends — Fixed Effects

Table D.1: Sentiments — End of Letters (200 words) — Candidate Institution FE — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	-0.0175	-0.0175	-0.0203	-0.0328	-0.0327	-0.0376	-0.0360
	(0.66)	(0.66)	(0.77)	(1.23)	(1.23)	(1.41)	(1.35)
Grindstone	0.0640	0.0640	0.0644	0.0573	0.0564	0.0502	0.0504
	(2.45)**	(2.45)**	(2.46)**	(2.17)**	(2.13)**	(1.88)*	(1.89)*
Recruitment	-0.0005	-0.0005	-0.0012	-0.0176	-0.0173	-0.0138	-0.0095
	(0.02)	(0.02)	(0.05)	(0.67)	(0.65)	(0.52)	(0.36)
Research	-0.0134	-0.0134	-0.0135	-0.0157	-0.0138	-0.0106	-0.0123
	(0.52)	(0.52)	(0.52)	(0.60)	(0.53)	(0.40)	(0.47)
Standout	0.0124	0.0124	0.0096	-0.0087	-0.0083	-0.0099	-0.0061
	(0.48)	(0.48)	(0.37)	(0.33)	(0.32)	(0.38)	(0.23)
Teaching &	-0.0117	-0.0117	-0.0108	-0.0177	-0.0182	-0.0277	-0.0264
Citizenship	(0.46)	(0.46)	(0.42)	(0.69)	(0.71)	(1.08)	(1.03)
FE/Variables absorbed	229	229	229	232	232	238	238
Additional covariates			1	1	5	6	7
Number of Letters	7453	7453	7453	7453	7453	7453	7453
dto for females	2360	2360	2360	2360	2360	2360	2360
Number of candidates	2266	2266	2266	2266	2266	2266	2266
dto female	713	713	713	713	713	713	713
Number of writers	3828	3828	3828	3828	3828	3828	3828
dto female	618	618	618	618	618	618	618
Letters by fem writers	1059	1059	1059	1059	1059	1059	1059
Year FE	yes	yes	yes	yes	yes	yes	yes
Writer FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	yes	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Figure 9 in the maintext.

Table D.2: Sentiments — End of Letters (200 words) — Writer FE for 2+ papers — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	-0.0382	-0.0371	-0.0400	-0.0394	-0.0375	-0.0375	-0.0367
	(1.16)	(1.12)	(1.21)	(1.19)	(1.13)	(1.13)	(1.11)
Grindstone	0.0584	0.0568	0.0536	0.0517	0.0521	0.0521	0.0527
	(1.73)*	(1.68)*	(1.58)	(1.53)	(1.53)	(1.53)	(1.55)
Recruitment	0.0291	0.0284	0.0286	0.0240	0.0273	0.0273	0.0284
	(0.99)	(0.97)	(0.97)	(0.81)	(0.93)	(0.93)	(0.97)
Research	-0.0173	-0.0146	-0.0124	-0.0117	-0.0123	-0.0123	-0.0133
	(0.53)	(0.45)	(0.38)	(0.36)	(0.38)	(0.38)	(0.41)
Standout	0.0052	0.0035	-0.0011	-0.0054	-0.0061	-0.0061	-0.0044
	(0.16)	(0.11)	(0.03)	(0.16)	(0.18)	(0.18)	(0.13)
Teaching &	-0.0234	-0.0239	-0.0253	-0.0225	-0.0273	-0.0273	-0.0264
Citizenship	(0.76)	(0.78)	(0.82)	(0.72)	(0.88)	(0.88)	(0.85)
FE/Variables absorbed	1040	1045	1045	1048	1048	1048	1048
Additional covariates			1	1	5	6	7
Number of Letters	3812	3812	3812	3812	3812	3812	3812
dto for females	1524	1524	1524	1524	1524	1524	1524
Number of candidates	2008	2008	2008	2008	2008	2008	2008
dto female	688	688	688	688	688	688	688
Number of writers	1035	1035	1035	1035	1035	1035	1035
dto female	150	150	150	150	150	150	150
Letters by fem writers	500	500	500	500	500	500	500
Year FE	no	no	no	no	no	no	no
Writer FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively. The sample includes only those letters from writers with two or more references for at least one male and one female candidate (gender mix).

Return to Figure 9 in the maintext.

Table D.3: Sentiments — End of Letters — Writer FE for 2+ papers, writers 'more familiar' with female candidates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	-0.0167	-0.0135	-0.0178	-0.0186	-0.0195	-0.0195	-0.0239
	(0.35)	(0.28)	(0.37)	(0.38)	(0.40)	(0.40)	(0.49)
Grindstone	0.0155	0.0080	0.0055	0.0072	0.0093	0.0093	0.0063
	(0.32)	(0.17)	(0.11)	(0.15)	(0.19)	(0.19)	(0.13)
Recruitment	0.0274	0.0316	0.0310	0.0267	0.0314	0.0314	0.0257
	(0.65)	(0.73)	(0.72)	(0.62)	(0.73)	(0.73)	(0.60)
Research	-0.0111	-0.0078	-0.0020	0.0016	0.0027	0.0027	0.0081
	(0.23)	(0.16)	(0.04)	(0.03)	(0.06)	(0.06)	(0.17)
Standout	-0.0948	-0.0848	-0.0871	-0.0880	-0.0890	-0.0890	-0.0951
	(1.95)*	(1.74)*	(1.78)*	(1.80)*	(1.81)*	(1.81)*	(1.94)*
Teaching & Citizenship	-0.0574	-0.0519	-0.0546	-0.0525	-0.0559	-0.0559	-0.0602
	(1.28)	(1.16)	(1.21)	(1.16)	(1.23)	(1.23)	(1.33)
FE/Variables absorbed	590	595	595	598	598	598	598
Additional covariates			1	1	5	6	7
Number of Letters	1655	1655	1655	1655	1655	1655	1655
dto for females	931	931	931	931	931	931	931
Number of candidates	1136	1136	1136	1136	1136	1136	1136
dto female	572	572	572	572	572	572	572
Number of writers	585	585	585	585	585	585	585
dto female	103	103	103	103	103	103	103
Letters by fem writers	292	292	292	292	292	292	292
Year FE	no	no	no	no	no	no	no
Writer FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively. Sample includes only those letters from writers with two or more references for at least one male and one female candidate (gender mix). We further limit the sample to those referees who have written the same number or more letters for females than for males.

Return to Figure 9 in the maintext.

Table D.4: Sentiments — End of Letters — Writer FE for 2+ papers, writers 'less familiar' with female candidates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	-0.0649	-0.0633	-0.0653	-0.0632	-0.0586	-0.0586	-0.0549
	(1.42)	(1.38)	(1.43)	(1.38)	(1.27)	(1.27)	(1.19)
Grindstone	0.1043	0.1050	0.1018	0.0961	0.0977	0.0977	0.1010
	(2.19)**	(2.19)**	(2.12)**	(2.00)**	(2.03)**	(2.03)**	(2.10)**
Recruitment	0.0315	0.0295	0.0304	0.0255	0.0287	0.0287	0.0340
	(0.76)	(0.72)	(0.74)	(0.62)	(0.69)	(0.69)	(0.82)
Research	-0.0219	-0.0193	-0.0197	-0.0238	-0.0276	-0.0276	-0.0326
	(0.49)	(0.43)	(0.44)	(0.52)	(0.61)	(0.61)	(0.72)
Standout	0.0979	0.0933	0.0877	0.0820	0.0824	0.0824	0.0929
	(2.16)**	(2.05)**	(1.92)*	(1.79)*	(1.81)*	(1.81)*	(2.07)**
Teaching &	0.0059	0.0025	0.0029	0.0069	0.0020	0.0020	0.0061
Citizenship	(0.14)	(0.06)	(0.07)	(0.16)	(0.05)	(0.05)	(0.14)
FE/Variables absorbed	455	460	460	463	463	463	463
Additional covariates			1	1	5	6	7
Number of Letters	2157	2157	2157	2157	2157	2157	2157
dto for females	593	593	593	593	593	593	593
Number of candidates	1423	1423	1423	1423	1423	1423	1423
dto female	388	388	388	388	388	388	388
Number of writers	450	450	450	450	450	450	450
dto female	47	47	47	47	47	47	47
Letters by fem writers	208	208	208	208	208	208	208
Year FE	no	no	no	no	no	no	no
Writer FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively. Sample includes only those letters from writers with two or more references for at least one male and one female candidate (gender mix). We limit the sample to those referees who have written fewer letters for females than for males.

Return to Figure 9 in the maintext.

D.2 Additional results — Letter ends — Candidate Research Fields

Table D.5: Sentiments — End of Letters (200 words) — By Candidate Research Field

Broad Research Field	(1) All Fields	(2) Theory	(3) Applied Micro	(4) Macro
Female Candidate Coefficient				
Ability	-0.0101	-0.0680	-0.0322	0.0339
	(0.43)	(1.31)	(0.82)	(0.64)
Grindstone	0.0746	0.0421	0.0812	0.0904
	(3.12)***	(0.77)	(1.97)**	(1.68)*
Recruitment	-0.0162	-0.0071	-0.0237	-0.0323
	(0.69)	(0.13)	(0.59)	(0.64)
Research	-0.0356	-0.0275	-0.0307	0.0549
	(1.52)	(0.52)	(0.75)	(1.10)
Standout	-0.0111	-0.0034	-0.0342	0.0701
	(0.48)	(0.06)	(0.85)	(1.41)
Teaching & Citizenship	0.0004	-0.0771	-0.0086	0.0398
	(0.02)	(1.46)	(0.21)	(0.78)
FE/Variables absorbed	23	20	20	20
Additional covariates	7	7	7	7
Number of Letters	9154	2146	2704	2089
dto for females	2695	489	1020	554
Number of candidates	2791	674	780	625
dto female	817	157	295	161
Number of writers	4705	1436	1844	1186
dto female	778	162	364	166
Letters by fem writers	1315	206	532	274
Year FE	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes
Institution Rank FE	yes	yes	yes	yes
Years since PhD	yes	yes	yes	yes
Research Field FE	yes	n/a	n/a	n/a
Publications	yes	yes	yes	yes
Writer characteristics	yes	yes	yes	yes
Letter length	yes	yes	yes	yes

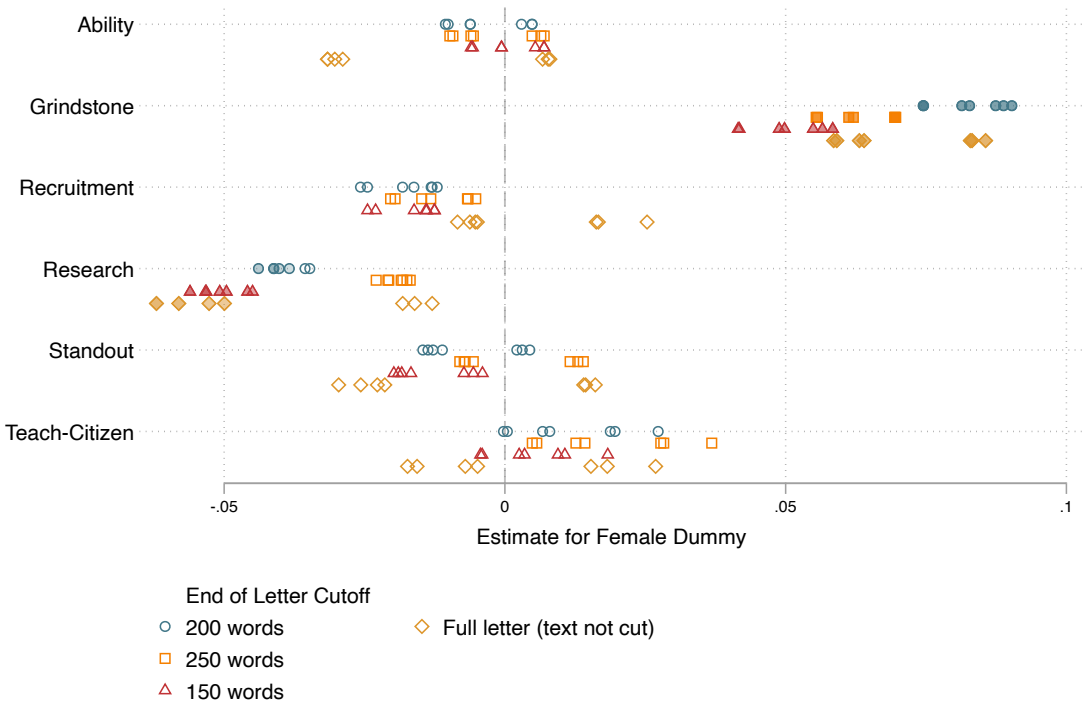
Notes: In the results presented in this table we split the sample by the institutional ranking of the candidate's PhD institution. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. See Table **??** for all other details. Sample restricted to the respective candidate research field (based on cluster analysis).

Return to Figure 11 in the maintext.

E Robustness checks

E.1 Robustness checks — Letter ends — Different letter end lengths

Figure E.1: Regression results, different letter lengths



Notes: This figure shows the coefficient estimates for the regressions specified in 6, estimated separately for different letter lengths (end of letter lines). We show the three most demanding specifications. The symbol's filling permit visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution, letter-writer's institutions, and letter writer), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. See Table overleaf for more information on sample and results for the letter-writer clustering. Return to Section 6.3 in the maintext.

Table E.1: Sentiments — End of Letters (150, 200 or 250 words) — All Writers — 2 Models

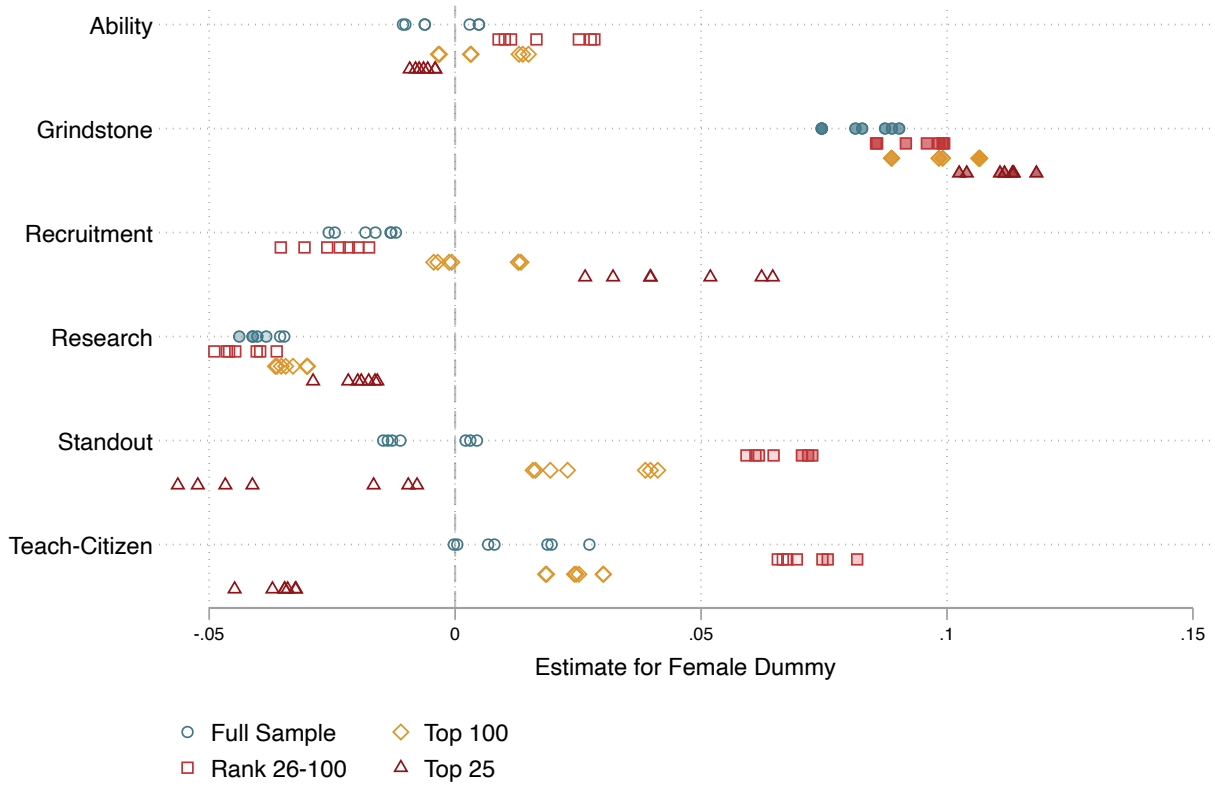
	(1)	(2)	(3)	(4)	(5)	(6)
Letter length	150 words		200 words		250 words	
Ability	0.0134	0.0055	0.0087	-0.0019	0.0091	-0.0027
	(0.57)	(0.23)	(0.37)	(0.08)	(0.39)	(0.11)
Grindstone	0.0634	0.0537	0.0930	0.0841	0.0715	0.0633
	(2.66)***	(2.24)**	(3.92)***	(3.52)***	(3.04)***	(2.65)***
Recruitment	-0.0086	-0.0161	-0.0100	-0.0192	-0.0046	-0.0156
	(0.36)	(0.69)	(0.42)	(0.82)	(0.19)	(0.66)
Research	-0.0528	-0.0472	-0.0412	-0.0368	-0.0183	-0.0193
	(2.25)**	(1.99)**	(1.77)*	(1.57)	(0.78)	(0.82)
Standout	0.0049	-0.0072	0.0073	-0.0076	0.0160	-0.0024
	(0.21)	(0.31)	(0.31)	(0.32)	(0.66)	(0.10)
Teaching & Citizenship	0.0226	0.0081	0.0307	0.0109	0.0393	0.0158
	(0.93)	(0.33)	(1.26)	(0.45)	(1.62)	(0.66)
FE/Variables absorbed	9	17	9	17	9	17
Additional covariates	1	7	1	7	1	7
Number of Letters	9213	9213	9196	9196	9146	9146
dto for females	2704	2704	2700	2700	2686	2686
Number of candidates	2791	2791	2791	2791	2791	2791
dto female	817	817	817	817	817	817
Number of writers	4727	4727	4717	4717	4686	4686
dto female	780	780	779	779	776	776
Letters by fem writers	1319	1319	1318	1318	1312	1312
Year FE	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	no	yes	no	yes
Years since PhD	no	yes	no	yes	no	yes
Research Field FE	no	yes	no	yes	no	yes
Publications	no	yes	no	yes	no	yes
Writer characteristics	no	yes	no	yes	no	yes
Letter length	no	yes	no	yes	no	yes

Notes: This table presents results for the analysis of three different letter end cut-offs: 150 words, 200 words or 250 words. For each category we present the most parsimonious and the most elaborate regression model.

Return to Section 6.3 in the maintext.

E.2 Robustness checks — Letter ends — Full Letter

Figure E.2: Regression results, all letter writers combined



Notes: This figure shows the coefficient estimates for the regressions specified in 6. We compare different type of specifications, from baseline ones only with candidate controls $\mathbf{X}_i'$ and fixed effects, to the ones with all the controls. The symbol's filling permit visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution, letter-writer, and letter-writer's institution), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error. See overleaf for additional information on the sample and results for the letter-writer clustering of standard errors. Return to Section 6.3 in the maintext.

Table E.2: Sentiments — Full Letters — All Writers — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0077	0.0067	0.0080	-0.0303	-0.0289	-0.0316	-0.0289
	(0.33)	(0.29)	(0.34)	(1.31)	(1.24)	(1.35)	(1.24)
Grindstone	0.0856	0.0832	0.0829	0.0631	0.0640	0.0585	0.0646
	(3.62)***	(3.52)***	(3.51)***	(2.65)***	(2.68)***	(2.45)**	(2.72)***
Recruitment	0.0163	0.0167	0.0253	-0.0084	-0.0062	-0.0049	-0.0067
	(0.67)	(0.68)	(1.05)	(0.35)	(0.26)	(0.20)	(0.28)
Research	-0.0161	-0.0182	-0.0130	-0.0621	-0.0581	-0.0527	-0.0548
	(0.67)	(0.76)	(0.54)	(2.71)***	(2.54)**	(2.31)**	(2.49)**
Standout	0.0140	0.0144	0.0161	-0.0296	-0.0257	-0.0227	-0.0241
	(0.60)	(0.61)	(0.69)	(1.27)	(1.10)	(0.97)	(1.04)
Teaching &	0.0269	0.0153	0.0182	-0.0049	-0.0070	-0.0156	-0.0092
Citizenship	(1.13)	(0.66)	(0.78)	(0.21)	(0.31)	(0.68)	(0.40)
FE/Variables absorbed	9	14	14	17	17	23	17
Additional covariates			1	1	5	6	7
Number of Letters	9228	9228	9228	9228	9228	9228	9228
dto for females	2708	2708	2708	2708	2708	2708	2708
Number of candidates	2791	2791	2791	2791	2791	2791	2791
dto female	817	817	817	817	817	817	817
Number of writers	4734	4734	4734	4734	4734	4734	4734
dto female	781	781	781	781	781	781	781
Letters by fem writers	1321	1321	1321	1321	1321	1321	1321
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the robustness analysis for the full reference letter. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Section 6.3 in the maintext.

Table E.3: Sentiments — Full Letters — By Writer Institutional Rank

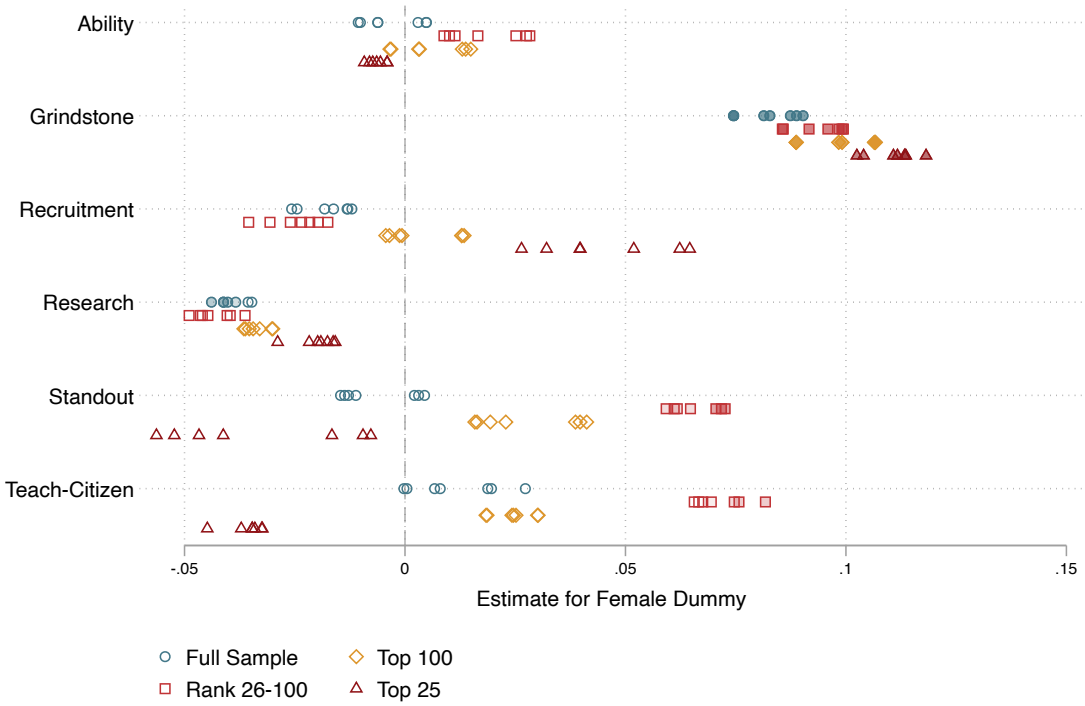
Writer RePEc Rank	(1) Top-25	(2) Top-50	(3) Top-100	(4) 26-100
Female Candidate Coefficient				
Ability	-0.0726	0.0121	0.0107	0.0644
	(1.46)	(0.29)	(0.31)	(1.39)
Grindstone	0.0144	0.0586	0.0574	0.0898
	(0.27)	(1.45)	(1.67)*	(1.97)**
Recruitment	0.0749	0.0168	0.0437	0.0201
	(1.36)	(0.39)	(1.25)	(0.44)
Research	-0.0846	-0.0511	-0.0500	-0.0298
	(1.64)*	(1.29)	(1.56)	(0.72)
Standout	-0.0625	-0.0239	-0.0083	0.0155
	(1.12)	(0.57)	(0.25)	(0.37)
Teaching & Citizenship	-0.0019	0.0256	0.0227	0.0404
	(0.04)	(0.63)	(0.71)	(0.97)
FE/Variables absorbed	18	19	21	19
Additional covariates	7	7	7	7
Number of Letters	1866	3026	4552	2686
dto for females	551	891	1325	774
Number of candidates	871	1278	1781	1215
dto female	257	363	511	348
Number of writers	823	1310	2009	1186
dto female	118	204	315	197
Letters by fem writers	225	389	583	358
Year FE	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes
Institution Rank (Band) FE	n/a	yes	yes	yes
Years since PhD	yes	yes	yes	yes
Research Field FE	yes	yes	yes	yes
Publications	yes	yes	yes	yes
Writer characteristics	yes	yes	yes	yes
Letter length	yes	yes	yes	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns indicate the different samples related to RePEc ranking of the writer's institution. This is the robustness analysis for the full reference letter. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Section 6.3 in the maintext.

E.3 Robustness Checks — Letter ends — Different Field Definitions

Figure E.3: Regression results, EJM fields as controls



Notes: This figure replicates Figure 7, but using EJM fields as controls rather than our broad research groupings. We show the three most demanding specifications. The symbol's filling permit visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution, letter-writer's institutions, and letter writer), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol the more often they are significant. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. See overleaf for more information on sample and results for the letter-writer clustering. Return to Section 6.3 in the maintext.

Table E.4: Sentiments — End of Letters (200 words) — EJM Field Dummies — 7 Models

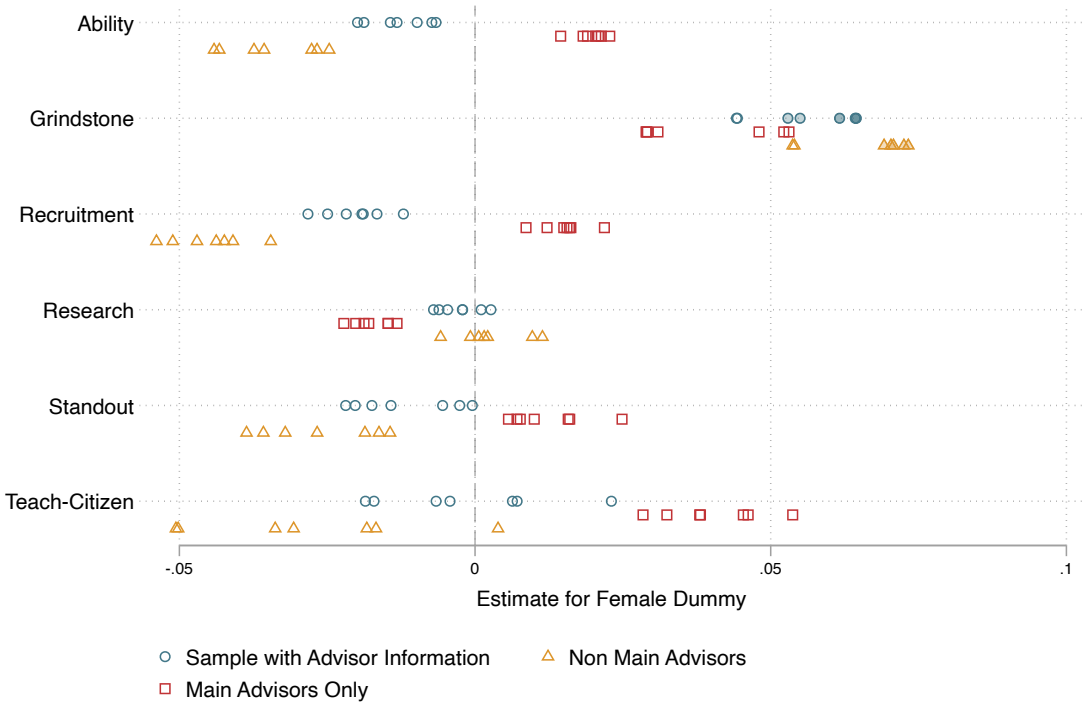
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0032	0.0032	0.0012	-0.0070	-0.0071	-0.0114	-0.0101
	(0.14)	(0.14)	(0.05)	(0.30)	(0.30)	(0.48)	(0.43)
Grindstone	0.0911	0.0883	0.0898	0.0828	0.0817	0.0753	0.0754
	(3.82)***	(3.70)***	(3.76)***	(3.40)***	(3.35)***	(3.10)***	(3.10)***
Recruitment	-0.0107	-0.0105	-0.0095	-0.0251	-0.0240	-0.0183	-0.0119
	(0.44)	(0.44)	(0.40)	(1.04)	(1.00)	(0.75)	(0.50)
Research	-0.0445	-0.0418	-0.0418	-0.0402	-0.0391	-0.0356	-0.0376
	(1.90)*	(1.78)*	(1.78)*	(1.70)*	(1.65)*	(1.50)	(1.59)
Standout	0.0055	0.0070	0.0046	-0.0011	-0.0002	0.0006	0.0058
	(0.23)	(0.30)	(0.19)	(0.05)	(0.01)	(0.03)	(0.25)
Teaching &	0.0267	0.0192	0.0181	0.0087	0.0078	0.0012	0.0030
Citizenship	(1.09)	(0.79)	(0.74)	(0.35)	(0.31)	(0.05)	(0.12)
FE/Variables absorbed	9	14	14	38	38	44	44
Additional covariates			1	1	5	6	7
Number of Letters	9091	9091	9091	9031	9031	9031	9031
dto for females	2667	2667	2667	2654	2654	2654	2654
Number of candidates	2775	2775	2775	2751	2751	2751	2751
dto female	810	810	810	804	804	804	804
Number of writers	4682	4682	4682	4657	4657	4657	4657
dto female	778	778	778	774	774	774	774
Letters by fem writers	1311	1311	1311	1303	1303	1303	1303
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively. We use alternative candidate research field dummies from EJM. The sample excludes candidates who specified 'any field'.

Return to Section 6.3 in the maintext.

E.4 Robustness Checks — Letter ends — vs Other Letter Writers

Figure E.4: Regression results, main advisor vs other letter writers



Notes: This figure shows the coefficient estimates for the regressions specified in 6, estimated separately for letters written by the main advisor and by others. We show the three most demanding specifications. The symbol's filling permit visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution, letter-writer's institutions, and field), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol the more often they are significant. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. See overleaf for information on sample and results tables for clustering by letter writer. Return to Section 6.3 in the maintext.

Table E.5: Sentiments — End of Letters (200 words) — Main Advisors only — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0145	0.0190	0.0209	0.0214	0.0228	0.0183	0.0204
	(0.28)	(0.37)	(0.41)	(0.41)	(0.44)	(0.35)	(0.39)
Grindstone	0.0522	0.0480	0.0531	0.0309	0.0291	0.0290	0.0289
	(1.08)	(0.99)	(1.09)	(0.64)	(0.61)	(0.60)	(0.60)
Recruitment	0.0155	0.0162	0.0150	0.0086	0.0122	0.0160	0.0219
	(0.30)	(0.31)	(0.28)	(0.16)	(0.23)	(0.30)	(0.42)
Research	-0.0147	-0.0132	-0.0147	-0.0202	-0.0188	-0.0180	-0.0222
	(0.29)	(0.27)	(0.30)	(0.41)	(0.38)	(0.36)	(0.44)
Standout	0.0249	0.0158	0.0160	0.0057	0.0071	0.0076	0.0100
	(0.50)	(0.32)	(0.32)	(0.11)	(0.14)	(0.15)	(0.20)
Teaching &	0.0537	0.0454	0.0462	0.0380	0.0381	0.0284	0.0324
Citizenship	(1.05)	(0.90)	(0.91)	(0.75)	(0.75)	(0.56)	(0.65)
FE/Variables absorbed	9	14	14	17	17	23	23
Additional covariates			1	1	5	6	7
Number of Letters	1948	1948	1948	1948	1948	1948	1948
dto for females	578	578	578	578	578	578	578
Number of candidates	1561	1561	1561	1561	1561	1561	1561
dto female	454	454	454	454	454	454	454
Number of writers	1383	1383	1383	1383	1383	1383	1383
dto female	191	191	191	191	191	191	191
Letters by fem writers	246	246	246	246	246	246	246
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute *t*-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Section 6.3 in the maintext.

Table E.6: Sentiments — End of Letters (200 words) — Exclude Main Advisors — 7 Models

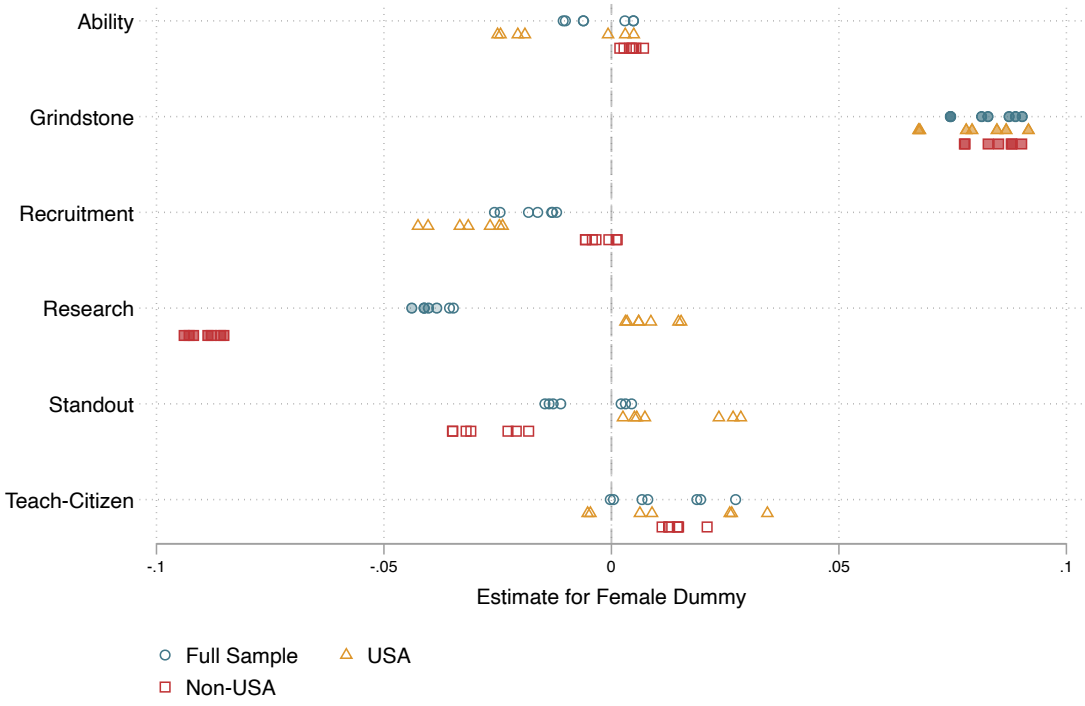
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	-0.0276	-0.0247	-0.0267	-0.0374	-0.0357	-0.0441	-0.0432
	(0.71)	(0.63)	(0.68)	(0.96)	(0.91)	(1.12)	(1.10)
Grindstone	0.0733	0.0708	0.0725	0.0703	0.0692	0.0537	0.0540
	(1.82)*	(1.75)*	(1.79)*	(1.72)*	(1.69)*	(1.31)	(1.32)
Recruitment	-0.0470	-0.0438	-0.0424	-0.0539	-0.0511	-0.0409	-0.0345
	(1.14)	(1.07)	(1.03)	(1.31)	(1.24)	(0.99)	(0.86)
Research	-0.0058	0.0015	0.0021	-0.0008	0.0006	0.0114	0.0097
	(0.15)	(0.04)	(0.05)	(0.02)	(0.02)	(0.28)	(0.24)
Standout	-0.0162	-0.0144	-0.0186	-0.0386	-0.0358	-0.0321	-0.0267
	(0.41)	(0.36)	(0.47)	(0.98)	(0.91)	(0.81)	(0.69)
Teaching &	0.0039	-0.0167	-0.0183	-0.0307	-0.0338	-0.0506	-0.0502
Citizenship	(0.10)	(0.42)	(0.46)	(0.77)	(0.84)	(1.27)	(1.26)
FE/Variables absorbed	9	14	14	17	17	23	23
Additional covariates			1	1	5	6	7
Number of Letters	3306	3306	3306	3306	3306	3306	3306
dto for females	953	953	953	953	953	953	953
Number of candidates	1590	1590	1590	1590	1590	1590	1590
dto female	459	459	459	459	459	459	459
Number of writers	2309	2309	2309	2309	2309	2309	2309
dto female	381	381	381	381	381	381	381
Letters by fem writers	526	526	526	526	526	526	526
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Section 6.3 in the maintext.

E.5 Robustness Checks — Letter ends — Location of PhD-granting institution

Figure E.5: Regression results, by location of letter writer institution



Notes: This figure shows the coefficient estimates for the regressions specified in 6, estimated separately for letter writers based in the US and in all other countries. We show the three most demanding specifications. The symbol's filling permit visualizing significance. Using 4 levels of possible standard error clustering (none, candidate's institution, letter-writer's institutions, and field), we flag significance at 3 different levels (10%, 5%, and 1%). We thus flag 12 possible significance indicators. Then, for each level of clustering, the symbol in the graph is shadowed with a 9% ($\approx 100/12$) opacity when it reaches significance at each possible level. The darker the symbol the more often they are significant. The darker the symbol, the more often it is significant. Fully filled symbols are significant at 1% level across all possible clustering. Hollow symbols do not reach significance for any level of standard error clustering. Return to Section 6.3 in the maintext.

Table E.7: Sentiments — End of Letters (200 words) — US-based candidates — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0090	0.0074	0.0037	-0.0162	-0.0147	-0.0204	-0.0211
	(0.28)	(0.23)	(0.12)	(0.50)	(0.45)	(0.63)	(0.65)
Grindstone	0.0944	0.0898	0.0879	0.0822	0.0809	0.0705	0.0702
	(2.81)***	(2.68)***	(2.62)***	(2.40)**	(2.36)**	(2.07)**	(2.06)**
Recruitment	-0.0211	-0.0231	-0.0204	-0.0390	-0.0367	-0.0277	-0.0296
	(0.62)	(0.69)	(0.61)	(1.16)	(1.09)	(0.82)	(0.90)
Research	0.0068	0.0124	0.0099	0.0076	0.0101	0.0187	0.0193
	(0.20)	(0.37)	(0.30)	(0.23)	(0.31)	(0.57)	(0.58)
Standout	0.0326	0.0311	0.0281	0.0067	0.0093	0.0111	0.0092
	(0.98)	(0.94)	(0.85)	(0.20)	(0.28)	(0.33)	(0.28)
Teaching &	0.0378	0.0303	0.0298	0.0124	0.0098	-0.0014	-0.0020
Citizenship	(1.10)	(0.90)	(0.89)	(0.37)	(0.29)	(0.04)	(0.06)
FE/Variables absorbed	9	14	14	17	17	23	23
Additional covariates			1	1	5	6	7
Number of Letters	4563	4563	4563	4563	4563	4563	4563
dto for females	1380	1380	1380	1380	1380	1380	1380
Number of candidates	1381	1381	1381	1381	1381	1381	1381
dto female	416	416	416	416	416	416	416
Number of writers	2294	2294	2294	2294	2294	2294	2294
dto female	420	420	420	420	420	420	420
Letters by fem writers	729	729	729	729	729	729	729
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Section 6.3 in the maintext.

Table E.8: Sentiments — End of Letters (200 words) — non US-based candidates — 7 Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ability	0.0077	0.0106	0.0088	0.0078	0.0074	0.0046	0.0055
	(0.23)	(0.31)	(0.26)	(0.23)	(0.22)	(0.14)	(0.16)
Grindstone	0.0907	0.0907	0.0928	0.0877	0.0852	0.0799	0.0798
	(2.73)***	(2.73)***	(2.80)***	(2.65)***	(2.57)**	(2.39)**	(2.39)**
Recruitment	-0.0006	0.0037	0.0020	-0.0032	-0.0034	-0.0021	0.0035
	(0.02)	(0.11)	(0.06)	(0.10)	(0.10)	(0.06)	(0.10)
Research	-0.0913	-0.0924	-0.0904	-0.0874	-0.0845	-0.0838	-0.0866
	(2.80)***	(2.82)***	(2.75)***	(2.64)***	(2.55)**	(2.52)**	(2.63)***
Standout	-0.0185	-0.0142	-0.0170	-0.0278	-0.0313	-0.0316	-0.0275
	(0.55)	(0.42)	(0.51)	(0.83)	(0.94)	(0.95)	(0.83)
Teaching &	0.0245	0.0181	0.0177	0.0144	0.0158	0.0155	0.0176
Citizenship	(0.72)	(0.53)	(0.52)	(0.42)	(0.46)	(0.45)	(0.52)
FE/Variables absorbed	9	14	14	17	17	23	23
Additional covariates			1	1	5	6	7
Number of Letters	4633	4633	4633	4633	4633	4633	4633
dto for females	1320	1320	1320	1320	1320	1320	1320
Number of candidates	1410	1410	1410	1410	1410	1410	1410
dto female	401	401	401	401	401	401	401
Number of writers	2748	2748	2748	2748	2748	2748	2748
dto female	391	391	391	391	391	391	391
Letters by fem writers	589	589	589	589	589	589	589
Year FE	yes	yes	yes	yes	yes	yes	yes
Ethnicity/Race FE	yes	yes	yes	yes	yes	yes	yes
Institution Rank FE	no	yes	yes	yes	yes	yes	yes
Years since PhD	no	no	yes	yes	yes	yes	yes
Research Field FE	no	no	no	yes	yes	yes	yes
Publications	no	no	no	no	yes	yes	yes
Writer characteristics	no	no	no	no	no	yes	yes
Letter length	no	no	no	no	no	no	yes

Notes: The table shows OLS regression results of the letter-specific sum of tf-idf statistics related to bag of expressions (dependent variable) mentioned in the row label, regressed on a female candidate dummy as well as controls indicated in the lower part of the table: a negative (positive) coefficient implies that on average fewer (more) expressions from the respective bag are used for female candidates relative to their male peers. Standard errors are clustered at the letter writer level, we report the absolute t-statistics in parentheses. Each pair of results (estimates, standard errors) is from a *separate* regression for the dependent variables in the row label, the columns refer to more and more additional control variables. This is the benchmark analysis for a letter end of 200 words. The coefficients are standardised and are reported in terms of standard deviations of the dependent variable (e.g. ability, grindstone, etc). *, ** and *** indicate statistical significance at the 10%, 5% and 1% level, respectively.

Return to Section 6.3 in the maintext.