Casamonti, Matilde; Zinovyeva, Natalia

Working Paper

Gendered Language in Academic Evaluations: Evidence from the Italian University System

IZA Discussion Papers, No. 17240

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Casamonti, Matilde; Zinovyeva, Natalia (2024) : Gendered Language in Academic Evaluations: Evidence from the Italian University System, IZA Discussion Papers, No. 17240, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/305682

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

DISCUSSION PAPER SERIES

IZA DP No. 17240

Gendered Language in Academic Evaluations: Evidence from the Italian University System

Matilde Casamonti
Natalia Zinovyeva

AUGUST 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17240

Gendered Language in Academic Evaluations: Evidence from the Italian University System

Matilde Casamonti
PwC Middle East

Natalia Zinovyeva
University of Warwick and IZA

AUGUST 2024

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

Gendered Language in Academic Evaluations: Evidence from the Italian University System*

We analyze the impact of evaluator and candidate gender on the language used in academic evaluations using data on 295,000 evaluation reports for applicants seeking professorial promotion across all academic fields in Italy. In this context, candidates are assessed by a national-level committee composed of five randomly selected evaluators from the corresponding field. We observe that the language used in evaluation reports varies significantly with applicants' productivity and professional ties to evaluators, but we find no indication that the language of the assessments depends on the gender of either the candidates or the evaluators.

JEL Classification: I23, J16, J71, M51

Keywords: academic evaluations, women in academia, gendered language

Corresponding author:
Natalia Zinovyeva
University of Warwick
Coventry CV4 7AL
Great Britain

E-mail: natalia.zinovyeva@warwick.ac.uk

* We would like to thank Arianna Ornaghi, Manuel Bagues and participants of Warwick-St Andrews Workshop of Women in Political Economy for their helpful comments and suggestions. We also extend our thanks to Milan Makany for his outstanding research assistance and valuable feedback.

1 Introduction

Women are underrepresented in higher academic positions. While there are as many female PhD graduates as men, women are less likely to advance up the career ladder, and it takes them longer to achieve promotion (Ceci et al., 2014; Weisshaar, 2017; Lundberg and Stearns, 2019; Directorate-General for Research and Innovation (European Commission), 2021).

One potential factor contributing to the slower progression of women in academia may be gender discrimination at various stages of their careers.[1] The evidence on the persistence of such discrimination is mixed. Some studies analyzing the presence of discrimination in academia provide evidence consistent with its existence (Moss-Racusin et al., 2012; Card et al., 2020; Hospido and Sanz, 2021; Sarsons et al., 2021; Koffi, 2021; Hengel, 2022)[2], while other studies challenge this view (Williams and Ceci, 2015; Carlsson et al., 2021; Card et al., 2022, 2023).[3] Ceci et al. (2023) conducted a meta-analysis on gender bias in six domains of academic science: tenure-track hiring, grant funding, teaching ratings, journal

1. Several alternative explanations have been proposed as potential causes of the "leaky pipeline." On the supply side, laboratory experiments and other studies have indicated that women often engage in activities less conducive to promotion (Babcock et al., 2017; Azmat and Ferrer, 2017). Women also tend to shy away from competition (Niederle and Vesterlund, 2007), are less likely to negotiate better compensation (Small et al., 2007; Leibbrandt and List, 2015), and are less inclined to apply for promotions (De Paola et al., 2017). Additionally, the lack of research networks and mentors may also hinder the success of women (Blau et al., 2010; Hilmer and Hilmer, 2007; Zinovyeva and Bagues, 2015).

2. Moss-Racusin et al. (2012) conducted a correspondence study where faculty members in Biology, Chemistry, and Physics ranked applicants for a laboratory manager position, finding that women were perceived as less competent and less "hirable" than male applicants. Card et al. (2020) analyzed data from referees' recommendations and editors' decisions at top Economics journals, finding that evaluators impose a higher bar for women than that implied by future citation maximization. Hospido and Sanz (2021) reported that women are less likely to have their papers accepted at an Economics conference, and Sarsons et al. (2021) found that women in Economics receive relatively less credit for co-authored work when evaluated for promotion. Koffi (2021) develops a prediction model for the probability that a given paper is included among the references in another paper, conditional on several measures of topic proximity, and finds that female-authored papers are more likely to be omitted from the references. Hengel (2022) suggests that female authors face higher standards during peer review in top Economics journals, as evidenced by the enhanced readability of paper abstracts from the working paper stage to publication.

3. In a correspondence study, Williams and Ceci (2015) demonstrated that faculty members in Biology, Engineering, Economics, and Psychology rated the hypothetical profiles of female applicants for tenure-track assistant professorships more favorably. Carlsson et al. (2021) run a large-scale experiment in Iceland, Sweden, and Norway asking 775 faculty members to access CVs of hypothetical candidates for Associate Professor positions with a randomly assigned male or female name and found a significant pro-female advantage. Card et al. (2022) and Card et al. (2023) found that women experience a premium in peer recognition: they are more likely to be nominated and elected as Fellows of prestigious academic societies such as the Econometric Society, American Academy of Arts and Sciences, and the National Academy of Sciences.

acceptances, salaries, and recommendation letters. Their comprehensive review reveals a small but discernible gender gap against women only in salaries and potentially in teaching evaluations. They highlight that this gap varies by discipline and evaluation context, with Economics and scenarios characterized by significant information asymmetries often having the largest gaps.

There is also a growing body of literature analyzing the language used in evaluations instead of just evaluation outcomes. Focusing on language has several advantages. First, it can be useful when evaluation outcomes are not observable, such as in assessments by reference letter writers. Second, analyzing language can help uncover the nature of stereotypes. Stereotypes may potentially balance each other out in evaluations, so examining gender differences in the way women and men are assessed can provide insights into when biases are likely to emerge. Third, language analysis offers more direct evidence of the presence of explicit or implicit stereotypes, which may or may not be relevant to evaluation outcomes in a specific context but could manifest in other settings. There is evidence suggesting that such stereotypes may exist. Two recent studies by Eberhardt et al. (2023) and Baltrunaite et al. (2024), which collectively analyzed over 30,000 reference letters for Ph.D. graduates applying for assistant professorships in two leading Economics departments in Europe, indicate that letters for women more frequently emphasize attributes such as diligence and hard work over ability and brilliance.

In this paper, we use data from a large-scale natural randomized experiment to examine whether the language used in the assessment of research performance varies with the gender of candidates and evaluators. We analyze 295,000 written assessments of 59,000 candidates applying for a national qualification certifying their eligibility for professorial promotion at Italian public universities. The candidates come from all academic disciplines and are applying for either associate or full professorships. The evaluations are based on CVs and publications submitted by the applicants. Evaluation committees are randomly formed from a pool of eligible evaluators within the corresponding academic field, effectively eliminating any risk of correlation between the characteristics of candidates or evaluators and the gender composition of evaluator-candidate pairs. A notable feature of Italian national evaluations is their transparency: individual CVs, evaluators' votes, and assessment reports are made

publicly accessible.

We characterize the writing style and effort put by evaluators in these assessments using measures such as word count, vocabulary richness, and readability. We assess the sentiment of evaluations with a lexicon-based approach that assigns a sentiment polarity score to each word. We identify the most predictive words used in evaluations of female and male candidates using a logistic classifier. Finally, we employ a dictionary approach to explore gender differences in how evaluators emphasize the standard dimensions of research productivity, such as quality, quantity, impact, and creativity.

To validate our measures, we examine how the writing style, sentiment of reports, and emphasized themes correlate with candidates' research productivity and their professional ties to committee members. We find that evaluators tend to be significantly more positive and produce more comprehensive reports — characterized by increased originality and technical complexity — when reviewing the work of their co-authors, colleagues, peers in the same subfield, and researchers with higher observable productivity. We also observe a significant correlation between the usage of words related to quantity, quality, impact, originality, and co-authorship and the bibliometric indicators derived from candidates' CVs, which capture these dimensions.

At the same time, we find no statistically significant differences in the writing style or sentiment based on the gender of the candidates. Similarly, our analysis of the typical words used to describe women and men reveals no systematic differences. We only observe a difference in mentions of maternity leave, which was relevant for determining the length of the period over which research productivity was assessed. When comparing men and women with similar observed productivity, as measured by various bibliometric indicators, we also find no gender differences in the emphasis placed by evaluators on describing the standard dimensions of researchers' productivity.

Finally, we investigate whether female and male evaluators systematically write different reports. When comparing assessments written for the same candidate by different, randomly assigned evaluators, we observe no differences in writing style between female and male evaluators. Additionally, no differences are observed when evaluators assess candidates of the same gender as themselves.

One could argue that the transparency of the Italian evaluation context may discourage explicit gender discrimination against female applicants (van den Brink et al., 2010). However, we believe that our results cannot be attributed solely to the suppression of gender stereotypes due to transparency. In fact, transparency has not entirely eliminated other biases in this context, such as those related to favoritism. For example, candidates who were randomly assigned to be evaluated by a Ph.D. advisor, a colleague, or a coauthor experienced a 13% increase in the likelihood of obtaining a national qualification (Bagues et al., 2019). Our findings also reveal that language use correlates with nearly every other observable characteristic of candidates, aside from gender. Nevertheless, it remains possible that gender biases in evaluations are more pronounced in less transparent settings. At the very least, our evidence suggests that gender stereotypes – whether explicit or implicit – regarding academics' research potential are not strong enough to be detectable within an open evaluation process.

Our results have important implications for designing effective measures to combat discrimination and prevent the unnecessary discouragement of women from pursuing research careers. While it is possible that various gender stereotypes affect the productivity of female researchers, our results indicate that gender stereotypes are unlikely to strongly influence the process of research evaluation.

Our paper makes several contributions to the literature. First, we contribute to the extensive body of research on gender discrimination in academia, particularly the studies analyzing the differences in language used in the assessment of female and male academics. Unlike Eberhardt et al. (2023) and Baltrunaite et al. (2024), which focus on specific disciplines or institutions, our study provides systematic evidence on gender differences in assessments across a wide range of academic fields and institutions. Moreover, the context we consider allows for the isolation of research assessments from other domains, as candidates are evaluated almost exclusively on their research productivity. In contrast, reference letters typically assess candidates across multiple dimensions, such as research, teaching, collegiality, and citizenship. By focusing on research assessments at advanced stages of academic careers, we also benefit from more accurate measures of observable productivity. Our overall conclusion is that there is no generalized evidence supporting the existence of widespread

gender stereotypes in the assessment of academics' research productivity.

Second, we contribute to the literature on whether female and male evaluators differ in their susceptibility to gender stereotypes. In many contexts, there is the risk that the observed gender mix of evaluator-candidate pairs is influenced by researchers' professional or personal proximity, potentially biasing the estimates of gender differences in stereotypes. Our study, based on a natural randomized experiment with random assignment of candidates to evaluators, avoids these selection issues. Our evidence suggests that neither male nor female evaluators exhibit gender stereotypes in their assessments.

Our study is closely related to Bagues et al. (2017), who used data from the same context to analyze the impact of female presence on evaluation committees on female candidates' success rates and found no positive effect. By focusing on the language used in evaluations rather than on evaluation outcomes, we are able to detect potential explicit or implicit gender stereotypes, determine if women and men are judged by different criteria, investigate whether evaluators' effort in writing reports varies by candidate gender, explore whether these differences depend on the gender of the evaluator, and assess if biases are present across the entire productivity distribution rather than at specific evaluation margins. Aside from minor gender differences in evaluation language for high-productivity candidates in the Social Sciences and Humanities, we find no indication of gender stereotypes.

The rest of the paper is organized as follows. In Section 2, we describe the institutional background of Italian national evaluations, and in Section 3, we summarize descriptive information on the data and text-based measures. Section 4 outlines the main aspects of our empirical strategy, and Section 5 presents the results. Finally, in Section 6, we discuss the results and conclude.

2 Institutional background

Women in Italy face significant challenges in the labor market compared to their counterparts in other European countries. In 2023, the labor force participation rate for women in the 15-64 age group was 58%, well below the EU average of 70% (OECD, 2024a). Gender stereotypes are also more pervasive in Italy than in many other European countries. According to OECD

(2024b), Italians are more likely than other Europeans to believe that men should have greater job rights, are better suited to business leadership, and make better political leaders than women.

These labor market challenges are reflected in academia as well. Similar to trends observed across Europe and the US, women in Italy are persistently underrepresented in top academic positions. While women constitute the majority of university graduates, their share among associate professors is below 40%, and only about one in four full professors are women, a figure slightly lower than the EU average of 26.2% in 2018 (Directorate-General for Research and Innovation (European Commission), 2021).

Since 2012, promotions in Italy are organized as a two stage process. All Italian academics applying for an Associate or a Full Professor position at a public university must first obtain a qualification certifying their research quality, granted by a national-level committee in the corresponding field.[4] Applicants seeking promotion must submit their CV and their recent publications through the website of the Ministry of Education and Research (MIUR). Once the deadline for applying is passed, evaluation committees are formed by the random draw. Each committee is composed of five members: four of them are Full Professors from Italian public universities, while one is a professor based in a foreign university from an OECD country. Evaluators are randomly chosen from the corresponding list of eligible evaluators who volunteered to participate and met specific requirements on research productivity.[5] The only restriction on the randomization procedure is that all members of the committee must come from different universities. Evaluators remain in their roles for the following two years and they cannot take part in national evaluations for the following three years after their mandate is over. Italian evaluators work *pro bono*, while evaluators based in foreign universities are paid 16,000 Euro.

Committees agree upon the evaluation criteria in their first meeting and make them

4. The system of national evaluations, Abilitazione Scientifica Nazionale (ASN), was introduced by law 240/2010 as part of the Gelmini Reform.

5. In STEMM disciplines (Science, Technology, Engineering, Mathematics, and Medicine) and psychology, evaluators must reach a minimum number of publications in scientific journals, citations, and the H-index. In SSH disciplines (Social Science and Humanities), they need to have a minimum number of journal articles, articles published in high-quality scientific journals, and books and book chapters. According to Bagues et al. (2017), about 40% of Italian full professors volunteered and were considered eligible.

public. While committees have full autonomy in setting up the criteria, the Ministry provides a nudge about the desired threshold. It asks committees to comment in their evaluation reports on whether the candidate is above the median in their field, according to three bibliometric indicators computed based on the research output over the ten years prior to the evaluation: in STEMM disciplines, these are the number of journal articles, the number of received citations, and the H-index and, in SSH disciplines, they are the number of journal articles, the number of articles in high-impact journals, and the number of books. As a result, some committees explicitly link their evaluation criteria to these bibliometric indicators. For individuals who took a leave during the last ten years, such as a maternity or parental leave, the period for considered production output is extended accordingly. Applicants can withdraw their applications within two weeks after the publication of the document with committee evaluation criteria.

Decisions on each candidate are reached by a qualified majority rule with a minimum of four positive votes out of five. Evaluators are required to accompany each of their individual votes with a written evaluation. The names of candidates and evaluators, individual evaluation reports and committees' final decisions and all published online after the evaluation.

Qualifications granted by the national committee are valid for six years. Unsuccessful applicants cannot re-apply for two years.

3 Data and Measures

3.1 Italian National Evaluations

The dataset contains information on all the evaluations conducted during the first edition of the Abilitazione Scientifica Nazionale (ANS) held between 2012 and 2014. As mentioned earlier, the ASN process is known for its high level of transparency, and the data used in this paper are sourced from the website of the Italian Ministry of Education and Research.[6] The dataset contains information on 58,948 applications and 39,496 individuals who were

6. The data on candidates, including their CVs, bibliometric indicators, and evaluation reports, were made available for six months after the completion of the evaluations on the following website: `https://abilitazione.mur.gov.it`. The data on evaluators and their CVs are available on this website as of August 20, 2024.

evaluated by 184 discipline-specific committees. About 14% of initial applicants withdrew their applications after the evaluation criteria of committees were made public. As shown in Bagues et al. (2017), while women were more likely to withdraw, the gender composition of committees didn't affect this decision. Each application consists of CVs and ten most relevant publications in the previous ten years. Similarly, we can observe CVs and publications of all eligible and selected evaluators. About 8% of initially drawn evaluators resigned and were substituted with another randomly drawn evaluator. We observe the outcome of the initial draw and the final committees. The Ministry provides information on individual evaluation reports and collective decision reached by the committees. About 43% of applicants received a positive evaluation (see Table A2 in the Appendix.).

The gender of researchers is determined using a binary classification based on their names. 43.6% of men obtained qualifications versus 41.5% of women (see Panel A in Table A3 in the Appendix). Women account for 19% of evaluators and 37% of applicants, however, there are substantial variations across fields (see Table A4 in the Appendix).

Using information from applicants' CVs, we create several bibliometric indicators to proxy for the quantity and the quality of publications. We construct measures of the number of publications overall and by type – articles, books, book chapters, proceedings, patents, and other publications. The average applicant has 66.7 publications, 39.1 of them being journal articles, 10.1 books or chapters of books, 10.1 conference proceedings, and 0.3 patents (see Table A2 in the Appendix). To account for research quality, we compute the number of 'top articles', i.e., the number of A-journal articles as classified by the Italian Agency for Quality Assessment and Accreditation (ANECA) in SSH fields and the number of articles in the top quartile journals according to their Article Influence Score (AIS) in STEMM fields.[7] We also computed the average AIS of articles published in the journals listed in the Web of Science. On average, 15 of applicants' journal articles are categorized as top articles. The average AIS of the Web of Science publications is 1.25, above the normalized average of 1 across all journals indexed in the Web of Science. The average publication has six coauthors; in 22% of cases the applicant is the first author and in 12% the last author of the article.

7. In Economics, journals included in the list of A-journals by ANECA are roughly the same journals as top quartile journals according to the AIS.

We also observe *mediane*, the indicators for whether the candidate is above the median in the respective field and category in terms of the number of publications, citations, and h-index in STEMM fields and the number of publications, A-journal articles, and books in SSH fields. These indicators are computed and provided by the Ministry for each candida. 84% of applicants are above the median according to at least one indicator, and 38% are above the median according to all three indicators. For those who work in academia when applying for the evaluation, we also observe the position held, the type of contract (fixed-term or tenured), and whether the position is in the field of the evaluation.[8]

Female applicants have significantly lower productivity than male applicants in the same field and category (see Panel A in Table A3 in the Appendix).

There are 294,740 individual evaluation reports, five per candidate. Figure A2 in the Appendix shows a typical-length, anonymized evaluation report as an example. Information on 84 individual votes of evaluators is missing and cannot be inferred from reports' texts.

3.2 Textual characteristics

We measure the length of each evaluation as the number of words in the corresponding text. To create the rest of our measures, we process the data by removing all punctuation, numerals and the so-called "stop-words", i.e. very frequent words such as articles and conjunctions. We remove first names from all reports. Given that Italian is a gendered language, we also applied a stemming algorithm, a pre-processing technique to reduce words to their root or base form.

We measure sentiments polarity by a polarity score based on *Sentix* (Sentiment Italian Lexicon), a set of large multilingual and English-language lexical databases containing words annotated with their semantic orientation (Basile and Nissim, 2013). For each word, Sentix contains the polarity (positive/negative) of the emotions associated. The polarity score ranges from -1 (totally negative emotions) to 1 (totally positive emotions). We define the

8. Applicants' CVs included information not only on affiliations and publications but also on other research-related activities such as participation in externally funded projects, editorial work, and international visits. According to committee reports, evaluations were primarily based on publications. We also expect that project leadership and participation, editorial work, and internationalization are strongly correlated with our publication-based measures, such as the number of publications in top journals, the average article influence score, and the share of first- and last-authored articles.

Total Polarity Score of each evaluation report as the sum of polarity scores across all adjectives and adverbs in the text. The *Average Polarity Score* is the average polarity score of adjectives and adverbs used in the text.

We measure the evaluators' writing readability with the Gulpease index, which is designed specifically for the Italian language by the Linguistic and Pedagogic University Group (in Italian Gruppo Universitario Linguistico Pedagogico – GULP) at the University of La Sapienza (Lucisano and Piemontese, 1988). For each evaluation i, the *Gulpease index* is:

$$Gulpease\ index_i = 89 + \frac{300 * sentences_i - 10 * letters_i}{words_i}. \tag{1}$$

This index takes into account the length of a word in characters rather than in syllables, which is more reliable for evaluating the readability of Italian writings. The index ranges from 0 (lowest readability) to 100 (maximum readability). A text with a score below 80 is considered to be hard to read for people with elementary education, while a text that scores below 60 is hard for those who attended middle school, and a text under 40 is hard to read for people with a high school diploma.

Another commonly used readability index for the Italian language is the Vacca-Flesh index (Franchina and Vacca, 1986). For each evaluation i, the *Vacca-Flesh index* is:

$$Vacca\ index_i = 206 - 0.65\frac{syllables_i}{words_i} - \frac{words_i}{sentences_i}. \tag{2}$$

Similarly to the Gulpease index, the Vacca index ranges from a minimum of 0 to a maximum of 100. When assessing the readability of texts, we exclude evaluations by foreign professors. In this case, we use information from 241,744 reports.

We measure the relative originality of the evaluator's vocabulary used in each evaluation as compared to the rest of reports written by the same evaluator. We first compute *evaluator-specific* inverse document frequency IDF_{ti} for each word t in document i as the log of one over the share of documents written by the corresponding evaluator containing t. We then define *Document originality* as the average, and *Total IDF* as the sum, of IDF_{ti} weights for all words in the document.

Finally, we extract the most typical words used in evaluations of men and women as

measured by Term-Frequency-Inverse-Document-Frequency (TF-IDF) scores:

$$TF\text{-}IDF_{ti} = TF_{ti}\ IDF_{ti}. \tag{3}$$

where the term frequency TF_{ti} of a word t in evaluation i is the count of occurrences of t in i, and the inverse document frequency IDF_{ti} is the log of one over the share of *all documents* containing t. The advantage of this approach over a simple word count is that frequently used words provide less information for the analysis.

Finally, we also employ a theory-based approach by hypothesizing which dimensions of research productivity were likely assessed by committees. We create dictionaries of words indicative of these dimensions and then explore whether these words were equally likely to be mentioned in evaluations of women and men. Specifically, we create indicators for the words 'quantity,' 'quality,' 'impact,' 'median,' 'coauthor,' and 'creativity,' including their synonyms (see Table A1 in the Appendix for the complete dictionaries).

3.3 Descriptive statistics on textual characteristics

An average evaluation report consists of 179 words but there is significant variation in the length of reports (see Table A2 in the Appendix). An average evaluation tends to express a positive sentiment. It typically scores high on readability but it also does not use very original applicant-specific vocabulary.

Female evaluators write slightly more positive reports than men in the same committee, and female applicants receive evaluations with less positive words than men (see Panels B and C of Table A3 in the Appendix).

Female evaluators write longer reports than men, but use less original language with lower readability/higher complexity. On average, there are no gender differences in the length of evaluations received by candidates.

When looking at particular words used in reports, we find that 32% of reports use the word 'quantity' (or synonyms) and 57% mention 'quality' (or synonyms). A third of reports mention the word 'median', which is likely to indicate the reference to bibliometric criteria provided by the Ministry. About 36% of assessments describe the 'impact' of applicants'

research, 25% refer to 'originality' or 'creativity', and only 2% explicitly mention the word 'co-author'. Female evaluators are relatively less likely to talk about 'quantity', 'quality', and 'impact', but are more likely to mention 'medians', 'coauthors', and 'creativity'. The assessments for female applicants are more likely to refer to 'medians'.

We then extract the most frequent words used in reports and check whether there are any noticeable gender differences in those. Panels A in Figures 1 and A3 in the Appendix show the top 20 most frequent (stemmed) words used in evaluations for respectively female and male applicants. The two rankings of words appear to be remarkably similar. Panels B of these figures report the most used (stemmed) adjectives and adverbs. The most frequent words used for female and male candidates appear to be roughly the same.

4 Methodology

4.1 Empirical strategy

When assessing whether there are systematic gender differences in evaluators' votes or the written language of evaluations, we compare male and female applicants with similar observable characteristics. We control for the measures of quantity and quality of publications described in Section 3, indicators for the number of medians satisfied by the candidate, the type and field of contract if working for university, and university dummies. We standardize all productivity indicators at the exam (committee x type of position) level and control for exam and evaluator fixed effects. We also cluster standard errors at the exam level.

To examine whether the gender of evaluators and candidates affects the voting and the language of the individual evaluation reports, we exploit the random assignment of members of the academic boards. We estimate the following equation:

$$\begin{aligned} Outcome_{ije} = \beta_0 + \beta_1\, Female\ Applicant_i + \beta_2\, Female\ Evaluator_j + \\ \beta_3\, Female\ Applicant_i\ * Female\ Evaluator_j + X_i'\phi + \mu_e + \mu_{ie} + \mu_{je} + \epsilon_{ije} \end{aligned} \tag{4}$$

where $Outcome_{ije}$ represents either an individual vote or a textual feature of the evaluation (word length, richness of vocabulary, readability score and polarity score) for the individual evaluations of application i by the evaluator j in evaluation e. $Female\ Applicant_i$ is a

dummy variable which indicates the gender of the candidate and $Female\ Evaluator_j$ is a dummy variable for the gender of the evaluator. X_i is a vector of productivity controls described above. μ_e, μ_{ie}, and μ_{je} are dummies for committees, applicants, or evaluators, which we gradually include in the estimation.

Committee fixed effects allow us to take into account systematic differences across fields in the share of women. However, a positive or a negative estimate of β_3 conditional on committee fixed effects may theoretically reflect both the fact that female evaluators give different evaluations to female candidates or a composition effect whereby, in committees that have more female evaluators, male evaluators tend to change their assessment of female candidates. Including evaluator fixed effects permits us obtaining a within-evaluator differences in evaluations of women and men. The application fixed effects assures that the estimates are not affected by the remaining differences in the quality of applications.

4.2 Logistic classifier

To explore if evaluators use words differently when assessing male and female applications, we perform both a gender classification and a sentiment classification. With the former, we can detect the gender of the candidate given the texts of the feedback that they received. With the latter, we can study the sentiment orientation of evaluations by classifying successful and rejected applications given the texts of evaluation reports. By applying this last sentiment classification method on two distinctive corpora, one for female applicants and one for males, we can display and compare the most predictive features for the classification decision for both men and women.

To perform these classifications, we used a logistic classifier, a basic supervised machine learning algorithm for classification. The goal of using a binary logistic regression is to train the logistic classifier so that it can predict the probability distribution over a set of labels and make a decision about the labels of new input of observations. Let T_i be a word vector for the evaluation i with TF-IDF scores as entries and X_i be a set of control variables, and let the following logistic functions describe the posterior probability of this evaluation i being

written for a female ($Female_i = 1$) and a male candidate ($Female_i = 0$):

$$\begin{aligned} P(Female_i = 1|T_i, X_i) &= \frac{e^{(T_i{}'w + X_i{}'v + \mu_e + b)}}{1 + e^{(T_i{}'w + X_i{}'v + \mu_e + b)}} \\ P(Female_i = 0|T_i, X_i) &= \frac{1}{1 + e^{(T_i{}'w + X_i{}'v + \mu_e + b)}} \end{aligned} \tag{5}$$

where w and v are vectors of weights associated with each word and individual characteristic according to their importance for the classification decision, μ_e captures the differences in the share of women across evaluations, and b is the intercept. The parameters are estimated to maximize the likelihood of getting the correct gender labels in the training data given the observations. Since there are only two outcomes for the gender of the applicants, this is a Bernoulli distribution, and the likelihood produced by the logistic classifier for a single observation can be written as:

$$P(Female_i|T_i, X_i) = P(Female_i = 1|T_i, X_i)^{y_i} P(Female_i = 0|T_i, X_i)^{1-y_i} \tag{6}$$

where y_i is the indicator for whether the i-th applicant in the training database is actually a female.

If there are m independent observations in the training set, the log-likelihood function for the whole dataset is:

$$\begin{aligned} \mathcal{L}(w, v, b) &= \log \prod_m P(Female_i|T_i, X_i) \\ &= \sum_m (y_i(T_i{}'w + X_i{}'v + \mu_e + b) - log(1 + e^{(T_i{}'w + X_i{}'v + \mu_e + b)})). \end{aligned} \tag{7}$$

As a convention, the log-likelihood function is transformed into a negative average loss function by first applying a negative transformation and secondly dividing the negative log-likelihood function by the overall number of observations. In this way, the maximization problem becomes a minimization of the probability of getting incorrect labels:

$$\hat{\theta} = \arg\min_\theta \{-\frac{1}{m}\mathcal{L}(\theta)\},\ with\ \theta = \{w, v, b\} \tag{8}$$

$$\hat{\theta} = \arg\min_\theta \{-\frac{1}{m}\sum (y_i(T_i{}'w + X_i{}'v + b) - log(1 + e^{(T_i{}'w + X_i{}'v + b)}))\} \tag{9}$$

The optimization algorithm used to perform this computation is stochastic gradient descent. This iterative method finds the gradient of the loss function and moves in the direction of the minimum.

There are 241,744 individual evaluations written by Italian evaluators in the ASN dataset, and 80% of them are used for the above training process. The rest are assigned to the test set and used to evaluate the accuracy of the classifier. We assess the residual accuracy of the prediction provided by the estimates of w net of the impact of individual characteristics X_i and committee indicators μ_e.

5 Results

5.1 Validation of text-based measures

To validate that our textual characteristics – word count, document originality, and language complexity (the negative of Gulpease index and Vacca index) – are indicative of the evaluators' effort in their assessments, we would like to establish that they tend to increase in scenarios where a higher evaluator effort is anticipated. We expect connections between the candidate and the evaluator, such as co-authorship, collegial relationships, or shared research interests within the same subfield, positively correlate with evaluators' effort. We do observe a positive correlation between the indicators of connections and the (standardized at the exam level) measures of word count, text originality, and text complexity, even after accounting for fixed effects associated with both candidates and evaluators (Table 1). This positive correlation supports the premise that these textual characteristics are indeed reflective of the evaluators' effort.

5.2 Evaluators' writing styles and candidates' gender

Table 2 shows how the writing style and the sentiment of assessments changes with applicants' characteristics, including gender and productivity. All outcome variables representing textual features as well as productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam. We exploit only within-committee variations to exclude the influence of field-specific differences that may similarly influence

textual features of all evaluations. For comparison, column 1 shows the impact of the same characteristics on individual evaluators' votes.

Generally, characteristics that are positively correlated with the individual vote (e.g., the number of articles in high-impact journals, percent of first- and last-authored articles, average Article Influence Score) are also positively correlated with the length of reports, originality of vocabulary, complexity of reports, and polarity scores.

While the writing style of evaluators and the sentiment they express in reports change in relation to productivity indicators and other characteristics of the applicants, they do not change in relation to the candidates' gender.

Note that men in our data have more publications, and candidates with more extensive publication records tend to receive longer, more original, and generally more positive reports. If our controls do not fully account for individual productivity, and if women are relatively less productive also in unobserved dimensions, the omitted variable bias could drive our estimate of the coefficient for the female dummy down. The absence of a correlation between textual features and the female dummy, conditional on productivity controls, reinforces the interpretation that the evaluators' writing style and language do not, on average, show a bias against women.

One potential concern could be that, in the context of Italian National Evaluations, some evaluators may have simply copy-pasted their reports for different candidates introducing only very minor modifications.[9] This can introduce a measurement error in our outcome variables, which may potentially lead to the lack of precision of our estimates. In Appendix Table A5, we repeat the analysis on the sub-sample of evaluators with below median copy-pasting behavior in each committee as measured by the share of unique words over the total words used by the evaluator across all reports. The results stay the same.[10]

We also re-run the analysis on the relationship between candidates' characteristics and textual features controlling for the evaluators' individual vote and the results are unchanged (see Appendix Table A6).

The advantage of the use of textual features over the voting is that it allows us exploring

9. Marzolla (2016) provides examples of such cases in Italian national evaluations.

10. Similarly, all remaining results in this paper hold on the sub-sample of evaluations written by not copy-pasting evaluators.

how evaluators' language changes outside the evaluation margin relevant for this particular context. We run the same analysis on the sub-samples of candidates with below and above mean number of publications in high impact journals. We classify journals as 'high-impact' if they fall within the top quartile based on their Article Influence Score (AIS) in the STEMM fields, and as A-journals as per ANECA's categorization in the SSH fields. Appendix Table A7 shows that there are no statistically significant gender differences in the reports received by candidates with relatively low productivity. However, in the case of highly productive candidates, there is a notable difference: women receive reports characterized by significantly lower sentiment polarity scores. This result is driven by evaluations in the Social Sciences and Humanities, where the gender gap in the average sentiment polarity of adjectives used in evaluation reports is about 0.03 standard deviations in favor of men. We explore whether the reduced support for highly productive women in SSH is fully explained by evaluators' votes. We find that this result is unaffected by the inclusion of a control for the evaluator vote, suggesting that it is attributable to infra-marginal candidates (see Appendix Table A8).

To further explore this regularity, we use as an alternative measure of research productivity the number of median criteria provided by the Italian Ministry that the candidate satisfies. We confirm that, among candidates with all three median criteria satisfied, women receive slightly less favorably written evaluations in SSH (see Panel A in Appendix Table A9).

While one potential interpretation of this result is that, in SSH fields, evaluators are biased against high-performing women, an alternative is that there are some remaining gender differences in unobserved research quality. In SSH fields, a significant portion of the output is published in books and volumes, and our measures may be less effective at capturing the quality of research in these outlets. Consistently with this possibility, in Economics, where most research is published in journals, we do not observe any significant gender differences in the sentiment polarity of reports.

5.3 Writing styles of female and male evaluators

We then assess whether female and male evaluators tend to write systematically different reports. Within committees, we do not observe substantial difference in the writing style between female and male evaluators (Table 3). If anything, female evaluators tend to write slightly more technical evaluation reports then men (column 4). We do not observe any statistically significant differences in reports depending on whether the evaluator assesses the candidate of the same gender.

Women do not seem to use more positively toned language when assessing female candidates in any of the considered sub-samples, based on candidates' productivity and field (see Appendix Table A10). However, in STEMM fields, female evaluators are significantly more likely to cast a positive vote for female candidates when assessing highly productive candidates, with an increased probability of 2 percentage points from the baseline of 63%.[11] They also tend to use relatively more complex language for female candidates when assessing low-productivity candidates. This suggests that, in STEMM fields, the gender of evaluators may somewhat influence assessments.

5.3.1 Dictionary-based approach

Our results so far suggest that, on average, the writing style and sentiment of the assessments are not affected by the gender of candidates. However, it might still be possible that the gender of candidates impacts the choice of topics and words that evaluators use when writing reports. If the weights evaluators assign to the various dimensions of research productivity differ across genders, it would indicate that women and men may face differential treatment in other evaluation contexts.

Using a dictionary approach, we explore whether evaluators are more likely to mention specific dimensions of productivity when describing the work of women, conditional on observable measures of candidates' productivity. We define indicators for the presence of stemmed words related to quantity (or number), quality, impact, medians, co-authors, and creativity (or originality and novelty).

11. This effect does not depend on the share of women among researchers in the corresponding field.

We regress indicators for the usage of certain words on the female dummy and all the controls included in Table 2. Results are presented in Table 4. First of all, it is striking how strongly our controls are correlated with usage of keywords. Indicators for the number of publications are strongly correlated with the usage of word 'quantity'. 'Quality' is more likely to mentioned for candidates with many top-ranked publications. The word 'coauthors' are more likely to appear in reports for candidates with many coauthors. However, we do not find evidence of any gender differences in the frequency of mentions of quantity and quality themes by evaluators, conditional on observable dimensions of candidates' productivity. We find no evidence of gender difference in dimensions described in any of the sub-samples depending on candidates' productivity or research field (see Appendix Table A11).[12]

5.3.2 Logistic classifier

Finally, we implement a logistic classifier model, as described in Section 4.2, to predict whether a report is written for a woman or a man based on the words used in the evaluation text. We train the logistic classifier model on 80% of all evaluation from evaluators based in Italy. The words used in evaluation texts are stemmed and vectorized before the estimation procedure using global TF-IDF weights. To assure that we compare evaluations of similar women and men, we include all individual characteristics from Table 2 and committee fixed effects as controls.

Table 5 presents the performance metrics for the logistic classifier model. We include the ROC AUC score, along with the precision and recall for each gender. The ROC AUC score measures the area beneath the receiver operating characteristic (ROC) curve. This curve shows how the true positive rate changes in relation to the false positive rate. The score ranges from 0 to 1, where 0.5 represents random guessing and 1 means perfect accuracy. Precision is the ratio of true positives among all predicted positives, while recall is the ratio of true predicted positives out of all actual positives.

The ROC AUC score for the full model, which includes both control variables and exam dummies, is 0.67. This score indicates that the model has relatively low predictive accuracy.

12. We also do not observe differences between female and male evaluators in the propensity to discuss different dimensions of research productivity.

Specifically, precision of the prediction that a report is written for a woman is 67%. However, the model only correctly identifies 47% of the actual reports written for women, suggesting less than half of all female candidates are accurately predicted. When the control variables and exam dummies are replaced with average values from the sample, the ROC AUC score decreases to 0.58, suggesting that a significant portion of the model's predictive ability is due to gender differences accounted for by the controls. Predictions based on texts only have a 23% recall rate.

Figure 1 displays the logistic classifier model's point estimates for the 25 stemmed words with the highest predictive power for each gender. The blue bars represent the coefficients for words most predictive of evaluations for men, while the red bars are for women. Notably, terms like *congedo* (the Italian term for leave of absence, typically maternity leave) and maternity (stemmed from *maternitá*) are prominent in evaluations of women. These terms likely reflect the committees' effort to comply with legal requirements and correctly identify the relevant assessment period for research output. For men, the most predictive word is 'born' (stemmed from *nato*), suggesting that evaluators may use age as a proxy for men's overall time in research, while they are less explicit about the age of women.

Other terms that are predictive of women include feminine (stemmed from *femminile*), family (stemmed from *famiglia*), children (stemmed from *bambini*), and gender. These terms are not so much linked to discussions on maternity leaves but rather suggest a gender segregation in research topics, with women more likely to engage in studies related to gender and family, especially in the Social Sciences and Humanities. Indeed, mentions of leaves of absence and maternity are almost equally common in evaluations of women across all fields: 1.2% in STEMM fields and 2.4% in non-STEMM fields (corresponding number for men are 0.8% and 0.4%). At the same time, terms like 'feminine,' 'family,' 'children,' and 'gender' are common in evaluations in non-STEMM fields – 5.8% for women and 1.9% for men – but are almost non-existent in STEMM fields for both genders.[13]

13. Note that Figure 1 presents the estimates but does not indicate their statistical significance. Some words with large estimates, particularly surnames, occur very rarely in the reports, making their corresponding estimates clearly not statistically significant. For example, the word 'Schopenhauer' is more likely to be found in reports for male candidates, but it appears in only 123 out of nearly 300,000 reports. Similarly, the word 'diligent' is more likely to be observed in reports for women, but it is found in only 125 reports. In contrast, 'maternity leave' is mentioned in 2,728 reports.

Note that the inclusion of exam dummies in the logistic regression already helps account for gender segregation across fields, thereby reducing the impact of field-specific words. For instance, in an extreme scenario where only men conduct research in Mathematics and only women in Linguistics, with these subjects perfectly aligned with specific fields, any gender differences in vocabulary related to Mathematics or Linguistics would be fully accounted for by field or exam dummies. However, if this alignment is not perfect, field-specific vocabulary may still have predictive power for gender. By applying field-specific TF-IDF weights – which reduce the importance of words that are frequent within specific fields – rather than global TF-IDF weights – which reduce the importance of words frequent across all fields – we can further minimize the influence of field-specific vocabulary and better highlight themes that transcend multiple fields.

The performance of the logistic classifier, when field-specific TF-IDF weights are used along with controls and exam dummies, is slightly worse, with the ROC AUC score being 0.61 compared to 0.68 when global TF-IDF weights are used. Interestingly, the residual predictive power, net of the contribution of controls and exam dummies, is essentially zero, with the ROC AUC score being 0.50. The recall rate for women is 1%, suggesting that almost no women are correctly predicted to be women based on texts.[14] In other words, once we account for gender segregation across research topics in the logistic classifier, the words used in evaluations become non-predictive of gender.

6 Conclusion

This paper contributes to the debate on the under-representation of women in top academic positions exploring whether women experience a disadvantage in academic evaluations. We utilize information from nearly 300,000 individual evaluation reports on candidates seeking promotion to Associate and Full Professor positions across all academic fields in Italian academia to study whether the evaluators' written language depends on the gender of both the candidates and the evaluators.

We find that, on average, the language used in evaluations is not affected by the gen-

14. As there are more male than female candidates in the sample, the default prediction, in the absence of other information, is that the candidate is male.

der of the candidates and evaluators. Conversely, characteristics that measure candidates' productivity, as well as their connections and research proximity to evaluators, are strongly correlated with the length of the reports, the originality and complexity of the vocabulary, as well as the sentiment of adjectives and adverbs used in the text.

In addition to exhibiting similar writing styles, individual evaluations for female and male candidates also utilize similar vocabulary and tend to discuss similar dimensions of research productivity. The words with the greatest predictive power for gender primarily reflect the higher likelihood of women to take maternity leave, which extends the period over which their research productivity is assessed.

Overall, these findings indicate that gender biases do not represent a major concern in the context of national academic evaluations for professorial promotions. Moreover, the underlying gender stereotypes are not strong enough to be detectable in evaluation reports.

While the overall sample reveals no indication of gendered language use, a heterogeneity analysis uncovers some minor variations across different productivity margins and fields of research. Specifically, in the Social Sciences and Humanities, high-productivity female candidates are assessed with a less positive tone than their male counterparts, regardless of the evaluator's gender. The magnitude is about 3% of the standard deviation in the corresponding measure of sentiment polarity. This estimate remains robust even when controlling for actual votes in favor of or against granting qualifications. In contrast, in STEMM fields, no gender differences are observed in evaluations, whether for marginal or infra-marginal candidates. These observations suggest that women in the Social Sciences and Humanities might potentially face a greater disadvantage in more competitive settings compared to their counterparts in STEMM. An alternative interpretation is that we fail to capture some of the remaining gender differences in unobserved research performance in Social Sciences and Humanities.

Also, in STEMM fields, we observe that female evaluators are significantly more likely to cast a positive vote for female candidates when evaluating relatively more productive candidates. The magnitude of this effect is small, and we do not observe any other gender differences in language use, either in measures reflecting the effort evaluators put into assessing candidates or in the themes emphasized. The effect does not depend on the share of

women in the corresponding field. Taken together, this evidence may suggest the presence of minor, non-stereotype-based gender dynamics in STEMM fields.

While our results do not rule out the possibility that gender stereotypes regarding certain dimensions of performance or individual preferences may impede women in academic careers, they also suggest that there are no widespread gender biases among evaluators when it comes to assessing research performance. These findings indicate that policy efforts to address gender inequality in academia should focus on areas outside of research evaluations, such as addressing the factors contributing to lower research productivity among women, the allocation of administrative tasks, and other barriers to women's progress, rather than on research performance assessment itself.

References

Azmat, G. and R. Ferrer (2017): "Gender Gaps in Performance: Evidence from Young Lawyers," *Journal of Political Economy*, 125, 1306–1355.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017): "Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability," *American Economic Review*, 107, 714–747.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017): "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review*, 107, 1207–1238.

——— (2019): "Connections in Scientific Committees and Applicants' Self-Selection: Evidence from a Natural Randomized Experiment," *Labour Economics*, 58, 81–97.

Baltrunaite, A., A. Casarico, and L. Rizzica (2024): "Women in Economics: The Role of Gendered References at Entry in the Profession," .

Basile, V. and M. Nissim (2013): "Sentiment Analysis on Italian Tweets," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 100–107.

Blau, F. D., J. M. Currie, R. T. A. Croson, and D. K. Ginther (2010): "Can Mentoring Help Female Assistant Professors? Interim Results from a Randomized Trial," *American Economic Review*, 100, 348–352.

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

——— (2022): "Gender Differences in Peer Recognition by Economists," *Econometrica*, 90, 1937–1971.

——— (2023): "Gender Gaps at the Academies," *Proceedings of the National Academy of Sciences*, 120, e2212421120.

Carlsson, M., H. Finseraas, A. H. Midtbøen, and G. L. Rafnsdóttir (2021): "Gender Bias in Academic Recruitment? Evidence from a Survey Experiment in the Nordic Region," *European Sociological Review*, 37, 399–410.

Ceci, S. J., D. K. Ginther, S. Kahn, and W. M. Williams (2014): "Women in Academic Science: A Changing Landscape," *Psychological Science in the Public Interest*, 15, 75–141.

Ceci, S. J., S. Kahn, and W. M. Williams (2023): "Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration," *Psychological Science in the Public Interest*, 24, 15–73.

De Paola, M., M. Ponzo, and V. Scoppa (2017): "Gender Differences in the Propensity to Apply for Promotion: Evidence from the Italian Scientific Qualification," *Oxford Economic Papers*, 69, 986–1009.

Directorate-General for Research and Innovation (European Commission) (2021): *She Figures 2021: Gender in Research and Innovation : Statistics and Indicators*, LU: Publications Office of the European Union.

Eberhardt, M., G. Facchini, and V. Rueda (2023): "Gender Differences in Reference Letters: Evidence from the Economics Job Market," *The Economic Journal*, uead045.

Franchina, V. and R. Vacca (1986): "Adaptation of Flesh Readability Index on a Bilingual Text Written by the Same Author Both in Italian and English Languages," *Linguaggi*, 3, 47–49.

Hengel, E. (2022): "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review," *The Economic Journal*, 132, 2951–2991.

Hilmer, C. and M. Hilmer (2007): "Women Helping Women, Men Helping Women? Same-Gender Mentoring, Initial Job Placements, and Early Career Publishing Success for Economics PhDs," *American Economic Review*, 97, 422–426.

Hospido, L. and C. Sanz (2021): "Gender Gaps in the Evaluation of Research: Evidence from Submissions to Economics Conferences*," *Oxford Bulletin of Economics and Statistics*, 83, 590–618.

Koffi, M. (2021): "Gendered Citations at Top Economic Journals," *AEA Papers and Proceedings*, 111, 60–64.

Leibbrandt, A. and J. A. List (2015): "Do Women Avoid Salary Negotiations? Evidence from a Large-Scale Natural Field Experiment," *Management Science*, 61, 2016–2024.

Lucisano, P. and M. E. Piemontese (1988): "Gulpease: Una Formula per La Predizione Della Leggibilita Di Testi in Lingua Italiana," *Scuola e Citta*, 110–124.

Lundberg, S. and J. Stearns (2019): "Women in Economics: Stalled Progress," *Journal of Economic Perspectives*, 33, 3–22.

Marzolla, M. (2016): "Assessing Evaluation Procedures for Individual Researchers: The Case of the Italian National Scientific Qualification," *Journal of Informetrics*, 10, 408–438.

Moss-Racusin, C. A., J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman (2012): "Science Faculty's Subtle Gender Biases Favor Male Students," *Proceedings of the National Academy of Sciences*, 109, 16474–16479.

Niederle, M. and L. Vesterlund (2007): "Do Women Shy Away From Competition? Do Men Compete Too Much?*," *The Quarterly Journal of Economics*, 122, 1067–1101.

OECD (2024a): "Employment Indicators 2023," .

——— (2024b): "Gender, Institutions and Development Database (GID-DB) 2023," .

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021): "Gender Differences in Recognition for Group Work," *Journal of Political Economy*, 129, 101–147.

Small, D. A., M. Gelfand, L. Babcock, and H. Gettman (2007): "Who Goes to the Bargaining Table? The Influence of Gender and Framing on the Initiation of Negotiation," *Journal of Personality and Social Psychology*, 93, 600–613.

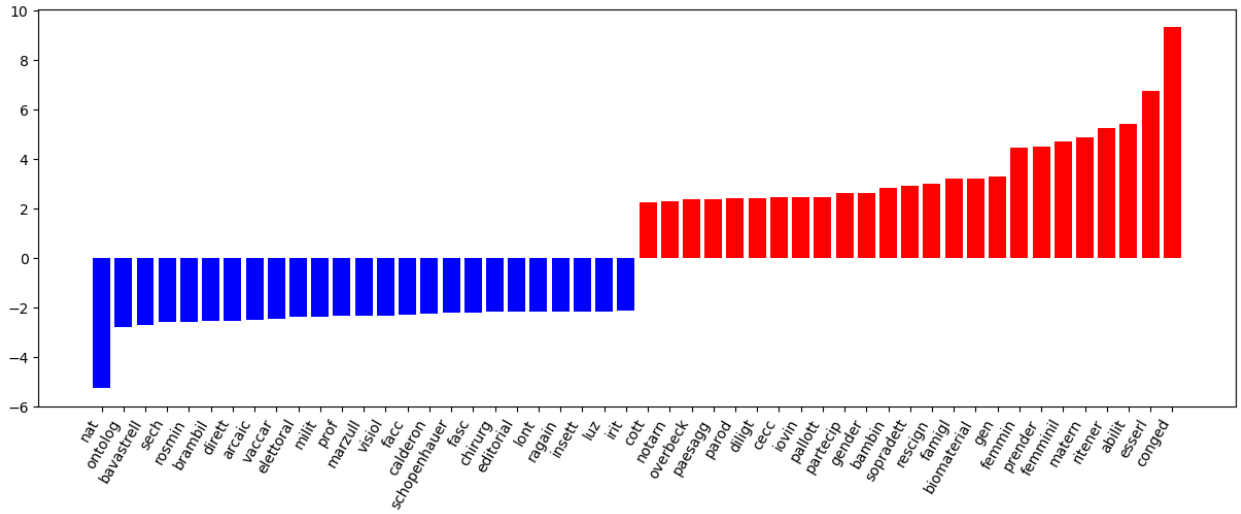
van den Brink, M., Y. Benschop, and W. Jansen (2010): "Transparency in Academic Recruitment: A Problematic Tool for Gender Equality?" *Organization Studies*, 31, 1459–1483.

Weisshaar, K. (2017): "Publish and Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia," *Social Forces*, 96, 529–560.

Williams, W. M. and S. J. Ceci (2015): "National Hiring Experiments Reveal 2:1 Faculty Preference for Women on STEM Tenure Track," *Proceedings of the National Academy of Sciences*, 112, 5360–5365.

Zinovyeva, N. and M. Bagues (2015): "The Role of Connections in Academic Promotions," *American Economic Journal: Applied Economics*, 7, 264–292.

Figure 1: Words with the highest predictive power by gender



Note: Estimates are negative (in blue) for words more predictive of evaluations for men and positive (in red) for words more predictive of evaluations for women. The logistic regression, from which these estimates are derived, includes global TF-IDF weights of words, individual controls, and indicators for exams as predictors.

Many of the word stems appear to be truncated surnames, including 'bavastrell', 'sech', 'rosmin', 'brambil', 'vaccar', 'marzull', 'visiol', 'calderon', 'schopenhauer', 'lont', 'ragain', 'luz', 'notarn', 'overbeck', 'cecc', 'iovin', 'palott', 'rescign'.

Remaining word stems with negative estimates: 'nat' for 'born'; 'ontolog' for 'ontology'; 'dirett' for 'director'; 'arcaic' for 'archaic'; 'elettoral' for 'electoral'; 'milit' for 'military'; 'prof' for 'professor'; 'facc' for 'face'; 'fasc' for 'promotion level'; 'chirurg' for 'surgery'; 'editorial' for 'editorial'; 'insett' for 'insects'; and 'irit' for iritis' (inflammation of the iris) or irritation'.

Remaining word stems with positive estimates: 'cott' for 'baked'; 'paesagg' for 'landscape'; 'parod' for 'parody'; 'diligt' for 'diligence'; 'particip' for 'participate'; 'gender'; 'bambin' for 'children'; 'sopradett' for 'above-mentioned'; 'famigl' for 'family'; 'biomaterial'; 'gen' for 'gender'; 'femmin' for 'female'; 'prender' for 'taking'; 'femminil' for 'femininity'; 'matern' for 'maternity'; 'ritener' for 'consider'; 'abilit' for 'qualified'; 'esser' for 'being'; 'conged' for (maternity) 'leave'.

Table 1: Textual features and connections

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Coauthors	0.02***	0.15***	0.04***	0.12***	−0.06***	−0.04**	0.12***	0.02
	(0.004)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
Same subfield	0.02***	0.17***	0.03***	0.08***	−0.10***	−0.06***	0.17***	0.05***
	(0.003)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Same university	0.04***	0.11***	0.02	0.07***	−0.04***	−0.02*	0.10***	0.01
	(0.004)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Application FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Evaluator FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. The table shows estimates from a regression of a given textual feature on the indicators for the candidate and the evaluator being coauthors, belonging to the same subfield, and coming from the same university, along with controls for application and evaluator fixed effects. All textual features are standardized for candidates applying to the same exam.

Document originally is the average Inverse Document Frequency (IDF) weight of words in the document, with this word varying across evaluators. The total IDF is the sum of all IDF weights for a given evaluation. Gulpease and Vacca scores are readability indexes taking lower values for less readable texts. Total polarity measures the sentiment direction of each evaluation based on the polarity score of all adjectives and adverbs. Average polarity is the average polarity score of all adjectives and adverbs used in the text.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2: Candidates' characteristics and Textual features

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female applicant	−0.0005	0.01*	0.003	0.01	−0.01	0.001	−0.01	−0.01
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Above 1 median	0.19***	0.25***	0.07***	0.12***	−0.01	−0.01	0.30***	0.20***
	(0.01)	(0.03)	(0.02)	(0.02)	(0.04)	(0.03)	(0.02)	(0.02)
Above 2 medians	0.22***	0.13***	0.04***	0.08***	0.04*	0.02	0.20***	0.16***
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
Above 3 medians	0.12***	0.002	0.0001	0.01	0.002	−0.03***	0.13***	0.11***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Top articles	0.05***	0.02***	0.02***	0.03***	−0.02***	−0.01***	0.05***	0.04***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)
Other articles	−0.01***	−0.01**	−0.01***	−0.01***	−0.001	−0.004	−0.01***	−0.01***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Books	−0.005**	0.02***	0.002	0.01**	−0.01*	−0.01	−0.001	−0.01**
	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Chapters	0.02***	0.03***	0.002	0.01***	−0.01**	−0.01***	0.02***	0.01***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Proceedings	0.01*	0.01**	−0.01***	−0.01*	−0.01	−0.002	0.01*	0.01*
	(0.003)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Patents	−0.001	−0.01	−0.004	−0.01*	0.004	−0.001	−0.004	−0.001
	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Other publications	−0.01**	0.01**	−0.001	0.0004	−0.01**	−0.003	−0.004	−0.002
	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Coauthors per paper	−0.03***	−0.01	−0.01	−0.01*	−0.01***	0.0000	−0.04***	−0.03***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)
Percent first-authored	0.02***	0.01***	0.02***	0.02***	−0.01***	−0.01***	0.02***	0.01***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Percent last-authored	0.03***	0.01***	0.01***	0.02***	−0.005	−0.01**	0.03***	0.02***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
AIS	0.01*	0.01**	0.02***	0.02***	−0.01*	−0.01**	0.03***	0.02***
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Tenured, same field	0.23***	0.20***	0.01	0.09**	−0.004	0.02	0.26***	0.20***
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)
Tenured, different field	−0.10***	0.09**	−0.03	−0.01	−0.08	0.01	−0.01	−0.01
	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
Tenured, same field x Associate professor exam	0.02	−0.03	−0.002	−0.02	−0.001	−0.02	−0.01	−0.003
	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Tenured, different field x Associate professor exam	0.06***	−0.02	0.01	0.01	0.02	−0.01	0.02	0.02
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	294,656	294,740	294,740	294,740	252,781	252,781	294,740	294,740
Adjusted R^2	0.36	0.50	0.20	0.12	0.34	0.30	0.40	0.27

Notes: Controls also include university dummies for candidates with a known affiliation. All textual features and productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3: Evaluators' gender and Textual features

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female applicant	0.0004	0.01	0.01	0.01	−0.01	−0.002	−0.01	−0.01
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Female evaluator	0.004	0.11*	−0.05	0.003	−0.12**	−0.04	0.06	0.01
	(0.01)	(0.06)	(0.04)	(0.03)	(0.05)	(0.05)	(0.05)	(0.04)
Female applicant X Female evaluator	−0.005	0.02	−0.03*	−0.03	−0.01	0.01	0.01	0.02
	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes. Controls include all variables presented in Table 2. All textual features and productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam.
Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Candidates' characteristics and the use of selected words

	Quantity	Quality	Impact	Median	Coauthor	Creativity
	(1)	(2)	(3)	(4)	(5)	(6)
Female applicant	0.001	0.002	0.0003	0.001	−0.001	0.002
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
Above 1 median	0.035***	0.066***	0.026***	−0.029***	0.004***	0.032***
	(0.007)	(0.008)	(0.009)	(0.011)	(0.001)	(0.005)
Above 2 medians	0.006	0.041***	0.002	−0.013***	0.003***	0.033***
	(0.005)	(0.007)	(0.006)	(0.005)	(0.001)	(0.004)
Above 3 medians	−0.005	0.013**	−0.010**	0.001	0.002***	0.009***
	(0.004)	(0.005)	(0.004)	(0.004)	(0.001)	(0.003)
Top articles	0.007***	0.004**	0.004***	−0.001	−0.00002	0.006***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.0004)	(0.001)
Other articles	0.003***	−0.005***	0.001	0.0001	−0.001*	−0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.001)
Books	0.002**	0.001	−0.0001	−0.001	0.001**	0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0004)	(0.001)
Chapters	0.004***	0.003***	0.0003	−0.0002	0.001***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0003)	(0.001)
Proceedings	0.002**	−0.0001	0.001*	0.0005	0.00003	−0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0003)	(0.001)
Patents	−0.0004	−0.004***	−0.002**	−0.001	0.002***	−0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Other publications	0.001	0.0005	−0.001	−0.002***	−0.001**	−0.0003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0003)	(0.001)
Coauthors per paper	0.002*	−0.009***	−0.005***	0.0001	0.005***	−0.007***
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Percent first-authored	0.003**	0.004***	0.001	0.00003	0.002**	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Percent last-authored	0.0003	0.003**	0.002**	−0.0004	0.002***	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
AIS	0.003***	0.005**	0.002	−0.002	0.001***	0.004**
	(0.001)	(0.002)	(0.001)	(0.001)	(0.0005)	(0.002)
Tenured, same field	0.018***	0.043***	0.039***	0.010**	0.002	0.051***
	(0.006)	(0.009)	(0.007)	(0.004)	(0.002)	(0.008)
Tenured, different field	−0.015***	−0.016*	0.003	−0.004	−0.002	−0.001
	(0.005)	(0.009)	(0.007)	(0.004)	(0.002)	(0.006)
Tenured, same field x	0.001	0.004	−0.019**	−0.007	−0.001	−0.013
Associate professor exam	(0.008)	(0.012)	(0.008)	(0.005)	(0.003)	(0.010)
Tenured, different field x	0.012*	0.013	−0.009	−0.001	0.003	−0.004
Associate professor exam	(0.006)	(0.010)	(0.007)	(0.005)	(0.002)	(0.007)
Committee FE	Yes	Yes	Yes	Yes	Yes	Yes
Evaluator FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.534	0.557	0.695	0.817	0.328	0.548

Notes: The outcome variables are indicators for whether a given (stemmed) word, or close synonyms, are found in the text. Controls also include university dummies for candidates with a known affiliation. All productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

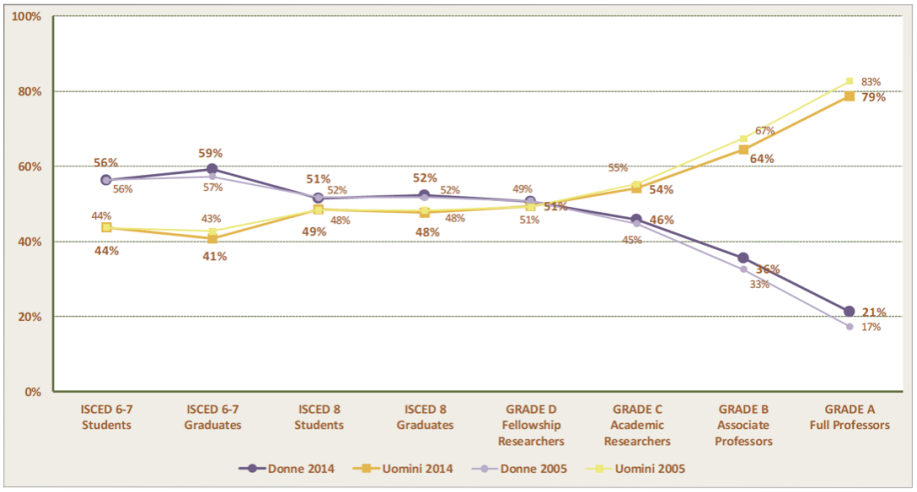
Table 5: Performance of the Logistic Classifier

		Global TF-IDF weights		Field-specific TF-IDF weights	
		Full model	Text only	Full model	Text only
ROC AUC score		0.67	0.58	0.61	0.50
Precision	Men	0.73	0.67	0.70	0.63
	Women	0.67	0.67	0.59	0.40
Recall	Men	0.86	0.93	0.84	0.99
	Women	0.47	0.23	0.38	0.01

Notes: The table provides performance statistics for the Logistic Regression. The full model includes both TF-IDF weights for words in evaluation texts and individual controls. The model was trained on 80% of the reports by Italian evaluators and then applied to a test sample.
ROC AUC score stands for the Area Under the Receiver Operating Characteristic Curve. Precision is the ratio of true positives to predicted positives. Recall is the ratio of true positives to actual positives.

Appendix

FIGURE A1: PROPORTION OF WOMEN AND MEN BY STEPS IN ACADEMIC CAREER (2004-2015)



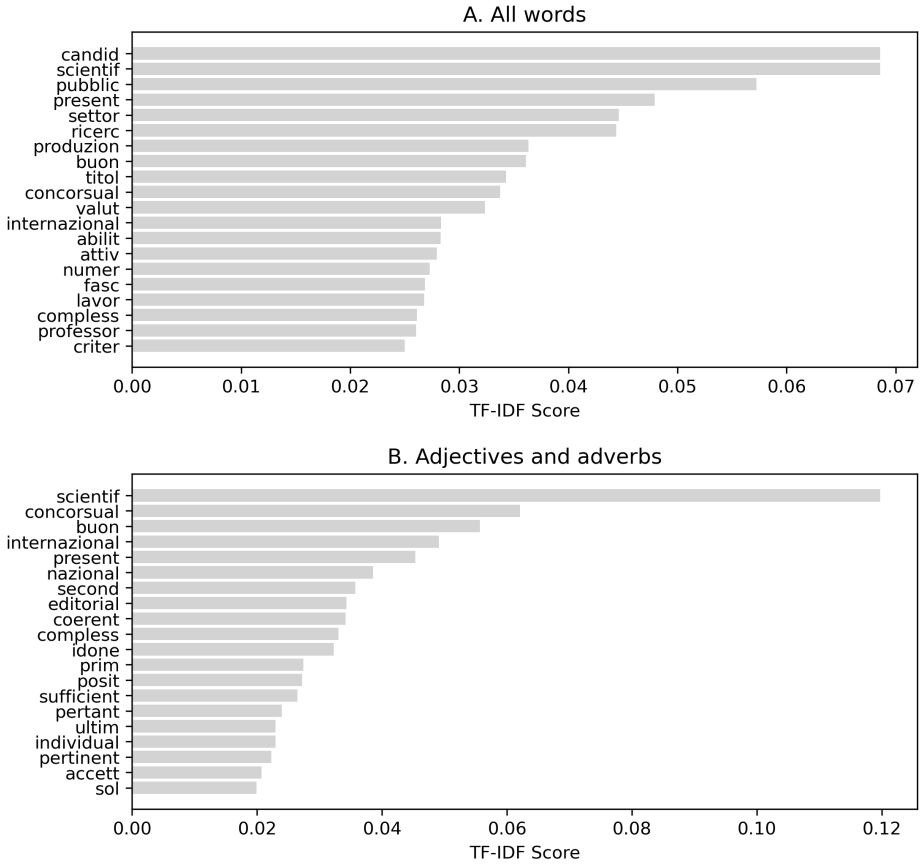
Notes. According to the 2011 UN International Standard Classification of Education (ISCED) which belongs to the United Nations International Family of Economic and Social Classifications, the ISCED level 6 corresponds to Bachelor's or equivalent level, while the ISCED level 7 to Master's or equivalent level and the ISCED level 8 to a Doctoral or equivalent degree. The grade classification for the academic positions comes from the Manuale di Frascati (2015), also used in the MIUR publication Focus Le cariche femminili in ambito accademico (2016) and the EU publication She Figures (2018). The dark lines represent the Proportion (%) of women and men in typical academic with data on 2014, violet for females and yellow for males. The lighter lines are computed with data on 2005.
Source: MIUR (2016). Focus Le cariche femminili in ambito accademico. Ministero Dell' Universita' e della Ricerca.

FIGURE A2: ANONYMIZED SAMPLE EVALUATION

DOE John

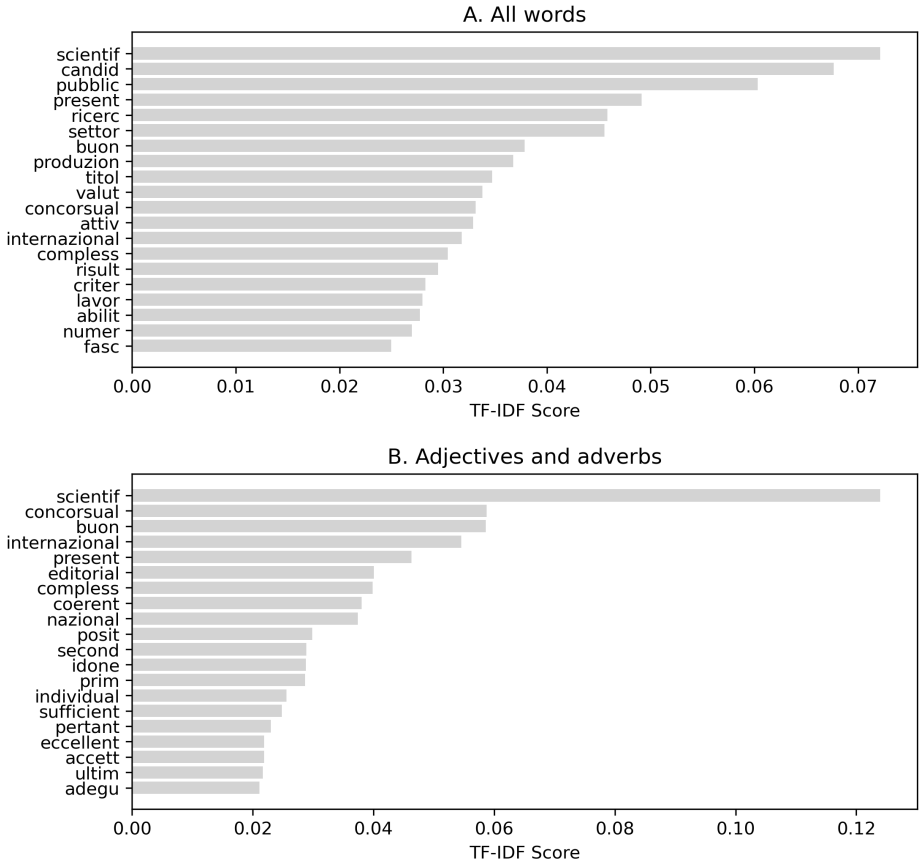
The candidate PINCO PALLO has been Ricercatore universitario at the Università di […] since […]. His scientific work is concerned with the development of democracy, including a monograph on the role of public opinion in political thought and a series of contributions concerning English and Anglo-American thought and developments from the 17th through 19th centuries, with special reference to […]. The candidate is a member of the […] project, based at the University of Oxford. The candidate has a significant number of international conference participations, among which those in which the English have invited him to speak about […] are perhaps the most indicative of a strong international reputation. In terms of specific contributions, the […] metaphor is particularly significant in explaining how […] plays out in the history of Italian political thought. The candidate scores above the median on two of the three indicators of impact and has substantial relevant teaching experience. On the basis of the application submitted, the candidate merits approval of the request for the abilitazione scientifica.

Figure A3: Frequency plot of the top 20 (stemmed) words for female applicant



Notes. The top frequent words and adjectives are in Italian. Starting from the top, *candidat* stands for "candidate", *scientif* for "scientific", *pubblic* for "publish", *present* for "show", *settor* for "sector", *ricerc* for "research", *produzione* for "production", *buon* for "good", *titol* for "title", *concorsual* for "of the competition", *valut* for "evaluate", *internazional* for "international", *abilit* for "abiliity", *attiv* for "active", *numer* for "number", *fasc* for "level", *lavor* for "work", *compless* for "complex", *professor* for "professor", *criter* for "criterio", *nazional* for "national", *second* for "according", *coerent* for "coherent", *prim* for "first", *posit* for "positive", *sufficient* for "sufficient", *partant* for "therefore", *ultim* for "last", *individual* for "individual", *pertinent* for "pertinent", *accett* for "accettable", *sol* for "only".

Figure A4: Frequency distribution of the top 20 (stemmed) words for male applicant



Notes. The top frequent words and adjectives are in Italian. Starting from the top, *scientif* stands for "scientific", *candidat* for "candidate", *pubblic* for "publish", *present* for "show", *ricerc* for "research", *settor* for "sector", *buon* for "good", *produzione* for "production", *titol* for "title", *valut* for "evaluate", *concorsual* for "competitive",*internazional* for "international", *compless* for "complex", *risult* for "is", *criter* for "criteria", *lavor* for "work", *numer* for "number", *fasc* for "level", *editorial* for "editorial", *coerent* for "coherent", *nazional* for "national", *posit* for "positive", *second* for "secondo", *ideone* for "idoneo", *prim* for "first", *individual* for "individual", *sufficient* for "sufficient", *pertant* for "therefore", *eccellent* for "eccelent", *accett* for "accettable", *ultim* for "last", and *adegua* for "adequat".

Table A1: Dictionaries for selected dimensions of research productivity

Keyword	Word stems:		Occurrence in reports:	
	English	Italian	Men	Women
Quantity	Quantit, amount, number, total	Quantit, numero, total	0.330	0.313
Quality	Qualit, excellenc, level, merit	Qualit, eccelenz, livello, merit, valore	0.588	0.528
Impact	Impact, influenc, citat, referenc, mention	Impatto, influenz, citazion, referenz, menzion	0.390	0.321
Median	Median	Median	0.340	0.345
Coauthor	Coaut, co-aut	Coaut, co-aut	0.023	0.022
Creativity	Creativ, inventiv, innovat, original, ingenuiti, novel, uniqu, new, pioneering, cutting-edge, forward-thinking, groundbreaking	Creativ, inventiv, innovat, original, ingegnno, nuovo, unico, pioneristico, avanguardia	0.264	0.231

Notes: The table shows dictionaries of (stemmed) words, in English and Italian language, characterizing selected dimensions of research productivity and the share of reports written for women and men that contain corresponding keywords.

Table A2: Descriptive statistics of numerical variables

Statistic	Mean	St. Dev.	Min	Max
PANEL A: evaluators				
Female evaluator	0.19	0.39	0	1
Italian evaluator	0.86	0.35	0	1
Evaluator based in Italy	0.82	0.38	0	1
PANEL B: applicants				
Female applicant	0.37	0.48	0	1
Qualified	0.43	0.49	0	1
Individual vote	0.45	0.50	0	1
Total publications	66.74	69.16	1	999
Top articles	15.14	26.60	0	573
Other articles	24.00	34.29	0	933
Books	2.58	4.68	0	103
Chapters	7.56	12.50	0	270
Proceedings	10.07	20.49	0	382
Patents	0.26	1.68	0	108
AIS	1.25	0.97	0.00	16.03
Coauthors per paper	6.27	19.00	1.00	526.23
Share of first-authored	0.22	0.19	0.00	1.00
Share of last-authored	0.12	0.16	0.00	1.00
Above one median	0.84	0.36	0	1
Above two medians	0.64	0.48	0	1
Above three medians	0.38	0.48	0	1
Fixed term contract	0.02	0.15	0	1
Same field	0.40	0.49	0	1
Applicant for associate professorship	0.69	0.46	0	1
PANEL C: evaluation reports				
Number of words	178.83	282.46	0	16,457
Document originality	0.01	0.02	0.00	1.22
Total IDF	0.40	1.81	0.00	784.78
Gulpease score	55.11	9.06	0.00	100.00
Vacca score	64.42	19.94	0.00	100.00
Total polarity	2.58	4.32	−208.64	233.18
Average polarity	0.12	0.15	−0.88	1.00
Word 'Quantity'	0.32	0.47	0	1
Word 'Quality'	0.57	0.50	0	1
Word 'Impact'	0.36	0.48	0	1
Word 'Median'	0.34	0.47	0	1
Word 'Coauthor'	0.02	0.15	0	1
Word 'Creativity'	0.25	0.43	0	1

Notes: The table shows descriptive statistics on the full sample of 294,740 evaluations. AIS is the average Article Influence Score of articles published in the journals indexed in the Web of Science. Top articles are the number of articles in the top quartile journals according to their AIS in STEMM fields and the number of A-journal articles as classified by ANECA in SSH fields.

Document originally is the average IDF weight of words in the document, where IDF weight for each word varies across evaluators. The total IDF is the sum of all IDF weights for a given evaluation. Gulpease and Vacca scores are readability indexes taking lower values for less readable texts. Total polarity measures the sentiment direction of each evaluation based on the polarity score of all adjectives and adverbs. Average polarity is the average polarity score of all adjectives and adverbs used in the text.

Table A3: T-test results by gender of the applicants and evaluators

Variable	Mean (men)	Mean (women)	T-Statistics	P-Value
PANEL A: outcomes & controls (applicants)				
Qualified	0.436	0.415	5.138	0.000
Position	0.657	0.758	-26.516	0.000
Individual vote	0.454	0.433	4.974	0.000
Total publications	0.036	-0.061	11.820	0.000
Top articles	0.037	-0.063	12.372	0.000
Other articles	0.053	-0.090	17.734	0.000
Books	0.049	-0.083	16.255	0.000
Chapters	0.010	-0.016	3.117	0.002
Proceedings	-0.010	0.016	-3.026	0.002
Patents	0.021	-0.035	8.020	0.000
AIS	0.019	-0.033	7.601	0.000
Coauthors per paper	-0.023	0.040	-7.328	0.000
Share of first-authored	0.001	-0.001	0.261	0.794
Share of last-authored	0.023	-0.038	7.210	0.000
Above one median	0.848	0.831	5.465	0.000
Above two medians	0.657	0.613	10.722	0.000
Above three medians	0.399	0.338	14.965	0.000
Applicant for associate professorship	0.657	0.758	-26.516	0.000
Fixed-term contract	0.023	0.025	-0.969	0.332
Same field	0.391	0.413	-5.239	0.000
PANEL B: textual features (by evaluators' gender)				
Number of words	-0.016	0.073	-19.761	0.000
Document originality	0.009	-0.039	10.177	0.000
Total IDF	0.001	-0.005	1.383	0.167
Gulpease score	0.017	-0.071	18.167	0.000
Vacca index	0.004	-0.017	4.206	0.000
Total polarity	-0.008	0.037	-9.507	0.000
Average polarity	-0.002	0.011	-2.971	0.003
Word 'Quantity'	0.004	-0.016	4.299	0.000
Word 'Quality'	0.005	-0.021	5.688	0.000
Word 'Impact'	0.015	-0.068	17.665	0.000
Word 'Median'	-0.008	0.037	-10.459	0.000
Word 'Coauthor'	-0.002	0.008	-2.347	0.019
Word 'Creativity'	-0.009	0.039	-9.923	0.000
PANEL C: textual features (by applicants' gender)				
Number of words	-0.002	0.004	-1.514	0.130
Document originality	0.001	-0.002	0.999	0.318
Total IDF	0.001	-0.002	0.809	0.419
Gulpease score	0.002	-0.003	1.142	0.254
Vacca index	-0.002	0.004	-1.431	0.153
Total polarity	0.007	-0.012	5.008	0.000
Average polarity	0.005	-0.009	3.825	0.000
Word 'Quantity'	-0.000	0.000	-0.077	0.939
Word 'Quality'	-0.002	0.004	-1.615	0.106
Word 'Impact'	-0.001	0.003	-1.112	0.266
Word 'Median'	-0.004	0.006	-2.891	0.004
Word 'Coauthor'	0.001	-0.002	0.835	0.403
Word 'Creativity'	-0.001	0.001	-0.496	0.620

Notes: All productivity indicators, with the exception of medians indicators, and all textual features are standardized at the exam (committee x position) level.

Table A4: Proportion of women by exam field

Exams	Female applicants (%)	Female evaluators (%)
Mathematics and information science	25.90	20.00
Physics	21.96	0.00
Chemistry	44.91	17.50
Geosciences	26.61	20.00
Biology	52.43	20.00
Medicine	32.51	6.92
Agricultural science and veterinary	40.89	17.14
Architecture and civil engineering	37.68	10.00
Industrial and computer engineering	20.23	5.00
Literature and art	51.72	40.00
History, philosophy, pedagogy and psychology	41.47	29.41
Law	34.60	21.25
Business, economics and statistics	34.76	26.67
Political and social sciences	38.20	22.86

Notes: The table is based on the full sample of 294,740 evaluations.).

Table A5: Candidates' characteristics, textual features, and individual vote: Excluding potentially copy-pasting evaluators

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female applicant	−0.0000	0.01*	0.004	0.01	−0.01	0.001	−0.001	−0.01
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Above 1 median	0.19***	0.19***	0.06**	0.10***	0.01	0.01	0.26***	0.21***
	(0.01)	(0.03)	(0.02)	(0.02)	(0.04)	(0.04)	(0.02)	(0.02)
Above 2 medians	0.21***	0.10***	0.03**	0.07***	0.06**	0.03	0.18***	0.17***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)
Above 3 medians	0.12***	−0.01	−0.01	−0.01	−0.003	−0.04***	0.12***	0.13***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Top articles	0.05***	0.02***	0.02***	0.03***	−0.02***	−0.01**	0.05***	0.04***
	(0.004)	(0.004)	(0.005)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Other articles	−0.01***	−0.01*	−0.01***	−0.01***	−0.003	−0.003	−0.01***	−0.02***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Books	−0.01***	0.02***	0.004	0.01**	−0.01	−0.01	−0.001	−0.01**
	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Chapters	0.02***	0.02***	0.003	0.01***	−0.01**	−0.01***	0.02***	0.01***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)
Proceedings	0.01*	0.01*	−0.01***	−0.01	−0.01*	−0.005	0.01**	0.01
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Patents	−0.002	−0.005	−0.01*	−0.01**	0.0004	−0.004	−0.003	−0.001
	(0.002)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Other publications	−0.004**	0.004	−0.002	−0.001	−0.01**	−0.001	−0.003	−0.001
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Coauthors per paper	−0.03***	−0.005	−0.01	−0.01	−0.01***	0.002	−0.04***	−0.03***
	(0.003)	(0.004)	(0.005)	(0.004)	(0.01)	(0.005)	(0.005)	(0.005)
Percent first-authored	0.02***	0.01***	0.02***	0.02***	−0.01***	−0.01**	0.02***	0.02***
	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Percent last-authored	0.03***	0.01***	0.01***	0.02***	−0.004	−0.01**	0.02***	0.02***
	(0.002)	(0.003)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.01)
AIS	0.01*	0.01	0.01**	0.02**	−0.01	−0.01	0.02***	0.02***
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Tenured, same field	0.23***	0.17***	0.02	0.09**	0.01	0.02	0.23***	0.19***
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)	(0.06)
Tenured, different field	−0.10***	0.08*	−0.04	−0.005	−0.04	0.03	−0.01	0.001
	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
Tenured, same field	0.02	−0.03	−0.01	−0.02	0.01	−0.02	−0.004	0.003
x Associate professor exam	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Tenured, different field	0.06***	−0.03	0.01	0.002	0.01	−0.02	0.02	0.02
x Associate professor exam	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	176,799	176,844	176,844	176,844	149,542	149,542	176,844	176,844
Adjusted R^2	0.36	0.42	0.20	0.12	0.32	0.29	0.37	0.26

Notes. Estimates are based the sample of individual reports written by evaluators with the committee-level median or higher ratio of the number of different words over the total number of words they used in all evaluations.

Controls also include university dummies for candidates with a known affiliation. All textual features and productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

TABLE A6: CANDIDATES' CHARACTERISTICS AND TEXTUAL FEATURES, CONDITIONAL ON INDIVIDUAL VOTE

	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female applicant	0.01*	0.003	0.01	−0.01	0.001	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Individual vote	0.12***	0.08***	0.14***	0.09***	−0.02	0.58***	0.52***
	(0.02)	(0.01)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Above 1 median	0.23***	0.05***	0.10***	−0.03	−0.01	0.19***	0.10***
	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
Above 2 medians	0.10***	0.03**	0.05***	0.02	0.02	0.08***	0.04***
	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)
Above 3 medians	−0.01	−0.01	−0.01	−0.01	−0.03***	0.06***	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Top articles	0.01***	0.01***	0.02***	−0.02***	−0.01**	0.02***	0.01**
	(0.004)	(0.004)	(0.004)	(0.01)	(0.01)	(0.004)	(0.005)
Other articles	−0.01	−0.01**	−0.01***	−0.0002	−0.004	−0.005	−0.01*
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Books	0.02***	0.002	0.01**	−0.01*	−0.01	0.001	−0.01*
	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Chapters	0.03***	0.001	0.01**	−0.01***	−0.01***	0.01**	−0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Proceedings	0.01**	−0.01***	−0.01*	−0.01	−0.002	0.005	0.004
	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)
Patents	−0.01	−0.003	−0.01*	0.004	−0.001	−0.003	−0.001
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)
Other publications	0.01**	−0.001	0.001	−0.01**	−0.003	−0.001	0.0004
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Coauthors per paper	−0.002	−0.003	−0.002	−0.01**	−0.001	−0.02***	−0.01***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Percent first-authored	0.01**	0.02***	0.02***	−0.01***	−0.01**	0.01**	0.003
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Percent last-authored	0.01***	0.01***	0.01***	−0.01*	−0.01**	0.01***	0.01*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
AIS	0.01*	0.02***	0.02***	−0.01**	−0.01**	0.02***	0.01***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.005)
Tenured, same field	0.17***	−0.004	0.06	−0.02	0.02	0.13***	0.08*
	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)
Tenured, different field	0.10**	−0.02	0.01	−0.07	0.005	0.05	0.04
	(0.04)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
Tenured, same field	−0.04	−0.01	−0.03	0.01	−0.01	−0.02	−0.02
x Associate professor exam	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)
Tenured, different field	−0.03	0.005	−0.01	0.02	−0.002	−0.01	−0.01
x Associate professor exam	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Observations	294,656	294,656	294,656	252,706	252,706	294,656	294,656
Adjusted R^2	0.50	0.20	0.12	0.34	0.30	0.45	0.31

Notes. Controls also include university dummies for candidates with a known affiliation. All textual features and productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A7: Candidates' gender and Textual features, by the number of high-impact journal articles and field

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PANEL A: Highly productive candidates								
ALL								
Female applicant	-0.01^{*}	0.01	0.005	0.01	-0.01	0.01	-0.02^{*}	-0.02^{**}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH								
Female applicant	-0.02^{*}	0.03^{**}	0.01	0.04^{**}	0.003	0.02^{*}	-0.03^{***}	-0.03^{***}
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
STEMM								
Female applicant	-0.01	0.0003	-0.005	-0.003	-0.02	-0.004	-0.005	-0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
PANEL B: Less productive candidates								
ALL								
Female applicant	0.01	0.01	0.001	0.001	-0.01	-0.002	-0.01	-0.005
	(0.004)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH								
Female applicant	0.004	0.01	0.0003	0.001	0.004	0.01	-0.001	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
STEMM								
Female applicant	0.01	0.01	-0.0004	-0.001	-0.02^{**}	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes. Controls include all variables presented in Table 2. All textual features are standardized for candidates applying to the same exam.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A8: Candidates' gender and Textual features conditional on Individual Vote, by the number of high-impact journal articles and the field of study

	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PANEL A: Highly productive candidates							
ALL							
Female applicant	0.01*	0.01	0.02	−0.01	0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH							
Female applicant	0.04**	0.02	0.04**	0.002	0.02*	−0.03**	−0.03**
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
STEMM							
Female applicant	0.001	−0.004	−0.002	−0.01	−0.004	0.002	−0.001
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
PANEL B: Less productive candidates							
ALL							
Female applicant	0.01	0.001	−0.0001	−0.01*	−0.002	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH							
Female applicant	0.01	0.001	0.0002	0.003	0.01	−0.002	0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
STEMM							
Female applicant	0.01	−0.001	−0.003	−0.02**	−0.01	−0.01*	−0.02*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes. Controls include all variables presented in Table 2 plus individual vote of the evaluator.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A9: Candidates' gender and Textual features by the number of median criteria satisfied

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PANEL A: 3 median criteria (110,794 obs.)								
ALL								
Female applicant	−0.01	−0.004	0.004	0.005	−0.01	0.003	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH								
Female applicant	−0.01	0.02	0.02	0.02	−0.01	−0.0003	−0.03**	−0.03
	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)
STEMM								
Female applicant	−0.01	−0.01	0.002	0.001	−0.01	0.003	−0.01	−0.004
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
PANEL B: 2 median criteria (77,913 obs.)								
ALL								
Female applicant	−0.002	0.01	−0.001	0.003	−0.01	−0.01	−0.01	−0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH								
Female applicant	−0.004	0.01	−0.003	0.02	0.01	0.02	−0.02	−0.003
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
STEMM								
Female applicant	0.001	0.01	−0.002	−0.01	−0.03**	−0.03**	−0.01	−0.02
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
PANEL C: 0-1 median criteria (57,225 obs.)								
ALL								
Female applicant	0.01	0.02**	0.002	0.01	−0.01	0.0004	−0.002	−0.01
	(0.005)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH								
Female applicant	−0.001	0.02*	0.003	0.01	0.01	0.02	0.004	0.01
	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
STEMM								
Female applicant	0.01**	0.02*	−0.002	0.01	−0.03**	−0.02	−0.01	−0.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes. Controls include all variables presented in Table 2. All textual features and productivity indicators, apart from indicators for median performance, are standardized for candidates applying to the same exam.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

TABLE A10: EVALUATORS' GENDER AND TEXTUAL FEATURES, BY CANDIDATES' PRODUCTIVITY AND FIELD

	Individual vote	Number of words	Document originality	Total IDF	Gulpease index	Vacca index	Total polarity	Average polarity
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PANEL A: Highly productive candidates								
ALL								
Female applicant	0.01^{*}	0.02	−0.002	0.01	−0.03	−0.01	0.01	0.01
X Female evaluator	(0.004)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
SSH								
Female applicant	−0.001	0.04^{**}	−0.0002	0.01	−0.02	−0.03	0.01	0.01
X Female evaluator	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
STEMM								
Female applicant	0.02^{***}	0.003	−0.004	0.004	−0.04	−0.001	0.01	0.01
X Female evaluator	(0.005)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)
PANEL B: Less productive candidates								
ALL								
Female applicant	0.001	0.004	0.005	0.01	-0.04^{***}	-0.03^{**}	0.004	0.02
X Female evaluator	(0.003)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
SSH								
Female applicant	0.01	0.0004	−0.01	−0.003	−0.02	−0.02	0.001	0.03^{*}
X Female evaluator	(0.004)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
STEMM								
Female applicant	−0.01	0.01	0.03	0.02	-0.06^{**}	-0.04^{**}	0.01	0.01
X Female evaluator	(0.004)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Applicant FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes. Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A11: Candidates' gender and dimensions of productivity, by the number of high-impact journal articles and field

	Quantity	Quality	Impact	Median	Coauthor	Creativity
	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: Highly productive candidates						
ALL						
Female applicant	0.003	0.003	0.001	0.001	−0.001	0.002
	(0.003)	(0.004)	(0.002)	(0.002)	(0.001)	(0.003)
SSH						
Female applicant	0.005	0.001	0.002	−0.0003	−0.002	−0.004
	(0.004)	(0.01)	(0.003)	(0.003)	(0.002)	(0.01)
STEMM						
Female applicant	0.001	0.004	0.001	0.002	−0.001	0.01
	(0.004)	(0.004)	(0.003)	(0.003)	(0.001)	(0.004)
PANEL B: Less productive candidates						
ALL						
Female applicant	0.0003	0.001	−0.0004	0.001	−0.001	0.003
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
SSH						
Female applicant	0.01*	0.01	0.0001	−0.002	0.001	0.003
	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)	(0.004)
STEMM						
Female applicant	−0.004	−0.002	−0.001	0.003*	−0.001	0.002
	(0.003)	(0.004)	(0.002)	(0.002)	(0.001)	(0.002)
Committee FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Evaluator FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes. The outcome variables are indicators for whether a given (stemmed) word, or close synonyms, are found in the text. Controls include all variables presented in Table 2.

Standard errors are clustered at the exam level. Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.