nature geoscience

LETTERS

PUBLISHED ONLINE: 3 OCTOBER 2016 | DOI: 10.1038/NGEO2819

Gender differences in recommendation letters for postdoctoral fellowships in geoscience

Kuheli Dutt[1]*, Danielle L. Pfaff[2], Ariel F. Bernstein[2], Joseph S. Dillard[2] and Caryn J. Block[2]

Gender disparities in the fields of science, technology, engineering and mathematics, including the geosciences, are well documented and widely discussed[1,2]. In the geosciences, despite receiving 40% of doctoral degrees, women hold less than 10% of full professorial positions[3]. A significant leak in the pipeline occurs during postdoctoral years[4], so biases embedded in postdoctoral processes, such as biases in recommendation letters, may be deterrents to careers in geoscience for women. Here we present an analysis of an international data set of 1,224 recommendation letters, submitted by recommenders from 54 countries, for postdoctoral fellowships in the geosciences over the period 2007–2012. We examine the relationship between applicant gender and two outcomes of interest: letter length and letter tone. Our results reveal that female applicants are only half as likely to receive excellent letters versus good letters compared to male applicants. We also find no evidence that male and female recommenders differ in their likelihood to write stronger letters for male applicants over female applicants. Our analysis also reveals significant regional differences in letter length, with letters from the Americas being significantly longer than any other region, whereas letter tone appears to be distributed equivalently across all world regions. These results suggest that women are significantly less likely to receive excellent recommendation letters than their male counterparts at a critical juncture in their career.

Under-representation of women in science, technology, engineering and math (STEM) disciplines, including the geosciences, is a well-documented phenomenon. Women occupy only 24% of STEM postdoctoral positions at federally funded R&D centres[5], despite being awarded 41% of STEM doctoral degrees[6]. Explanations for such under-representation range from implicit gender bias to historical, social and institutional factors to the 'leaky pipeline'—that is, women leave scientific fields at higher rates than males[1,2,7]. Of particular relevance to this study is the implicit gender bias framing of this issue. Research has shown that, compared to female candidates, equivalent male candidates in STEM fields are rated more highly, given higher starting salaries and greater mentoring[8], perceived as more competent[9], and twice as likely to be hired[10,11]. While gender disparities are observed across the entire scientific academic trajectory, postdoctoral years are associated with the largest leak in the pipeline for female scientists, with women 35% less likely to get a tenure-track position than men[4].

Specific to the geosciences, women hold fewer than 10% of full professor positions, despite holding around 40% of all geoscience doctoral degrees[3], so a deeper examination of how females are perceived compared to males at the postdoctoral stage is important. Recommendation letters play a key role in academic selection processes, as they contribute to the overall perception of a candidate's 'fit' for a position and often provide the first impression of the applicant[12,13]. Further, recommendation letters offer personal information about the candidate, and due to the subjective nature of these letters, the biases of the writer are more likely to surface[11,14,15]. Implicit biases can surface via the way applicants are described in recommendation letters, with women being described as less confident and forceful, and more nurturing and helpful than men[12], and receiving fewer 'standout' adjectives such as superb and brilliant, and more 'grindstone' adjectives such as hardworking and diligent[13,14]. Also, women are under-represented in fields where raw, innate intellectual talent is considered a requirement for success, since women are stereotyped as not possessing such talent[16].

Table 1 | Recommendation letters by gender.

	Female applicant	Male applicant	Total
Female recommender	67	81	148
Male recommender	295	781	1,076
Total	362	862	1,224

Thus, there is evidence of qualitative differences in recommendation letters written for male versus female applicants. However, past research has several limitations, including: lack of an international data set and/or limited statistical ability to explore regional differences[12–14]; use of descriptive, rather than inferential statistics[13]; inclusion of letters for only selected candidates, as opposed to letters for all applicants[13]; software and coding limitations due to an inability to account for the context in which various words and phrases are used[12,14], and failure to examine the overall letter tone, which may play an important role in evaluators' overall impressions of applicants. This present study addresses these limitations by examining recommendation letters submitted for highly selective postdoctoral fellowships in the geosciences (acceptance rate of 3.8%) at a competitive university in the Northeast US.

To our knowledge, this is the only research study ever published on gender bias in recommendation letters in the geosciences—a field strongly dominated by males. This is also, to our knowledge, the single largest study of gender bias in recommendation letters in any STEM field so far. Further, our sample allows us to expand upon the work of prior researchers[13] via robust statistical analyses of potential regional differences in the tone and length of letters. The international nature of our sample is of particular significance, given the steadily increasing rates of graduate school applications from international students across the globe, particularly in the STEM fields[17], implying increasing globalization of the workforce.

[1]Lamont-Doherty Earth Observatory of Columbia University, New York 10964, USA. [2]Teachers College, Columbia University, New York 10027, USA.
*e-mail: kdutt@ldeo.columbia.edu

Table 2 | Summary of coding scheme.

	Overall tone	Individual comments
Excellent	Reflected the applicant's potential as a top-notch scientist; stated that the applicant was superior to other students; and/or praised the applicant for conducting novel or groundbreaking research, and/or being a scientific leader and role model	Examples: 'scientific leader'; 'brilliant scientist'; 'one of the best students I've ever had'; 'trailblazer'; and 'role model'. Also, references to accomplishments, such as publications, conference presentations, and awards/honours
Good	Provided clear praise and portrayed the applicant as a solid scientist doing good/very good work, yet were less likely to declare the applicant as comparatively superior to others or praise the applicant's potential to become a scientific leader or role model	Examples: 'highly intelligent'; 'very productive'; 'thorough understanding of the subject matter'; 'very knowledgeable'; and 'very good skill set'. Also, comments that serve as an acknowledgement of the applicant's knowledge/familiarity with the subject matter, for example, 's/he worked on X project'; 's/he has taken courses in Y subject'
Doubtful	Questioned the applicant's calibre as a scientist, and expressed uncertainty that the applicant would become a successful scientist	Examples: 'I haven't worked directly with him/her'; 'I haven't seen any leadership skills'; 'I don't think s/he will make a top-notch scientist'; 'I don't know him/her very well'.

Besides, the findings from a large, international data set such as ours are relevant to institutions, scientists and policymakers all over the world. The global applicability of our results strengthens our conclusions and findings.

Our research questions examine the relationship between applicant gender and two outcomes of interest: letter length and letter tone. Earlier studies show that letter length is positively associated with overall letter quality[18] and that women tend to receive fewer long letters (letters over 50 lines) compared to men, and more short letters (10 lines or less) compared to men[13]. Therefore, we explore letter length as both an outcome variable, and a control variable in predicting overall letter tone. We included recommender region as a control variable to allow for the possibility that region might account for letter length and tone differences. Finally, we included recommender gender as a control variable to examine whether male and female recommenders write differently. Specifically, this study seeks to identify whether male and female applicants receive similar letters of recommendation and asks the following questions: First, does applicant gender influence letter length, after controlling for the effects of recommender gender, and recommender region? Second, does applicant gender influence letter tone, after controlling for the effects of recommender gender, recommender region, and letter length?

Our sample comprised 1,224 letters, written for 452 applicants (averaging 2.71 letters per applicant) by 1,101 recommenders from 54 countries (entire list given in the Methods) over the period 2007–2012. There were 137 female applicants (30.31%) with a total of 362 letters (averaging 2.64 letters per applicant) and 315 male applicants (69.69%) with a total of 862 letters (averaging 2.74 letters per applicant). There were 133 female recommenders (12.09%) and 967 male recommenders (87.91%). Of note, 105 recommenders wrote letters for more than one applicant, but because the overall proportion of recommenders who wrote multiple letters was so small (9.54%) we chose to treat each recommender as independent in later analyses. Table 1 summarizes the recommendation letters by gender.

Countries were grouped according to the United Nations classification system, coupled with the classification scheme of Trix and Psenka[13]. The resulting categories included Africa and Middle East (3.75% of letters), Australia, Europe, and New Zealand (20.7% of letters), East Asia and Pacific (9.9% of letters), South Asia (9.0% of letters), and the Americas (all of North, Central, and South America; 56.7% of letters). A coding manual (available from the first author upon request) was developed, and letters were coded into one of three tones: excellent, good, or doubtful. This was based on the overall content of each letter, and whether the applicant was portrayed as having the potential to become a successful scientist. The coding scheme is explained in greater detail in the Methods. Excellent letters comprised 20.9% of the data set ($n = 256$), good letters comprised 76.6% ($n = 937$), and doubtful letters 2.5% ($n = 31$). Table 2 summarizes the coding scheme, while Tables 3 and 4 provide the mean word length of letters, and letter tone by applicant gender, respectively.

Table 3 | Mean letter length by region.

Region	*N*	Mean	s.d.	Minimum	Maximum
Africa and Middle East	46	304.76	238.96	98	1,074
Australia, Europe, and New Zealand	253	345.05	187.42	60	986
South Asia	110	274.56	127.64	52	745
East Asia and Pacific	121	319.64	133.92	101	858
The Americas	694	561.06	311.49	37	2,444
Total	1,224	457.16	286.36	37	2,444

s.d., standard deviation.

Two fixed-effects hierarchical linear models (HLMs) were used to examine the relationship between applicant gender and letter length. Doubtful letters were excluded from both models, due to their small number. First, the level 2 variable (applicant gender) predicted the intercept of the level 1 variable, letter length. Results were nonsignificant, $t(438.14) = -0.02$, $p > 0.05$, indicating that applicant gender was not a significant predictor of letter length. Next, a more robust model was utilized wherein the level 2 variable (applicant gender) predicted the intercept of the level 1 variable, letter length, with recommender gender and recommender region as control variables. Males are the reference group for applicants and recommenders; the 'Americas' group is the reference group for region. Table 5 depicts the results of the second model.

Consistent with the initial model, in the second model, applicant gender was not a significant predictor of letter length ($p = 0.22$). Recommender gender was also not significant ($p = 0.17$). That is, female and male applicants did not receive letters of differing lengths, nor did male recommenders write significantly longer or shorter letters than female recommenders. This is unsurprising, given the huge variation in letter length and the regional diversity in the data set.

If regional differences are significant, then excellent, good and doubtful letters will likely be consistently longer (or shorter) for

Table 4 | Letter tone by applicant gender.

	Excellent	Good	Doubtful	Total
Female applicant	53 (15%)	302 (83%)	7 (2%)	362
Male applicant	203 (24%)	635 (73%)	24 (3%)	862
				1,224

some regions, regardless of applicant or recommender gender. This was confirmed when we controlled for regional differences; results showed statistically significant differences in letter length. That is, after controlling for applicant and recommender gender, when compared to letters written in the Americas, letters written in all other regions were significantly shorter (all p values < 0.001).

We then examined the relationship between applicant gender and letter tone with three fixed-effects HLMs. All models excluded doubtful letters. Letter length was standardized via grand-mean centring and z-scoring for ease of interpretation[19]. First, the level 2 variable (applicant gender) predicted the odds ratio of the level 1 variable (letter tone). Results indicated that in comparison to male applicants, female applicants were significantly less likely to receive an excellent versus good letter ($\beta = -0.69$, OR $= 0.58$, $p = 0.009$). To further examine the relationship between applicant gender and letter tone, a second HLM analysis was conducted with the addition of recommender gender, recommender region, and letter length as control variables, in addition to applicant gender as a predictor. Males are the reference group for applicants and recommenders; the 'Americas' group is the reference group for region; and 'good' is the reference group for the categorical variable, tone. Table 6 depicts the results of this model.

Finally, an exploratory third model tested for the possibility of an interaction between recommender gender and applicant gender in predicting letter tone, while accounting for the effects of applicant gender, recommender gender, recommender region and word count. The interaction term was not significant ($\beta = -0.67$, OR $= 0.51$, $p = 0.25$), and applicant gender and word count remained significant predictors in the third model (p values < 0.01).

The results show that after controlling for recommender region, recommender gender, and letter length, female applicants are only half as likely to receive excellent letters versus good letters compared to male applicants. Also, after controlling for recommender region, recommender gender, and applicant gender, longer letters were more likely to be excellent. Statistically, with every one standard deviation increase in word count, the likelihood of receiving an excellent letter compared to a good letter increased by more than two times.

Neither recommender gender nor the interaction of recommender and applicant gender were significant in predicting letter tone. In other words, there were no statistical differences in letter tone for letters written by male and female recommenders, nor was there evidence that male and female recommenders differed in their likelihood to write excellent versus good letters for male or female applicants. Also, no particular region had significantly stronger letters than any other region, that is, letter tone appears to be equivalently distributed across all regions. Letters written by recommenders in the Americas were significantly longer than those written in any other region. There were no statistical differences in letter length between female and male applicants, which is consistent with earlier findings[13,14].

Our results show that at a critical career juncture (that is, at the postdoctoral stage), women are only half as likely to receive excellent letters of recommendation, regardless of recommender gender or region. The large sample size and geographical diversity in recommenders strengthens our findings and conclusions. Given the gender disparity in full professorships and the necessity of obtaining a postdoctoral fellowship en route to professorships in the geosciences, these findings are especially important.

Table 5 | Results of HLM predicting letter length, with applicant gender as predictor, and recommender gender and recommender region as controls.

Parameter	Estimate	Std. error	t
Applicant gender	−23.70	19.34	−1.27
Recommender gender	31.38	22.82	1.38
Recommender region (Middle East, Africa)	−241.79	43.44	−5.57*
Recommender region (Australia, Europe, New Zealand)	−205.72	21.09	−9.75*
Recommender region (South Asia)	−276.67	29.83	−9.28*
Recommender region (East Asia and Pacific)	−225.88	28.35	−7.97*

Std. error, standard error. *$p < 0.001$.

This study advances our understanding of gender bias in geoscience recommendation letters; however, there are important limitations that set the stage for future research. Due to the nature of the archival data, we were unable to control for applicant qualifications, so we were statistically unable to rule out the possibility that male applicants may have been better qualified than females. However, since letters came in from all over the world, it was highly unlikely that there is a systemic deficit in the quality of only the female applicants worldwide. This assertion is strengthened by the fact that our results are consistent with previous research on gender bias in recommendation letters that were able to control for applicant qualifications[12,14].

We were unable to control for the relationship between the recommender and the applicant, such that the quality of a letter may have depended on how well a recommender knew an applicant. Evidence of gender disparities in access to social networks[20] necessitates further research on the recommender–applicant relationship as it affects letters of recommendation. Moreover, differing levels of familiarity with male versus female applicants may well be another source of bias in STEM.

The small proportion of doubtful letters precludes a more advanced analysis. It is interesting to note that male and female applicants had roughly the same proportion of doubtful letters. This seems paradoxical, since females were half as likely to receive excellent letters; however, we have too few letters to do a detailed analysis. For the purposes of this paper, it is likely that selection committees focus on differentiating excellent from good candidates, therefore differentiation between these two categories is most meaningful.

As discussions around diversity and implicit bias gain prominence in national-level conversations, studies such as this one advance our understanding of the subject. The postdoctoral stage is a critical career juncture, and the writing and reviewing of recommendation letters are an integral part of women entering and advancing within these fields.

A possible area for future research is to examine specific words and phrases that comprise excellent versus good letters, via detailed linguistic analysis. Prior work has shown gender-related differences in applicant descriptions[12,14], thus a thorough examination of gendered words and phrases is warranted in a data set of this size.

Finally, it is important to note that recommendation letters may just be one way in which gender biases emerge; even after selection, women may face various gender-related obstacles[21]. It is important for institutions to foster women's academic success and create environments that benefit everyone[22]. Our results strike at the heart

Table 6 | Results of HLM predicting letter tone, with applicant gender as predictor, and recommender gender, recommender region, and letter length as controls.

Parameter	Coefficient (β)	Std. error	*t*	OR* (95% CI for OR)†
Applicant gender	−0.69	0.22	−3.10‡	0.50 (0.32–0.78)
Recommender gender	−0.29	0.23	−1.24	0.75 (0.48–1.18)
Region (Africa and Middle East)	0.17	0.44	0.38	1.18 (0.50–2.80)
Region (Australia, Europe, and New Zealand)	−0.38	0.27	−1.40	0.68 (0.40–1.16)
Region (South Asia)	−0.32	0.38	−0.84	0.73 (0.35–1.53)
Region (East Asia and Pacific)	−0.30	0.37	−0.82	0.74 (0.36–1.52)
Word count	1.05	0.10	10.26‡	2.87 (2.34–3.51)

*OR, odds ratio. In this table, for categorical variables, an OR refers to the comparison between the reference group in each categorical parameter to the non-reference group(s) in the likelihood of receiving a letter categorized as good, compared to excellent. For word count, a continuous variable, an OR refers to the average increase in odds in receiving an excellent versus good letter, per one-unit increase in word count. †95% CI, confidence interval for odds ratios. A CI that does not contain the null value (1) indicates either higher or lower odds of an outcome than what would be expected due to chance. ‡$p < 0.001$.

of the problem, that is, that women are disadvantaged right from the beginning of their geoscience careers because they are possibly perceived as not contributing as much as their male colleagues; and this only worsens along their career trajectory, with leaks in the pipeline at every stage[2,4]. These results are relevant to people in STEM fields, policymakers, institutional leaders, department chairs, and the general public at large. We hope that studies such as this one will spread awareness of the differences in how men and women in the geosciences are perceived worldwide, and that institutions will use this information to develop initiatives to recruit, retain and advance women in STEM fields.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of this paper.

Received 30 April 2016; accepted 31 August 2016; published online 3 October 2016

References

1. Valian, V. *Why So Slow? The Advancement of Women* (MIT Press, 1998).
2. Hill, C., Corbett, C. & St. Rose, A. American Association of University Women. *Why So Few? Women in Science, Technology, Engineering, and Mathematics* (AAUW, 2010).
3. Holmes, M. A., O'Connell, S. & Dutt, K. *Women in the Geosciences: Practical, Positive Practices Towards Parity* (AGU Special Publication Series, Wiley, 2015).
4. Goulden, M., Frasch, K. & Mason, M. A. *Staying Competitive: Patching America's Leaky Pipeline in the Sciences* (University of California at Berkeley; Berkeley Center on Health, Economic and Family Security; and the Center for American Progress, 2009).
5. *Postdocs at Federally Funded R&D Centers* (National Science Foundation, 2015); http://www.nsf.gov/statistics/2015/nsf15312
6. *Women, Minorities and Persons with Disabilities in Science and Engineering* (National Science Foundation, 2015); http://www.nsf.gov/statistics/2015/nsf15311/tables.cfm
7. National Academy of Sciences *Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering* (National Academies, 2007).
8. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. Science faculty's subtle gender biases favor male students. *Proc. Natl Acad. Sci. USA* **109,** 16474–16479 (2012).
9. Wenneras, C. & Wold, A. Nepotism and sexism in peer review. *Nature* **387,** 341–343 (1997).
10. Reuben, E., Sapienza, P. & Zingales, L. How stereotypes impair women's careers in science. *Proc. Natl Acad. Sci. USA* **111,** 4403–4408 (2014).
11. Sheltzer, J. M. & Smith, J. C. Elite male faculty in the life sciences employ fewer women. *Proc. Natl Acad. Sci. USA* **111,** 10107–10112 (2014).
12. Madera, J., Hebl, M. & Martin, R. Gender and letters of recommendation for academics: agentic and communal differences. *J. Appl. Psychol.* **94,** 1591–1599 (2009).
13. Trix, F. & Psenka, C. Exploring the color of glass: letters of Recommendation for female and male medical faculty. *Discourse Soc.* **14,** 191–220 (2003).
14. Schmader, T., Whitehead, J. & Wysocki, V. H. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles* **57,** 509–514 (2007).
15. Morgan, W. B., Elder, K. B. & King, E. B. The emergence and reduction of bias in letters of recommendation. *J. Appl. Psychol.* **43,** 2297–2306 (2013).
16. Leslie, S.-J., Cimpian, A., Meyer, M. & Freeland, E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347,** 262–265 (2015).
17. Allum, J. *Phase II: Final Applications and Initial Offers of Admission. Findings from the 2014 CGS International Graduate Admissions Survey* (Council of Graduate Schools, 2014).
18. Judge, T. A. & Higgins, C. A. Affective disposition and the letter of reference. *Organ. Behav. Hum. Dec.* **75,** 207–221 (1998).
19. Hofmann, D. A. & Gavin, M. B. Centering decisions in hierarchical linear models: theoretical and methodological implications for organizational science. *J. Manage.* **24,** 623–641 (1998).
20. Ibarra, H. Paving an alternative route: gender differences in managerial networks. *Soc. Psychol. Q.* **60,** 91–102 (1997).
21. Heilman, M. E. & Wallen, A. S. Wimpy and undeserving of respect: penalties for men's gender-inconsistent success. *J. Exp. Soc. Psychol.* **64,** 664–667 (2010).
22. Stewart, A. J., Malley, J. E. & LaVaque-Manty, D. *Transforming Science and Engineering: Advancing Academic Women* (Univ. Michigan Press, 2007).

Acknowledgements

K.D. would like to thank S. Pfirman for the discussion and guidelines surrounding the initial stages of this study. D.L.P. would like to thank J. Boyce for assistance with statistical modelling. This paper is contribution number 8044 from Lamont-Doherty Earth Observatory of Columbia University.

Author contributions

K.D. initiated this study; K.D. coded the letters and did a preliminary descriptive analysis; D.L.P. analysed the data; A.F.B. coded a subset of the letters; J.S.D. assisted with statistical analysis; C.J.B. served in an advisory capacity; K.D. and D.L.P. co-wrote the paper, with all authors contributing towards discussing and interpreting the results and refining the paper.

Additional information

Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to K.D.

Competing financial interests

The authors declare no competing financial interests.

Methods

This study used archival letters of recommendation, gathered during a five-year period at the sponsoring institution. There were a total of 1,224 letters submitted by recommenders from 54 countries. These countries were: Albania, Argentina, Armenia, Australia, Austria, Bangladesh, Botswana, Brazil, Bulgaria, Cameroon, Canada, Chile, China, Czech Republic, Denmark, Egypt, France, Germany, Greece, Honduras, Hungary, Iceland, India, Iran, Ireland, Israel, Italy, Japan, Kuwait, Libya, Malaysia, Mexico, Morocco, Netherlands, Nepal, New Zealand, Nigeria, Norway, Panama, Portugal, Russia, Singapore, South Africa, South Korea, Spain, Sri Lanka, Sweden, Switzerland, Taiwan, Thailand, Turkey, United Kingdom, United States, and Vietnam.

Prior to coding, letters were redacted for gender (of the applicant and the recommender) and all identifying information by an assistant. The assistant assigned each letter a unique serial number and then noted this serial number and key information from the letter—such as region and gender—in a spreadsheet. The assistant then redacted the letters by blacking out any identifying and gender-related information (for example, name and address of recommender and/or applicant; pronouns such as 'his', 'her', 'she' and 'he'). Coders were unaware of gender or any other identifying information for either the applicant or the recommender at the time of coding the letters.

Letter length was measured by counting the number of words in the body of the letter (that is, excluding the header and signature). To determine letter tone, a coding scheme was created by surveying 18 senior scientists/faculty (11 female and 7 male, representing 7 countries). These senior scientists are individuals who serve or have served on postdoctoral selection committees at the hiring institution. That is, they play a key role in the decision-making process of selecting postdoctoral scientists, so their opinions were directly relevant to the development of a coding scheme. The coding scheme also drew from the work of Trix and Psenka[13] and Schmader, Whitehead and Wysocki[14]. Samples of comments were randomly selected from approximately 100 letters in the data set and were provided to the senior scientists/faculty for evaluation. They categorized each of the anonymized comments into one of three categories: 'excellent'; 'good'; or 'doubtful'. These comments were then listed as exemplars of each of these three categories. Based on the responses of the senior scientists, and following the guidelines outlined in Riffe, Lacy & Fico (2005)[23], a coding scheme was developed. To ensure consistency in evaluation across all 1,224 letters, the coding scheme clearly defined the three categories and the specific content that fit each of these categories. This coding scheme is summarized in Table 2 of the main manuscript, and is available upon request from the first author.

A letter rated 'excellent' had an overall tone that was excellent or outstanding, and focused on the candidate's superior scientific ability. Such letters typically had comments that described the candidate in terms such as 'brilliant', 'rising star', 'pioneer', 'genius' and 'trailblazer'; and/or praised the candidate's ability to conduct groundbreaking and novel research, and/or reflected the candidate's potential to become a scientific leader and role model. Such letters also contained language that described ways in which the candidate was superior to others—such as the candidate's academic scholarship being the best that the recommender had seen in several years, and/or the candidate having exceptional scientific ability and/or impressive leadership skills usually not seen in graduate students. Some letters described the candidate's awards/honours in detail, sometimes noting that it was very rare for a student to receive such honours. Excellent letters could contain some doubtful comments; however, any doubts expressed were more than offset by the strong positivity of other comments. An example was a letter that commented on a candidate's outstanding academic performance and his/her tremendous potential as a top-notch scientist and a leader in the field, but also said that the recommender was unfamiliar with some aspect of the candidate, such as the candidate's teaching skills.

A letter rated 'good' had an overall tone that was positive and solid. Such letters mentioned things such as the candidate's strong knowledge of the subject, their very good track record, their intelligence and aptitude for learning new topics, and/or praised their skill set. Further, good letters did not contain language that clearly rated the applicant as superior to others; nor did they indicate that the candidate had the potential to be a scientific leader or role model. Although good letters could have instances of excellent and doubtful comments, the overall tone of the letter was good/solid. An example was a letter that mentioned a candidate's excellent academic record (for example, straight As or excellent GPA), and praised the candidate's knowledge, background, skills, and so on, and stated that the candidate would make a good postdoctoral fellow, but did not offer any praise about his/her potential as an outstanding scientist, scientific leadership, and/or the ability to do novel or groundbreaking work. Such a letter could also mention that the recommender was not familiar with some aspect of the candidate, such as the candidate's teaching potential.

A letter rated 'doubtful' had an overall tone that was either negative or doubtful and did not provide evidence of the candidate's scientific ability and potential. Such letters typically questioned the candidate's calibre as a scientist, or included language that described the candidate's failings, be it in academic performance, scientific ability, or personal failings such as a lack of leadership skills or displaying poor judgement. There could be instances of excellent and good comments within the letter, but the overall tone was negative or doubtful. An example was a letter that referred to a candidate's outstanding programming skills, and mentioned that s/he had experience working on a certain project, but also said that the candidate was more of a lab technician than a scientist, and/or would likely be a middle-of-the-pack scientist rather than an excellent one.

Following the creation of a coding manual, each letter was assigned an overall code for tone (excellent, good or doubtful) by the first author, based on the overall content of each letter. In line with the suggestion of Lombard, Snyder-Duch and Bracken (2002, 2010)[24,25], one of the co-authors coded 180 randomly selected letters (that is, 15% of all letters). Thirty letters were used for coder training and the remaining 150 letters were then coded after the training period to establish inter-rater reliability. Cohen's κ was used to determine if there was agreement between the two raters' judgements of letter tone (excellent, good or doubtful) in the 150 letters examined after the training period. Substantial agreement was found, $\kappa = 0.79$, $p < 0.001$ as per Landis and Koch[26]. After establishing substantial inter-rater reliability, the ratings from the first author were used for the analysis.

Letters in our data set ranged in length from 37 words to 2,444 words, with a mean of 457.11 (s.d. = 286.07). The length of the doubtful letters ranged from 37 words (the lowest value in the data set) to 871 words, with a mean of 473.58 words (s.d. = 228.90). The good letters ranged from 52 words to 1,640 words, with a mean of 392.27 (s.d. = 226.43). Finally, the excellent letters ranged from 119 words to 2,444 (the highest value in the data set) with a mean of 693.18 (s.d. = 356.39). The mean word count for letters written for female applicants was 457.34 words (s.d. = 283.84) and the mean word count for letters written for male applicants was 457.01 (s.d. = 287.17).

Of the letters categorized as excellent ($n = 256$), approximately 79% were for male applicants and 21% for female applicants. For letters categorized as good ($n = 937$), approximately 68% were for male applicants and 32% for female applicants. For letters categorized as doubtful ($n = 31$), approximately 77% were for male applicants and 23% for female applicants. However, the letters in the doubtful category were too few to do a meaningful statistical analysis. When grouped by gender, 24% of the letters for male applicants were excellent compared to 15% for female applicants; 73% of the letters for male applicants were good compared to 83% for female applicants; and 3% of letters for male applicants were doubtful compared to 2% for female applicants. Doubtful letters were excluded from the HLMs because of their small number.

All statistical analyses were performed with SPSS, version 21. For a study such as this, where a majority of the applicants received two or more letters, the appropriate statistical method is to use HLMs. These models were utilized to examine the two dependent variables of interest, letter length and letter tone, to account for the relatedness of letters for any given applicant[12].

Of note, in this model, 85.2% of letters were classified correctly. To address this, two alternate fixed-effects HLMs were also examined wherein applicants with only two or more letters of recommendation were examined; and wherein applicants with only three or more letters of recommendation were examined. Classification only improved marginally, thus the original model including all applicants was utilized, as diminishing our sample size did not significantly improve predictive accuracy, nor did the statistical significance of our findings and the odds ratios shift after restricting the sample size. These additional analyses are available from the second author upon request.

There were 37 recommenders who wrote letters for both male and female applicants, and an overwhelming majority of these recommenders wrote longer letters for males than for females. However, these letters were too few in number to allow for controlling of recommender region, and statistical tests revealed no systematic bias in the quality of letters across genders, a result consistent with our original finding that applicant gender was not a significant predictor of letter word length.

Data availability. The original letters of recommendation that support the findings of this study are available on request from the corresponding author K.D. in anonymized form, with references to names, scientific disciplines, and project descriptions redacted before sharing. These letters are not publicly available because they contain information that could compromise the privacy and anonymity of recommenders and/or applicants.

References

23. Riffe, D., Lacy, S. & Fico, F. G. *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (Lawrence Erlbaum Associates, 2005).
24. Lombard, M., Snyder-Duch, J. & Bracken, C. C. Content analysis in mass communication: assessment and reporting of intercoder reliability. *Hum. Commun. Res.* **28,** 587–604 (2002).
25. Lombard, M., Snyder-Duch, J. & Bracken, C. C. Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects (2010); http://matthewlombard.com/reliability/index_print.html
26. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–174 (1977).