**Title:** Assessing Gender Bias in Physics and Social-Science Recommendations for Academic Jobs (90 characters including spaces)

**Authors:** R. H. Bernsteina,b, M. W. Macyb, C. Cameronb, S. Williams-Cecib, W. M. Williamsb, S.J. Cecib

**Affiliations:**

aFermi National Accelerator Laboratory, Batavia IL 60510, USA.

bCornell University, Ithaca, NY 14853, USA.

\*Direct Correspondence to R.H. Bernstein: [rhbob@fnal.gov](mailto:rhbob@fnal.gov)

M.W. Macy: [mwmacy@cornell.edu](mailto:mwmacy@cornell.edu)

C. Cameron: [cjc73@cornell.edu](mailto:cjc73@cornell.edu)

S. Williams-Ceci: [scw222@cornell.edu](mailto:scw222@cornell.edu)

W.M. Williams: [wendywilliams@cornell.edu](mailto:wendywilliams@cornell.edu)

S. Ceci: [sjc9@cornell.edu](mailto:sjc9@cornell.edu)

**Abstract:** We analyzed 2,206 letters of recommendation written for positions at the assistant professor level in two fields differing dramatically in women’s representation: experimental particle physics and social science. Using larger samples and more measures of potential gender differences than in previous research (including length, content, and authorship of letters), we found few statistically significant differences, regardless of discipline or gender of writer. Of these, more differences favored women than those that favored men. We conclude that policies to correct gender imbalances in math-intensive fields may be more effective if they target barriers other than bias in letters of recommendation, such as how letters are evaluated by search committees and obstacles that discourage women from choosing STEM fields and academic careers. (120 words of 125 allowed)

**Short Title:** AssessingGender Bias in Recommendations (36 characters of 40 characters allowed)

**One Sentence Summary**:

Analysis of 2206 letters of recommendation in particle physics and social science reveal little systematic evidence of gender bias against women. (125 characters of 125 allowed)

**Main** **Text: (2,895 words including references and captions for SM; would be 2,654 without SM references)**

The underrepresentation of women in math-intensive fields is a problem that is historically persistent and extensively studied (1-6). Possible causes include hiring and promotion biases (7-8), leaky-pipeline issues (9-11), differences in career preferences (12-14), and differential persistence/retention (15-18). Recent studies have examined possible gender bias in letters of recommendation (19-25), noting that “*there is little research that addresses whether letters of recommendation for academia are written differently for men and women and whether potential differences influence selection decisions in academia*” (21).

This study investigates the first of these two concerns -- gender bias in letters of recommendation -- using 2,206 letters of recommendation for positions at the assistant-professor level in two fields differing dramatically in women’s representation: social science and experimental particle physics (EPP). Women are well-represented among PhDs in the two social science disciplines in this study -- psychology and sociology -- with 71.4% and 62.6% of PhDs, respectively. In contrast, women remain significantly underrepresented in EPP (only 13.4% of PhDs). The letters were accumulated over multiple job searches between 2011 and 2017 for EPP positions at Fermi National Accelerator Laboratory (963 letters for 206 men; 198 letters for 39 women) and for social science positions at Cornell University (440 letters for 163 men; 605 letters for 222 women; see Supplemental Materials SM1 for details).

Previous research has found that the length and enthusiasm of letters written for women are less than in letters for men: “…*female candidates are half as likely as male candidates to receive an excellent letter compared to a good letter or to have ‘standout’ adjectives like ‘excellent’, ‘outstanding’ or ‘amazing*’” (20). Studies have shown that letters for women contain more “grindstone” words, like “hard-working” and “reliable”(24-25), and more “communal” or other-oriented words, like “cooperative” and “nurturing” (21,26,30). However, these studies were unable to rule out applicant characteristics as an alternative explanation.

This inability to control for possible differences in accomplishments that might be reflected in the length and content of letters is a potential limitation of our study as well. Our study goes beyond previous research in that we address this limitation in two ways. First, we compared letters written for candidates in two disciplines differing dramatically in the underrepresentation of women If we assume women’s accomplishment relative to men’s is about the same in EPP and social science, then a larger gender difference in EPP letters compared to letters in social science should not be attributed to unobserved gender differences in accomplishments.

Second, we tested for differences between male and female writers using a restricted sample of 918 letters for 234 candidates with letters from both female and male non-advisors. This allowed direct comparisons of differences in how men and women recommenders depicted the same candidates when neither was the advisor. We found no greater gender differences between writers for the restricted sample than for the entire sample of 2,206 letters, and we therefore report results for the entire sample.

In addition, our study uses larger samples with a wider array of measures than in previous research (SM8), including not only the content of letters but also their length and status of the author. Past research shows that candidates benefit from longer letters *(20)*. We also included the same three measures of letter content derived from the Linguistic Inquiry and Word Count (LIWC) used in many previous studies *(20,21,24,25)*: the proportion of total words in the letter that appear in the lists of “posemo,” “negemo,” and “drives” words. “Posemo” and “negemo” convey affective attraction and aversion, while “drives” consists of the LIWC sub-lists for “achievement” words and “power” words connoting high-achievers with important roles.

We also looked for gender differences not captured by LIWC measures, based on the frequency of words identified in previous research *(20,21,24)* as important in hiring and as potential sources of gender bias: so-called “agentic,” “standout,” “grindstone,” “communal,” and “ability” words. “Agentic” words include active, take-charge, and leadership references; “communal” words include references to nurturing, helpful, other-oriented, and collaborative traits; “standout” and “ability” words express exceptional talent and promise; and “grindstone” words include “hard-working,” “conscientious,” and “diligent,” with the implication that effort compensates for deficiencies in ability*(30)*.

In addition to the word counts obtained from these eight word lists (including the three LIWC lists), we developed other measures of gender bias not used in previous research. First, we reversed the word-list methodology by empirically identifying words and phrases most predictive of candidate gender in the letters. This open-ended measure addresses concerns that pre-existing word-lists may not contain terms that disfavor women. Second, we analyzed two measures of a letter’s authorship that might also influence hiring: the gender and academic rank of the recommender (see SM2 and SM3 for details). The authorship measures address the possibility that recommendations may disadvantage women not only by their length and content but also by the gender and rank of the recommender.

*Letter content*. Figure 1 reports results for the eight lexical analyses and word count. Panels A and B show that, when writing for women, female physicists used more positive affect words (*t*=2.41, *p*=0.017) while male physicists used fewer negative affect words (*t*=2.18, *p*=0.03). Panels C and F show that male social scientists used more “drives” words (*t* = 2.31, *p* = 0.021) and male physicists used more “grindstone” words (*t*=2.25, *p*=0.025) when writing for women. However, regardless of discipline or gender of writer, men were not depicted as more “agentic” or as “standouts,” nor were women depicted as more “communal.”

We defined “drives” as the sum of LIWC “achievement” and “power” with a small (~10%) correction described in SM4. There were no significant differences in EPP. In social science, there was a significant difference: men writers used fewer drive words for men than they did for women (*t* = -2.08, *p* = 0.038).

Female social scientists also used “communal” words (Panel H) more frequently than did male writers in the full sample (*t* = 3.21, *p* < 0.001) but the gender difference was not statistically significant among the subset of candidates with letters from both genders (*t* = 1.66, *p* = 0.098) and women used communal terms equally for both male and female candidates (*t* = 0.09, *p* = 0.93).

Physicists used more grindstone, communal, and agentic terms than did social scientists, but female and male writers in physics were equally likely to use these terms. (See SM5 for robustness checks.)

*Letter Length*. Figure 1 (Panel I) shows that letters for women candidates in social science were 65.4 words longer than letters for men (*t* = 2.33, *p* = 0.02), while in EPP, letter length did not differ significantly by gender. Across both disciplines, women recommenders wrote longer letters than did men, a difference of 63.4 words (*t* = 2.67, *p* = 0.008).

*Letter authorship*. Figure 2 also reports gender differences in authorship of letters, given the gender and academic rank of the writer (non-tenure track, assistant or equivalent, associate or equivalent, full professor or equivalent, and chaired professor or equivalent). In both EPP and social science, male authors have higher academic rank but the difference is statistically significant only for social science (*t* = 3.07, *p* = 0.002). A two-category χ2 test yields χ2 =9.95/3 dof, *p* = 0.02 for social science and χ2 = 2.70/3 dof, *p* =0.44 for EPP. Tables S2 and S3 in SM2 and Fig. 2 report the F/M ratio among letter-writers for men and women. In physics the F/M ratio is 0.14 for both men and women, with an expected value of 0.16 ± 0.022; in social science, the ratios are 0.46 ± 0.047 for men and 1.03 ± 0.084 for women, with an expected value of 0.67 ± 0.04. This difference in authorship in social science (“gender homophily”) could reflect gender bias or it may instead be a by-product of differences in gender composition across field specializations within social science (SM2). Whatever the explanation, in contrast to social science letters, there were no statistically significant gender differences in letter authorship in EPP (*t* = -0.57, *p* = 0.571).

In addition to gender, the impact of a letter might also depend on the author’s academic rank. However, Figure 2 reports the absence of statistically significant differences between male and female candidates in the academic rank of their letter writers in EPP but

*Validation.* The similarity of content in letters for men and women candidates when measured using pre-existing word lists may reflect the inability of the measures to capture gender bias. We addressed this possibility using an open-ended analysis that reversed the word-list methodology (SM7). Instead of starting with words that might indicate enthusiasm and then measuring their gender distribution, we started with words highly correlated with gender and looked for those indicating the strength of recommendation. The results revealed gender differences in topics more likely to be studied (for example, “family” among female social scientists vs. “neuroimaging” among males), but there were no gender-correlated words likely to signal enthusiasm for a candidate in either discipline.

*Discussion.* In social-science letters, only three differences among candidates were statistically significant: letter length, “drives” words, and author gender, but only the latter might be interpreted as favoring men. In EPP, there were also three significant differences; two favored women (more positive affect and less negative affect) and one favored men (fewer “grindstone” words). However, positive affect only favored women among letters by female physicists, who wrote fewer than 10% of letters. Other than more “grindstone” words in EPP letters by men for women, we did not observe differences found in previous studies showing that women are less likely to be depicted as “agentic” or “standouts” *(31)*. In sum, we found few statistically significant differences, and of these, more differences could be interpreted as an advantage to women than men, regardless of discipline or gender of writer.

These results should not lead us to conclude that recommendation letters are free of gender bias. We may have under-estimated the number of significant differences due to insufficient statistical power based on 1045 letters in social science and 1161 in EPP. However, these samples are much larger than those of previous analyses of gender differences in letters of recommendation (see SM8).

It is also possible that we over-estimated the number of significant differences. We tested for gender differences on ten letter attributes (letter length, author rank, and eight word counts), broken down by discipline and writer gender, plus author gender broken down by discipline. Out of 42 significance tests using a conventional benchmark of *p* < 0.05, we should expect two false positives if the null hypothesis were true. However, given the persistent gender imbalance in many math-intensive disciplines, a false negative in tests for gender bias might be equally serious. We therefore recommend conditionally assuming the gender differences we observed are true positives until proven otherwise. (See SM6 for results using Bonferroni correction.)

We cannot rule out the possibility that female candidates were superior to male candidates and deserved stronger letters than those they received. However, we controlled for candidate qualities when assessing gender differences between writers for candidates with letters from both genders. Assuming EPP and social science differ little in the level of women’s accomplishment relative to men’s, the similarity between EPP and social science in the number and magnitude of gender differences suggests the current findings might change very little were we able to completely control for candidate qualifications.

Another possibility is that gender differences would have been more evident had we used more sophisticated lexical measures. However, our open-ended search for gender-biased expressions revealed no additional evidence of bias against female applicants.

Caution is also needed when generalizing these results from entry-level to senior positions, from EPP to all of physics, from a national lab to academia, and from two high-profile institutions to all of science. Nevertheless, it is also important to note that candidates typically apply to dozens of positions and letters for a given candidate rarely differ substantively from one search to another. Therefore, the letters in our samples are likely to resemble those submitted by these applicants to other searches beyond these two institutions.

Finally, this study focused on the attributes of letters that might influence hiring decisions, not on how decision-makers *respond* to these attributes. Future research should manipulate the apparent candidate gender of identical letters *(8,27)* to test the possibility that search committee deliberations favor men in the evaluation of letters, despite the similarity of letters for women and men applicants.

We close with a note of optimism: Extreme gender disparities in nearly every academic discipline are being rapidly attenuated, through concerted efforts at institutional change guided by results of careful research. However, pockets of extreme disparity remain, the causes of which are not yet fully understood.

**References**

1. C. Hill, C. Corbett, A. St. Rose, *Why So Few?* AAUW (2010).
2. V. Valian, *Why So Slow?* MIT Press (1998).
3. National Academy of Sciences. *Beyond Bias and Barriers.* Washington, DC: NAS (2007).
4. National Research Council, *Gender Differences at Critical Transitions*. Washington DC: NAS (2009).
5. Y. Xie, Y.K.A. Schauman, *Women in science*. Cambridge, MA: Harvard (2003).
6. C.A. Ross-Macusin, J.F. Dovidio, V.L. Brescoll, M.J. Graham, J. Handelsman, *Proc. Natl Acad. Sci. USA* **109**, 16474-16479 (2012).
7. J.M. Sheltzer, J.C. Smith, *Proc. Natl Acad. Sci. USA* **111**, 10107-10112 (2014).
8. A.A. Eaton, J.F. Saunders, R.K. Jacobson, K. West, *Sex Roles 1-15.* <https://doi.org/10.1007/s11199-019-01052-w>
9. S.A. Adamo, *Bioscience,* **63**(1), 43–48 (2013).
10. M. Goulden, K. Frasch, M.A. Mason, University of California at Berkeley; the Center for American Progress (2009).
11. R. Skibba, *Nature Reviews* **1**, 298-300 (2019).
12. R.A Lippa, K. Preston, J. Penner, *PloS ONE*, **9** (2014).
13. R. Su, R.J. Rounds, *Frontiers in Psychology* (2015).
14. M.T. Wang, J.S. Eccles, S. Kenny, *Psychological Science,* **24**, 770-775 (2013).
15. S. Kelchtermans, R. Veugelers, *Review of Economics and Statistics* **95**(1),273-85 (2013).
16. D. Kaminski*,* B. Geisler, *Science,* **335**, 864 (2012). DOI: 10.1126/science.1214844
17. L. Martinez, K.O’Brien, M. Hebl, *Journal of Women’s Health,* **26** (2017).
18. S.J. Ceci, D.K. Ginther, S. Kahn, W.M. Williams, *Psychological Science in the Public Interest* (2014). <http://psi.sagepub.com/content/15/3/75.abstract?patientinform-links=yes&legid=sppsi;15/3/75>, <https://doi.org/10.1177/1529100614541236>
19. M. Biernat, S. Eidelman, *European Journal of Social Psychology* (2007).
20. K. Dutt, D.F. Pfaff, J.S. Bernstein, J.S. Dillard, C.J. Block, *Nature* *Geoscience 9*, 805 (2016).
21. J. Madera, M. Hebl, R.C. Martin, *Journal of Applied Psychology, 84*, 1391-1399 (2009).
22. W.B.Morgan, K.B.Elder, E.B.King, *J. Appl. Psychol.* **43**, 2297-2306 (2013).
23. T. Schmader, J. Whitehead, V.H. Wysocki, *Sex Roles* **57**, 509-514. <https://doi.org/10.1007/s11199-007-9291-4> (2007).
24. F. Trix, C. Psenka, *Discourse and Society,* **14**, 191–220 (2003). [https://doi.org/10.1177/0957926503014002277](file:///Users/rhbob/Desktop/scienceDrafts/1.%09https:/doi.org/10.1177/0957926503014002277)
25. S. Li, A. Fant, D. McCarthy (2017). <https://doi.org/10.1002/aet2.10057>.
26. A.H. Eagly, S.J. Karau, *Psychological Review*, **109**, 573– 598 (2002).
27. W.M. Williams, S.J. Ceci, *Proc.* *Natl Acad Sciences,* **112***,17, 5360.*10.1073/pnas.1418878112.
28. J.W. Pennebaker, NY: Bloomsbury Press (2011); J.W. Pennebaker, R.J. Booth, R.L. Boyd, and M.E. Francis LIWC2015 *User Manual*.
29. National Science Foundation (2019), “National Center for Science and Engineering Statistics,” NSF 20-300, Table 9-5.
30. L.L. Aull, David West Brown. “Fighting Words: A Corpus Analysis of Gender Representations in Sports Reportage.” Corpora 8, no. 1 (May 1, 2013): 27–52. <https://doi.org/10.3366/cor.2013.0033>.
31. Jack Hessel. Jmhessel/FightingWords.
32. <https://github.com/jmhessel/FightingWords>.
33. J.C. French, S.J. Zolin, E. Lampert, A. Aiello, K. Bencsath, K.A. Ritter, M. Valente, A.S. Probhu. “Gender and letters of recommendation: a linguistic comparison of the impact of gender on general surgery residency applicants.” Journal of Surgical Education, 76, 899-905 (2019). [https://doi.org/10.1016/j.jsurg.2018.12.007](https://doi-org.proxy.library.cornell.edu/10.1016/j.jsurg.2018.12.007)
34. S. Li, A. Fant, D. McCarthy, D. Miller, J. Craig, & A.Kontrick (2017). AEM Education and Training, 1, 4. DOI: <https://doi.org/10.1002/aet2.10057>.
35. J.M. McCarthy, R.D. Goffin. “Improving the validity of letters of recommendation: an investigation of three standardized reference forms.” Military Psychology, 13(4), 199-222 (2001).
36. A.H. Messner, E. Shimahara. “Letters of Recommendation to an Otolaryngology/Head and Neck Surgery Residency Program: Their Function and the Role of Gender.” The Laryngoscope, 118: 1335–1344 (2009). <https://doi.org/10.1097/MLG.0b013e318175337e>
37. J. Blue, A.L. Traxler, X.C. Cid (2019). Gender matters. Physics Today 71, 3, 40 (2018); <https://doi.org/10.1063/PT.3.3870>

**Acknowledgements**

The authors thank Fermi National Accelerator Laboratory and the Cornell University IRB for access to the letters and Jamie Pennebaker for helpful suggestions.











**Figure 1.** Nine lexical measures for male and female candidates, by discipline and gender of writer. Error bars are 95% confidence intervals. 844 letters were written by and for men in EPP, with 301 in social science; 176 letters by men for women in EPP and 298 in social science; 121 letters by women for men in EPP and 139 in social science; and 22 letters by and for women in EPP and 307 in social science. In EPP, women receive more positive affect words (A) than do men among letters written by women, while men receive more negative words (B). In social science, men receive fewer drive (achievement + power) words from men than women (C) and fewer grindstone words (F) than do women among letters written by men. Finally, women in social science write longer letters than men (I).



****

Figure 2. In panels A and B: Gender differences in proportion of letters by recommenders of different academic ranks. The academic ranks are (0-2) instructor/lecturer; 3 is assistant professor; 4 is associate professor/full professor; 5 is distinguished professor or named chair or their equivalents. EPP shows no male/female difference; social science reveals a significant difference between men and women. The third panel C) shows the F/M ratio for writers compared to predictions (shaded bands). All uncertainties including the predictions in C) are 95% confidence limits. EPP shows no significant difference; social science shows a significant discrepancy between men and women, with women favoring female writers and men favoring male writers.