

Novel Measures Reveal Subtle Gender Bias in Academic Job Recommendations

R.H. Bernstein^{a,b,1}, M.W. Macy^b, C.J. Cameron^b, S. Williams-Ceci^b, W.M. Williams^b, and S.J. Ceci^b

^aFermi National Accelerator Laboratory, Batavia IL 60510, USA; ^bCornell University, Ithaca, NY 14853, USA

This manuscript was compiled on November 22, 2021

Linguistic analysis of 2,206 letters of recommendation compared gender bias in two disciplines differing in women’s representation: experimental particle physics (EPP, <15% female) and social science (>60% female). Standard lexical measures (e.g., “communal,” “agentic,” and “standout”) did not show bias against women in either discipline. On the contrary, in letters about women, female physicists used more positive-affect words, while male physicists used fewer negative-affect words as well as more references to hard work/effort. Neither discipline showed gender differences in the rank of letter writers, but social scientists wrote longer letters about women and wrote more often for candidates of their own gender. However, standard lexical measures assess only *overt* expressions of bias and may miss more subtle gendered language. We therefore developed a novel, open-ended measure of gendered usages. In striking contrast to conventional measures, our open-ended analysis uncovered troubling gender disparities. Positive references (e.g., to talent, innovation, creativity, etc.) appeared more often in letters about men in EPP and about women in social science (although female EPP candidates were more likely to be characterized as “brilliant”). Two of the largest gender disparities were for references to “physicist” in EPP and “science” in social science. These terms were used in more letters for men in both disciplines, possibly indicating unconscious gender stereotypes, even in a majority-female discipline. We conclude that future studies of linguistic bias should include open-ended measures, and that policies to correct gender imbalances in physics should raise awareness of subtleties of potential bias in letters of recommendation.

women in science | gender bias in science | letters of recommendation | underrepresentation of women | hiring bias

1. Introduction

The underrepresentation of women in math-intensive fields — such as physics, engineering, computer science, economics, and mathematics — is a problem that is historically persistent and extensively studied (1–5). Possible causes include hiring and promotion biases (6–8), leaky-pipeline issues (9–11), differences in career preferences (12–14), and differential persistence/retention (15–17). Recent studies have examined possible gender bias in letters of recommendation (18–24), 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” (19). Specifically, there is no research on gender differences in recommendations that compares academic fields in which women are well-represented to fields in which they are not.

This study investigates gender bias in letters of recommendation by comparing two fields differing dramatically in women’s representation: experimental particle physics (EPP) and social science. Women are well-represented

Significance Statement

Recommendation letters are central to academic hiring; a single negative comment in a letter can derail a candidate. The underrepresentation of women in physics has been attributed to less enthusiastic depictions of their ability in letters. We investigated the possibility that letters depict women less favorably than men across two scientific disciplines, particle physics (<15% female) and social science (>60% female). We analyzed the largest sample of recommendation letters to date, using more measures than previous studies. Using standard lexical measures, gender disparities were no greater in physics than in social science, and several gender differences favored women. A novel open-ended search for all gender-differentiated terms of endorsement revealed language favoring men in physics and women in social science.

R.H.B., S.J.C., M.W.M., C.J.C., and W.M.W. designed the research; R.H.B., M.W.M. and C.J.C. analyzed data; R.H.B., C.J.C., and S.C.W. coded data and organized/maintained spreadsheets; R.H.B., M.W.M., W.M.W., S.C.W., C.J.C. and S.J.C. wrote the paper.

All of the authors declare there is no situation that could be perceived as exerting an undue influence on the presentation of their work. This body of work is not influenced by financial, professional, contractual, or personal relationships or situations.

¹ To whom correspondence should be addressed: rhbob@fnal.gov

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) (25). We collected 2,206 recommendation letters written on behalf of candidates for positions at the assistant professor level, over multiple job searches between 2011 and 2017. These letters were written 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; the breakdown of letters by the genders of writer and the candidate and other relevant information are given in Supplementary Section S1 and Tables S1-S4).

Previous research has found that recommendation letters written for women are shorter and less enthusiastic than letters for men. Dutt et al. (18) state “Our results reveal that female applicants are only half as likely to receive excellent letters versus good letters compared to male applicants.” Specifically, there has been extensive investigation into “standout” and “grindstone” adjectives. In a summary for physicists, Blue et al. (26) explain that standout adjectives are words such as “outstanding,” “amazing,” and “unmatched.” Drawing on Schmader et al. (21), they conclude:

Standout words, which portray a candidate as talented and exciting, are most often found in letters of recommendation for men. Grindstone words, which create the impression that a candidate works hard but is not intellectually exceptional, are more often used for women.

This interpretation of “hard-working” as a backhanded compliment is supported in an influential paper by Trix and Psenka (22):

Of the letters for female applicants, 34 percent included grindstone adjectives, whereas 23 percent of the letters for male applicants included them. There is an insidious gender schema that associates effort with women, and ability with men in professional areas. According to this schema, women are hard-working because they must compensate for lack of ability.

Another line of research examined “agentic” and “communal” terms. Madera et al. write (19):

Agentic behaviors at work include speaking assertively, influencing others, and initiating tasks. Communal behaviors at work include being concerned with the welfare of others (i.e., descriptions of kindness, sympathy, sensitivity, and nurturance), helping others, accepting others’ direction, and maintaining relationships. . .

The authors found that agentic terms were more frequent in letters for men and communal terms were more frequent in letters for women.

However, these studies were limited by their reliance on a small number of lexical measures that have two important limitations. First, lexical measures rely on lengthy pre-existing lists that may include some words used more for women and others used more for men, with offsetting differences that could obscure the use of gendered language. For example, compared to men, women might have more descriptions saying they are creative but fewer saying they are innovative, with the aggregate score showing little or no gender difference. Second, the lexical measures were developed in an earlier era to detect overt bias and may fail to detect new and potentially more subtle expressions of bias (Supplementary Section S2 reviews older studies that relied on lexical measures.)

Accordingly, we reversed the lexical word list methodology used in previous research. In addition to starting with a pre-existing list of words that express enthusiasm and testing for gender differences in their usage, we also worked backwards, beginning by identifying words associated with gender that were not necessarily contained in any of the word lists and then evaluating whether these gender-differentiated expressions communicated enthusiasm for the candidate. This bottom-up analysis addresses concerns that pre-existing word lists may not contain terms that disadvantage women at a time of greater sensitivity to gender bias. It also avoids the possibility that words from the same semantic group are nevertheless used differently for men and women.

Previous studies suffered from other limitations as well, including the absence of a baseline, small samples, few measures, and the inability to rule out applicant characteristics as an alternative explanation for gender differences in the strength of letters. Our study also addresses these limitations:

1. We incorporated a baseline by comparing letters written for candidates in two disciplines differing dramatically in the underrepresentation of women. When comparing across disciplines, we do not need to assume the absence of a gender gap in accomplishments to attribute a gender gap in the strength of recommendations as an indication of bias. We only need to assume that the gender gap is roughly similar in both disciplines. If we were to observe a larger gender gap in the strength of letters in EPP compared to social science, then assuming the gender gap in accomplishments is no larger in EPP, we can then attribute the disciplinary difference to gender bias.

2. We tested for differences between female and male writers using a restricted sample of 918 letters for 234 candidates with letters from both male and female non-primary advisors (excluding the PhD committee chair). This method allowed direct comparisons of differences in how women and men recommenders depicted the same candidates, thereby ruling out gender differences in candidate accomplishments as an explanation for differences in letters written by female versus male recommenders.
3. Our study used larger samples with a broader array of measures than those used in previous research (Supplementary Section S2). In addition to the four lexical measures found in nearly all previous research (19, 21, 22) (“agentic,” “communal,” “standout,” and “grindstone”), we also used the proportion of total words in the letter that appear in the lists of “achievement” words connoting accomplishment and success, words associated with “ability,” and words that convey affective attraction and aversion (“posemo” and “negemo”) (27).
4. In addition to these measures of the content of letters, we also measured the length of the letter and the status of its authors. Past research shows that candidates benefit from longer letters (18). Letters may also have a larger impact depending on the gender and academic rank of the recommender.

2. Results

Given the striking underrepresentation of women in EPP, we were surprised not to find weaker letters for women using the standard measures of form and content found in much previous research.

A. Letter Content. Figure 1 reports the percent of the total words that matched each of eight widely used lexical measures, averaged across all letters in four groups based on discipline and author gender (the numerical values required to perform t -tests and calculate p -values are given in Supplementary Table S5). Panels A and B show that (a) female physicists used more positive-affect words ($t = 2.41, p = 0.017$) and (b) male physicists used fewer negative-affect words ($t = 2.18, p = 0.030$) when writing about women than when writing about men. Consistent with previous studies, Panel F shows that male physicists used more “grindstone” words when writing about women ($t = 2.25, p = 0.025$). However, regardless of either the discipline or the gender of the writer, men were not depicted as more “agentic” or as “standouts,” nor were women depicted as more “communal.”

In social science, female writers used “communal” words (Panel H) more frequently than did male writers ($t = 3.88, p < 0.001$), but they did this equally for male and female candidates in both the full sample and for the subset of candidates with letters from both genders (Supplementary Table S2 provides the numbers of letters for both the full and restricted samples). Physicists used more grindstone, communal, and agentic terms than did social scientists, but female and male writers in EPP were equally likely to use these terms.

B. Letter Length. Figure 2 (Panel A) 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 candidate gender. Across both disciplines, women recommenders wrote longer letters than did men, a difference of 63.4 words ($t = 2.67, p = 0.008$). (A similar effect has been observed for surgical resident letters (28).)

C. Letter Authorship. Figure 2, Panel B also reports gender differences using two measures of the authorship of letters: the gender and academic rank of the writer (non-tenure track instructor/lecturer, assistant, associate, full, and chaired professor or the rank equivalents). We observe no decline in the proportion of total letters written for female candidates as the rank of the author increases from instructor to chaired professor. This is confirmed using Spearman’s rank-order correlation between the writer’s academic rank and candidate gender; we found no significant correlations in either social science or EPP (the largest of the four correlations was very small ($\rho = 0.031, p = 0.45$) for male-authored letters in social science).

Accordingly, in addition to academic rank, Figure 1, Panel C reports differences in the gender of authors of letters for male and female candidates. There were no statistically significant differences in the gender of authorship in EPP ($t = -0.57, p = 0.571$) but the differences were surprisingly large in social science. Among all letters written for social science men, 68.4% were male-authored, compared to 49.3% of letters for women, a difference of 19.1% ($t = 6.29, p < .001$). This “gender homophily” in letter authorship could reflect gender preferences of authors and/or candidates, and/or gender differences in subfield specializations (see Supplemental Section S4 for the expected rates. Supplemental Section S5 and Supplemental Figure S1 show subfield specialization by gender).

D. Open-ended Analysis. The similarity of content in letters for women and men in Figure 1 may reflect the inability of pre-existing word lists to capture gender bias for two reasons. First, each list contains many individual expressions, some of which may favor women and others favor men, such that the aggregate score obscures what may be important

gender disparities. Second, these lexical measures were developed and validated in an earlier era in which women were underrepresented even in social science disciplines and sensitivity to overt bias was relatively weak. Contemporary letters may reflect changes in cultural awareness and normative expectations that discourage the use of red-flag expressions, even though letters still contain subtle expressions of gender bias that are not included in, and therefore cannot be found with, the standard word lists used in all prior studies.

We addressed these concerns using an open-ended analysis that reverses the word list methodology used in previous studies. Instead of starting with words that might indicate enthusiasm and then measuring their gender distribution, we considered every word in every letter as a potential source of gender bias and then eliminated those without gendered usage or relevance to hiring. The remaining gender differences are potentially problematic, even if the influence of the content on hiring decisions is not always immediately apparent. Are letters of recommendation relatively androgynous? Are gender differences in word choice larger in EPP than in social science? Do the differences favor one gender over the other?

The screening began by algorithmically removing “stop words,” personal pronouns, and words with little or no difference in gendered usage, leaving a set of 405 words that were potential sources of gender bias. Two of the senior authors, a physicist and a social scientist, then independently (without knowing each other’s choices) eliminated words that did not signal recommendation without knowing the gender distribution. This included the elimination of terms like “family” and “neuroscience,” for which a gender difference might indicate gender-specific subfield specializations (e.g., social science women are overrepresented in the study of family and men are overrepresented in neuroscience).

We also eliminated words with overtly context-dependent meaning (e.g., “confident” could refer to the candidate or to the writer). The screening process left a list of 63 words as potential indicators of gender bias based on gendered usage and relevance to hiring. However, the direction and magnitude of bias were still unknown. We then stemmed the words on the list, replacing the original word with the entire “family” of semantically consistent words that shared a common root. Figure 3 reports the fraction of letters that contained one or more usages of a member of the word family.

Figures 1 and 3 use different measurement units, and these units need to be considered when interpreting the results of each figure and when comparing across the two figures. Figure 1 reports the fraction of words in a letter that appear in a lengthy word list (the “per-word” measure). Figure 3 reports the fraction of letters with one or more usages of a single word or word family (the “per-letter” measure). Figure 1 does not use the per-letter measure because nearly every recommendation letter can be expected to contain one or more terms that appear in a long list of superlatives. Hence the per-letter measure is likely to underestimate the level of gender bias using a lexical analysis, and all previous studies that used lexical analysis relied on this same per-word measure that we use in Figure 1. Figure 3 uses the per-letter measure because a word randomly chosen from a long letter has a near-zero probability to match a particular superlative that is being used to distinguish one candidate from another. Although this “per letter” measure is not useful for widely used terms like “research” that appear one or more times in nearly every letter, regardless of gender, these terms have almost no discriminating power and are therefore likely to understate the level of gender bias. For terms that letter writers use sparingly (such as “brilliant”), and which therefore effectively distinguish candidates, the per-letter measure is appropriate. (Supplementary Section S3 contains additional comments on the difference between per-word and per-letter measurements.)

The open-ended analysis revealed gender differences that were not evident in the lexical analyses in Figure 1 or in letter length and authorship in Figure 2. Gendered language favored women in social science, but in EPP, the disparity favored men. Of the 16 gender-differentiated terms of endorsement found in social science letters, 12 were more likely to be used to describe female candidates ($p = 0.027$), compared to only two out of eleven gender-differentiated terms in EPP ($p = 0.028$).

Importantly, the open-ended measure also revealed subtle bias in social science. References to contributions to knowledge (using terms like “science,” “discovery,” and technical expertise) were more likely to be emphasized in social science letters for men. In contrast, personal attributes (volunteering, delightfulness, initiative, leadership, ambition, accomplishment, success, and commitment) were emphasized more often for women. The largest disparities were in social science letters written by men. “Commit*” appears in 24.5% of letters about women, compared to 13.6% of letters about men, $p < 0.0008$), while “scienc*” is mentioned in 54.6% of letters for men but only 45.6% of letters for women ($p < .03$). Of the 12 terms used more for women, only one, “volunteer*,” is not used by male writers more than female writers. This reflects in part the greater statistical power for male-authored ($N = 600$) than female-authored ($N = 445$) letters). However, of the four terms used more for men, only one, “scienc*,” is not used by female writers more than by male writers.

In contrast to the disparity favoring women candidates in social science, the gender disparity in EPP strongly favored male candidates whose letters were more likely to contain nine out of eleven gender-differentiated terms of endorsement. As in social science, “technical*” and “discover*” (as references to contributions to knowledge) were

used more often for men, but unlike social science, male candidates were also more likely to be personally characterized as intellectual, creative, innovative, dedicated, and talented. Only two terms were used more for women: “notabl*” and “brilliant*.” The latter is surprising, given recent findings by Leslie et al. (29, 30) that “brilliance” was more highly regarded in STEM disciplines with a lower representation of women. Nonetheless, “brilliant*” was much more likely to be used in physics letters about women (8.6%) than men (3.4%) (Supplementary Table S6 supplies the raw counts and the counts for “brilliant” are further broken down by gender of writer and candidate in Supplementary Table S7.) Nearly all gender disparities in EPP were in letters by men, possibly due to very low statistical power for letters by and about women ($N = 22$). The two exceptions (“technical*” and “dedicat*”) both favored men.

Comparing the two disciplines, two of the largest gender disparities were for the terms “physicist*” in EPP and “scien*” in social science. These two expressions were used more often in letters for men than in letters for women (50.6% vs. 36.9% for “physicist*” ($p < 0.001$) and 61.6% vs. 55.0% for “scien*” ($p = 0.03$). This could reflect — and possibly reinforce — unconscious gender stereotypes in which physicists and scientists are imagined as being men. That might not be surprising in an overwhelmingly male discipline, but it is noteworthy that the gender disparity for “scien*” is one of the largest in a discipline that is majority female.

In sum, the open-ended analysis reveals a clear overall pattern of gender differences in the probability that a letter contains a term of endorsement, with the differences favoring men in EPP and women in social science. Moreover, Figure 3 omits four expressions whose high usage frequency causes substantively minor gender differences to be statistically significant. In social science, references to research and teaching appear in 97% and 75% of all letters, respectively. References to research occur once per letter for female candidates and 0.90 times per letter for male candidates ($p = 0.008$), and references to teaching occur 0.40 times per female letter and 0.30 times per male letter ($p < 0.001$). Other minor differences in usage of “standard language” ran counter to the overall pattern in Figure 3. In social science, “publish*” is used 0.12 times per female letter and 0.14 times per male letter ($p = 0.045$). In EPP, over half the letters contain references to “scien*,” with 0.2 usages per female letter and 0.15 per male letter ($p = 0.01$).

3. Discussion

The underrepresentation of women in math-intensive academic disciplines like particle physics is a persistent problem that has resisted remedial efforts. We investigated gender bias in letters of recommendation as a potential contributor, using a comparison to the social sciences in the same time period as a benchmark for what might be expected in a discipline with gender parity. Using measures similar to those in previous research, we found little evidence of systematic gender bias. In social science, letter length favored women candidates, but the higher proportion of male-authored letters for male candidates could favor men, given the historical legacy of gendered status inequality. In EPP, conventional lexical measures identified three significant differences, two of which favored women: letters for women candidates contained more positive affect in letters written by women and less negative affect in letters written by men. Women wrote fewer than 10% of the EPP letters, so less negative affect in the larger proportion of letters by men could be more consequential than the greater use of positive affect in the smaller proportion of EPP letters by women. Male physicists also used more “grindstone” words when writing for women, which is consistent with previous studies and has been interpreted as a backhanded compliment based on the assumption that “effort” and “hard-working” are gender-biased code-words implying pedestrian research by women (26). However, Panel F in Figure 1 raises the possibility of an alternative interpretation. Male and female physicists used “grindstone” words twice as often as did social scientists, for all candidates regardless of gender, suggesting the possibility that these references have a positive connotation in usage by physicists. Across both disciplines, we did not observe differences found in previous studies showing that women are less likely to be depicted as “agentic” or “standouts” (19, 21).

In sum, an exhaustive analysis of letter content using conventional lexical measures (Figure 1), as well as letter length and authorship (Figure 2), revealed gender disparities that, overall, were no greater in EPP than in social science, and some differences could be interpreted as representing an advantage for women over men. Simply put, using the standard lexical measures common in the literature, we did not find evidence of systematic gender bias in physics, nor evidence that gender bias is greater in physics compared to social science.

Our core argument, however, is that these results do not lead us to conclude that recommendation letters are free of gender bias. The standard lexical measures rely on lengthy pre-existing lists that may include some words used more for women and others used more for men, with offsetting differences that could obscure the use of gendered language. For example, compared to men, women might have more descriptions saying they are creative but fewer saying they are innovative, with the aggregate score showing little or no gender difference. There may also be gender-differentiated words not in the word lists but which confer enthusiasm, along with words that do not overtly express enthusiasm but may indicate the use of gendered language in letters of recommendation.

We therefore supplemented the standard lexical measures with an open-ended search that checked every word in every letter for gender differences in usage. This methodology has never been used in previous studies of gender bias in letters of recommendation. The analysis revealed gendered use of language favoring women candidates in social science (a discipline that is over 60% female) but favoring men in physics (a discipline that is under 15% female). For example, men in physics were significantly more likely than women to be described as talented, intellectual, innovative, and creative. In contrast, women were more likely to be described as “brilliant,” a more highly regarded term in STEM disciplines with a lower representation of women (29). Two of the largest gender disparities were for “physicist*” in EPP and “scien*” in social science, both used more often in letters for men than for women. Although other explanations are possible, future research should investigate the possibility that recommendations can reflect unconscious gender stereotypes in which physicists and scientists are imagined as men, not only in a discipline that is overwhelmingly male but also in a discipline that is majority female.

There are limitations of the open-ended methodology as well. First, our analysis focused on gendered language and tested to see who was favored by the disparity. Accordingly, Figure 3 did not include the much larger number of expressions of endorsement that were equally likely to be used in letters about women and men. Second, the interpretation of individual words depends on context, and even words that unambiguously express enthusiasm have unknown effects on hiring decisions. This study focused on the attributes of letters that might influence hiring decisions, not on how decision-makers respond to these attributes. Figure 3 reveals gender differences in how letters are written, not how they are read. We did not examine whether that disparity affected hiring decisions, an important problem that falls outside the scope of a lexical analysis. Although Figure 3 includes terms like “brilliant*,” “talent*,” and “creativ*” whose stem variants are included in word lists used for Figure 1 and have been validated as exerting an influence on hiring, other terms (e.g., “delightful” and “physicist*”) have not been shown to systematically influence hiring. Nonetheless, the gendered language is inconsistent with the assumption that letters for female candidates are indistinguishable from letters for males. Moreover, even if letters are written in an unbiased manner, they might still be read in a biased manner by search committees (6, 8, 31). Future research should use randomized trials to manipulate the apparent candidate-gender of identical letters 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.

All the methods we used share a common limitation: We cannot rule out the possibility that female candidates were in fact superior to male candidates and deserved stronger letters than those they received. However, we controlled for candidate qualities when assessing gender differences between recommenders for candidates with letters from both genders. In addition, we compared EPP to a benchmark discipline with gender parity. Assuming that gender differences in candidate qualifications are no greater in EPP than in social science, our results might change very little were we able to completely control for candidate qualifications. This applies as well to the differences we observed between the lexical and open-ended measures.

In addition, although our sample is much larger than those in previous analyses of gender differences in letters of recommendation, we may have underestimated the number of significant differences due to insufficient statistical power based on 1,045 letters in social science and 1,161 in EPP. It is also possible that we overestimated the number of significant differences by using the conventional 0.05 benchmark for statistical significance. Figures 1 and 2 tested for gender differences on ten letter attributes (letter length, author rank, and eight different lexical dimensions), broken down by discipline and writer gender, plus author gender broken down by discipline. Out of 44 significance tests using a conventional benchmark of $p < 0.05$, we should expect two false positives if the null hypothesis were true. Had we lowered α to 0.001, none of the observed lexical differences in Figure 1 would have been statistically significant, and only the differences associated with the gender of writers in social science (Figure 2) would remain significant.

The open-ended approach in Figure 3 is particularly vulnerable to false positives, and caution is therefore warranted in drawing inferences of gender bias in the use of any individual term. Of the 27 terms in Figure 3, only three gender differences would remain significant with $p < 0.001$: “physicist*” and “intellectu*” in EPP (both favoring men) and “commit*” in social science, favoring female candidates. However, the overall pattern in Figure 3 is unlikely to be random. The sign of the gender difference will be random if the difference is a false positive, given that the measure takes into account gender differences in the number and length of letters. If all 27 terms in Figure 3 were false positives, an equal number would be expected to favor each gender. The probability that only two out of eleven random differences would favor women in EPP is 0.027, which supports the rejection of the null hypothesis that the differences are false positives. Moreover, given the persistent gender imbalance in many math-intensive disciplines, a false negative in tests for gender bias might be equally serious if it were used to justify existing practices that systematically favor male candidates. Furthermore, using the same $p < 0.05$ benchmark as previous studies makes a direct comparison possible; using a more stringent criterion precludes an “apples-to-apples” comparison.

Caution is also needed when generalizing these results from entry-level to senior positions, from EPP to all of

physics, from two high-profile institutions to all of science, and from National Laboratories to universities. Similarly, we do not wish to overgeneralize from the samples in sociology and psychology to all of the social sciences, or from Cornell University to the entire set of universities. At the same time, 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. We therefore expect that the letters in our samples are likely to resemble those submitted by these applicants to other searches beyond these two institutions.

In closing, our study has vital practical implications. Future research on letter content should also include an open-ended search for all expressions of endorsement with gender disparities. Lexical analyses using word lists obtained from previous studies indicated gender disparities in EPP letters that were no greater than those observed in social science fields in which women are well-represented. However, our open-ended search of all gender-differentiated words revealed terms of endorsement that were used more often in EPP to depict men and in social science to depict women. Normative expectations about language evolve alongside changes in gender distribution. Word lists that were sensitive to overt expressions of bias in an earlier time period may need to be supplemented by open-ended measures.

Our study also has implications for hiring practices. Disciplines with extreme gender disparities need to proactively engage faculty to raise awareness of subtle as well as overt biases in how letters are written for job candidates and how they are read. At the same time, we also caution against over-reaction. It is highly unlikely that biased letters are the principal cause of women's underrepresentation in the workplace. Policies to correct gender imbalances in math-intensive fields may be more effective if they target barriers in addition to bias in letters of recommendation, such as how letters are evaluated, and obstacles that discourage women from pursuing math-intensive academic careers.

4. Acknowledgements

We thank Fermi National Accelerator Laboratory (Fermilab) and Cornell University for use of the letters. We would also like to thank the developer of LIWC, Prof. James W. Pennebaker of the University of Texas, Austin, for valuable discussions. M.W.M. thanks the National Science Foundation (SBR 2049207 and SBR 1756822) for funding during the time the research was conducted.

5. Competing Interest

Authors declare no competing interest. There is no situation that could be perceived as exerting an undue influence on the presentation of their work. This body of work is not influenced by financial, professional, contractual, or personal relationships or situations.

6. Data Sharing Plans

All data and code used in this study are available at <https://github.com/rhbob/genderDifferences>, in the Supplementary Materials, or can be provided by the authors without undue delay.

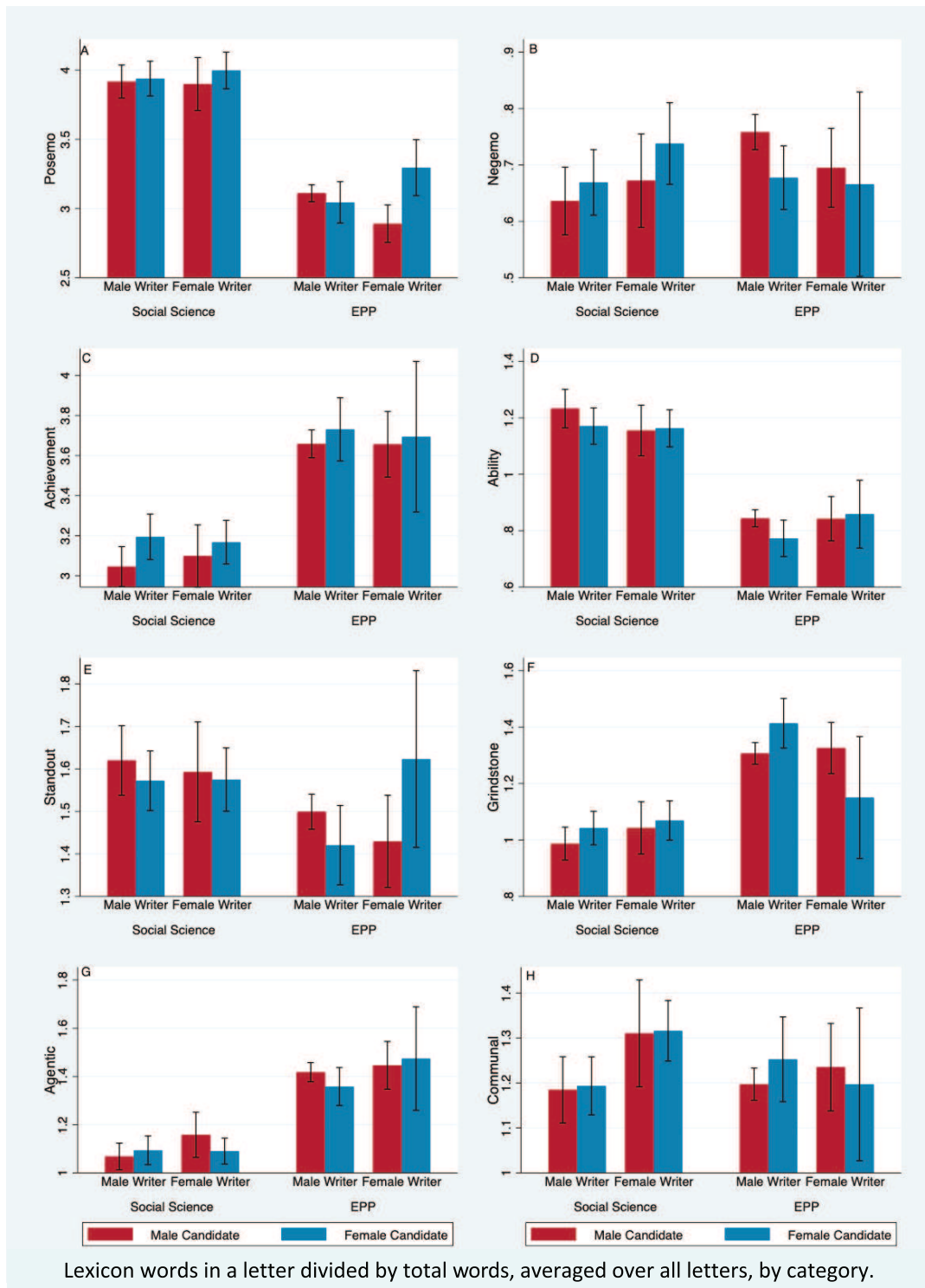


Fig. 1. Eight lexical measures for male and female candidates, by discipline and gender of writer. The Y-axis measures the percent of words in each letter that appear in the lexicon (*e.g.* "Posemo"), averaged over all letters in each of the four categories. Error bars are 95% confidence intervals. 844 letters were written by and for men in EPP and 301 were in social science; 176 letters were by men for women in EPP and 298 were in social science; 121 letters were by women for men in EPP and 139 were in social science; and 22 letters were by and for women in EPP and 307 were in social science. In EPP, women receive more positive affect words (A) than do men among letters written by women and fewer negative words (B) and more grindstone words (F) among letters written by men.

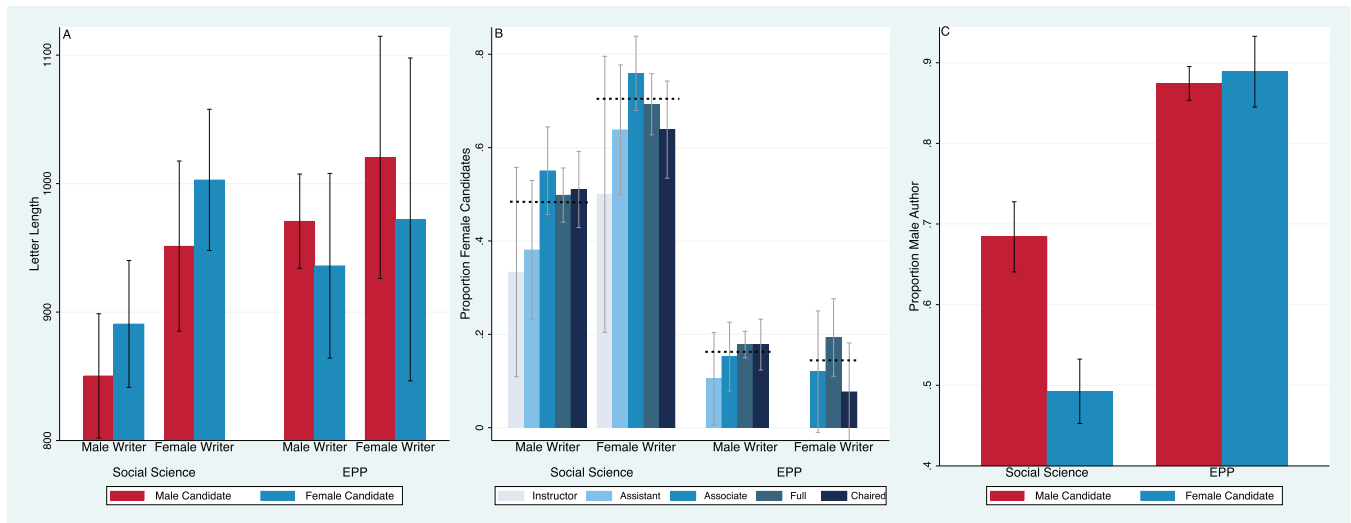


Fig. 2. Word count, proportion of letters for women across author academic ranks, and male authorship of letters for male and female candidates, by discipline and gender of writer. Error bars are 95% confidence intervals. Note the suppressed zero in the word count histogram. The dotted lines in Panel B show the expected values. In social science, letters for men are 19% more likely to be male-authored (Panel C). No other gender differences were statistically significant at the 0.05 level.

Gender Difference in Percent of Letters Containing Term

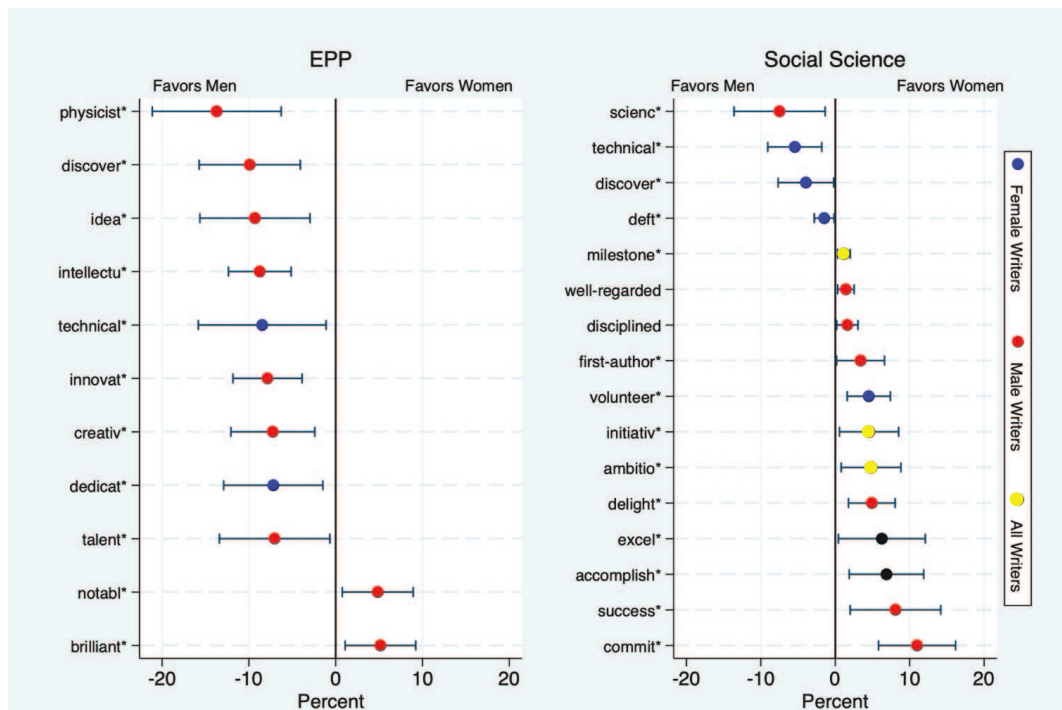


Fig. 3. Gender difference in percent of a recommendation letter containing a given term, by discipline. Error bars are 95% confidence intervals. We removed stop words, gender pronouns, references to gender differentiated sub-field specializations, and words with little or no gender difference in usage, leaving 405 words that were assessed by the senior physicist and a social scientist author, who agreed on 63 terms that signaled support for the candidate. These were expanded to 63 word families with a common stem and tested for gender differences in the percent of letters containing one or more occurrences of a family member. In EPP, nine terms were more likely to appear in letters for men compared to two for women. In social science, of 16 gender disparities, all but four favored women. Compared to letters about women, letters about men were more likely to contain references to science in social science and to physicists in EPP. Yet physicists also used "brilliant" in nearly three times as many letters about women as men. Supplementary Table S6 provides the numerical data.

Supplementary Information for

Novel Measures Reveal Subtle Gender Bias in Academic Job Recommendations

R.H. Bernstein, M.W. Macy, C.J. Cameron, S. Williams-Ceci, W.M. Williams, and S.J. Ceci

Robert H. Bernstein
email: rhbob@fnal.gov

S1. Descriptive Statistics for Gender Distributions in the Data Set

Table S1 contains the sample size for each category examined in this study. This study considered Fermilab searches for four years starting with the search starting in October 2014 and ending with the search starting in October 2017. As can be seen, the smallest category is that of female candidates with letters by female writers, $N = 22$. This reflects the phenomenon that motivated our analysis — the dearth of women in EPP. Table S2 gives the number of candidates with letters from writers of both genders used in our “matched-pair” analysis.

Our assignment of gender is binary: male or female depending on name and pronoun use. No candidates or writers supplied a preference for non-binary choices.

S2. Comparison of Current and Former Studies

How do the findings of the current study compare with those from prior studies? Here we examine all previous studies of academic letters, some of which have reported divergent results from the current study, which as was seen, found few gender differences, most of which favored women. In this section we highlight four possible reasons for any inconsistencies between these results and former ones.

1. Date: All but one of the prior studies that reported findings inconsistent with the current study are older than the current study, sometimes much older. This could result in disparities because of the dramatic attention in the last decade to gender bias in STEM (6), including massive media attention (#MeToo, #TimesUp) that has been recognized inside the academy. Behavior may have changed in recent years because of this growing awareness, and fewer gender gaps may exist today than 10-20 years ago when the letters were written for many of the contradictory findings, a point made by others recently (28).
2. Different fields: Some previous studies examined applicants for jobs in fields that are less math-intensive and have much greater female representations than EPP (for example, medicine, biology, and geoscience). *It is therefore even more curious that some of these studies found gender bias in letters of recommendation, while we found little evidence of gender bias in EPP.* We therefore urge extreme caution in generalizing our results beyond EPP and our two subfields of the social sciences, including even other subfields of physics such as theoretical particle physics. The core values for a field such as EPP, which requires large collaborative efforts (hence, a possibly greater emphasis on communal and grindstone traits), may differ from fields in which research is carried out by much smaller collaborative teams or even by lone individuals. Again, the Social Sciences are usually less collaborative and math-intensive than EPP and have dramatically higher representations of women than EPP so we expected the Social Science effects if anything to be markedly smaller, which they were not.
3. Different levels of jobs: Some prior studies that are seemingly contradictory with the current study involved letters for job levels that were lower-status than those in the current study, which were for elite positions at Fermilab and Cornell. In contrast, many but certainly not all past studies focused on candidates for normal post-doctoral fellowships, residencies, and internships rather than tenure-track-equivalent positions.
4. Nationality of writers: Some prior findings, particularly Dutt et al., that are inconsistent with the current findings contained a large proportion of letters written by recommenders from outside America. That study finds non-US letters are much shorter and they are less enthusiastic than letters written by American scientists (18). Thus, there is evidence that differences in letter length and tone vary significantly across cultures, with letters written by Americans being longer and more positive than those emanating from other regions. The vast majority of the letters in the present study were written by Americans; in contrast, in Dutt et al., 530 out of 1,224 letters were written by scientists outside the US. Thus, the current study clarifies what (until now) has been a set of seemingly contradictory findings that result from studies that are mostly small-scale and based on samples lacking important controls (such as the unavailability of letters for unsuccessful applicants, or too small subsamples of female writers to do gendered analyses). Some, but not all, of these earlier studies have reported evidence of gender bias; however, there are many exceptions and contradictions that we enumerate next.
5. Sample size: The current study is based on one of the largest sample of letters thus far examined, 2,206 letters. Prior studies varied from 237 letters (23) to 2,625 letters (32). This is important because the smaller samples precluded examining correlations between the gender of applicant and the gender of the writer.

As noted in elsewhere in this article, it is important to remember that candidates typically apply to many 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. Nonetheless, caution is needed when generalizing these results; for example, the internal culture of EPP, with large collaborations requiring communal behavior, may differ from other fields of physics.

6. Dependent variables: With the exception of McCarthy et al. (33), the current study employs the largest number of dependent variables: length, agentic, communal, grindstone, standout, achievement, positive affect, negative affect, ability, homophily, and writer status.
7. Control of background: Only a few prior studies attempted to control for applicants' personal characteristics (by using number of publications, conference presentations, class rank, and/or awards as covariates). These are less than ideal controls for other ways applicants can differ (e.g., status of their mentor, quality of the journals in which they publish, prestige of their university, their contribution to multi-authored studies). In contrast, the current study included an analysis of a subsample of 918 of the 2,206 letters for candidates with letters from both genders, neither of whom was the candidates primary advisor. Each candidate thereby contributed to the measures for each writer gender. This minimizes differences between male and female candidates caused by gender differences in applicants' personal characteristics. However, it does not address the possibility that differences in letters for male and female candidates might reflect unmeasured differences in candidate qualifications.
8. Comparing disciplines that differ in women's representation: The current study contrasted two disciplines in which women are disparately represented. Only one prior study examined a math-intensive field, the analysis of letters for postdoctoral positions in geoscience (18). This current study examined an even more male-dominated field, EPP (Elementary Particle Physics) and contrasted it with fields with high female representation, psychology and sociology. This contrast provided a principled basis for the expectation of larger correlations between gender and the dependent variables within the less female-represented field, EPP.

In sum, the current study was much larger and contained more measures than in most former studies.

S2A. Detailed description of prior studies. Trix and Psenka (22) analyzed 300 letters written on behalf of applicants for faculty positions at a single U.S. medical school. The letters were written in the mid-1990s, preceding the secular movements that occurred two decades later. Unfortunately, Trix and Psenka only had access to letters written on behalf of successful applicants, precluding a comparison with unsuccessful letters. They also were unable to analyze their data as a function of letter writer's gender, due to their small sample, rendering their results more narrative than quantitative. Like the current study, they found no differences in the frequency of letters that contained standout terms (58% for men vs. 63% for women). However, they found that letters written for men tended to repeat standout terms: on average, such letters contained 2.0 standout terms vs. only 1.5 for women's letters. Because of the unavailability of letters for unsuccessful candidates, there is no way of knowing how instrumental these features were in hiring decisions. Trix and Psenka also found that letters for women contained twice as many "doubt-raisers" as did letters for men. Finally, like Dutt et al., they found that letters by writers from Europe were shorter: "Even letters from Canada were less hyperbolic than those from the USA. But we did not have enough letters to make more than general observations." Thus, Trix and Psenka (22) study was limited by its small sample size, a single institution, and a single field (medicine), and the authors were unable to analyze letters for unsuccessful candidates or as a function of the gender of the writer, precluding many of the analyses in the current study. Despite these limitations, Trix and Psenka provided a rich narrative analysis that influenced the hypotheses in the current study.

Messner and Shimahara's (2008) study was twice as large as Trix and Psenka's, 763 letters vs. 300 letters, written on behalf of applicants for a 1-year residency in otolaryngology/head and neck surgery at Stanford University's Medical School (24). Only 8.8% of letters were written by women, which limited gender of applicant \times gender of writer analyses. They found that all letters were quite positive, which echoes Dutt et al.'s finding that roughly 98.5% of letters were either good or excellent. However, Messner and Shimahara found that letters written for women contained more communal terms (e.g., team player, compassionate), and male writers were more likely to mention a female applicant's physical appearance. They found that 86% of all letters contained standout terms (averaging 2.6 per letter). However, like the current findings — and unlike Trix and Psenka's — standout terms did not differ by gender of applicant. They also found that doubt raisers (present in 19% of letters) did not differ by gender of applicant, unlike Madera *et al.* and Trix and Psenka, both of which reported more "doubt-raisers" for women applicants (19, 22). Finally, unlike most studies (e.g., Trix and Psenka's), Messner and Shimahara did not find a difference in letter length as a function of gender of writer or gender of applicant, nor did they find a correlation between letter length and favorability. However, the mean length of their letters was less than half of the length in the current study: female writers' letters = 345 words, male letter writers' letters = 328 words (not statistically

significant); in contrast, the mean length of letters in the current study ranged between 915-960 words, which is considerably longer than letters written in other studies. Our much longer letters will be qualitatively different, with more depth and detail. Comparisons between these two sets are then complicated by the evident difference in the commitment of the writer.

In a larger study of letters written for geoscience postdocs, Dutt et al. (18) analyzed 1,224 recommendation letters, submitted by writers from 54 countries (43% were from outside the U.S.), for postdoctoral fellowships in a single field, geosciences, submitted between 2007 and 2012. Unlike Messner and Shimahara, Dutt and her colleagues found that letters written for women contained fewer words (24). However, like both Messner and Shimahara as well as the current study — but unlike Trix and Psenka — Dutt et al.'s letters contained similar numbers of standout words and more grindstone words. Although these researchers found that female applicants were only half as likely as men to receive excellent letters, they found no evidence that male and female recommenders differed in their likelihood to write stronger letters for male applicants. Like Trix and Psenka, they also found that letters from American writers were on average 561 words whereas those written by Africans, (305 words), South Asians (275 words) and East Asians (320 words) were notably shorter; even Europeans, New Zealanders and Australians wrote shorter letters (345 words) than American writers. In contrast to the current findings, Dutt et al. concluded: “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.”

Madera et al. (19) analyzed 624 letters from an earlier study that had been written for 174 applicants who had applied for positions in academic psychology at a single R1 university. Letters written for females contained more doubt-raisers, even after controlling for personal accomplishments (number of first-authored publications, honors, *etc.*) In this regard, their findings agreed with several of the above studies (e.g., Trix and Psenka (22)) but disagreed with several others.

Li et al. (23) analyzed 237 letters written on behalf of applicants to a four-year emergency medicine residency at Northwestern University. Of the fifteen dimensions they analyzed, only three revealed gender differences: letters written for female applicants were slightly longer, contained more ability terms that referred to expertise, competence, and intelligence, and also more affiliative/communal terms that referred to teamwork, helpfulness, communication, compassion, and empathy. Unlike other studies such as Trix and Psenka (22) they found no gender differences in doubt-raisers, grindstones, or standouts. Overall, they found little evidence of gender bias in letters, although the special nature of the application process may have influenced their findings (including the constraints that letters were limited to 250 words in length and only the top quarter of applicants were invited to apply to apply).

Schmader et al. (21) examined 886 letters written for 235 male and 42 female applicants for a chemistry/biochemistry faculty position at a single R1 university. Sample sizes were too small to analyze data by gender of writer \times gender of applicant, so many of the analyses in the current study were not possible. The word count of letters for women was 604 words vs. 555 for men. They found no gender differences in the frequency of grindstone words. However, unlike the current study and several others, recommenders used significantly more standout adjectives to describe male than female applicants. Letters containing more standout words also included fewer grindstone words, which runs counter to the current study's finding of weak statistical relationship ($r = -0.04$) between the co-occurrence of grindstones and standouts.

Interestingly, in the earlier analysis, Madera *et al.* (19) analyzed 624 letters written for 194 applicants in psychology and found that male recommenders wrote 262 letters for male applicants and 194 letters for female applicants; in contrast, female recommenders wrote 78 letters for male applicants and 109 letters for females. Hence, to some extent they resemble the current study's finding that male applicants submit more letters from male writers than from female writers and the reverse trend for females, which the current study did not find (that study found homophily only for males in the Social Sciences). Madera et al. also found that women were described as more communal and less agentic than men, neither of which was found in the current study. Finally, although they found more agentic adjectives in letters for males, there was no difference in “agentic orientation”, a summary of indices of how much writers referred to the applicants as active, dynamic, and achievers (using words such as “earn”, “insight”, “think”, “know”, and “do”).

The study of recommendation letters has continued to the present: we examine two we found of particular relevance. Powers et al. (32) studied a larger sample than ours for an orthopaedic residency program in 2018, examining 2625 letters for both race and gender. The reference letters were standardized to reduce potential bias, a relatively new idea. The researchers concluded (UiO indicating “underrepresented in orthopaedics”):

Small differences were found in the categories of words used to describe male and female candidates and white and UiO candidates. These differences were not present in the standardized LOR compared with traditional LOR. It is possible that the use of standardized LOR may reduce gender- and race-based bias in the narrative assessment of applicants.

The study was performed using LIWC 2015 for the standard categories: agentic/communal, standout/grindstone, and ability. Interestingly, the researchers concluded that standardized letters of reference may only produce a small effect. They also made an interesting speculation that a orthopaedics-focused word list may have obscured bias; our reverse methodology addresses this issue and is a powerful method for going beyond LIWC or other pre-defined lists. These authors also note:

A similar discrepancy has been noted in studies analyzing letters of recommendation for surgical residency and suggests that applicants preferentially ask men faculty over women faculty for letters of recommendation... If applicants believe that letters from writers of higher academic rank carry more weight, then the larger proportion of men at higher academic rank could be one explanation for this difference...

These authors also hypothesized a rank/weight correlation, but we observed no significant effect.

Kobayashi et al. (34) studied 2834 letters in another orthopaedic residency study. Their conclusions, again using LIWC 2015, were quite similar to ours:

Although there were some minor differences favoring women, language in letters of recommendation to an academic orthopaedic surgery residency program were overall similar between men and women applicants... Given the similarity in language between men and women applicants, increasing women applicants may be a more important factor in addressing the gender gap in orthopaedics.

The authors made some of the same points regarding the limitations of word lists made in Powers et al. (32) and in the current work.

To recap, amidst many similar findings, there were also many differences between the current study and former studies, any of which might be responsible for inconsistencies when they occurred. There are no studies directly comparable with the current study: they are either older (predating the recent focus on gender issues in STEM), not written on behalf of applicants for tenure-equivalent positions in STEM fields, and/or written for less math-intensive STEM fields that have higher representations of women, or written by writers from different cultures. These differences may partly explain why we found less evidence of gender bias against women candidates than might have been expected from statements made in articles such as the following (26):

Standout words in letters of recommendation... portray a candidate as talented and exciting, (and) are most often found in letters of recommendation for men. Grindstone words create the impression that a candidate works hard but is not intellectually exceptional, (and) are more often used for women... As a result of that discrepancy, female candidates seem both more boring and less intellectually promising than their male competitors.

This article appeared in *Physics Today*, a magazine for members of the American Physical Society, not a peer-reviewed journal; it was written for physicists who wanted to understand the general issue and accurately reflected a distillation of common sentiment. In contrast, the current study found that letters written for women and men in experimental particle physics (EPP) contained consistent occurrences of standout words.

S3. Differences Between Per-Word and Per-Letter Rates

Despite the effect sizes in many studies being about one word per letter, it has been argued that “only one statement can make a difference” (19). For terms that letter writers use sparingly, and which effectively distinguish candidates (such as “brilliant”), the per-letter measure is appropriate. “Research” is mentioned one or more times in nearly every letter, and hence there may be a strong signal in writing a letter that never mentions “research.” In contrast, if a writer uses “research” twice in one letter, that does not necessarily indicate that the writer was more enthusiastic than if they had used that same term only once. The measurement of the per-letter rate removes effects from repetition, allowing us to examine the “one statement” argument by using our open-ended analysis to isolate words and word families with a significant male-female difference without the need for the unblinded creation of new word categories.

S4. Differences Associated With the Gender of Writers: Gender Homophily Effects

We noted in the main text that men in the Social Sciences tended to receive letters from male writers. The existence of this gender homophily is independent of the actual ratio in the pool of possible writers: either the ratio is the same for male or female candidates or it is not. We stress that any expected value does not change the discrepancy between men and women applicants: a change in expectation only changes the relative degree of homophily assigned

to male or female candidates. The NSF’s 2017 Survey of Doctoral Recipients (25) tells us the social science fraction for women is just over 60%; in EPP, it gives 0.16 for all of EPP in 2017. A far more extensive demographic study would be required to determine a more precise expectation, but the large difference is established independent of that expectation.

S5. Research Words and Association With Gender

We use random permutation to generate a distribution of differences in word frequency under the assumption that word frequency is independent of gender. For each permutation, we shuffle the candidate gender labels and compute the t -score for the difference in frequency with pooled variance. We compare the t -score for the observed difference in word frequency against the distribution of the t scores across 100,000 permutations and compute the proportion of permutation t -scores that are less extreme than the observed difference. To limit the false discovery rate, we use a method akin to a Bonferroni correction: we compare the observed t -score for each word against the distribution of the most extreme t -scores observed in each of the sampled permutations (35). The proportion of permutations with more extreme values than the observed value gives a p -value for the observation. We also compute the uncorrected p -value by comparing the observed score for each word against the distribution of scores obtained for that word across the permutations.

We examined a subset of frequent words whose uncorrected p -values were significant at $\alpha = 0.95$. From the initial set of 646 words in Social Science and 473 words in physics, we selected words with an average word frequency exceeding 10^{-4} in letters for at least one gender, yielding 218 words for Social Science and 159 for physics. Since the average letter length was about 1000 words, a frequency of 10^{-4} translates into an average of approximately once per ten letters. For each word in these subsets, we identified words describing a research area by inspecting randomly selected examples of the word from one of these categories were dropped.

For example, the term “family” appears in the Social Science corpus about three times more often in letters for female candidates than in letters for male candidates. The t -score for the observed difference in average letter proportions is -4.47 (df=1039). The permutation test yields three two relevant comparison distributions generated under the independence assumption: (1) because the observed value is negative, the distribution of the minimal t -score observed for any word in each permutation and (2) the distribution of t -scores observed for “family” in each permutation. We find that only 176 of 100,000 permutations have a minimum t -score less than the observed t -score, which translates to a corrected p -value of 0.00176. The observed t -score is greater than any t -score obtained from permutation, so the non-corrected p -value is 1/100,001. Based on our examination of random samples drawn from the text, we determined that the use of the word “family” is overwhelmingly related to candidate research area (for example, “she does research on family processes”) rather than to the candidate’s personal life. We computed term frequency by averaging within-letter frequency across the letters which yields frequencies comparable to the LIWC scores used in the main analysis. Word counts were generated for each letter using the CountVectorizer from scikit-learn, with the tokenizer set to retain hyphens (36). We generated within-letter frequencies by dividing the within-letter word counts by the total word count for the letter (including non-vocabulary words).

Fig. S1. Research area terms, per-word, associated with candidate gender, in social science. Gendered research interests are apparent: note the clustering of words such as “child”, “parent”, “develop*” with women and “econom*”, “comput*”, or “model” with men. Some specific words such as “women/womens”, “family*” and “children/childrens”, in red, have Z -values < -4 for women; “computational” and “models”, in blue, have Z -values $> +4$ for men. The Z -score represents uncorrected p -values from the permutation test. Plot text color is a function of the corrected p -values which account for the expected false discovery rate when making multiple comparisons; any non-grey color indicates that the observed difference was among the most extreme values obtained in the permutation test but does not necessarily indicate significance (See Section S5). Words with $|Z| < 2$ are suppressed for clarity.

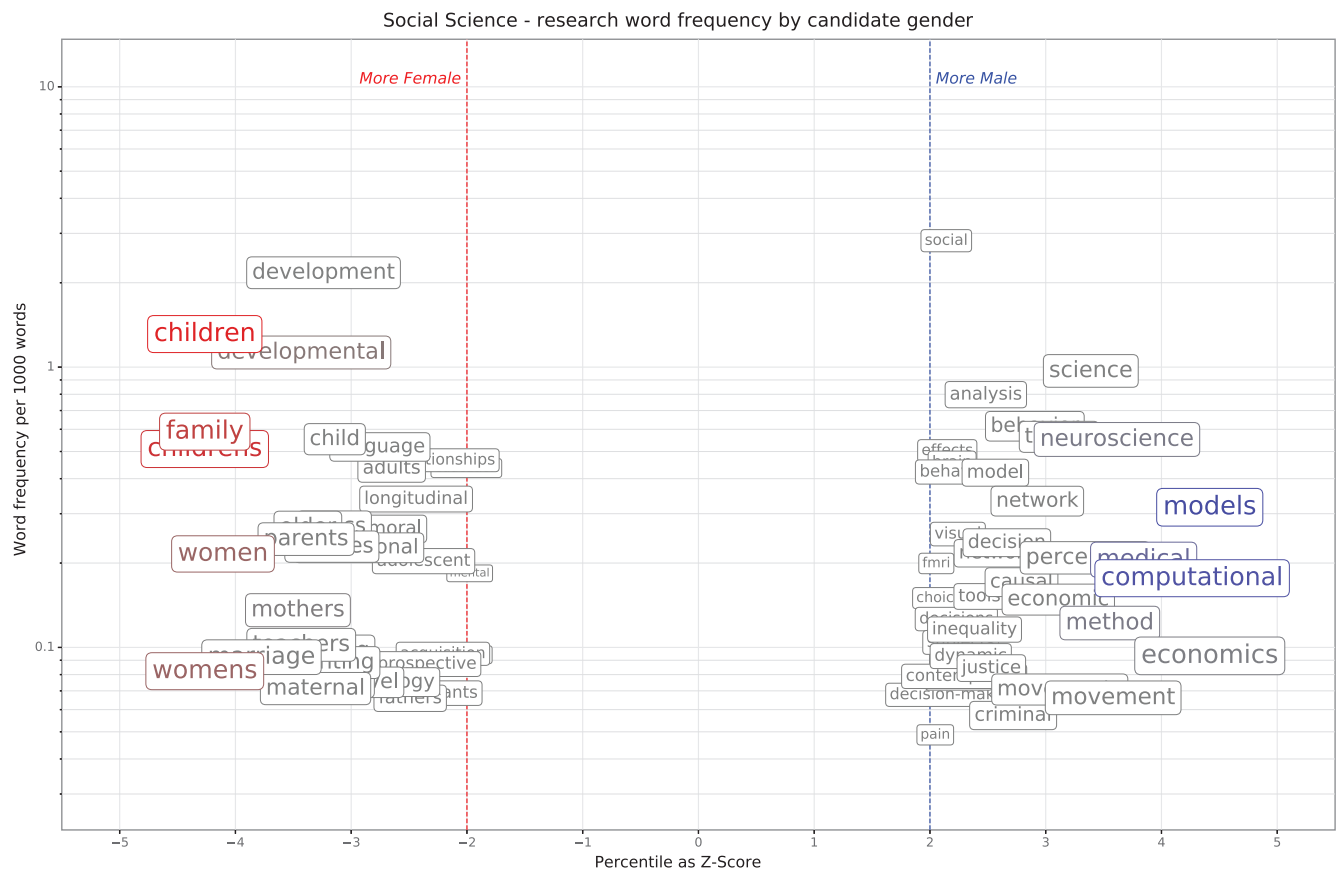


Table S1. Descriptive Statistics for sample sizes used in this study.

	Male Writer		Female Writer		Sum	Unique Male Candidates	Unique Female Candidates
	Male Candidate	Female Candidate	Male Candidate	Female Candidate			
EPP	842	176	121	22	1161	206	39
Social Science	301	298	139	307	1045	163	222

Table S2. Numbers of candidates with letters from both genders of writers.

	EPP	Social Science
Male	91	75
Female	12	141

Table S3. Demographics of writers and candidate choice. Uncertainties are 1σ statistical.

	EPP	Social Science
Letters/Male Candidate	4.67	2.70
Letters/Female Candidate	5.08	2.73
Female Writers/Male Candidate	0.59	1.34
Female Writers/Female Candidate	0.56	1.38
Male Writers/Male Candidate	4.09	1.85
Female Writers/Female Candidate	4.51	1.34
Female Writers/Male Writers		
Male candidates	0.14 ± 0.014	0.46 ± 0.04
Female Candidates	0.13 ± 0.028	1.03 ± 0.084
Expected from Pool of Writers	0.16 ± 0.022	0.67 ± 0.04

Table S4. Numbers of letters before and after requiring candidates have letters from both genders of writers.

	Male Writer		Female Writer	
	Male Candidate	Female Candidate	Male Candidate	Female Candidate
EPP Total	844	176	121	22
EPP Both Genders	302	47	121	22
Social Science Total	301	298	139	307
Social Science Both Genders	133	214	104	221

Table S5. Data for Word Counts and Fractions. These values are plotted in Figure 1 and 2 of the main text. Figure 1 errors at 95% CL are $\times 1.96$ larger than these.

	male writer		female writer	
	male candidate	female candidate	male candidate	female candidate
% of Posemo words				
SocSci	3.92 ± 0.0605	3.94 ± 0.064	3.90 ± 0.0974	4.00 ± 0.067
EPP	3.11 ± 0.031	3.04 ± 0.076	2.89 ± 0.101	3.30 ± 0.0687
% of Negemo words				
SocSci	0.636 ± 0.0305	0.667 ± 0.0420	0.668 ± 0.0295	0.740 ± 0.0367
EPP	0.758 ± 0.0159	0.678 ± 0.0287	0.695 ± 0.0355	0.665 ± 0.0814
% of Achievement words				
SocSci	3.04 ± 0.0891	3.19 ± 0.507	3.09 ± 0.0575	3.17 ± 0.793
EPP	3.66 ± 0.0354	3.73 ± 0.0802	3.66 ± 0.832	3.69 ± 0.187
% of Ability words				
SocSci	1.23 ± 0.0347	1.17 ± 0.0327	1.16 ± 0.0458	1.16 ± 0.0333
EPP	0.844 ± 0.015	0.772 ± 0.0327	0.843 ± 0.0398	0.859 ± 0.0599
% of Standout words				
SocSci(mean)	1.62 ± 0.042	$1.570.036$	1.60 ± 0.060	1.57 ± 0.038
EPP	1.50 ± 0.021	1.42 ± 0.048	1.43 ± 0.055	1.62 ± 0.104
% of Grindstone words				
SocSci	0.987 ± 0.030	1.04 ± 0.030	1.04 ± 0.047	1.05 ± 0.031
EPP	1.31 ± 0.020	1.41 ± 0.045	1.33 ± 0.046	1.15 ± 0.108
% of Agentic words				
SocSci	1.07 ± 0.028	1.09 ± 0.030	1.15 ± 0.048	1.09 ± 0.027
EPP	1.42 ± 0.020	1.36 ± 0.040	1.45 ± 0.050	1.47 ± 0.107
% of Communal words				
SocSci	1.17 ± 0.035	1.20 ± 0.033	1.29 ± 0.058	1.32 ± 0.035
EPP	1.20 ± 0.018	1.25 ± 0.048	1.24 ± 0.049	1.20 ± 0.085
LIWC wordcount				
SocSci	$850. \pm 24.6$	$891. \pm 25.0$	$952. \pm 27.9$	$1002. \pm 25.0$
EPP	$966. \pm 18.1$	$936. \pm 36.5$	$1020. \pm 62.6$	$972. \pm 36.5$

Table S6. The raw fractions, difference, and 95% CI for the terms in Figure 3. In EPP there were 198 letters about females and 963 about males; in social science, there were 605 letters about females and 440 about males. The Table shows the gender correlation in proportion of recommendation letters containing a given word, by discipline. Stems are shown for expressions with stem variations. All variations were checked for semantic consistency. We first identified all gender-correlated terms ($p < 0.05$) without considering the sign. We then removed gender pronouns, references to gender differentiated sub-field specializations, and terms with context-dependent meaning, leaving 16 unambiguous expressions of endorsement in social science, of which 12 were used more often about women. In EPP, nine terms were used more often about men and two for women.

Term	Use for Females	Use for Males	Male – Female	95% CI
EPP				
physicist*	0.3687	0.5057	-0.1370	±0.0743
discover*	0.1616	0.2606	-0.0990	±0.0583
idea*	0.2071	0.3001	-0.0930	±0.0634
intellectu*	0.0455	0.1329	-0.0875	±0.0361
technical*	0.3535	0.4382	-0.0847	±0.0736
innov*	0.0606	0.1391	-0.0785	±0.0398
creativ*	0.1010	0.1734	-0.0724	±0.0483
dedic*	0.1566	0.2285	-0.0719	±0.0571
talent*	0.2121	0.2825	-0.0703	±0.0636
notabl*	0.0859	0.0374	0.0485	±0.0408
brilliant*	0.0859	0.0343	0.0516	±0.0407
Social Science				
scienc*	0.4661	0.5409	-0.0748	±0.0612
technical*	0.0661	0.1205	-0.0543	±0.0363
discov*	0.0810	0.1205	-0.0395	±0.0374
deft*	0.0033	0.0182	-0.0149	±0.0133
mileston*	0.0116	0.0000	0.0116	±0.0085
well-regarded	0.0165	0.0023	0.0143	±0.0111
disciplined	0.0231	0.0068	0.0163	±0.0142
first-author	0.9455	0.9114	0.0341	±0.0321
volunteer*	0.0860	0.0409	0.0450	±0.0290
initiativ*	0.1455	0.100	0.0455	±0.0397
ambitio*	0.1504	0.1023	0.0481	±0.0402
delight*	0.0992	0.0500	0.0492	±0.0313
excel*	0.6876	0.6250	0.0626	±0.0584
accomplish*	0.2529	0.1841	0.0688	±0.0501
success*	0.4992	0.4182	0.0810	±0.0609
commit*	0.3008	0.1909	0.1099	±0.0518

Table S7. Count of letters with and without the term “brilliant.” If a letter contains “brilliant” more than once, it is only counted once. Because of low statistics (3 and 4 uses of “brilliant” for female/female and male/female candidate/writer pairs) the reader needs to take care with drawing conclusions from this Table. Specifically, the data do not reveal any statistically significant difference in the use of the term “brilliant” for (1) male versus female writers or (2) male writers recommending male candidates versus female writers recommending female candidates.

Candidate Gender	Writer Gender	“brilliant” in letter	total letters
Female	Female	3	22
Female	Male	13	176
Male	Female	4	121
Male	Male	24	842

Supporting information

All of the files below can be found in the Supplementary Materials submitted with this article or at the GitHub archive locations given below.

S1 File. Master spreadsheet for LIWC analysis. Contains all results from the LIWC analysis of letters. Includes both LIWC2015 words and non-LIWC2015 words gathered from the literature. It can be found in the Supplementary Information or at

<https://github.com/rhbob/genderDifferences/blob/master/scienceMasterSpreadsheet.xlsx>

S2 File. Dictionary of Words not in the LIWC2015 dictionary. The literature has many words associated with agentic/communal, standout/grindstone, etc. that we have captured in this dictionary. It can be found in the Supplementary Information or at

<https://github.com/rhbob/genderDifferences/blob/master/extraCategoriesDict.dic>

S3 File Stem mappings. Stem mappings used in the bottom-up analysis. It can be found at

https://github.com/rhbob/genderDifferences/blob/master/word_mappings_fig_3_final_v12.csv

S4 File STATA *t*-tests. STATA code used in *t*-tests. It can be found in the submitted article or at

<https://github.com/rhbob/genderDifferences/blob/master/ttestsForPaper.do>

S5 File STATA data for per-letter analysis in social science. STATA data file for the social science per-letter analysis. It can be found at

https://github.com/rhbob/genderDifferences/blob/master/soc_masterfile_stemmed2.dta

S6 File STATA data for per-letter analysis in EPP. STATA data file for the EPP per-letter analysis. It can be found at

https://github.com/rhbob/genderDifferences/blob/master/epp_masterfile_stemmed2.dta

S7 File STATA code for per-letter analysis. STATA code for the per-letter analysis used to create Fig. 3. It can be found at

https://github.com/rhbob/genderDifferences/blob/master/fig3_2color.do

S8 File Data for Rank Analysis. Data to generate Fig 2B. It can be found at

<https://github.com/rhbob/genderDifferences/blob/main/rank.xlsx>

S9 File Definitions for Rank Analysis. Definitions used to generate Fig 2B. It can be found at

<https://github.com/rhbob/genderDifferences/blob/main/rankDefinitions.txt>

1. C Hill, C Corbett, A St. Rose, Why so few? (<https://www.aauw.org/app/uploads/2020/03/why-so-few-research.pdf>) (2010) AAUW, Accessed: 2020-03-15.
2. V Valian, *Why So Slow? The Advancement of Women*. (The MIT Press, Cambridge, MA), (1998).
3. National Academy of Science, National Academy of Engineering, Institute of Medicine, *Beyond Bias And Barriers: Fulfilling the Potential of Women in Academic Science and Engineering*. (The National Academies Press, Washington, DC), (2007).
4. National Research Council, *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty*. (The National Academies Press, Washington, DC), (2010).
5. Y Xie, KA Shauman, *Women in Science: Career Processes and Outcomes*. (Harvard University Press, Cambridge, MA), (2003).
6. CA Moss-Racusin, JF Dovidio, VL Brescoll, MJ Graham, J Handelsman, Science faculty's subtle gender biases favor male students. *Proc. Natl. Acad. Sci.* **109**, 16474–16479 (2012).
7. JM Sheltzer, JC Smith, Elite male faculty in the life sciences employ fewer women. *Proc. Natl. Acad. Sci.* **111**, 10107–10112 (2014).
8. A Eaton, J Saunders, R Jacobson, K West, How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles* **82**, 127–141 (2020).
9. S Adamo, Attrition of women in the biological sciences. *Bioscience* **63**, 43–48 (2013).
10. M Goulden, K Frasch, M Mason, Keeping women in the science pipeline. *Annals Am. Acad. Polit. Soc. Sci.* **638**, 141–162 (2011).
11. R Skibba, Women in physics. *Nat. Rev. Phys.* **1**, 298–300 (2019).
12. R Su, R Rounds, All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields. *Front. Psychol.* **6**, 1–20 (2015).
13. M Wang, J Eccles, S Kenny, Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychol. Sci.* **24**, 770–775 (2013).
14. S Kelchtermans, R Veugelers, Top research productivity and its persistence: gender as a double-edged sword. *Rev. Econ. Stat.* **95**, 273–285 (2013).
15. D Kaminski, C Geisler, Survival Analysis of Faculty Retention in Science and Engineering by Gender. *Science* **335**, 864–866 (2012).
16. L Martinez, K O'Brien, M Hebl, Fleeing the ivory tower: Gender differences in the turnover experiences of women faculty. *J. Women's Heal.* **26**, 580–586 (2017).
17. SJ Ceci, DK Ginther, S Kahn, WM Williams, Women in Academic Science: A Changing Landscape. *Psychol. Sci. Public Interest* **15**, 75–141 (2014) PMID: 26172066.
18. K Dutt, D Pfaff, J Bernstein, J Dillard, C Block, Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nat. Geosci.* **9**, 805–808 (2016).
19. J Madera, M Hebl, R Martin, Gender and letters of recommendation for academia: Agent and communal differences. *J. Appl. Psychol.* **94**, 1391–1399 (2009).
20. W Morgan, K Elder, E King, The emergence and reduction of bias in letters of recommendation. *Journal of Applied Psychology* **43**, 2297–2306 (2013).
21. T Schmader, J Whitehead, V Wysocki, A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles* **57**, 509–514 (2007).
22. F Trix, C Psenka, Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society* **14**, 191–220 (2003).
23. S Li, et al., Gender Differences in Language of Standardized Letter of Evaluation Narratives for Emergency Medicine Residency Applicants. *AEM education training* **1**, 334–339 (2017).
24. AH 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 (2008).
25. National Science Foundation, Data Tables (2019) National Center for Science and Engineering Statistics, NSF 20-300 (2019), Table 9-5, <https://nces.nsf.gov/pubs/nsf19304/data>.
26. J Blue, AL Traxler, XC Cid, Gender matters. *Phys. Today* **71**, 40–46 (2018).
27. J Pennebaker, R Boyd, K Jordan, K Blackburn, *The Development and Psychometric Properties of LIWC 2015* (Univ. of Texas at Austin, Austin, TX), (2015).
28. JC French, et al., Gender and Letters of Recommendation: A Linguistic Comparison of the Impact of Gender on General Surgery Residency Applicants. *J. Surg. Educ.* **76**, 899 – 905 (2019) <http://www.sciencedirect.com/science/article/pii/S193172041830624X>.
29. SJ Leslie, A Cimpian, M Meyer, E Freeland, Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347**, 262–265 (2015).
30. M Meyer, A Cimpian, SJ Leslie, Women are underrepresented in fields where success is believed to require brilliance. *Front. Psychol.* **6** (2015).
31. WM Williams, SJ Ceci, National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proc. Natl. Acad. Sci.* **112**, 5360–5365 (2015).
32. A Powers, et al., Race- and gender-based differences in descriptions of applicants in the letters of recommendation for orthopaedic surgery residency." (2020). *JB & JS Open Access* **5**, 3 (2020) [10.2106/JBJS.OA.20.00023](https://doi.org/10.2106/JBJS.OA.20.00023)
33. J McCarthy, R Goffin, Improving the validity of letters of recommendation: an investigation of three standardized reference forms. *Mil. Psychol.* **13**, 199–222 (2001) https://doi.org/10.1207/S15327876MP1304_2.
34. A Kobayashi, et al., Are there gender-based differences in language in letters of recommendation to an orthopaedic surgery residency program? (2020) PMID: 31794493; PMCID: PMC7310286.
35. T Nichols, A Holmes, Chapter 46 - Nonparametric Permutation Tests for Functional Neuroimaging in *Human Brain Function*, eds. RS Frackowiak, et al. (Academic Press), Second edition, pp. 887 – 910 (2004).
36. F Pedregosa, et al., Scikit-learn: Machine learning in python (<https://arxiv.org/abs/1201.0490>) (2018).