Visual Salience to Mitigate Gender Bias in Recommendation Letters

Category: Research

ABSTRACT

Letters of recommendation (LOR) are an important and widely used evaluation criterion for hiring, university admissions, and many other domains. Prior work has identified that gender stereotypes can bias how recommenders describe female applicants compared to male applicants in the context of faculty positions, undergraduate research internships, residency candidates, etc. For example, female applicants are more likely to be described as communal (e.g., affectionate, warm) while male applicants are more likely to be described as agentic (e.g., confident, intellectual). In this paper, we investigate the extent to which these differences in language affect readers' impression of applicant competitiveness and explore the effectiveness of a mitigation strategy: visual highlighting. Our findings suggest that simple modifications to visual salience through highlighting more commonly-female language can emphasize biased language and has a negative effect on readers' evaluation of candidates, while highlighting more commonly-male language can reduce the effects of the bias.

1 INTRODUCTION

Letters of recommendation are a commonly used evaluation criterion for hiring, university admissions, and many other domains. In this work, *we study the evaluation of recommendation letters in the context of university admissions and specifically, how visualization might be used to mitigate readers' impressions of biased language*. University admissions committees must analyze heterogeneous data in each application, including numerical test scores, academic transcripts, applicant statements, and letters of recommendation to make admissions decisions. Many aspects of the admissions process can be susceptible to bias, compromising the objectivity and impartiality of the evaluation of applicants. We focus on recommendation letters that, while providing insight on an applicant's working style, personality, and motivations, may be fraught with biased language related to race [4] or gender [11] that can impact readers' perceptions. For instance, a previous study demonstrated that recommendation letters written on behalf of female applicants tend to more frequently feature communal qualities (e.g., affectionate, warm), whereas male applicants are more likely to be portrayed as possessing agentic traits (e.g., confident, intellectual) [20].

Building on prior work indicating that *visualization has potential to heighten awareness of biases* by revealing patterns [9], encouraging exploration [31], incorporating uncertainty [18] and providing alternative perspectives [27], we posit that **visual highlighting may help emphasize or de-emphasize gender biased language and mitigate biased perceptions of applicant competitiveness in LORs**. We emphasize that a major driving force behind our work is that we are not focusing on debiasing the letter *writers*. Some existing approaches attempt to address this process, e.g., providing a text analyzer to the letter writers that quantify gender bias in a letter [2, 25]. Instead, we are shifting the agency of de-biasing onto the *readers* of the letters, to provide novel mechanisms to mitigate the likelihood that students are evaluated in a biased manner, even if the letters composed on their behalf are fraught with biased language.

We contribute results of a crowdsourced experiment with 560 participants that investigates the efficacy of visual highlighting to emphasize or de-emphasize potentially biased language and reduce the effects of gender bias in reviewing application letters. Our findings (1) confirm that biased language can have a negative impact on readers' evaluations of candidates; (2) demonstrate that visual highlighting of biased language has potential to influence evaluation of candidates; and (3) indicate correlations between individuals' implicit biases and their perceived competitiveness of candidates who are described with more female-associated language. Namely, we found that highlighting female-associated language lead to a negative effect on readers' evaluation of candidates, while highlighting male-associated language could reduce the effects of bias. While the visualization technique itself is rudimentary, our findings indicate that changes to information salience through visual highlighting can have an effect to exacerbate or mitigate implicit gender bias.

2 RELATED WORK

Bias in Admissions. The admissions procedure can be complex and is susceptible to various forms of unconscious biases. For example, the under-representation of women in the domain of computer science is often perpetuated by gender-based discrimination during the admissions process [23]. Recommendation letters are a critical component of an applicant's portfolio, offering valuable perspectives on personal and professional attributes that might not be discernible from other application materials like curricula vitae or standardized test scores; however, these letters often contain biased language. Prior research has investigated the presence of implicit bias in recommendation letters by analyzing collections of letters and discovered that such correspondence tends to favor male candidates in aspects such as letter length (letters written for male applicants tend to be longer) [29], and overall quality (female applicants were significantly less likely than male applicants to receive an excellent versus good letter based on a coding scheme towards letter tone) [11].

Characterizing Gender Bias in LOR. Researchers have applied different methods to analyze gender bias in letters of recommendation such as qualitative content analysis [29, 34], word count-based analysis [12, 20, 25], and natural language processing (NLP) techniques [24]. Zhang, Neil, et al. applied a qualitative analysis on LOR for applicants to a cardiology fellowship program and found that underrepresented applicants were more likely to be described using communal language and doubt raising language (hedging, faint praise, and negative language) [34]. The Linguistic Inquiry and Word Count (LIWC) tool [7] was used to examine the word usage of the agentic word category (e.g., aggressiveness, assertiveness) and the communal word category (e.g., kindness, sympathy) in letters written for female and male applicants [20]. The results suggested that women were more likely to be described as communal and less agentic than men. Similarly, Filippou et al. used LIWC to characterize the vocabulary in recommendation letters [12] and discovered that male applicants were favored in terms of personal drive, work, and power, even among candidates with equivalent qualifications. Sarraf et al. utilized NLP techniques to extract the sentiment, emotion, and implicit tones conveyed by letter writers [24]. The findings revealed that LOR for male applicants exhibited considerably more positive sentiment among female authors. While there is an abundance of work analyzing document corpora to find biases in the way letters are written, to our knowledge, there has not been evaluation on the impact on readers' perception of the letters. Further, and central to our contribution, there has not yet been evaluation on the impact of real-time interventions to *reduce bias* in readers' impressions from letters with biased language in applicant descriptions.

Bias Mitigation. Traditionally, organizations have relied on diversity, equity, and inclusion (DEI) training to promote fair and

impartial procedures [5]. However, given that DEI training is often found to have minimal impact, e.g., w.r.t. augmenting the proportion of white women, black women, and black men occupying managerial positions [19], researchers have recognized the capacity of visualization to expose potential biases in the admissions process by depicting discrepancies in the distribution of applicants [28]. Other recent efforts in the visualization community have developed techniques to address biases [33] such as the attraction effect using highlighting and interaction [10] and political and gender biases [32] using techniques that encode previous interaction data e.g., in the color of data points [22].

Implicit Association Test. The Implicit Association Test (IAT) is a psychological assessment tool that measures implicit biases by gauging the strength of automatic associations between concepts and attributes [15]. The IAT has been applied across various domains including gender [14], race [15], and sexual orientation [8] by measuring reaction time to word and/or image pairs. For example, Greenwald et al. conducted a study on racial biases and found many participants were quicker to associate positive words with "white" names than with "black" names, indicating implicit racial bias [15].

3 Materials

In this section, we describe the core of the materials that will be used throughout the forthcoming pilot studies and main experiment.

Gendered Language Dictionary. Based on previous work [20, 25, 30], we created a dictionary with five categories including Grindstone, Ability, Standout [25, 30], Agentic, and Communal [7, 20]. Previous studies suggest that Communal words and Grindstone words are more often used in letters written for female applicants while Agentic words, Ability words and Standout words are more often used in recommendation letters written for male candidates. In our study, we define Communal and Grindstone related words as more commonly *female-associated words* and Agentic, Ability and Standout related words as more commonly *male-associated words*. Table 1 shows sample words in each category of our final dictionary. The dictionaries used in each study are included in supplemental materials [1].

Table 1: Sample words in each category in our dictionary.

Grindstone	Ability	Standout	Agentic	Communal
Dedicated	Adept	Amazing	Ambitious	Caring
Hardworking	Capable	Exceptional	Confident	Helpful
Organized	Talented	Superb	Independent	Warm

Recommendation Letters for Ph.D. Applicants. To select stimuli for pilot studies, we analyzed a set of recommendation letters for applicants applying for the Ph.D. program in Computer Science at the authors' university. There were 422 letters of recommendation written on behalf of 147 applicants (70% male). For each letter, all words are compared to our dictionary to obtain word counts for each category, which informed selection of letters that had comparable total word count and ratios of female- and male-associated language.

Recommendation Letters for University Applicants. Due to a relatively small participant pool of qualified reviewers for Computer Science Ph.D. applications, we switched to ChatGPT-generated letters for college admissions in later pilot studies and the main experiment. We used ChatGPT [1] to generate two letters for female applicants and two letters for male applicants using the following prompt: *"Pretend you are a high school teacher writing a moderately strong recommendation letter for a student who is applying to college. The letter is for a female/male student described as A, B, and C"* where A, B, and C are words in our dictionary (two female-, one male-associated word for the letters written for female applicants, and vice versa for male applicants). Details for the words we used and the letters are included in the supplemental materials. Multiple letters were generated and all the authors read the letters to select stimuli that were sufficiently different while maintaining comparable quality.

Customized Implicit Association Test (IAT). We created a customized IAT [15] to see if participants have an automatic gender-association for the words in our dictionary. Specifically, we asked participants to categorize common male names (e.g., Ben, Paul) and female names (e.g., Rebecca, Michelle) derived from the Gender-Career test from Project Implicit [3] as Male or Female, and to categorize the more frequently-male associated words (e.g., Leadership, Skillful) and the more frequently-female associated words (e.g., Pleasant, Warm) as Ability or Personality, respectively. The details of the test in described in supplementary materials.

4 Pilot Studies

We conducted several pilots to inform the design of the experiment described in Section 5. Full details are in supplemental materials.

Refining the Dictionary. Pilot 1 (N = 51) and 2 (N = 80) informed the final dictionary of more commonly female- v. male-associated language. These studies were used to understand if people's perceptions of gendered language aligned with our dictionary.

In these pilot studies, we selected recommendation letters written for Ph.D. applicants (see Section 3) that varied in language use (more female-associated (L_F) or more male-associated (L_M)) and other properties such as letter quality (Strong or Weak) and whether the letter mentioned the applicant had published. The selected letters were anonymized with sensitive information redacted. Additionally, gendered pronouns (he/him, she/her, etc.) were replaced with gender-neutral pronounces (they/them).

Participants were randomly assigned to read one of the recommendation letters to answer questions about 1) perceived gender of the applicant; 2) perceived gender of the recommender (letter writer); and 3) how competitive the applicant was, based only on the letter, on a scale of 1 (Extremely uncompetitive) to 7 (Extremely competitive). Participants also rated their confidence in each answer using a slider from 0 (Extremely Uncertain) to 100 (Extremely Certain) and either listed (pilot 1) or highlighted (pilot 2) the words/phrases from the letters that informed their inference of gender and competitiveness.

Our **results** showed that while participants struggled to accurately guess the gender of the applicant (average accuracy and confidence 0.47 and 55.80, respectively in pilot 1; and 0.54 and 53.93 in pilot 2), the specific language that led to individuals' inferences for female applicants consistently aligned with our dictionary including *cooperative, collaborative, hardworking, polite, and dedicated*. Participants had mixed gender perception of some of the male-associated language in our dictionary, including *excellent, intellectual, and skill*. Participants also consistently mentioned words that were not in our dictionary as indicators for male applicants (*initiative and leadership*) and female applicants (*enthusiasm and pleasure*). We used these findings to revise our dictionary to remove ambiguously perceived words and add missing words.

Testing Interventions & Determining Sample Size. Pilot 2 (N = 80) and 3 (N = 80) were used to assess the effect of interventions and determine the sample size for the final experiment.

In pilot 2, we selected two strong and two weak letters with primarily female-associated language to test interventions that included highlighting the gender-associated words in the letter, displaying the total count female- and male-associated words, and visualizing the counts in a bar chart. The procedure was similar to pilot 1 except that for the intervention group, we asked additional questions about the effect of the intervention features.

Our **results** suggested that the interventions can lead to a *negative* impact on participants' ratings of the applicants (in some cases, the

[1] https://github.com/CAV-Lab/Bias-in-LOR

intervention group rated applicants *lower* than control). This led us to evaluate *two additional interventions* where (i) only female-associated language was highlighted and (ii) only competitive-associated language was highlighted. Competitive-associated language consists of all male-associated words in our dictionary and additional words that were mentioned as indicators of competitiveness by participants from pilot 1 and 2 (e.g., enthusiasm, devotion). For all interventions, we only kept word highlighting, since word count and bar chart did not appear to have as much of an effect. In pilot 3, we ran all three interventions alongside a control using university admissions letters generated by ChatGPT as described in Section 3 to determine a sample size of N = 560 would be required (power analysis with $\beta = 0.05$, $\alpha = 0.8$).

5 Experiment Design

In this pre-registered[2] experiment, we test four visual salience modes (V_C, V_X, V_Y, V_Z; see *Stimuli* section) to explore the following hypotheses: **(H1)** Letters containing more female-associated words will be rated lower than those containing more male-associated words; **(H2)** For letters containing more female-associated words, V_X will lead to the lowest competitiveness ratings, followed by V_C, V_Y, then V_Z; and **(H3)** Participants with at least moderately positive IAT score (≥ 0.35) will rate letters with more female-associated words lower than those with more male-associated words.

Participants. We recruited 560 participants on Prolific who are fluent in English and possess a bachelor's degree or higher.

Stimuli. We generated four recommendation letters as described in Section 3. Each letter can be shown in four different visual presentation modes (V) as shown in Figure 1: (1) plain text (V_C), (2) highlight female-associated language (V_X), (3) highlight both female- and male-associated language (V_Y), and (4) highlight competitive language (V_Z). For V_Y, we provide participants context about gendered language and inform them that *"we highlighted more commonly female-associated words and more commonly male-associated words"*, while for V_X and V_Z, we informed the participants that *"we highlighted some salient words"*.

Conditions. Each participant was randomly assigned to one of eight conditions for a between-subjects design based on language in the letter (more female-associated (L_F) & more male-associated (L_M)) and visual presentation modes (V_C, V_X, V_Y, V_Z) Each participant completed two trials by rating two unique recommendation letters that used similarly gendered language (either both female-associated language L_F or both male-associated language L_M, order counterbalanced). The first letter was always shown as plain text (V_C) and the second with one of the four visual presentation modes to facilitate a within-subjects comparison of the intervention effect. Figure 1 summarizes the conditions in the study.

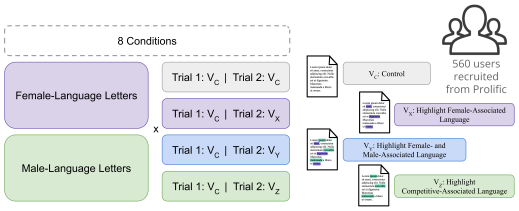


Figure 1: The conditions of the experiment.

Procedure. Participants accessed the experiment through a Qualtrics survey. Participants provided informed consent, answered demographic questions, then read two unique letters and rated the competitiveness of each applicant on a 7-point Likert scale (1 = Extremely uncompetitive, 7 = Extremely competitive) and their confidence (0-100) in the rating. For letters displayed with an intervention, participants were also asked to answer questions about whether and how the word highlighting influenced their rating of the applicant. Participants finished the study with a customized implicit association test (see Section 3).

[2] https://aspredicted.org/blind.php?x=H7Z_KNF

6 Results

We interpreted Likert responses on competitiveness rating as interval data [26] and performed parametric analysis to assess hypotheses.

Letter Language. Figure 2a shows the overall ratings for each of the letters from trial 1 (without intervention). We observe that letters (1 and 2) with more female-associated language were rated lower than letters (3 and 4) with more male-associated language, which *aligns with **H1***.

Interventions. Figure 2b shows the overall ratings for each letter from trial 2 grouped by intervention. We found that for the letters (1 and 2) with more female-associated language, V_X led to the lowest ratings (on average 5.21 and 5.08, respectively), while V_Y and V_Z led to higher ratings which *aligns with **H2***, except that V_Y led to the highest ratings instead of V_Z. For the letters (3 and 4) with more male-associated language, we also observed that V_X led to the lowest ratings. Different from female language letters, V_Y and V_Z did not lead to higher ratings compared with the control group.

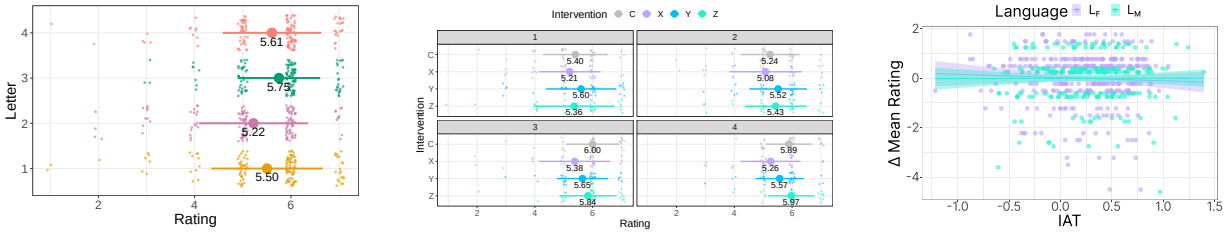
Statistical Analysis of H1 and H2. To validate the significance of the observed trends, we used a mixed-effect linear model to predict the competitiveness rating with language of the letter and intervention as fixed effects and participants as random intercepts. Dummy variables were created for each of the categorical predictors and female language (L_F) and control(V_C) were set as the reference levels. The results are summarized in Table 2. We observe that intervention V_X and language (L_M) are significant predictors while intervention V_Y and V_Z have no significant effects. The positive correlation between male language and competitiveness rating **supports H1**. The negative correlation between intervention V_X and competitiveness rating, and the positive correlation between intervention V_Z and competitiveness rating **partially supports H2**.

Table 2: Mixed-effect linear model results using the language of the letter and the intervention as predictors.

	Coefficient	Std. Error	t-value	p-value
V_X	-0.239	0.075	-3.205	0.001 **
V_Y	-0.003	0.075	-0.039	0.969
V_Z	0.134	0.074	1.805	0.072
L_M	0.329	0.079	4.161	3.67e-05 ***

IAT Score. This study employed the improved scoring algorithm for the Implicit Association Test (IAT) developed by Greenwald et al. [16] to determine the implicit gender biases of each participant. Scores range from -2 to 2, where a higher IAT score indicates a stronger inclination to perceive females as more personality-oriented. The results of the study showed that participants had an average IAT score of 0.240 (SD=0.394), indicating a slight implicit association for females with personality and males with ability. This was derived from a total of 553 participants (5 were excluded from this analysis due to rapid responses and 2 due to an inadequate number of trials).

We compared this score to the ratings from trial 1 (without intervention) to understand the relationship between implicit associations and the resulting perceived competitiveness of applicants. Figure 2c shows IAT score plotted against $\Delta_{rating} = p_{rating} - \mu_{rating}$, where p_{rating} is the rating from a participant p, and μ_{rating} is the average rating for that letter by all participants. Regression lines were fitted separately to the letters containing more female-associated language and more male-associated language. The regression lines



(a) (b) (c)

Figure 2: (a). Overall ratings per letter in trial 1. (b). Overall ratings in trial 2 per intervention. (c). IAT score analyzed against the distance of participants' ratings from the mean. Points are colored based on whether the participant rated female- or male-language letters.

exhibited opposing directions, with L_F declining and L_M slightly increasing, signifying that a higher IAT score (greater association of females with personality) negatively impacts letters containing female-associated language.

To understand the statistical magnitude of the trend, we conducted a linear regression analysis on two groups of participants: IAT < 0.35 and IAT ≥ 0.35. We utilized language ($L_F = 0$, $L_M = 1$) and IAT score as regressors for competitiveness rating. In both cases, language was a significant predictor of rating (low IAT group: $Coef. = 0.252, p = 0.029$; high IAT group: $Coef. = 0.404, p < 0.01$). The greater coefficient in the high IAT group suggests that participants with stronger implicit associations were more likely to rate letters with female-associated words lower. Overall, our results **support H3**.

Confidence in Ratings. We used a mixed-effects linear model to predict the confidence in competitiveness rating with the intervention as a fixed effect and participants as random intercepts. We observe that intervention V_Z has a positive correlation with the confidence ($Coef. = 2.107, p = 0.024$), meaning participants were more confident in their ratings when competitive words were highlighted. This is consistent with qualitative feedback that we discuss next.

Feedback on Interventions. We asked participants to indicate how the visual highlighting influenced their ratings of applicant competitiveness on a 7-point scale (-3: Much lower, 0: About the same, 3: Much higher). More than half of the participants (63%) indicated that the highlighting had no influence on their ratings. Among the participants who were influenced by the intervention, all the interventions had a slightly positive impact on the ratings (V_X: $\mu = 0.60$ (N = 62); V_Y: $\mu = 0.69$ (N = 38); V_Z: μ=1.06 (N = 61)), with V_Z showing the strongest influence.

Participants also provided qualitative feedback on how the visual highlighting influenced their ratings in a free-text question. We summarize some salient observations. For many participants, highlighting words *"brought attention to the key words that indicated the candidate strengths"* and influenced higher ratings or made some *"a bit more confident in [their] rating of "competitive," as the highlighted words did seem to emphasize the candidate's competence quite a bit."* Other participants found the opposite to be true – in particular, the highlighted words made some *"realize that there were not any specific examples attached to the salient words"* leading to skepticism or in some cases *"distracted from the letter because [their] attention was pulled to the highlighted words."* Apart from the effects on participants' ratings, we also observed some feedback to support increased *awareness* of implicit biases, *"It made me think about how I was perceiving the candidate/visualizing what they might look and act like on a day to day basis. It made me think more critically about the bias I attach to specific words."* However, this awareness was not salient enough to mitigate biases in many cases, *"They seemed more male and possibly more competitive."*

7 Discussion & Conclusion

Reading Between the Lines. Our qualitative analysis of pilot data illuminated additional nuance to the way people perceive language. Words that seem positive on the surface, e.g., "inspiring, important, impressive", were occasionally perceived as underwhelming to describe competitive candidates. One participant indicated that the phrase "*extremely impressed* at the candidate's intellect" was a strong indicator that the applicant was female, because it implied surprise, and the letter writer would only be surprised at the intellect if the applicant were female. These nuances highlight the difficulty to find universally effective mitigation strategies.

Implicit Association and Biased Outcomes. While IAT has been employed in various domains, its credibility remains a subject of debate, namely in whether implicit associations are meaningfully related to discriminatory behaviors. By reassessing the results of McConnell and Leibold [21, 35], Blanton et al. indicated that race only marginally influenced prediction errors and the IAT was incapable of predicting individual-level actions [6]. Given the uncertainty surrounding the method, additional studies are required to understand if there are meaningful correlations between implicit associations and biased outcomes.

Limitations. We note at least two primary limitations of our study. First, to maintain the independence of participants' evaluations and minimize the potential influence of confounding factors (e.g., anchoring [13], contrast effect [17], we chose not to allow participants to go back and revise their ratings. However, this approach deviates from most real-world application review scenarios, wherein decision-makers typically review multiple applicants and may revisit ratings after calibrating their judgments. Second, our analysis of H1 pre-supposes that the four recommendation letters are equal in quality. While we generated them to be as comparable as possible, there are many nuances in the language of letters that make it difficult to produce truly equivalent letters. For instance, one participant noted *"the letter-writer repeated some of the same words over and over again. I feel like if the candidate was very competitive, or truly exceptional, the letter writer would have found some more creative or descriptive terms to describe them"* – when highlighted, this repetition can become more apparent and may lead to unintended unequal perceptions of candidate competitiveness.

Conclusion. We reported results of a crowdsourced experiment with 560 participants on the effects of visual highlighting interventions for mitigating gender bias in recommendation letters. We found that letters containing more commonly-female language were rated as less competitive than letters containing more commonly-male language. Furthermore, we found that there is a negative correlation between scores in implicit association tests [15] and the competitiveness ratings for letters containing more female-associated words. Finally, we found that interventions that visually highlight female-associated language lead to lower competitiveness ratings, compared to control condition and interventions that highlighted male-associated language. This result further emphasizes recent work suggesting the power of visualizations to influence decision making and suggests a compelling possibility for visualizations to address implicit gender biases.

REFERENCES

[1] Chatgpt. https://openai.com/blog/chatgpt. Accessed: 2023-04-30. 2

[2] Gender bias calculator. https://tomforth.co.uk/genderbias/. Accessed: 2023-04-30. 1

[3] Project implicit. https://implicit.harvard.edu/implicit/index.jsp. Accessed: 2023-04-30. 2

[4] M. Bertrand and S. Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, 94(4):991–1013, 2004. 1

[5] K. Bezrukova, C. S. Spell, J. L. Perry, and K. A. Jehn. A meta-analytical integration of over 40 years of research on diversity training evaluation. *Psychological bulletin*, 142(11):1227, 2016. 2

[6] H. Blanton, J. Jaccard, J. Klick, B. Mellers, G. Mitchell, and P. E. Tetlock. Strong claims and weak evidence: reassessing the predictive validity of the iat. *Journal of applied Psychology*, 94(3):567, 2009. 4

[7] R. L. Boyd, A. Ashokkumar, S. Seraj, and J. W. Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pp. 1–47, 2022. 1, 2

[8] A. B. Breen and A. Karpinski. Implicit and explicit attitudes toward gay males and lesbians among heterosexual males and females. *The Journal of social psychology*, 153(3):351–374, 2013. 2

[9] M. Correll and J. Heer. Regression by Eye: Estimating Trends in Bivariate Visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1387–1396. ACM, Denver Colorado USA, May 2017. doi: 10.1145/3025453.3025922 1

[10] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri. Mitigating the attraction effect with visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):850–860, 2018. 2

[11] K. Dutt, D. L. Pfaff, A. F. Bernstein, J. S. Dillard, and C. J. Block. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience*, 9(11):805–808, Nov. 2016. doi: 10.1038/ngeo2819 1

[12] P. Filippou, S. Mahajan, A. Deal, E. M. Wallen, H.-J. Tan, R. S. Pruthi, and A. B. Smith. The presence of gender bias in letters of recommendations written for urology residency applicants. *Urology*, 134:56–61, 2019. 1

[13] A. Furnham and H. C. Boo. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42, 2011. 4

[14] A. G. Greenwald and M. R. Banaji. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4, 1995. 2

[15] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998. 2, 4

[16] A. G. Greenwald, B. A. Nosek, and M. R. Banaji. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197, 2003. 3

[17] P. M. Herr, S. J. Sherman, and R. H. Fazio. On the consequences of priming: Assimilation and contrast effects. *Journal of experimental social psychology*, 19(4):323–340, 1983. 4

[18] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, Jan. 2019. doi: 10.1109/TVCG.2018.2864889 1

[19] A. Kalev, F. Dobbin, and E. Kelly. Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies. *American Sociological Review*, 71(4):589–617, Aug. 2006. doi: 10.1177/000312240607100404 2

[20] J. M. Madera, M. R. Hebl, and R. C. Martin. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009. 1, 2

[21] A. R. McConnell and J. M. Leibold. Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of experimental Social psychology*, 37(5):435–442, 2001. 4

[22] A. Narechania, A. Coscia, E. Wall, and A. Endert. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1009–1018, Jan. 2022. arXiv:2108.02909 [cs]. doi: 10.1109/TVCG.2021.3114827 2

[23] E. S. Roberts, M. Kassianidou, and L. Irani. Encouraging women in computer science. *ACM SIGCSE Bulletin*, 34(2):84–88, June 2002. doi: 10.1145/543812.543837 1

[24] D. Sarraf, V. Vasiliu, B. Imberman, and B. Lindeman. Use of artificial intelligence for gender bias analysis in letters of recommendation for general surgery residency candidates. *The American Journal of Surgery*, 222(6):1051–1059, 2021. 1

[25] T. Schmader, J. Whitehead, and V. H. Wysocki. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles*, 57(7):509–514, 2007. 1, 2

[26] L. South, D. Saffo, O. Vitek, C. Dunne, and M. A. Borkin. Effective use of likert scales in visualization evaluations: a systematic review. In *Computer Graphics Forum*, vol. 41, pp. 43–55. Wiley Online Library, 2022. 3

[27] J. Sun, T. Wu, Y. Jiang, R. Awalegaonkar, X. V. Lin, and D. Yang. Pretty Princess vs. Successful Leader: Gender Roles in Greeting Card Messages. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, New Orleans LA USA, Apr. 2022. doi: 10.1145/3491102.3502114 1

[28] P. Talkad Sukumar, R. Metoyer, and S. He. Making a Pecan Pie: Understanding and Supporting The Holistic Review Process in Admissions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, Nov. 2018. doi: 10.1145/3274438 2

[29] F. Trix and C. Psenka. Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society*, 14(2):191–220, Mar. 2003. doi: 10.1177/0957926503014002277 1

[30] F. Trix and C. Psenka. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, 14(2):191–220, 2003. 2

[31] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert. Podium: Ranking Data Using Mixed-Initiative Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):288–297, Jan. 2018. doi: 10.1109/TVCG.2017.2745078 1

[32] E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):966–975, Jan. 2022. doi: 10.1109/TVCG.2021.3114862 2

[33] E. Wall, J. Stasko, and A. Endert. Toward a Design Space for Mitigating Cognitive Bias in Vis. In *2019 IEEE Visualization Conference (VIS)*, pp. 111–115. IEEE, Vancouver, BC, Canada, Oct. 2019. doi: 10.1109/VISUAL.2019.8933611 2

[34] N. Zhang, S. Blissett, D. Anderson, P. O'Sullivan, and A. Qasim. Race and gender bias in internal medicine program director letters of recommendation. *Journal of Graduate Medical Education*, 13(3):335–344, 2021. 1

[35] J. C. Ziegert and P. J. Hanges. Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. *Journal of applied psychology*, 90(3):553, 2005. 4