Visual Salience to Mitigate Gender Bias in Recommendation Letters

Yanan Da, Mengyu Chen, Ben Altschuler, Yutong Bu, and Emily Wall

Emory University, Atlanta GA 30322, USA
yanan.da@emory.edu, mengyu.chen@emory.edu, ben@altschuler.com,
audrey.bu@emory.edu, emily.wall@emory.edu

Abstract. Letters of recommendation (LORs) 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 contexts such as faculty positions and undergraduate research internships. For example, female applicants are more likely to be described using communal adjectives (e.g., affectionate, warm) while male applicants are more likely to be described using agentic adjectives (e.g., confident, intellectual). In this paper, we investigate (i) the extent to which these differences in language affect readers' impression of applicant competitiveness and (ii) the efficacy of a mitigation strategy: visual highlighting. Our findings suggest that simple changes in visual salience through highlighting language more commonly used to describe women can negatively affect readers' evaluation of candidates, while highlighting the language more commonly used to describe both men and women can reduce the effects of the bias.

Keywords: Visual salience · Gender bias · Bias mitigation.

1 Introduction

University admissions is a complex and often subjective decision-making process that can be susceptible to bias, compromising the objectivity and impartiality of the evaluation of applicants. Researchers in the HCI field have studied how to apply visualization techniques to support the admissions decision-making process and mitigate potential biases in the process [51,38]. In this work, we study the evaluation of letters of recommendation (LORs), 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 transcripts or standardized test scores. While LORs can offer valuable insights, previous research has revealed that they often contain biased language reflecting gender stereotypes [53,47,36,20,46]. These biases in language not only reflect stereotypes but may also influence evaluators' perceptions, potentially disadvantaging certain applicants [36].

We operationally define gender bias in LORs as the systematic differences in language and descriptions used by letter writers to describe female vs. male

applicants, which often reflect gender stereotypes. Such bias, in turn, can impact the evaluation of applicants. While many studies have focused on analyzing the gendered language patterns present in recommendation letters, there has been limited exploration into how these biased descriptions affect evaluators' perceptions and decision-making. Furthermore, there has been little exploration of potential interventions that could mitigate these biases in evaluators' judgments. In this work, we aim to fill this gap by investigating two research questions: (i) whether biased language in LORs influences evaluators' assessments of applicants, and (ii) whether visualization interventions can effectively mitigate the potentially negative impacts of this bias in the context of university admissions.

Our approach is grounded in prior research that suggests *visualization has potential to heighten awareness of biases* by revealing patterns [14], encouraging exploration [55], incorporating uncertainty [30] and providing alternative perspectives [50]. Specifically, we posit that **visual highlighting, by changing the salience of biased language, may help emphasize or de-emphasize gender biased language and reduce the likelihood of biased evaluations**. 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 correct bias during the writing process, e.g., providing text analyzers that quantify gender bias for self-correction [3,47]. Instead, the purpose of this research is to investigate the agency of *readers* of the letters, to provide novel mechanisms to mitigate the likelihood that applicants are evaluated in a biased manner, even if the letters composed on their behalf contain biased language.

We first created a gendered language dictionary based on prior work [53,47,36], refining it through multiple preliminary studies. Based on this dictionary, we created letters that vary in use of stereotypically gendered language and conducted a crowdsourced experiment with 560 participants to answer our research questions. Our findings indicate that: (1) Biased language can negatively affect candidate evaluations; (2) Visual highlighting of specific types of language in LORs has potential to influence evaluation of candidates; and (3) Evaluators' implicit bias correlates with their assessment of candidates.

Our results show that candidates described with more female-associated language are perceived as less competitive than those described with more male-associated language. Moreover, highlighting female-associated terms can amplify this bias, further lowering perceived competitiveness compared to plain text. In contrast, highlighting male-associated language can redirect attention away from female-associated terms, reducing bias effects. Our findings suggest that visual highlighting can either exacerbate or mitigate implicit gender bias, depending on how it alters information salience.

2 Background

2.1 Gender Bias in LORs

Recommendation letters are a key component of admissions and hiring processes but often reflect gender bias. Previous work has studied this bias across various

contexts, including undergraduate admissions [6], research internships [29], postdoctoral fellowships [17], academic faculty hiring [36], and medical fields [34]. Systematic differences in LORs for male and female applicants have been observed in various aspects such as letter length [53], linguistic style (e.g., sentiment, emotion) [46], and the use of gendered descriptors [36].

An analysis of LORs for medical faculty applicants showed that letters written for female applicants were generally shorter, lacked basic features, and were more likely to include "doubt raisers" such as negative or hedging language, faint praise, and irrelevant comments [53]. Additionally, women were described with "grindstone" adjectives (e.g., hardworking, conscientious), while men were more often described with "standout" adjectives (e.g., excellent, superb), reflecting gender schema that associates effort with women, and ability with men in professional areas. Madera et al. [36] explored language patterns in LORs through social role theory [18], which suggests that gendered language stems from societal expectations: men are typically perceived as agentic (e.g., assertive, competitive), while women are perceived as communal (e.g., friendly, unselfish). The study found that letters for female applicants more often included communal terms, while letters for male applicants emphasized agentic traits. These differences likely reflect the writers' perceptions, influenced by social role stereotypes or norms about appropriate descriptors. Importantly, communal characteristics were negatively associated with hiring decisions, as agentic traits align more closely with leadership and high-status academic roles.

While existing research has extensively analyzed gender bias in letters of recommendation, it has focused primarily on analysis of written content, rather than reader perception and has not yet considered interventions to reduce biases in reader perception.

2.2 Implicit Bias

Implicit bias is a form of unconscious bias shaped by cultural and societal norms or past experiences [23,24]. It can affect how people interpret information, interact with others, and make decisions without conscious awareness [23]. Implicit gender bias (unconscious associations between specific traits and gender) can influence how recommenders describe candidates. When evaluating candidates, even when their qualifications are similar, implicit bias can lead evaluators to favor certain gender groups. For instance, a laboratory experiment showed that science faculty rated male candidates as significantly more competent and hireable than equally qualified female candidates [40]. In our study, we explore how biased language in LORs affects the evaluation of candidates and how the reader's own implicit bias comes into play.

The implicit Association Test (IAT) [25] characterizes implicit biases by measuring the association that people hold between attributes and concepts. The test asks users to quickly and accurately categorize words or images and in turn measures reaction time, such that faster (correct) responses indicate stronger associations than slower responses, suggesting how implicit attitudes can influence cognitive processes and behaviors. Various IATs have been developed to measure

people's implicit attitudes and stereotypes toward different social groups including race, gender, age, etc. For example, the gender–science IAT measures gender stereotypes by asking participants to classify science and liberal arts terms (e.g., physics and literature) while classifying male and female terms (e.g., he and she). Data accumulated from an online IAT website [4] demonstrated robust associations of male with science and female with liberal arts – aligning with findings in laboratory studies [43].

2.3 Bias Mitigation

Diversity training [8] has been used to address implicit biases in organizational and educational settings (e.g., to improve attitudes toward women in STEM [31]). However, these trainings are often found to have minimal impact [16,33]. Recent attempts to mitigate bias in LORs have primarily focused on writers, such as text analyzers that quantify gender bias [3,47], guidelines for writing less-biased letters [42], and standardized letter formats [52]. While important, these approaches are implemented separate from the evaluation moment itself. In contrast, our work introduces an in-situ intervention that operates during the evaluation phase.

Recently, researchers in the HCI community have explored potential biases in undergraduate admissions and proposed techniques such as presenting alternative visual representations of application attributes, applying single-text visualization methods on letters of recommendation and students' essays to identify salient points, and integrating sensemaking and storytelling tools to mitigate potential biases [38,49,51]. Similarly, visualizations have been used to increase gender role awareness and prevent gender stereotypes in greeting card messages [50]. Other recent efforts have proposed computational metrics that can be applied to user interactions with data to quantify bias in real-time [19,22,54], and investigated methods to mitigate bias in data analysis [10,35,57] by altering the framing of the task [15], communicating bias metrics visually in real-time to increase the awareness of bias [41,56] or offering interaction-driven feedback and suggestions to promote more balanced exploration [32].

3 Experimental Design: Overview

We conducted three preliminary studies (Section 4) and a main study (Section 5) to understand how individuals interpret gendered language and the impact of visualization on their interpretations. This section outlines the general procedures for our studies, while details specific to individual studies are elaborated upon in the respective sections. Additional details are in supplemental materials[1].

3.1 Task and Procedure

The main task for the studies was to read an LOR written for an applicant and rate the competitiveness based on the letter. Participants completed the

[1] https://osf.io/ue7sd/

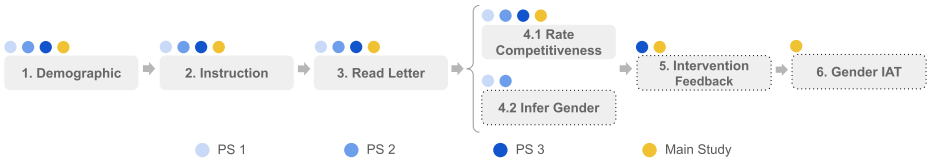


Fig. 1: Overall Procedure. Each study is represented by a colored dot which is placed around the steps that it involved. Steps 1 (provide demographic information), 2 (task instruction), 3 (read the recommendation letter), and 4.1 (rate the competitiveness of the applicant) were shared by all studies, while step 4.2 (infer the gender of the applicant) was only involved in PS 1 and 2, step 5 (provide feedback for the interventions) was only involved in PS 3 and the main study, and step 6 (take the gender IAT) was only involved in the main study.

Table 1: Sample words in each category in our dictionary. *Female-associated words* are colored in purple) and *male-associated words* are colored in green).

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

study as a Qualtrics survey. Each study consisted of 4-6 steps as summarized in Figure 1. We describe the common procedure (steps 1, 2, 3, 4.1) shared by all the studies and leave the variations in the procedure for individual studies in the respective sections. Participants first provided demographic information such as gender and age. The participants were then given the context of the recommendation letter and the explanation of gendered language (only in the intervention condition). After participants read the assigned letter, they were asked to rate the competitiveness of the applicant based on the letter on a 7-point Likert scale (1 = Extremely uncompetitive, 7 = Extremely competitive) and give their confidence (0-100).

3.2 Materials

In this section, we describe the materials used throughout the studies.

Gendered Language Dictionary Based on previous work [53,47,36], we created an initial dictionary with five categories of language including Grindstone, Ability, Standout [53,47], Agentic, and Communal [36,11] words. Previous studies [53,47,36] suggest that Communal words and Grindstone words are used more often in LORs written for female applicants while Agentic words, Ability words and Standout words are used more often in LORs written for male candidates. Based on these prior findings, we collectively refer to Communal and Grindstone words as *female-associated words* and Agentic, Ability and Standout words as *male-associated words*. Table 1 shows sample words in each category of our final dictionary. The dictionary was then used to select stimuli for our studies.

The final dictionary contained a total of 120 female-associated words and 162 male-associated words.

Letters of Recommendation Throughout our studies, we used the gendered language dictionary to select LORs that vary in language use – those containing more female-associated words and those containing more male-associated words. We refer to them as *female-language letters* and *male-language letters* respectively. This section describes how we obtained letters for each of the studies, which included (1) sourcing anonymized letters from a Ph.D. admissions season and (2) generating artificial letters.

LORs for Ph.D. Applicants. To select stimuli for PS 1 and 2, 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, we counted the number of total and unique female- and male-associated words using the gendered language dictionary. We selected letters for our preliminary studies that varied in language use (more female-associated (L_F) or more male-associated (L_M)) and letter length resulting in four conditions. We use a set of criteria to decide the candidate letters based on: 1) the total number of words of a letter (Length); 2) the unique number of female-associated words (FW); 3) the unique number of male-associated words (MW); and 4) the ratio between female-associated words and male-associated words. These criteria allowed us to select letters that had comparable word counts and ratios of female- and male-associated language (FW/MW). More details about the criteria are included in supplemental materials. The length of the letter, while not a sufficient proxy for letter strength, was used as an initial heuristic to screen the corpus. As noted in prior work [53], length is one of the simplest variables in evaluating recommendation letters. A brief letter may lack important elements typically expected in such letters. In addition to the criteria above, we also restricted the applicant pool to those with a master's degree to ensure comparable backgrounds of applicants. Two of the authors read the candidate letters that satisfied these criteria and selected one letter for each condition. This manual review ensured that the selected letters varied in both language patterns and perceived quality.

LORs for University Applicants. Because the pool of qualified reviewers for Computer Science Ph.D. applications is relatively small, we conducted subsequent studies for the context of undergraduate college admissions which allowed us to recruit from a larger population of qualified reviewers. For this context, we used ChatGPT (GPT-4) [2] to generate LORs due to its ability to produce customizable and diverse text output. This allowed us to create stimuli with controlled linguistic characteristics for our experiment. To ensure validity, all generated letters were carefully reviewed by the authors, one of whom has served on admissions committees for multiple years, to verify that they appeared realistic and comparable in quality.

Table 2: The number of unique words for each category in each letter. Letters 1 and 2 contain more female-associated words while letters 3 and 4 contain more male-associated words.

	Female-associated	Male-associated
Letter 1	8	5
Letter 2	7	6
Letter 3	4	7
Letter 4	5	8

The following prompt was used to generate two types of letters (i.e., female-language letters and male-language letters):

> *"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 gendered-language dictionary. For generating female-language letters, we incorporated two words from female-associated categories (one Communal and one Grindstone) and one from a male-associated category (Agentic). For male-language letters, we used two words from male-associated categories (one Ability and one Agentic) and one from a female-associated category (Communal). This allowed us to control the ratio of female vs. male-associated words, ensuring that female-language letters contained more female-associated words, and vice versa, while still balancing the presence of both types of language. It also enabled us to control the overall number of gender-associated words in each letter.

Multiple letters were generated for each type and all the authors read them to select two stimuli for each type that were sufficiently different in content yet comparable in quality. The four letters each contained 276 words on average (min = 273, max = 278). Table 2 summarizes the number of unique female-associated and male-associated words in each letter. All letters were anonymized where sensitive information, such as applicant and recommender names, university/high school names, and email addresses, was redacted. Additionally, gendered pronouns (he/him, she/her, etc.) were replaced with gender-neutral pronouns (they/them) to isolate the impact of linguistic characteristics on evaluations, independent of the applicants' gender.

Customized Implicit Association Test (IAT) We created a customized IAT [25] to see if participants have an automatic gender-association for the words in our dictionary. While there are established gender-based IATs, such as the gender-science and gender-career IATs that capture gender stereotypes [44], we opted to customize the IAT to incorporate the language in our dictionary for a more precise assessment. Specifically, we asked participants to categorize names (e.g., Ben, Paul, Rebecca, Michelle; derived from the Gender-Career test from Project Implicit [4]) as Male or Female, and to categorize a subset of words from our dictionary (e.g., Leadership, Skillful, Pleasant, Warm) as Ability or

Personality, respectively. The Ability words were selected from male-associated words, while the personality words were selected from female-associated words from our dictionary.

4 Preliminary Studies

We conducted three preliminary studies (PS), which had (in some cases) overlapping goals to (i) refine the gendered language dictionary, (ii) pilot test interventions, and (iii) determine the sample size for the main study. All participants were recruited via the Prolific crowdsourcing platform and were required to be fluent in English and based in the USA. The following sections outline these key goals and their corresponding preliminary studies.

4.1 Refining the Dictionary

We conducted two preliminary studies (N = 51 and N = 80, respectively, for PS 1 and PS 2) to inform the final gendered language dictionary of more commonly female- vs. male-associated words (as described in Section 3.2). We selected recommendation letters written for Ph.D. applicants (see Section 3.2) that varied in language use (more female-associated (L_F) or more male-associated (L_M)) and other properties such as letter length and whether the letter mentioned research publication. Our goal was to validate whether these words were in fact perceived as associated with male and female applicants, and we also sought to generate a category of competitive-associated words, independent of gender. Thus in addition to the common procedures (as described in Section 3), participants were also asked to indicate words/phrases that informed their judgment of the applicant's competitiveness, the perceived gender of the applicant, confidence (0-100) about the inferred gender, and words/phrases that informed their inference of gender (step 4.2 in Figure 1).

Our results showed that while participants struggled to accurately guess the gender of the applicant based on language alone (average accuracy and confidence 0.47 and 55.80, respectively in PS 1; and 0.54 and 53.93 in PS 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*, so we removed these ambiguities from our dictionary. 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*), which we then added to our dictionary. The final refined dictionary consists of 120 female-associated words and 162 male-associated words. Additionally, we introduced a new category named Competitive, which includes 30 words that were frequently identified as indicators of competitiveness by participants (e.g., enthusiasm, devotion). Out of these 30 words, 10 are already present in our female-associated word list, four are present in our male-associated word list, and the remaining words were not previously included in our dictionary.

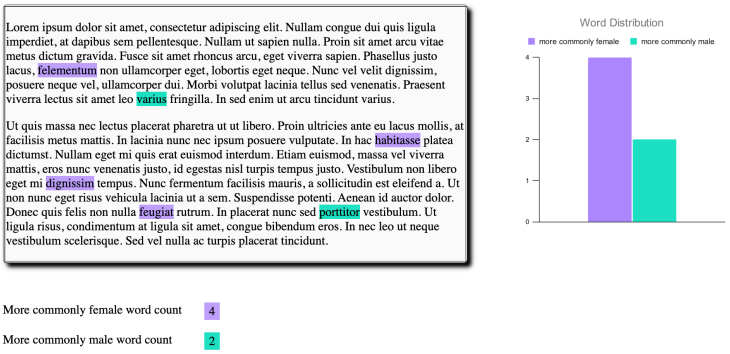


Fig. 2: Interventions used in PS 2: female- and male-associated words were highlighted in purple and green respectively. The total count of gendered words was displayed below the letter and visualized in a bar chart alongside the letter.

4.2 Testing Interventions

In PS 2 (N = 80), we tested interventions on two strong and two weak letters with primarily female-associated language. The interventions included highlighting gender-associated words, displaying word counts, and visualizing counts in a bar chart (see Figure 2). Each participant was randomly assigned to either the plain-text condition or the intervention condition where the letter was displayed with the intervention features. In the intervention group, participants were first provided with an explanation of the context of gendered language and informed that *"in the letter you will be reading, more commonly female words are highlighted in purple, and more commonly male words are highlighted in green."* Purple and green were chosen for the intervention design because they offer an alternative to traditional gender colors like pink and blue, which can reinforce stereotypes [1]. After reading and rating the letter, participants were asked questions about each of the intervention features including whether the view influenced their rating of the candidate, whether the view was useful in increasing their awareness of gendered language in the letter, whether they liked the view on a scale of 1 - 7 (1: Strongly disagree, 7: Strongly agree), and an open-ended question about how the view influenced their rating of the applicant (step 5 in Figure 1). We hypothesized that the interventions could increase readers' awareness of gendered language in the letter and potentially lead to a higher competitiveness rating compared with the plain-text condition.

Contrary to our hypotheses, the interventions sometimes led to lower ratings of applicants compared to the plain-text group. For one of the weak letters, the average competitiveness rating was 5.77 in the plain-text group and 4.80 in the intervention group. Similarly, for one of the strong letters, the average competitiveness rating was 6.09 in the plain-text group and 5.40 in the intervention group. We observed that some female-associated words (e.g., cooperative, polite) were often correlated with uncompetitiveness. This led us to evaluate *two additional interventions* in PS 3 (N = 80): (i) highlighting only female-

associated language and (ii) highlighting only competitive-associated language. Competitive-associated language consists of all male-associated words in our dictionary and the Competitive words in the dictionary (derived from PS 1 and PS 2).

For all intervention features (word highlighting, word count, and bar chart), participants rated word highlighting the highest in terms of influencing their competitiveness rating of the candidate ($M = 4.11, 2.92, 3.08$ respectively), while there were no significant differences in participants' ratings for the usefulness in increasing awareness of gendered language ($M = 5.16, 4.57, 4.22$ respectively) or whether they liked the feature ($M = 4.59, 4.43, 4.00$ respectively). Therefore, we only kept word highlighting for subsequent studies.

4.3 Determining Sample Size

In PS 3 (N = 80), we tested all three interventions (highlighting female-associated language, male- and female-associated language, and competitive-associated language) alongside a plain-text condition using university admissions letters generated by ChatGPT as described in Section 3.2. Each participant was randomly assigned to one of the eight conditions based on letter language (2) and intervention (4) in a 2 x 4 design. A power analysis using this preliminary study data suggested that a target sample size of N = 560 would be required in the main experiment to detect an overall difference between interventions with 80% power at an alpha level of 0.05.

5 Main Study Design

Built upon the preliminary studies, we conducted a pre-registered[2] experiment to test four visual salience modes (V_P: **plain** text, V_F: highlighting **female**-associated language, V_{FM}: highlighting both **female**- and **male**-associated language, V_C: highlighting **competitive** language) 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_F will lead to the lowest competitiveness ratings, followed by V_P, V_{FM}, then V_C.
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.

The first hypothesis aims to examine how bias embedded in the language of LORs influences candidate evaluations. As discussed in Section 2, although extensive research has analyzed gendered language in LORs, relatively little work has investigated how these linguistic differences actually affect readers' perceptions of candidates. We contribute to this underexplored area by moving beyond

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

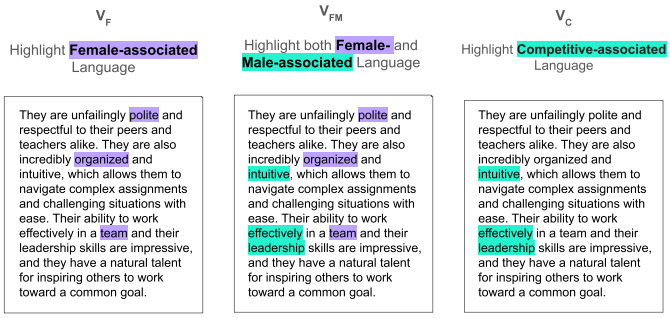


Fig. 3: Examples of four visual presentation modes on a paragraph of one of the letters used in the study.

content analysis to conduct an experimental evaluation of how readers interpret and respond to gendered language. Prior research [53] suggests that while some stereotypically female traits may be perceived positively, they often provide less constructive feedback on a candidate's academic qualifications compared to stereotypically male traits. As a result, letters containing more stereotypical female descriptors may be perceived as less competitive (**H1**). We hypothesized that highlighting only female-associated language could draw readers' attention to these stereotypically undervalued traits and potentially activate readers' implicit biases, leading to lower competitiveness ratings compared to the plain-text group. Conversely, highlighting competitive-associated language could shift focus away from undervalued traits, thus increasing competitiveness ratings (**H2**). Furthermore, our third hypothesis investigates the role of the reader's own implicit gender bias in evaluating candidates. We hypothesize that the bias reflected in LORs could activate and reinforce the reader's gender stereotypes, leading them to perceive women as less competitive than men. This dynamic may result in even lower competitiveness ratings for female candidates compared to those given by readers who do not possess strong implicit biases (**H3**).

5.1 Participants

Since we required a large sample size, the only inclusion criteria were that participants be fluent in English, based in the USA, and have at least a bachelor's degree (so that they were in principle familiar with university admissions). We ultimately recruited 560 participants (291 men, 261 women, and 8 non-binary) on Prolific based on the power analysis from PS 3 (Section 4.3). Of these, 137 participants indicated prior experience in university admissions (undergraduate/master's/Ph.D.).

5.2 Stimuli and Conditions

We generated four LORs as described in Section 3.2. Each letter is shown in four different visual presentation modes (V) as depicted in Figure 3: (1) plain text

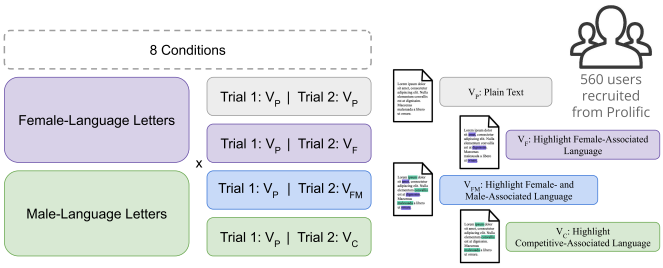


Fig. 4: The conditions of the experiment.

Table 3: Number of unique and total highlighted words per letter for each intervention. Values in parentheses indicate total counts, including repeated words.

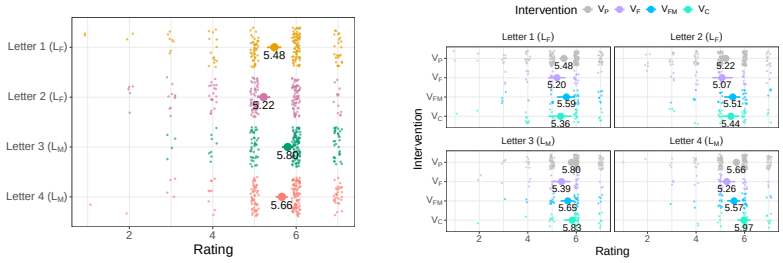
	V_F	V_{FM}	V_C
Letter 1	8 (9)	13 (13)	9 (11)
Letter 2	7 (9)	13 (16)	10 (12)
Letter 3	4 (5)	11 (18)	11 (18)
Letter 4	5 (7)	13 (19)	11 (15)

(V_P), (2) highlighting female-associated language (V_F), (3) highlighting both female- and male-associated language (V_{FM}), and (4) highlighting competitive language (V_C). The highlighting procedure was implemented using string matching against our predefined dictionary (see Section 3.2). Table 3 reports the number of unique and total highlighted words in each letter across intervention conditions. To enhance readability and minimize visual clutter, we varied the opacity of highlights for repeated words within a letter — using the darkest shade for the first occurrence and progressively lighter shades for subsequent instances. For V_{FM}, we provided participants with the context about gendered language and informed them that *"we highlighted more commonly female-associated words and more commonly male-associated words"*, while for V_F and V_C, we informed the participants that *"we highlighted some salient words."*

Each participant was randomly assigned to one of eight conditions based on language in the letter $\in$ {more female-associated (L_F) & more male-associated (L_M)} x visual presentation modes $\in \{V_P, V_F, V_{FM}, V_C\}$. Each participant completed two trials by rating two unique LORs that used similarly gendered language (either both female-associated language L_F or both male-associated language L_M, with letter order counterbalanced). The first letter was always shown as plain text (V_P) and the second with one of the four visual presentation modes to facilitate a within-subjects comparison of the intervention effect. Figure 4 summarizes the conditions in the study.

5.3 Procedure

The procedure of the study is summarized in Figure 1. Participants provided informed consent, answered demographic questions, then read two unique letters



(a) (b)

Fig. 5: (a). Overall ratings per letter in the control condition with bootstrapped 95% confidence intervals. The X-axis represents competitiveness ratings and the Y-axis represents different letters. (b). Overall ratings for each letter per intervention with bootstrapped 95% confidence intervals. The X-axis represents competitiveness ratings and the Y-axis represents the interventions.

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 (Figure 1, step 5). Participants finished the study with a customized implicit association test (see Section 3.2).

6 Results

We interpreted Likert responses on competitiveness rating as interval data [48] and performed parametric analysis to assess hypotheses.

6.1 Effects of Letter Language and Interventions

We observed that letters with more female-associated language (L_F) were generally rated lower than letters with more male-associated language (L_M) when displayed in the plain text mode (5.353 vs. 5.730), *aligning with* ***H1***. Figure 5a provides a detailed breakdown of the ratings per letter in the plain text condition. The figure reveals that the two letters (1 and 2) featuring more female-associated language consistently received lower ratings compared to the letters (3 and 4) including more male-associated language.

For the interventions, we found that for the letters with more female-associated language, V_F led to the lowest ratings ($M = 5.141$), while V_{FM} and V_C led to higher ratings ($M = 5.561$ and 5.394 respectively) which *partially aligns with* ***H2***. Notably, V_{FM} unexpectedly resulted in the highest ratings, deviating from our initial prediction that V_C would lead to the highest ratings. For the letters with more male-associated language, we also observed that V_F led to the lowest ratings ($M = 5.319$). However, different from the letters with more female-associated language, V_{FM} did not lead to higher ratings compared to the plain

Table 4: Mixed-effect linear model results using the language of the letter and the intervention as fixed effects. Language (L_M) has a significant positive effect on competitiveness rating. Highlighting female-associated language (V_F) has a significant negative effect, while highlighting competitive language (V_C) has a marginally positive effect.

	Coef.	Std. Error	t-value	p-value	95% CI
Intercept	5.369	0.059	91.461	<0.001 ***	[5.254, 5.484]
V_F	-0.239	0.075	-3.205	0.001 **	[-0.385, -0.093]
V_{FM}	-0.003	0.075	-0.039	0.969	[-0.149, 0.143]
V_C	0.134	0.074	1.805	0.072	[-0.011, 0.279]
L_M	0.329	0.079	4.161	<0.001 ***	[0.174, 0.484]

text group. Figure 5b shows a further breakdown of the overall ratings for each letter grouped by intervention. Interestingly, we noted that the effects of interventions V_{FM} and V_C differ depending on the letter language condition (as demonstrated in the interaction plot shown in Figure 6).

To validate the significance of the observed trends, we used a mixed-effects linear model to analyze our data. Given the repeated measures and the incomplete within-subject design of the intervention (i.e., subjects were not exposed to every intervention), a linear mixed-effects model was utilized to account for the non-independence of observations within participants and to handle missing observations. The fixed factors included the language of the letter and the type of intervention, while participant IDs were treated as a random factor. The dependent variable for the model was the competitiveness rating. Dummy variables were created for each of the categorical predictors, and female language (L_F) and plain text (V_P) were set as the reference levels. The results are summarized in Table 4. We observed that language (L_M) has a significant positive effect on competitiveness ratings ($Coef. = 0.329,\ p < 0.001$), **supporting H1**. Furthermore, our results indicate that intervention V_F has a significant negative effect ($Coef. = -0.239,\ p < 0.01$) on competitiveness ratings, while intervention V_C shows a marginally significant positive effect ($Coef. = 0.134,\ p < 0.1$). However, intervention V_{FM} does not significantly impact competitiveness ratings. Collectively, these findings provide **partial support for H2**.

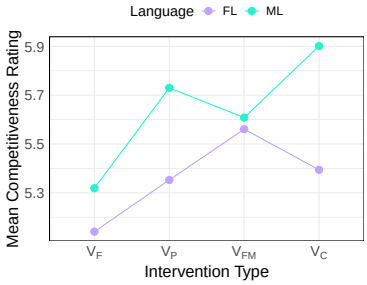


Fig. 6: The interaction between letter language and interventions.

To further investigate potential interaction effects as observed in the previous section (see Figure 6), we added interaction terms between letter language and intervention to the model. However, we found no significant interactions.

We also 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_C 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.

6.2 Association between IAT Score and Evaluation Behavior

We used the scoring algorithm for the IAT developed by Greenwald et al. [26] to determine the implicit gender biases of each participant. IAT scores range from -2 to 2, where a positive score indicates an inclination to perceive females as more personality-oriented and males as more ability-oriented, while a negative score indicates the opposite. The results 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 with 7 participants excluded (5 were excluded from this analysis due to rapid responses and 2 due to an inadequate number of trials). The excluded participants were retained in earlier analyses (Section 6.1), as they provided valid responses for the letter evaluation task.

To test H3, we divided the participants into two groups based on their IAT scores: those with IAT ≤ 0.35 (low IAT) and those with IAT > 0.35 (high IAT). The high IAT group represents participants with at least a moderate automatic association of females with personality [4]. Figure 7 shows the interaction between letter language and IAT score group. Across both groups, letters with more male-associated words were rated higher than letters with more female-associated letters. However, the difference in ratings between different letters was more pronounced in the high IAT group. To understand the statistical magnitude of the trend, we conducted a linear regression analysis on these two groups of participants. We used the language ($L_F = 0$, $L_M = 1$) and the IAT score as predictors for the 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** – participants with higher implicit biases are more susceptible to being influenced by gendered language.

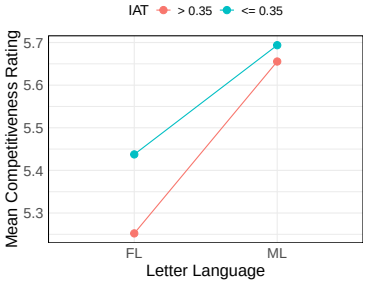
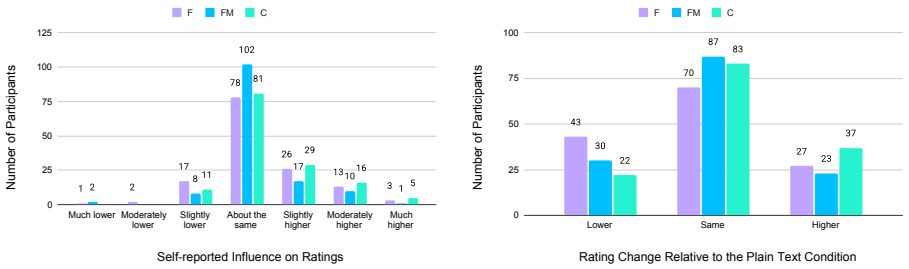


Fig. 7: The interaction between letter langue and IAT score.

6.3 Feedback on Interventions.

We asked participants in the intervention conditions (N = 422) 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). The majority of participants (62%) indicated that the highlighting had no impact on



(a) (b)

Fig. 8: (a). Distribution of participants by their Likert response on how the visual highlighting influenced their ratings grouped by intervention. (b). Distribution of participants by change in competitiveness ratings, grouped by intervention. Each participant's change was calculated by comparing their rating in the intervention trial to the plain text trial.

their ratings, while 10% indicated a negative influence, and 28% reported a positive influence. Figure 8a shows the distribution of the Likert influence score across interventions. Intervention V_F (highlighting female-associated language) had the highest proportion of negative ratings, intervention V_{FM} (highlight both female- and male-associated language) had the most neutral responses, while intervention V_C (highlight competitive associated language) had the most positive ratings. In addition to self-reported influence, we examined behavioral changes in participants' ratings between the plain text and intervention trials. For each participant, we calculated the difference between the competitiveness rating for the second letter (with a highlighting intervention) and the first letter (plain text). Figure 8b shows the distribution of these rating differences between interventions, where 'Lower' indicates that the intervention led to a lower rating compared to the plain text condition, 'Same' indicates no change and 'Higher' indicates higher rating. The distribution showed a trend similar to that of the self-reported data.

Participants also provided qualitative feedback on how visual highlighting influenced their ratings through a free-text question. To better understand the impact of the interventions, we conducted a thematic analysis [13] of these responses. The analysis followed an inductive coding approach. One of the authors conducted the coding and analysis that involved an iterative process of generating initial codes, grouping them into potential themes, and refining these themes through multiple rounds of review to ensure internal consistency and coherence within the data. The findings revealed that the same aspects of the intervention were perceived differently among the participants, highlighting its nuanced effect on evaluative judgments. Below, we summarize key themes from the analysis, capturing both positive and negative perspectives on similar aspects of the intervention.

Impact on Competitiveness/Positiveness Perception. Many participants (26) mentioned that word highlighting helped them capture important traits of

the candidate (*"visually seeing these words helped me determine their strengths and values"*) and reinforced a positive perception of the candidate. In contrast, some participants (7) in the V_F condition noted that the highlighted words *"did not indicate that the applicant was competitive"*. More specifically, some participants felt that highlighted characteristics such as teamwork and collaboration did not align with their criteria for a competitive candidate and negatively affected their evaluation. For example, one participant commented, *"It highlighted phrases that were based on teamwork and collaboration, so they were contrary to my attempts to gauge their competitiveness, thus I rated them slightly lower."* Another noted, *"it made me think the candidate is not competitive and prefers working with others to solve problems"*.

Enhanced Focus and Readability vs. Distraction. Many participants (31) appreciated how the highlighting enhanced their focus on the candidate's qualifications (e.g., *"They make the candidate's qualities stand out"*). A few participants (7) noted that the highlighting improved the readability of the letter. On the other hand, some participants (9) found the highlighted words as a distraction (*"It distracted me from the other aspects of the letter which outlined what the student has actually accomplished up to this point."*), and skewed their judgment (*"The highlighting drew my attention to specific words, which reduced the weight I placed on non-highlighted traits such as tenacity"*).

Increased Awareness of Specific Traits or Bias. A few participants (9) noted that the highlighting made them *"think and be more aware of the terms used"*, including recognition of gendered terms in the letter. However, in some cases, this awareness can unexpectedly bias participants' evaluation. For example, the highlighted words made one participant *"think of the candidate as more masculine because there were more masculine words and this made me think he was more competitive."* Conversely, another participant noted that the highlighting *"made me see that there were biases present and it was going with more gentle language"*, which led them to perceive the candidate as less competitive.

6.4 Exploratory Analyses

In addition to the pre-registered analysis, we performed exploratory analyses to understand the influence of readers' gender on competitiveness ratings in relation to gendered language. Prior work suggests that in-group cues can shape evaluative judgments, as shown by Park et al. [45], who demonstrated that same-race endorsements reduced racial discrimination in online marketplaces. We explored whether similar in-group dynamics might be observed in our setting: do readers rate applicants more favorably when the language in recommendation letters aligns with their own gender identity? We conducted a two-way ANOVA to assess the impact of the readers' gender and the language of the letter on the competitiveness ratings in trial 1 (without interventions). The results showed that readers' gender did not have a significant effect on the competitiveness ratings ($F = 0.285, p = 0.594$). The interaction effect between readers' gender and letter language was not statistically significant either ($F = 0.356, p = 0.551$). In

other words, we did not observe evidence of in-group preference in participants' ratings.

It is important to note that applicant gender was not explicitly disclosed in our study; instead, we manipulated only the presence of gendered language. The lack of a direct identity cue may have reduced the salience of in-group dynamics. Future work could explore whether in-group preferences are more likely to emerge when applicant gender is explicitly stated. Further research could also examine how reader gender interacts with bias mitigation strategies, for example, by varying not only gendered language but also the salience of the recommender's identity. Such work could shed light on whether in-group cues can be leveraged to reduce bias in evaluative contexts. Beyond gender, future investigations could examine how racial identity cues in LORs influence evaluative judgments (as discussed in Section 7).

7 Discussion

Reading Between the Lines. Our qualitative analysis of preliminary study 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 of finding universally effective mitigation strategies given the multitude of interpretations of language.

Effects of the Interventions and Design Implications. We found that V_F (highlighting female-associated language) has a negative impact on the evaluation of applicants, suggesting that merely exposing readers to biased language without providing context may activate their own bias and amplify the impact of embedded biases in the text. This aligns with prior findings that awareness-based interventions can sometimes exacerbate conscious biases rather than mitigate them [56]. On the other hand, by providing context for gendered language and highlighting those languages, V_{FM} effectively increased readers' awareness of bias, prompting more critical interpretations. As Figure 6 shows, V_{FM} leads to the highest ratings for letters with more female-associated words, demonstrating its potential as a bias mitigation strategy. We also observed evidence of increased awareness of bias from qualitative feedback (e.g., *"It made me think more critically about the bias I attach to specific words."*). In addition, highlighting both types of language offered a more balanced and comprehensive portrayal of the applicant, which could lead to more informed decision-making as demonstrated by previous work [32]. However, this awareness may not necessarily counteract biases and can unexpectedly amplify bias in some cases (as discussed in Section 6.3). This suggests the need for bias-mitigation strategies

that go beyond passive exposure. Furthermore, our results indicate that shifting readers' focus away from biased language (V_C) has the potential to reduce its negative effect (see Section 6.1). This implies that a strategic redirection of attention from biased elements can be an effective approach to bias mitigation, aligning with established practices of technology-mediated nudges in HCI research [12]. As discussed in Section 6.3, we also observed divergent perceptions of visual highlighting – while some found the highlighting beneficial in enhancing focus and improving readability, others perceived it as distracting. These findings provide key insights for designing future bias mitigation interventions for decision-support systems:

- **Supporting Critical Reflection.** Instead of merely pointing out potentially biased elements, we could offer meaningful context and promote critical reflection [7]. This could include displaying aggregate statistics of language usage across similar applications, explaining why certain terms are considered gendered, or incorporating prompts that encourage evaluators to articulate their reasoning. For example, pairing bias awareness features (such as visual highlighting) with prompts can remind readers to interpret the letters through the lens of bias. Asking readers to elaborate on their evaluation, for example, writing out their rationale for their ratings, could also help to reduce bias [39].
- **Interactive and Adaptive Bias Mitigation Interfaces.** Given the variability in cognitive load and user preferences, systems designed to surface bias-related cues can benefit from adaptive approaches. Systems could offer customizable levels of support, allowing users to toggle features, adjust the intensity of visual cues, or request additional information when needed. Future research can explore interactive designs where highlighting is activated based on user engagement, minimizing cognitive overload and distraction. For instance, similar to the interventions tested in PS 2, future designs could include visual aids (e.g., charts or textual summaries) displaying gendered language distribution alongside the letter content and interactive elements where highlighting is activated only when users engage with the visual aids, ensuring that awareness enhances evaluation rather than distracting from it.

Implicit Association and Biased Outcomes. While IAT has been employed in various domains, its credibility remains debated, particularly regarding whether implicit associations reliably predict discriminatory behaviors. For example, Blanton et al. [9] reassessed results from previous studies [37,59], concluding that IAT scores failed to predict individual-level discrimination behavior. Despite such critiques, the reliability of IAT has been validated across multiple disciplines such as job hiring [5] and healthcare [27]. Nonetheless, given the uncertainty surrounding the method, additional studies are required to understand if there are meaningful correlations between implicit associations and biased outcomes in this context of university admissions.

Limitations and Future Work. Our work has several limitations. First, to maintain the independence of participants' evaluations and minimize the potential influence of confounding factors (e.g., anchoring [21], contrast effect [28]),

we did not allow participants to go back and revise their ratings. However, this approach differs from most real-world application review scenarios, where 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, subtle nuances in language make it difficult to produce truly equivalent letters. For instance, a few participants noted repeated usage of certain words – when highlighted, this repetition can become more apparent and may lead to unintended unequal perceptions of candidate competitiveness. In addition, although our intervention focused the type of highlighted language, the number of highlighted words may have also influenced participants' ratings. As shown in Table 3 and Figure 5b, the V_{FM} and V_C conditions consistently had more highlighted words across all letters and received higher competitiveness ratings compared to the V_F condition. This suggests that a greater number of visual cues might have implicitly led participants to perceive candidates as more competitive. We acknowledge that word count was not independently controlled in our study and should be explored further in future work. Third, our study focused on binary male and female language differences. This suggests critical next steps to explore language and intervention in the context of gender as a fluid rather than a binary construct. Another limitation relates to the background of our raters. While all participants had at least a college degree, not all had direct experience with the admissions process. We deemed having a college degree a reasonable proxy for understanding the nuances of recommendation letters. However, this may not match the expertise of admissions professionals, potentially impacting the evaluation of letters and our findings. Future studies should include experienced admissions raters to better reflect real-world review processes and understand their impact on letter evaluations. In addition to the existence of gender bias in letters of recommendation, studies have also found differences in language used to describe applicants of different racial demographics in recommendation letters [29,58,42]; future work can explore how visual salience can be used to mitigate the effects of racial bias in recommendation letters.

8 Conclusion

We reported the 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 female-associated language were rated as less competitive than letters containing more male-associated language, and that the perceived competitiveness of an applicant was correlated with the rater's implicit gender associations [25]. Finally, we found that highlighting female-associated terms can amplify bias, further decreasing perceived competitiveness compared to plain text while highlighting male-associated language can shift evaluators' focus away from female-associated descriptors, helping to mitigate bias effects. Overall, our findings suggest a compelling possibility for visualizations to address implicit gender biases.

References

1. An alternative to pink & blue: Colors for gender data. https://blog.datawrapper.de/gendercolor/, accessed: 2023-04-30
2. Chatgpt. https://openai.com/blog/chatgpt, accessed: 2023-04-30
3. Gender bias calculator. https://tomforth.co.uk/genderbias/, accessed: 2023-04-30
4. Project implicit. https://implicit.harvard.edu/implicit/index.jsp, accessed: 2023-04-30
5. Agerström, J., Rooth, D.O.: The role of automatic obesity stereotypes in real hiring discrimination. Journal of Applied Psychology **96**(4), 790–805 (2011). https://doi.org/10.1037/a0021594
6. Akos, P., Kretchmar, J.: Gender and ethnic bias in letters of recommendation: considerations for school counselors. Professional School Counseling **20**(1), 1096–2409 (2016). https://doi.org/10.5330/1096-2409-20.1.102
7. Bentvelzen, M., Woźniak, P.W., Herbes, P.S., Stefanidi, E., Niess, J.: Revisiting reflection in HCI: Four design resources for technologies that support reflection. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **6**(1), 1–27 (2022). https://doi.org/10.1145/3517233
8. Bezrukova, K., Jehn, K.A., Spell, C.S.: Reviewing diversity training: Where we have been and where we should go. Academy of Management Learning & Education **11**(2), 207–227 (2012). https://doi.org/10.5465/amle.2008.0090
9. Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., Tetlock, P.E.: Strong claims and weak evidence: reassessing the predictive validity of the iat. Journal of applied Psychology **94**(3), 567–582 (2009). https://doi.org/10.1037/a0014665
10. Borland, D., Zhang, J., Kaul, S., Gotz, D.: Selection-bias-corrected visualization via dynamic reweighting. IEEE Transactions on Visualization & Computer Graphics **27**(2), 1481–1491 (2020). https://doi.org/10.1109/TVCG.2020.3030455
11. Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W.: The development and psychometric properties of liwc-22. Austin, TX: University of Texas at Austin pp. 1–47 (2022)
12. Caraban, A., Karapanos, E., Gonçalves, D., Campos, P.: 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. p. 1–15. CHI '19, ACM (2019). https://doi.org/10.1145/3290605.3300733
13. Clarke, V., Braun, V.: Thematic analysis. The Journal of Positive Psychology **12**(3), 297–298 (2017). https://doi.org/10.1080/17439760.2016.1262613
14. Correll, M., Heer, J.: 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 (2017). https://doi.org/10.1145/3025453.3025922
15. Dimara, E., Bailly, G., Bezerianos, A., Franconeri, S.: Mitigating the attraction effect with visualizations. IEEE Transactions on Visualization and Computer Graphics **25**(1), 850–860 (2018). https://doi.org/10.1109/TVCG.2018.2865233
16. Dobbin, F., Kalev, A.: Why doesn't diversity training work? The challenge for industry and academia. Anthropology Now **10**(2), 48–55 (2018). https://doi.org/10.1080/19428200.2018.1493182
17. Dutt, K., Pfaff, D.L., Bernstein, A.F., Dillard, J.S., Block, C.J.: Gender differences in recommendation letters for postdoctoral fellowships in geoscience. Nature Geoscience **9**(11), 805–808 (2016). https://doi.org/10.1038/ngeo2819
18. Eagly, A.H., Wood, W., Diekman, A.B.: Social role theory of sex differences and similarities: A current appraisal. In: Eckes, T., Trautner, H.M. (eds.) The developmental social psychology of gender, p. 123–174. Psychology Press (2000)

19. Feng, M., Peck, E., Harrison, L.: Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. IEEE Transactions on Visualization and Computer Graphics **25**(1), 501–511 (2018). https://doi.org/10.1109/TVCG.2018.2865117
20. Filippou, P., Mahajan, S., Deal, A., Wallen, E.M., Tan, H.J., Pruthi, R.S., Smith, A.B.: The presence of gender bias in letters of recommendations written for urology residency applicants. Urology **134**, 56–61 (2019). https://doi.org/10.1016/j.urology.2019.05.065
21. Furnham, A., Boo, H.C.: A literature review of the anchoring effect. The Journal of Socio-Economics **40**(1), 35–42 (2011). https://doi.org/10.1016/j.socec.2010.10.008
22. Gotz, D., Sun, S., Cao, N.: Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In: Proceedings of the 21st International Conference on Intelligent User Interfaces. pp. 85–95. ACM (2016). https://doi.org/10.1145/2856767.2856779
23. Greenwald, A.G., Banaji, M.R.: Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychological Review **102**(1), 4 (1995). https://doi.org/10.1037/0033-295x.102.1.4
24. Greenwald, A.G., Krieger, L.H.: Implicit bias: Scientific foundations. California Law Review **94**(4), 945–967 (2006). https://doi.org/10.2307/20439056
25. Greenwald, A.G., McGhee, D.E., Schwartz, J.L.: Measuring individual differences in implicit cognition: the implicit association test. Journal of Personality and Social Psychology **74**(6), 1464–1480 (1998). https://doi.org/10.1037//0022-3514.74.6.1464
26. Greenwald, A.G., Nosek, B.A., Banaji, M.R.: Understanding and using the implicit association test: I. An improved scoring algorithm. Journal of Personality and Social Psychology **85**(2), 197–216 (2003). https://doi.org/10.1037/0022-3514.85.2.197
27. Hall, W.J., Chapman, M.V., Lee, K.M., Merino, Y.M., Thomas, T.W., Payne, B.K., Eng, E., Day, S.H., Coyne-Beasley, T.: Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. American Journal of Public Health **105**(12), e60–e76 (2015). https://doi.org/10.2105/AJPH.2015.302903
28. Herr, P.M., Sherman, S.J., Fazio, R.H.: On the consequences of priming: Assimilation and contrast effects. Journal of Experimental Social Psychology **19**(4), 323–340 (1983). https://doi.org/10.1016/0022-1031(83)90026-4
29. Houser, C., Lemmons, K.: Implicit bias in letters of recommendation for an undergraduate research internship. Journal of Further and Higher Education **42**(5), 585–595 (2018). https://doi.org/10.1080/0309877X.2017.1301410
30. Hullman, J., Qiao, X., Correll, M., Kale, A., Kay, M.: In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. IEEE Transactions on Visualization and Computer Graphics **25**(1), 903–913 (Jan 2019). https://doi.org/10.1109/TVCG.2018.2864889
31. Jackson, S.M., Hillard, A.L., Schneider, T.R.: Using implicit bias training to improve attitudes toward women in stem. Social Psychology of Education **17**(3), 419–438 (2014). https://doi.org/10.1007/s11218-014-9259-5
32. Jasim, M., Collins, C., Sarvghad, A., Mahyar, N.: Supporting serendipitous discovery and balanced analysis of online product reviews with interaction-driven metrics and bias-mitigating suggestions. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, ACM (2022). https://doi.org/10.1145/3491102.3517649

33. Kalev, A., Dobbin, F., Kelly, E.: Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies. American Sociological Review **71**(4), 589–617 (Aug 2006). https://doi.org/10.1177/000312240607100404
34. Khan, S., Kirubarajan, A., Shamsheri, T., Clayton, A., Mehta, G.: Gender bias in reference letters for residency and academic medicine: a systematic review. Postgraduate Medical Journal **99**(1170), 272–278 (2023). https://doi.org/10.1136/postgradmedj-2021-140045
35. Law, P.M., Basole, R.C.: Designing breadth-oriented data exploration for mitigating cognitive biases. In: Cognitive Biases in Visualizations, pp. 149–159. Springer (2018). https://doi.org/10.1007/978-3-319-95831-6_11
36. Madera, J.M., Hebl, M.R., Martin, R.C.: Gender and letters of recommendation for academia: agentic and communal differences. Journal of Applied Psychology **94**(6), 1591 (2009). https://doi.org/10.1037/a0016539
37. McConnell, A.R., Leibold, J.M.: Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. Journal of Experimental Social Psychology **37**(5), 435–442 (2001). https://doi.org/10.1006/jesp.2000.1470
38. Metoyer, R.A., Chuanromanee, T., Girgis, G.M., Zhi, Q., Kinyon, E.C.: Supporting storytelling with evidence in holistic review processes: A participatory design approach. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW1), 1–24 (2020). https://doi.org/10.1145/3392870
39. Morgan, W.B., Elder, K.B., King, E.B.: The emergence and reduction of bias in letters of recommendation. Journal of Applied Social Psychology **43**(11), 2297–2306 (2013). https://doi.org/10.1111/jasp.12179
40. Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., Handelsman, J.: Science faculty's subtle gender biases favor male students. Proceedings of the national academy of sciences **109**(41), 16474–16479 (2012). https://doi.org/10.1073/pnas.1211286109
41. Narechania, A., Coscia, A., Wall, E., Endert, A.: Lumos: Increasing awareness of analytic behavior during visual data analysis. IEEE Transactions on Visualization and Computer Graphics **28**(1), 1009–1018 (2021). https://doi.org/10.1109/TVCG.2021.3114827
42. Newkirk-Turner, B.L., Hudson, T.K.: Do no harm: Graduate admissions letters of recommendation and unconscious bias. Perspectives of the ASHA Special Interest Groups **7**(2), 463–475 (2022). https://doi.org/10.1044/2021_PERSP-20-00117
43. Nosek, B.A., Banaji, M.R., Greenwald, A.G.: Math = male, me = female, therefore math ≠ me. Journal of Personality and Social Psychology **83**(1), 44 (2002)
44. Nosek, B.A., Smyth, F.L., Hansen, J.J., Devos, T., Lindner, N.M., Ranganath, K.A., Smith, C.T., Olson, K.R., Chugh, D., Greenwald, A.G., et al.: Pervasiveness and correlates of implicit attitudes and stereotypes. European Review of Social Psychology **18**(1), 36–88 (2007). https://doi.org/10.1080/10463280701489053
45. Park, M., Yu, C., Macy, M.: Fighting bias with bias: How same-race endorsements reduce racial discrimination on Airbnb. Science Advances **9**(6), eadd2315 (2023)
46. Sarraf, D., Vasiliu, V., Imberman, B., Lindeman, B.: 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). https://doi.org/10.1016/j.amjsurg.2021.09.034
47. Schmader, T., Whitehead, J., Wysocki, V.H.: A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex Roles: A Journal of Research **57**(7), 509–514 (2007). https://doi.org/10.1007/s11199-007-9291-4

48. South, L., Saffo, D., Vitek, O., Dunne, C., Borkin, M.A.: Effective use of likert scales in visualization evaluations: a systematic review. In: Computer Graphics Forum. vol. 41, pp. 43–55. Wiley Online Library (2022). https://doi.org/10.1111/cgf.14521
49. Sukumar, P.T., Metoyer, R.: A visualization approach to addressing reviewer bias in holistic college admissions. Cognitive Biases in Visualizations pp. 161–175 (2018). https://doi.org/10.1007/978-3-319-95831-6_12
50. Sun, J., Wu, T., Jiang, Y., Awalegaonkar, R., Lin, X.V., Yang, D.: Pretty princess vs. successful leader: Gender roles in greeting card messages. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–15. ACM (2022). https://doi.org/10.1145/3491102.3502114
51. Talkad Sukumar, P., Metoyer, R., He, S.: 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 (2018). https://doi.org/10.1145/3274438
52. Tavarez, M.M., Baghdassarian, A., Bailey, J., Caglar, D., Eckerle, M., Fang, A., McVety, K., Nagler, J., Ngo, T.L., Rose, J.A., et al.: A call to action for standardizing letters of recommendation. Journal of Graduate Medical Education **14**(6), 642–646 (2022). https://doi.org/10.4300/JGME-D-22-00131.1
53. Trix, F., Psenka, C.: Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse & Society **14**(2), 191–220 (2003). https://doi.org/10.1177/0957926503014002277
54. Wall, E., Blaha, L.M., Franklin, L., Endert, A.: Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In: 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 104–115. IEEE (2017). https://doi.org/10.1109/VAST.2017.8585669
55. Wall, E., Das, S., Chawla, R., Kalidindi, B., Brown, E.T., Endert, A.: Podium: Ranking Data Using Mixed-Initiative Visual Analytics. IEEE Transactions on Visualization and Computer Graphics **24**(1), 288–297 (Jan 2018). https://doi.org/10.1109/TVCG.2017.2745078
56. Wall, E., Narechania, A., Coscia, A., Paden, J., Endert, A.: Left, right, and gender: Exploring interaction traces to mitigate human biases. IEEE Transactions on Visualization and Computer Graphics **28**(1), 966–975 (2021). https://doi.org/10.1109/TVCG.2021.3114862
57. Wall, E., Stasko, J., Endert, A.: Toward a design space for mitigating cognitive bias in vis. In: 2019 IEEE Visualization Conference (VIS). pp. 111–115. IEEE (2019). https://doi.org/10.1109/VISUAL.2019.8933611
58. Zhang, N., Blissett, S., Anderson, D., O'Sullivan, P., Qasim, A.: Race and gender bias in internal medicine program director letters of recommendation. Journal of Graduate Medical Education **13**(3), 335–344 (2021). https://doi.org/10.4300/JGME-D-20-00929.1
59. Ziegert, J.C., Hanges, P.J.: Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias. Journal of Applied Psychology **90**(3), 553 (2005). https://doi.org/10.1037/0021-9010.90.3.553