Journal of Business and Psychology (2019) 34:287–303
https://doi.org/10.1007/s10869-018-9541-1

ORIGINAL PAPER

Raising Doubt in Letters of Recommendation for Academia: Gender Differences and Their Impact

Juan M. Madera[1] · Michelle R. Hebl[2] · Heather Dial[2] · Randi Martin[2] · Virgina Valian[3]

Published online: 26 April 2018

Abstract

The extent of gender bias in academia continues to be an object of inquiry, and recent research has begun to examine the particular gender biases emblematic in letters of recommendations. This current two-part study examines differences in the number of doubt raisers that are written in 624 authentic letters of recommendations for 174 men and women applying for eight assistant professor positions (study 1) and the impact of these doubt raisers on 305 university professors who provided evaluations of recommendation letters (study 2). The results show that both male and female recommenders use more doubt raisers in letters of recommendations for women compared to men and that the presence of certain types of doubt raisers in letters of recommendations results in negative outcomes for both genders. Since doubt raisers are more frequent in letters for women than men, women are at a disadvantage relative to men in their applications for academic positions. We discuss the implications and need for additional future research and practice that (1) raises awareness that letter writers are gatekeepers who can improve or hinder women's progress and (2) develops methods to eliminate the skewed use of doubt raisers.

Keywords Letters of recommendation · Gender schemas · Discrimination · Sex roles · Academia

Gender equity in all fields in academia has progressed over the past several decades, but data from the National Science Foundation (2004) and the U.S. Department of Commerce (2011) suggest that women continue to be less likely than men to access academic careers, to attain full-time positions, and to be promoted and tenured in the natural and social sciences, engineering, and mathematics disciplines. Dubbed the "pipeline problem," women enter graduate school at about the same frequency as do men, but are less likely to enter and succeed in academia than are their male counterparts (Aiston, 2014; Deo, 2014; Ding, Murray, & Stuart, 2013; Ellemers, Heuvel, Gilder, Maas, & Bovini, 2004; Taylor, 2007; Yost, Winstead, Cotten, & Handley, 2013). In addition, once hired, women leave academia at slightly higher rates than their male counterparts across various disciplines (e.g., Adamo, 2013; Easterly & Ricard, 2011; Kaminski & Geisler, 2012; Levine, Lin, Kern, Wright, & Carrese, 2011; National Academy of Sciences et al., 2007).

One of the limitations in this literature is that the majority of the research focuses on the selection rates of women versus men in specific fields of academia and how women experience bias in their academic careers (i.e., after selection decisions are made) (e.g., Aguirre, 2000; Howe-Walsh & Turnbull, 2016; Lee & Won, 2014; Lerback & Hanson, 2017; Settles, Cortina, Malley, & Stewart, 2006). This is an important limitation because it has left a gap in understanding how bias manifests in the early stages of the selection process. In the current studies, we address this limitation by examining letters of recommendation, one of the most important early-stage selection tools used in academia (Abbott et al. 2010; Sheehan, McDevitt, & Ross, 1998). A growing body of literature has shown how bias can influence the manner in which letters of recommendation are written. Specifically, gender biases arising from perceived gender differences can lead to differences in how letters are written for men and women (Dutt, Pfaff, Bernstein, Dillard, & Block, 2016; Isaac, Chertoff, Lee, & Carnes, 2011; LaCroix, 1985; Madera, Hebl, & Martin, 2009; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Rubini & Menegatti, 2014; Schmader, Whitehead, & Wysocki, 2007; Shen, 2013).

✉ Juan M. Madera
jmmadera@uh.edu

[1] University of Houston, Houston, TX, USA

[2] Rice University, Houston, TX, USA

[3] Hunter College and CUNY Graduate Center, New York, NY, USA

The current studies draw from the literature on gender schemas, which are mental models summarizing implicit beliefs and expectations of male and female roles (Crockett, 1988; Fiske & Linville, 1980; Valian, 1998), and the literature on gender linguistic bias (Maass & Arcuri, 1996; Rubini & Menegatti, 2014) to examine doubt raisers (i.e., phrases or statements that question an applicant's aptness for a job) in letters of recommendation (Trix & Psenka, 2003). Examples of doubt raisers include statements like "somewhat challenging personality," "might make a good colleague," and "in view of the difficulties [being experienced],. . performance was especially impressive." Though they may vary in the degree of negativity and subtleness, they all potentially raise doubts for the evaluator because they indicate that the writer is uncertain about the applicant or does not have an entirely positive impression of the applicant.

The first aim of the current studies is to determine if letters of recommendations for academic positions include more doubt raisers for women than for men. In study 1, we examine gender differences in letters of recommendation using objective methods (i.e., language content analysis) and statistical procedures appropriate for nested data. In addition, because there are well-known gender differences for several job predictor domains, such as various measures of cognitive, personality, and vocational interests (Hough, Oswald, & Ployhart, 2001; Su, Rounds, & Armstrong, 2009), we include measures of academic performance as control variables. Specifically, we use several variables that reflect objective measures of academic performance (e.g., number of publications and number of courses taught) to examine gender differences in academic performance and control for any potential differences that could be related to the use of doubt raisers.

The second aim of the current studies is to determine if doubt raisers actually affect how applicants are evaluated. Even if more doubt raisers are used for women than men in letters of recommendations, such subtleties in language may not matter. In study 2, we use experimental methods and an academic sample to examine if doubt raisers in letters of recommendation negatively affect how applicants are evaluated.

By examining gender differences in the use of doubt raisers in letters of recommendations (study 1) and how doubt raisers negatively affect applicant evaluations (study 2), the current studies will provide a better understanding of how gender schemas affect women in the early stages of the selection process in academia. By examining doubt raisers in letters, the current studies contribute to understanding how gender schemas influence the manner in which men and women are described differently in letters, even after accounting for various indicators of productivity. Research suggests that bias against women might be reduced when women are described as highly qualified because it reduces the uncertainty of whether an applicant will be successful (Heilman, Wallen, Fuchs, & Tamkins, 2004) and offsets gender schema stereotypes that work against women in occupations or roles that are often related to male gender norms (Heilman, 2012). Therefore, it is important to examine if more doubt is raised for women than men in the letters of recommendation, because doubt raisers lead to questions regarding the potential for success of an applicant by introducing uncertainty.

This research also contributes to our understanding of how gender schemas can affect women even before the selection process begins. That is, gender schemas can influence how letters of recommendation are constructed, before they are even used to evaluate an applicant, potentially biasing evaluations for women in the earliest stages of selection. This is particularly important to examine because recent research suggests a new trend for female applicants in academia; namely, selection rates for women in academia seem to be substantially improving in some STEM-related fields (Ceci, Ginther, Kahn, & Williams, 2014a, 2014b; National Research Council, 2009). A series of studies show that women were preferred over men, but only when they were described as equally and not less qualified than men (e.g., Williams & Ceci, 2015; Ceci & Williams, 2015). Despite this encouraging progress, what this research ignores is the possible bias women face at earlier stages of the selection process, before final selection decisions are made. The results of our current research represents a particularly important contribution to this literature, considering that so much of the research on gender bias in academia has focused on what occurs after selection decisions are made.

Background

Letters of Recommendations in Academia

Although they are only one of numerous factors that are considered in evaluating and selecting applicants for jobs, letters of recommendation are an important tool used to screen graduate students, medical school applicants, and faculty in academic settings (Johnson et al., 1998; Landrum, Jeglum, & Cashin, 1994; Nicklin & Roch, 2009; Sheehan et al., 1998) and are valid predictors of undergraduate performance, graduate performance, and professional school performance (Kuncel, Kochevar, & Ones, 2014). Letters of recommendation are tools that screen candidates in the selection process (Guion, 1998; Morgan, Elder, & King, 2013) because they verify information provided by applicants and offer information about applicants' past performance (Aamodt, Nagy, & Thompson, 1998; Gatewood & Feild, 2001; McCarthy & Goffin, 2001).

Both quantitative and qualitative research have identified the use and strong importance of letters of recommendation in academia. First, letters are critical determinants of who gets academic-based internships. That is, Mittenberg, Peterson, Cooper, Strauman, and Essig (2000) found that letters of recommendation and personal interviews were considered more important than grades or work samples. Similarly, in a study of predoctoral internships, 82% of internship selection members from the Association of Psychology Postdoctoral and Internship Centers ranked letters of recommendation as "important" to "very important" in their selection process (APPIC, 2005).

Second, letters are important in assessing teaching abilities of academicians. For example, in a study of how search committee chairs in psychology evaluate applicants' teaching, Benson and Buskist (2005) found that letters of recommendation were the second most used criteria (after student evaluations), and were more important than previous teaching experience, statement of teaching philosophy, and the applicant's job talk. In a similar qualitative study of how search committees in academia evaluate teaching ability, Meizlish and Kaplan (2008) examined a sample of 457 surveys from various departments, including English, history, political science, psychology, biology, and chemistry. They found that search committees put more weight on letters of recommendation to assess faculty applicants than any other criteria and that CVs, cover letters, and letters of recommendation were the three most commonly requested materials for open positions.

Third, letters of recommendation are important for inviting applicants in academia for an interview. A study of the hiring process from 368 English departments (Broughton & Conlogue, 2001) found that letters of recommendation were ranked among the top four application materials in terms of importance when screening candidates for on-campus interviews. Letters of recommendation ranked higher than other metrics, such as the number of teaching awards and course evaluations. A similar study of search committee chairs from psychology (Landrum & Clump, 2004) found that letters of recommendation were ranked higher in screening applicants than quality of graduate school, grant activity or potential, and transcripts. Most of the literature on the use letters of recommendation to assess applicants in academia has been either survey-based or qualitative in nature. However, an experiment using a sample of professors who evaluated a hypothetical applicant for an academic job in an experiment found that a strong letter of recommendation (versus a weak letter) had a significant effect on the likelihood of inviting an applicant for an on-campus interview (Applegate, Cable, & Sitren, 2009). Not only do professors use letters of recommendation to select candidates for interviews, but academic administrators also value letters of recommendation as important and useful. For example, a study of political science department chairs from 231 universities (Fuerstman & Lavertu, 2005) found that letters of recommendation were among the top three factors in inviting applicants to campus interviews across all types of universities (e.g., liberal arts colleges, doctoral-granting institutions). They found that letters of recommendation outranked a variety of other factors.

Fourth, letters of recommendation are important for the actual selection of applications for academic positions. Showing the importance of letters of recommendation for selection purposes, Nicklin and Roch (2009) found that letters of recommendation are particularly used and relied upon more in selecting candidates by those in academics than those in applied professions outside of academia. Additionally, the more that faculty wrote letters themselves, the more likely they were to rely on others' letters when making selection decisions. Provosts, department heads, and other administrators also use letters of recommendations for hiring and promoting faculty (Abbott et al., 2010). In fact, decision-makers in academic administration positions rely on letters of recommendation, particularly from outside experts, more heavily than impact factor, citations, and other metrics available. They reasoned that the best applicants have similar impact factors and citation counts, so letters help distinguish applicants more.

Several conclusions emerge from examining the literature focusing on the use letters of recommendation to assess applicants in academia. First, letters are among the most commonly requested materials for the academic selection process. Second, letters are used to evaluate applicants for both specific (e.g., teaching) and general abilities. Third, letters are often used in the early stage of the hiring process to make decisions for campus visits, so their weight and use are important to advance further in the selection process. Thus, any potential bias in letters can hinder applicants from being hired, not only because they are used to make hiring decisions, but also because they are used when selecting applicants for a campus interview.

Letters of Recommendations and Gender Differences

Despite the frequent use of letters of recommendation in academia, the instructions for how to write those letters are often ambiguous and open-ended (Morgan et al., 2013). Further, the way in which letters of recommendation are used to evaluate candidates usually lacks structure (Liu, Minsky, Ling, & Kyllonen, 2009). The ambiguity and lack of structure of letters of recommendations can lead to biases in how letters are written for men and women (Dutt et al., 2016; LaCroix, 1985; Madera et al., 2009; Schmader et al., 2007). Gender schemas, mental models summarizing beliefs about what it means to be male or female (Crockett, 1988; Fiske & Linville, 1980), provide a theoretical framework for gender biases in letters of recommendation. Gender schemas can be both descriptive and prescriptive (Burgess & Borgida, 1999; Heilman, 2001; Rudman & Glick, 2001), and are implicit, mostly nonconscious beliefs and expectations that can lead to different interpretations of the same behavior in men and women (Valian, 1998). These differences are due, at least partially, to a perceived lack of fit between the stereotypes about and the positions held by men and women (Heilman, 1983; Heilman, 2012).

Central to understanding how gender schemas can affect women in academia is the gender-typing of work through two conditions. First, the distribution of men and women in an occupation is used to stereotype an occupation as either a male or female occupation (Cejka & Eagly, 1999). Men are disproportionately highly represented in academia: women enter graduate school at about the same rate as men, but have a lower percentage of staying in academia (Aiston, 2014; Deo, 2014; Ding et al., 2013; Ellemers et al., 2004; Taylor, 2007; Yost et al., 2013). Many academic departments, such as the natural sciences, engineering, and mathematics, remain male-dominated, whereas other departments, such as education and social work, remain female-dominated (Bailyn, 2003; Eveline, 2005; Pyke, 2013; Van den Brink & Benschop, 2012; Westring et al., 2012). In fact, the majority (86%) of full professors at American institutions are men (U.S. Department of Education, National Center for Education Statistics, 2015).

Second, the responsibilities of the job are tied to gender norms (Heilman, 2001). For example, management roles traditionally have been considered to be male gender-typed because of the importance of traits (e.g., agency) that comprise the male gender schema (Eagly & Johannesen-Schmidt, 2001; Eagly & Karau, 2002; Ragins & Sundstrom, 1989; Ragins, Townsend, & Mattis, 1998. Job advertisements for male- (versus female-) dominated areas of employment use more masculine wording, thereby enhancing the belongingness that men versus women will experience when reading the ads (Gaucher, Friesen, & Kay, 2011). Responsibilities of academics have been based historically on masculine traits, such as being assertive, competitive, authoritative, independent, and experts in their field (Bailyn, 2003). All of these traits are tied to agency, which are a set of traits that men, but not women, are expected to hold (Eagly & Johannesen-Schmidt, 2001). Women, in contrast, are expected to be communal, which includes being concerned with the welfare of other people, affectionate, kind, sensitive, and nurturing.

One example of how gender schemas influence expectations in academia comes from a study of the awarding of endowed professorships at a sample of business schools at tier 1 American research universities. Treviño, Gomez-Mejia, Balkin, and Mixon (2015) found that female professors were less likely to be awarded named professorships than male professors were, even after controlling for years of experience, research productivity, and other performance factors. The disparity was even wider when the endowed chair was awarded to an internal candidate. Female professors had to meet a higher bar for recognition, as shown by the fact that women with endowed chairs scored significantly higher on performance measures than did men. Treviño et al. (2015) argued that these results were partly due to the facts that the majority (86%) of full professors at American institutions are men, and men make up the majority of gatekeepers for hiring and promoting in universities, which develops a work environment based on male gender norms. As such, a masculine-gendered work environment is incongruent with female gender norms.

Because what is required for success in many academic departments may be based on norms of masculinity (Bailyn, 2003; Van den Brink & Benschop, 2012; Westring et al., 2012), a potential bias against female faculty can arise when writing letters of recommendation. Letter writers may have sex-related stereotypes about women that are incongruent with the attributes that are believed to be required for success in a particular job (Eagly & Karau, 2002; Heilman, 2001), such as academia. The language used to describe men and women in work domains also may be related to gender schemas (Maass & Arcuri, 1996; Rubini & Menegatti, 2014). For example, letters of recommendations for medical school residency show gender differences in the language used to describe the applicants (Isaac et al., 2011). Specifically, letters for female (versus male) applicants contained more "tentative" words (e.g., "she might," "it is possible she could").

In chemistry and biochemistry faculty positions, letters of recommendations for male versus female applicants were found to contain more standout adjectives, such as "superb," "outstanding," "remarkable," and "exceptional" (Schmader et al., 2007). Similarly, in psychology, male applicants for faculty positions were described as more agentic and less communal than female applicants (Madera et al., 2009). In addition, communal descriptions were negatively related to the hireability of the applicants. Such studies suggest that (1) language in letters of recommendation may be biased unintentionally by gender schemas and (2) male and female writers are equivalent in their attribution of traits to male and female candidates.

Standout adjectives are not the only domain in which writers can describe job candidates. A qualitative study conducted by Trix and Psenka (2003) examined over 300 letters of recommendations that were written for medical school faculty at a large American medical school. Letters for women tended to contain more doubt raisers than letters for men, with no difference between male and female writers. The authors described four sets of doubt raisers: negativity, faint praise, hedges, and irrelevant information. For example, one might describe an applicant as someone who "does not have much teaching experience" (negativity), who "needs only minimum supervision" (faint praise), who "might not be the best…" (hedging), or who "is active in church" (irrelevancy). Doubt raisers vary in how negative and subtle they are and may not have an equivalently pernicious impact. Negativity may tend to be the most obvious and negative doubt raiser, because it points out an overt weakness of the applicant. Irrelevancy is typically the least negative and most subtle, but because they are not related to the essential functions of a job, the reader wonders why they are present at all, making them a doubt raiser. Hedging is less negative than negativity, but is still a forthright doubt raiser, because the writer directly admits uncertainty. Lastly, faint praise is a something of a backhanded compliment.

In general, the majority of letter content was very positive, so the inclusion of a single doubt raiser questions an applicant's aptness for a job in a manner that is not necessarily direct and apparent (Trix & Psenka, 2003). Letter writers may not have intended to put female applicants at a disadvantage, but may have done so nevertheless if they included doubt raisers more frequently in letters for women versus men.

The current studies build on Trix and Psenka's (2003) preliminary evidence of gender differences in doubt raisers by using different methodological and statistical procedures. For example, they scored letters of recommendations without removing information about the gender of the applicant. Thus, the possibility of confirmation bias might have been present—coders (who were the authors themselves) were not blind to the applicant gender and were coding for gender differences. Additionally, Trix and Psenka (2003) did not use inferential statistics, nor did they control for the fact that letters of recommendations were nested within applicants. Given these potential limitations, it is important to establish whether doubt raisers really do appear more in letters of recommendation written for women than men.

Study 1

Overview and Hypothesis

To examine gender differences in how men and women are described in letters of recommendation, we analyzed letters of recommendation written for applicants for faculty positions in a psychology department at a university that is classified as having a very high research activity level (Carnegie Classification of Institutions of Higher Education, n.d.). Because academic positions, particularly at elite research institutions, tend to be more male gendered, and because gender schemas portray men as more agentic, task-oriented, and instrumental than women (Burgess & Borgida, 1999; Rudman & Glick, 2001; Valian, 1998), we expected that men would be described more positively in letters of recommendation than would women, even after controlling for ten indicators of academic achievement (e.g., number of publications). Based on the studies by Trix and Psenka (2003), Schmader et al. (2007), and Madera et al. (2009), we specifically examined gender differences in doubt raisers.

Hypothesis 1 Letters of recommendation written for women are more likely to include doubt raisers than are letters of recommendation written for men.

Method

Sample

We examined letters of recommendation for psychology junior-faculty job applicants (collected and reported by Madera et al., 2009) and analyzed letter content that has not been reported previously (see Appendix 1 for data transparency). The sample consisted of 624 letters of recommendations for 174 applicants applying for eight assistant-level faculty positions at a university in the southern USA. In regard to applicant and recommender sex, 49% ($n = 85$) of the applicants were female and 51% ($n = 89$) were male; 29% ($n = 179$) of the recommenders were female and 69% ($n = 430$) were male (the sex for 2% could not be identified). Applicants' ages ranged from 26 to 40 years, with a mean of 32 (SD = 3.69). The mean number of letters per applicant was 3.59.

Procedure

Three trained research coders rated the extent to which letters contained doubt raisers. Through a redaction procedure in which all information about the gender of the applicant and letter writers was removed, we kept coders blind to the purpose of the study and also to the gender of both the applicant and the letter writer. The anonymity of the applicants also was preserved by removing identifying information, such as the name of the applicants, letter writers, institutions, and research labs. The coders were provided with the definitions and examples of each of the four different types of doubt raisers.

Measures

Doubt Raisers To measure doubt raisers, the coders used a 9-point Likert-type scale anchored at 1 (not at all) and 9 (very much) on four items assessing the extent to which letters contained (a) negativity, (b) hedging, (c) faint praise, and (d) irrelevant information. The coders also recorded the frequency of doubt raisers using a free-response format by responding to the following items: (a) How many instances of negative language did the letter contain? (b) How many hedging comments did the letter contain? (c) How many times did the letter contain faint praise? (d) How many times did the letter contain irrelevant information? The eight items were standardized because they were rated on different scales. These items represent the four doubt raiser types: negativity, hedging, faint praise, and irrelevancy.

Following the recommendations from LeBreton and Senter (2007), we used a two-way mixed-effects intraclass correlation ($\text{ICC}_{A,1}$) and the group mean intraclass correlation ($\text{ICC}_{A,K}$) to measure coder agreement and coder consistency. The results showed sufficient individual coder reliability,

$ICC_{A,1} = 0.86$, and group mean reliability $ICC_{A,K} = 0.94$. On the basis of these indexes, ratings were combined by averaging within and then across the coders. The alpha coefficient for the measure was 0.79. A principal components factor analysis revealed one meaningful factor that accounted for 71% of the variance. All four items representing negativity, hedging, faint praise, and irrelevant information were retained.

Gender Gender for both applicants and recommenders was coded separately female (1) or male (2).

Control Variables We used ten control variables to assess applicant performance on the basis of curriculum vita (CV) information. These were the number of first-author publications, the number of honors, the number of post-doc years, the number of courses taught, the ranking of the applicants' school, the highest journal impact factor by the applicant, the number of total publications, the position applied for, number of years in graduate school, and the length of the letters measured as the number of words in each letter. The number of first-author publications, the number of honors, the number of post-doc years, the number of courses taught, the ranking of the applicants' school, the highest journal impact factor by the applicant, and the number of total publications are direct indicators of productivity. We also controlled for the position applied for because applicants from certain backgrounds, such as industrial/organizational psychology, might have more publications; those with cognitive backgrounds might have more post-doc years. The other two control variables are not necessarily objective measures of productivity, but they might influence perceptions of productivity. For example, years in graduate school was controlled for because letter readers might adjust their estimation of productivity by taking into account number of years (i.e., divide productivity by number of years). For example, 3 publications in 5 years would be equivalent to 4.2 in 7 years. Lastly, we controlled for the letter length because past research suggests that longer letters are seen as more positive when assessing applicants in general (Liu et al., 2009; Trix & Psenka, 2003), even if they do not necessarily reflect an applicant's productivity. In addition, longer letters might provide more opportunity for doubt raisers.

Results

Descriptive statistics and intercorrelations for all of the variables are reported in Table 1. Table 2 shows the descriptive statistics for the variables by the gender of the applicants. For exploratory purposes, we conducted a multivariate analysis of variance (MANOVA) with the objective measures of applicant performance from their CVs (i.e., control variables as the dependent variables and applicant gender as the independent variable to examine if male and female applicants differed in the measures of applicant performance). The omnibus MANOVA result was not significant for gender, Wilk's $\Lambda = 0.86$, $F(10, 50) = 0.81$, $p > 0.05$, $\eta_p^2 = 0.12$, suggesting no differences by gender emerged among the control variables. Because doubt raisers are aggregated data, nested within applicants, they were not included in this initial test.

Since letters of recommendations were nested within applicants, we used the HLM6 program (Raudenbush, Bryk, Cheong, & Congdon, 2004) to analyze the data. We used full maximum likelihood estimation procedures and included random effects. For the analyses, the intercepts of the level 1 variables (doubt raisers) were predicted by the level 2 variable (gender of the applicant). That is, we predicted the content of the letters of recommendation (level 1 variables, which were nested within applicants) by the gender of the applicant (level 2 variable). For exploratory purposes, we also included the gender of the letter writer and the interaction of the gender of applicant and letter writer in the analyses (level 2 variables). Before testing the hypotheses, we investigated whether systematic within- and between-applicant variance existed in the hypothesized dependent variable (i.e., doubt raisers). The results of the unconditional (null) models indicated that there was significant between-applicant variance in the dependent variable; 14% of doubt raiser variance was accounted for by differences between applicants. Thus, there is substantial between and within variance that warrants the use of HLM to examine level 1 and level 2 variables.

Test of Hypothesis

We first tested the standardized measures of doubt raisers as a whole. As shown in Table 3, applicant gender significantly predicted doubt raisers (estimate = − 0.11, $p < 0.05$). Letters for women contained significantly more doubt raisers ($M = 0.12$, SD = 0.69) than letters for men ($M = -0.05$, SD = 0.51). Using the frequency items of the doubt raiser measure (i.e., the raw sum of the times the letter had negativity, hedges, faint praises, and irrelevant information), the letters for female applicants had an average of 0.69 (SD = 0.96) doubt raisers and the letters for male applicants had an average of 0.55 (SD = 0.71) doubt raisers. Across gender, 52% of the letters had at least one doubt raiser in the letter, 10% had at two or more doubt raisers, and 48% of the letters had no doubt raisers (ranging from 0 to 4.5 doubt raisers). For female applicants, 54% had at least one, 13% of the letters had two or more, and 46% had no doubt raisers. For male applicants, in contrast, 51% had at least one, 7% had two or more, and 49% had no doubt raisers. Neither the main effect of the letter writer gender nor the interaction between the applicant and writer gender was significant.

When broken down by type of doubt raiser, across gender, 12% of the letters had at least one negativity, 18% had a hedging, 27% had a faint praise, and 14% had an irrelevancy. For female applicants, 14% had at least one negativity, 20%

Table 1 Means, standard deviations, and correlations for level 1 variables in study 1

	M	SD	1	2
1. Length of letters	722	403	–	
2. Doubt raisers	0.00	0.59	0.18*	–

Means, standard deviations, and correlations for level 2 and aggregated level 1 variables.

	M (SD)	1	2	3	4	5	6	7	8	9	10	11	12
1. Years in graduate school	4.16 (2.02)	–											
2. Total publications	4.23 (3.56)	0.06	–										
3. First author publications	1.93 (2.16)	0.01	0.75*	–									
4. Number of honors	0.91 (1.39)	− 0.01	0.06	0.06	–								
5. Post-doc years	1.09 (1.53)	0.06	0.39*	0.44*	0.18*	–							
6. Number of courses taught	5.45 (3.34)	0.12	− 0.10	− 0.13	− 0.03	− 0.02	–						
7. Applicant gender	1.51 (0.50)	0.16	0.09	0.14	− 0.11	0.15*	− 0.09	–					
8. Writer gender	1.71 (0.28)	0.16	0.10	0.11	− 0.11	− 0.01	− 0.01	0.23*	–				
9. Length of letters[a]	698 (214)	− 0.16	0.27*	0.18*	0.03	0.02	0.07	− 0.05	0.08	–			
10. Doubt raisers[a]	0.03 (0.36)	− 0.17	− 0.06	− 0.07	− 0.08	− 0.08	0.01	− 0.12	0.14	0.10	–		
11. School ranking	2.02 (1.02)	0.06	− 0.01	− 0.06	− 0.24*	− 0.04	0.07	− 0.01	0.08	− 0.13	− 0.04	–	
12. Highest impact factor	2.03 (1.39)	0.13	0.29*	0.27*	0.15	0.27*	− 0.05	− 0.03	0.18	0.13	0.05	− 0.12	–

Gender was coded as female = 1, male = 2

*$p < 0.05$

[a] Means and correlations are based on aggregated data

had a hedging, 30% had a faint praise, and 12% had an irrelevancy in their letters. For male applicants, in contrast, 10% had at least one negativity, 15% had at a hedging, 24% had a faint praise, and 16% had an irrelevancy in their letters.

We next examined the effect of applicant gender on each individual doubt raiser and using the same set of control variables (see Table 4 for a summary of the results). For three of the four types, there was a significant effect of applicant gender. Letters for women contained significantly more negativity ($M = 0.18$, SD = 1.21) than letters for men ($M = -0.06$, SD = 0.87; estimate = − 0.12, $p < 0.05$). Letters for women contained significantly more hedging ($M = 0.13$, SD = 1.09) than letters for men ($M = -0.04$, SD = 0.86; estimate = − 0.14, $p < 0.05$). Letters for women contained significantly more faint praises ($M = 0.15$, SD = 1.14) than letters for men ($M = -0.04$, SD = 0.90; estimate = − 0.15, $p < 0.05$). But there was no effect of applicant gender on irrelevant information doubt raisers (estimate = − 0.05, $p = 0.30$). Neither the main

Table 2 Descriptive statistics for level 2 variables and aggregated level 1 variables by applicant gender for study 1

	Female applicants		Male applicants		Total means		*F*	η_p^2
	Means	SD	Means	SD	Means	SD		
Number of years in graduate school	3.84	1.62	4.48	2.34	4.18	2.05	2.99	0.02
Number of total publications	3.95	3.46	4.53	4.11	4.25	3.82	0.71	0.01
Number of first author publications	1.61	1.73	1.96	2.56	1.80	2.21	0.77	0.01
Number of honors	1.18	1.78	0.72	1.14	1.07	1.53	2.66	0.02
Number of post-doc years	0.72	1.14	1.39	1.76	1.07	1.54	5.97*	0.05
Number of courses taught	5.78	3.28	5.11	3.27	5.43	3.28	1.29	0.01
School ranking	1.93	0.99	1.88	1.04	2.02	1.02	0.03	0.00
Highest impact factor	1.98	1.26	1.85	1.67	2.03	1.39	0.12	0.00
Length of letters[a]	706	211	691	218	698	214	0.15	0.00
Doubt raisers[a]	0.11	0.42	− 0.03	0.28	0.03	0.36		

The scores for doubt raisers are standardized *z*-scores

*$p < 0.05$

[a] Means are based on aggregated data

Table 3 Hierarchical linear modeling results with applicant gender, writer gender, and their interaction as predictors for study 1

	Doubt raisers estimate	*t*
Control variables		
Years in graduate school	− 0.02 (0.02)	− 1.57
Number of total publications	0.01 (0.02)	0.03
Number of first author publications	− 0.01 (0.03)	− 0.31
Number of honors	− 0.03 (0.03)	− 1.27
Number of post-doc years	0.01 (0.03)	0.34
Number of courses taught	0.01 (0.01)	0.03
Position: applied experimental	0.19 (0.34)	0.55
Position: applied psychology	− 0.10 (0.16)	− 0.63
Position: cognitive	0.14 (0.16)	0.87
Position: health	− 0.12 (0.16)	− 0.78
Position: industrial/organizational	− 0.04 (0.16)	− 0.29
Position: social	0.07 (0.31)	0.22
Position: cognitive/neuroscience developmental	0.07 (0.24)	0.31
Length of letters	0.003* (0.01)	3.08*
School ranking	0.071 (0.05)	1.39
Highest impact factor	0.008 (0.04)	0.22
Predictors		
Applicant gender	− 0.11* (0.05)	− 2.24*
Writer gender	0.01 (0.04)	0.10
Interaction	0.01 (0.04)	0.34

Gender was coded as female = 1, male = 2. Applicant position was dummy coded with cognitive/neuroscience as the reference category. Standard errors are in parentheses

*$p < 0.05$

effect of the letter writer gender nor the interaction between the applicant and writer gender was significant for each individual doubt raiser.

In addition to our quantitative analysis of the data, we provide coded examples of actual doubt raisers from the letters of recommendation to provide contextual information. Examples of doubt raisers in letters for women include the following: "She is unlikely to become a superstar, but she is very solid," "She is not the brightest, the most creative, the most independent, or original or productive, the most likely to be an outstanding teacher, or the most "anything" of her peers," "A look at [applicant's] publication record will show that she has not published a huge amount....," "Although she has a number of papers in preparation and one under review, I think it would be fair to say that her record on paper would not place her among the top echelon of candidates for first rate programs," "At first, despite truly spectacular GRE scores, she seemed quite unsure of herself," "I assume she will be a relatively good teacher of undergraduate and graduate students," and "She may not be the strongest student we've ever put out in any one aspect of academic excellence, but her profile of talents is unique."

Table 4 Descriptive statistics for doubt raisers by applicant gender for study 1

Type of doubt raiser	Female applicants		Male applicants	
	Mean	SD	Mean	SD
All doubt raisers	0.12	0.69	− 0.05	0.51
Negativity	0.18	1.21	− 0.06	0.87
Hedging	0.13	1.09	− 0.04	0.86
Faint praises	0.15	1.14	− 0.04	0.90
Irrelevant information	0.10	0.93	0.05	0.84

The scores for all measures of doubt raisers are standardized *z*-scores, and the means are adjusted for the covariates

Examples of doubt raisers in letters for men include the following: "I know that first-author publications are priceless for job applicants. Although [applicant] doesn't have any as of yet, that should not be a concern for you....," "Instead he chose to apply what he had learned to a venture that involved web-based monitoring of internal states—a great idea, but one that unfortunately coincided with the bottom falling out of the dotcoms, so [applicant] is back on the academic market, somewhat poorer but hopefully wiser," "His speaking style is fairly slow, and his ideas do not always spring forth into words without a bit of a struggle," "He has always been passionate about developing himself and improving our program. At times, this has meant that he has not followed through on lower priority projects...," "[Applicant] was dividing himself among an unusual number of projects and, although each was interesting and important, and all were inter-related, nonetheless his projects seemed stuck approximately 90% of the way to publication," and "I no longer need to make any major corrections on his manuscripts with regards to grammar and usage. And although he has an accent, I would say it is less thick than many others from a similar background." Thus, these exemplary doubt raisers show that doubt raisers are mostly related to potential research productivity or their overall ability. Specifically, 66% of the coded doubt raisers were related to research productivity; only 17% were related to teaching. This pattern was found for both men and women.

Discussion

As predicted, letters of recommendation for female applicants for faculty positions contained more doubt raisers than letters for male applicants. In regard to the type of doubt raiser, letters for women contained more negativity, hedging, and faint praises than the letters for the men. Although irrelevant information did not reach statistical significance, the directions of the means of irrelevant information were consistent with the means for negativity, hedging, and faint praises. These differences were obtained even though we controlled for objective measures of applicant performance from their CVs. Given that we included these control variables, we can conclude that the differences in doubt raisers were not due to these specific objective aspects of candidates' performance.

Study 2

Overview and Hypotheses

Study 1 showed that, as predicted, letters for women contain more doubt raisers than do letters for men, but it leaves open whether doubt raisers influence how applicants are evaluated. It is possible that letter readers are not affected by doubt raisers. To test that possibility, using a sample of university professors, study 2 examines the influence of doubt raisers on evaluations. One reason to think that doubt raisers will have an effect is that in the sea of positive comments that make up most letters of recommendation (Knouse, 1983; Ralston & Thameling, 1988), even small numbers of doubt raisers may stand out and be disadvantageous to applicants. Although doubt raisers are not necessarily directly or overtly negative, they question an applicant's aptness for a job, suggesting that the applicant may not be the strongest candidate (Trix & Psenka, 2003). We thus predicted the following:

Hypothesis 2 Applicants for academic job positions whose letters of recommendation contain (versus do not contain) doubt raisers will be evaluated more negatively by actual faculty members.

Method

Sample

The sample consisted of 305 university professors from various universities across the USA (46% men, 54% women). In regard to their discipline, 43% were from psychology and 57% were from various disciplines, such as sociology, engineering, neuroscience, and business departments. The majority of respondents were full professors (39%), followed by associate professors (25%), assistant professors (26%), and lecturers (10%). In regard to racial/ethnic identity, 83% of the participants identified themselves as White/Caucasian, 1.4% as African-American/Black, 7.6% as Asian/Asian-American, 3.4% as Hispanic, and 4.6% as other/mixed.

Procedure and Experimental Manipulations

The authors sent an email with the study link to a convenience sample of faculty members, who were also requested to forward the study to their colleagues. After consenting to participate for a study called "Letter of Recommendation," participants were presented with written instructions indicating that they were going to read a letter of recommendation for a junior faculty position at a tier 1 research institution. Participants were informed that the letter they were going to read had been redacted to remove identifying information. Embedded in the four-paragraph, one-page letter of recommendation was a doubt raiser manipulation that immediately followed the introductory paragraph (see Appendix 2 for the script of the letter).

The doubt raiser manipulation was based on the four doubt raisers from study 1 and related to research productivity as shown in study 1: negativity, faint praise, hedging, and irrelevant information. Participants in the doubt raiser condition read one of the following: (1) "*I can say with certainty that AA does not have the skills to be the best researcher you have ever seen, but she/he does have the potential to become* successful in developing an independent research program at your institution" (negativity) or (2) "*I have confidence that AA will become better than average at being* successful in developing an independent research program at your institution" (faint praise) or (3) "*I am uncertain that AA has the potential to become one of the best researchers but I believe she/he could be a solid independent researcher at your institution and be* successful in developing an independent research program at your institution" (hedging) or (4) "*Also impressively, AA is an avid skier and enjoys photography—two tasks that we share in common.* I believe she/he can be successful in developing an independent research program at your institution" (irrelevant information). The manipulations from these conditions were derived from earlier work by Trix and Psenka

(2003) and measured in study 1. Participants in the control condition read: "I believe that AA will be a solid independent researcher at your institution." See Table 5 for the manipulated statements. After reading the letter of recommendation, the participants evaluated the applicant. We also manipulated the gender of the applicant to examine whether doubt raisers are equally damaging to male and female candidates.

Measures

Teaching competence We developed a measure of professional teaching competence based on the dictionary from Trix and Psenka (2003) and Schmader et al. (2007). Participants evaluated the applicant on five items using a Likert-type scale from 1 ("I strongly disagree") to 7 ("strongly agree"). These items assessed whether the applicant (a) had teaching competence, (b) had professionalism, (c) had teaching skills, (d) had teaching potential, and (e) had mentoring skills ($\alpha = 0.85$).

Research competence We developed a measure of research competence also based on the dictionary from Trix and Psenka (2003) and Schmader et al. (2007). The participants evaluated applicants on the five items using a Likert-type scale from 1 ("I strongly disagree") to 7 ("strongly agree"). These items included (a) research skills, (b) research potential, (c) external funding potential, (d) being a top-notch researcher, and (e) excellence in research ($\alpha = 0.91$).

Table 5 Manipulation of doubt raiser and control statements in study 2

Doubt raiser
Negativity: "I can say with certainty that AA does not have the skills to be the best researcher you have ever seen, but she/he does have the potential to become..."
$N = 45$
Faint praise: "I have confidence that AA will become better than average at being..."
$N = 59$
Irrelevancy: "Also impressively, AA is an avid skier and enjoys photography—two tasks that we share in common. I believe she/he can be..."
$N = 71$
Hedging: "I am uncertain that AA has the potential to become one of the best researchers but I think he could be a solid independent researcher at your institution and be..."
$N = 49$
Control: "I believe that AA will be a solid independent researcher at your institution and be..."
$N = 57$
All five statements were followed up with "…successful in developing an independent research program at your institution."

Manipulation check Participants were asked to identify the gender of the applicant using a three-option response: male, female, I do not remember. Two participants did not correctly identify the gender, but their inclusion in the analysis did not change the results. Thirty-six (12%) respondents indicated not remembering the gender, but their inclusion also did not change the pattern of the results.

Results

Psychometric Analyses

A CFA on the teaching and research items demonstrated adequate fit: $\chi^2 = 82.39$, df = 34, $p < 0.05$; CFI = 0.97; IFI = 0.97; RMSEA = 0.074; all loadings were statistically significant and were higher than 0.5 (they varied from 0.55 to 0.91), indicating convergent validity (Hair, Black, Babin, & Anderson, 2010; Anderson & Gerbing, 1988). The AVE was 0.54 for the teaching competence measure and 0.64 for the research competence measure, both greater than the 0.50 cutoff (Bagozzi & Yi, 1988). The squared correlation between the measures ($r^2 = 0.25$) was lower than each AVE, demonstrating discriminant validity (Fornell & Larcker, 1981). This two-factor model was compared to a one-factor-model, which demonstrated poor fit and did not significantly improve the fit: $\chi^2 = 636.07$, df = 35, $p < 0.05$; CFI = 0.68; IFI = 0.68; RMSEA = 0.24 ($\Delta\chi^2 = 553.68$; Δdf = 1; $p < 0.05$).

Test of Hypothesis

Table 6 shows the descriptive statistics for study 2 dependent variables by experimental conditions. A 5 × 2 MANOVA with the teaching and research competence as the dependent variables and the doubt raisers and applicant gender as the independent variables revealed a significant main effect for doubt raiser (Wilks's $\Lambda = 0.85$, $F(8, 582) = 5.91$, $p < 0.05$), but not for applicant gender (Wilks's $\Lambda = .99$, $F(2, 291) = 1.27$, $p > 0.05$); the interaction was not significant (Wilks's $\Lambda = 0.95$, $F(8, 582) = 1.79$, $p > 0.05$).

The main effect of doubt raisers on the research competence measure was significant, $F(4, 292) = 7.39$, $p < 0.01$, $\eta_p^2 = 0.09$. Tukey HSD and Scheffe's post hoc tests showed that the applicants with the negativity and hedging doubt raisers were evaluated significantly lower than the applicants in the other conditions, whereas the control and the other doubt raiser conditions were not significantly different from each other. The main effect of doubt raiser on teaching competence was not significant, $F(4, 292) = 1.38$, $p > 0.05$, $\eta_p^2 = 0.02$. The univariate main effects of applicant gender were not significant

Table 6 Descriptive statistics dependent variables by experimental condition for study 2

Dependent variable	Applicant gender	Doubt raiser	Mean	SD	Total	SD
Teaching competence	Male	Control	5.05	0.83		
		Irrelevant	4.79	0.94		
		Faint praise	4.58	1.22		
		Negativity	4.72	0.78		
		Hedging	5.01	0.86	4.83	0.96
	Female	Control	5.16	0.87		
		Irrelevant	4.93	1.01		
		Faint praise	4.76	1.22		
		Negativity	5.26	1.13		
		Hedging	4.75	1.44	4.96	1.16
	Total	Control	5.10	0.87		
		Irrelevant	4.86	1.01		
		Faint praise	4.66	1.22		
		Negativity	5.02	1.13		
		Hedging	4.87	1.44	4.89	1.07
Research competence	Male	Control	3.81	1.22		
		Irrelevant	3.78	1.19		
		Faint praise	3.66	1.25		
		Negativity	2.56	1.16		
		Hedging	3.15	0.83	3.46	1.22
	Female	Control	3.76	1.24		
		Irrelevant	4.12	1.20		
		Faint praise	3.34	1.14		
		Negativity	3.21	1.27		
		Hedging	3.67	1.05	3.64	1.23
	Total	Control	3.78	1.23		
		Irrelevant	3.96	1.22		
		Faint praise	3.52	1.14		
		Negativity	2.89	1.27		
		Hedging	3.43	1.05	3.53	1.22

The scale was rated from 1 to 7

for either teaching competence, $F(1, 292) = 1.45$, $p > 0.05$, $\eta_p^2 = 0.01$, or research competence, $F(1, 292) = 2.33$, $p > 0.05$, $\eta_p^2 = 0.01$. Similarly, the interaction univariate effects were not significant for either teaching, $F(4, 292) = 0.85$, $p > 0.05$, $\eta_p^2 = 0.01$, or research competence, $F(4, 292) = 1.57$, $p > 0.05$, $\eta_p^2 = 0.02$.

Discussion

Using experimental methods and an academic sample, the results from study 2 show that doubt raisers in letters of recommendation do indeed influence how applicants are evaluated. The applicant whose letter contained negativity ("… does not have the skills …") was evaluated lower on research skills than the otherwise identical applicant in the other conditions. In addition, hedging ("I am uncertain …") also led to lower evaluations on the research skills. But doubt raisers did not affect the ratings of teaching skills, probably because they were specifically related to research and not teaching. That suggests that faculty evaluate applicants based on the specific content of the doubt raiser (e.g., research) without generalizing to other domains (e.g., teaching). Further, the effects of doubt raisers were equally detrimental for both female and male applicants. Even a small island of negativity in an otherwise positive letter (Liu et al., 2009; Morgan et al., 2013) stands out and reduces an applicant's standing.

General Discussion

Study 1 showed that letters of recommendation for women, compared to letters for men, contain more doubt raisers, specifically, negativity, hedges, and faint praise.

This result held despite controls for productivity, such as number of publications and teaching experience. Thus, objective gender differences in productivity do not appear to be the reason that more women than men receive doubt raisers in their letters of recommendation. Differences in doubt raisers are more likely due to gender schemas than to systematic differences in the preparedness or quality of male versus female applicants.

Study 2 showed that both negativity (i.e., a type of doubt raiser that points out weaknesses) and hedging (i.e., a forthright admission of uncertainty) in letters of recommendation lead to lower evaluations of applicants, regardless of the gender of the applicant. Taken together, the key contribution of these studies is the clear illustration that doubt raisers in letters of recommendation do indeed hurt women more than men, but only because doubt raisers are more frequent in letters for women. In other words, evaluators treat doubt raisers equally seriously whether they are provided for a woman or a man (study 2), but because doubt raisers are more often used for women than for men (study 1), women are more likely to be negatively affected by them.

The combined findings are particularly interesting because the lack of evidence of gender bias when doubt raisers are presented in letters of recommendation potentially obscures the gender bias that has occurred at an earlier point, namely, when a recommender is writing the letter. Doubt raisers are a minus for everyone, but letter writers assign that minus more often to women than to men. If search committees ignored letters of recommendation, that asymmetry would not matter. But letters of recommendation are commonly used as selection tools in academia (Nicklin & Roch, 2009; Kuncel et al., 2014). The data have important implications for women in academia, particularly because women face biases early in the selection process (Bailyn, 2003; Eveline, 2005; Pyke, 2013; Van den Brink & Benschop, 2012; Westring et al., 2012; cf. Ceci et al., 2014a, b).

The current research makes important contributions to the literature on the effects of gender schemas on workplace outcomes. Our studies reveal how gender schemas can negatively affect women through the use of doubt raisers in letters of recommendations. That is to say, the letters in our sample contained more phrases that doubt the female (versus male) applicants' ability to be successful. Letters of recommendation can be ambiguous and unstructured, which allows for biases stemming from gender schemas to play a role. For example, Heilman et al. (2004) argued that biases are more likely in situations that are ambiguous. Because instructions for what should be included in letters of recommendations are often ambiguous and open to interpretation, letter writers may depend on heuristics and stereotypes when writing letters and describing women; these biased descriptions (including doubt raisers) are negatively related to applicant evaluations, as shown in study 2.

The phenomenon that we have reported is not propagated more by male versus female letter writers (study 1) or evaluators (study 2). There were no main effects of the gender of the letter writer and letter writer gender did not interact with applicant gender to predict doubt raisers. The female letter writers (in study 1) wrote letters similarly to their male counterparts and were just as likely as men to describe female applicants with more doubt raisers than male applicants. This provides some support for the universality of gender schemas and the manner in which men and women are described. Similarly, the evaluators (in study 2) interpreted doubt raisers (negativity and hedging) and rated letters containing them more negatively than they rated letters that did not have doubt raisers. The lack of gender differences in how doubt raisers affect an applicant are consistent with the broader literature on stigma (Crocker, Major, & Steele, 1998; Hebl, Tickle, & Heatherton, 2000) and more specifically the literature on sex bias in the workplace (e.g., see Heilman et al., 2004; Heilman & Okimoto, 2007).

The results showed that the inclusion of even a single doubt raiser—particularly negativity or hedging—was enough to lead to statistically lower evaluations of the applicant (study 2). This finding is of particular interest because study 1 showed that 14 and 20% of the letters for female applicants had at least one negativity and hedging doubt raiser, respectively, compared to 10 and 15% of the letters for the male applicants. Although these gender differences, while reliable, are small, the results from study 2 showed that only one statement can make a difference for an applicant.

The results of the current studies also offer important implications for the use of letters of recommendation outside of academia. Although professionals outside of academia rely on letters of recommendation less than academics (Nicklin & Roch, 2009), there are reasons to expect that gender schemas can also influence the development of letters of recommendation outside of academia. As shown in study 1, letters written for women had more doubt raisers than letters for men, even after controlling for objective measures of research productivity. We argue that this occurs partly because of how gender schemas can influence what is expected from men and women and how they are described, particularly in occupations that have norms related to one sex. In particular, we argue that, because what is required for success in many academic departments may be based on norms of masculinity (Bailyn, 2003; Van den Brink & Benschop, 2012; Westring et al., 2012), a potential bias against female faculty can arise when developing letters of recommendation. Letter writers can have sex-related stereotypes of women that are incongruent with the attributes that are believed to be required for success in a particular job (Eagly & Karau, 2002; Heilman, 2001), such as academia. Likewise, gender schemas can also influence the development of letters of recommendation, particularly in male-dominated occupations. For example, extant research shows how gender schemas influence the evaluations and stereotypes of managers and leaders, such that management and leadership qualities are still perceived to be more masculine than feminine (Duehr & Bono, 2006; Heilman, 2012; Koenig, Eagly, Mitchell, & Ristikari, 2011).

Thus, we would expect that if occupations (e.g., accounting positions in Big 8 firms) or positions (e.g., management roles) are related to masculine schemas (e.g., agentic qualities), then letter writers for applicants might be influenced by schemas when developing these letters, despite real or perceived gender differences. Again, we want to highlight that study 1 shows gender differences in doubt raisers even after controlling for productivity. Because the male and female applicants did not differ in the number of publications, impact factor, and teaching experience, gender schemas might provide a reason for why letters for women contained more doubt raisers than letters for the men.

Organizational Implications

Our research has important implications for academic institutions and for organizations that do rely on letters of recommendation. Our findings show that the gender disparity in doubt raisers found in study 1 is related to selection decisions, as shown in study 2. One obvious implication for academic institutions and organizations is that they should adopt strategies that can help identify such biases (see also Kervyn, Bergsieker, & Fiske, 2012) and then work to reduce those biases in the selection process. For example, universities can give less weight to letters of recommendation, or they can wait to collect letters of recommendation until they have reviewed an applicant's work, or they can provide letter writers prompts so that recommenders are less likely to include doubt raisers in the letters. For instance, recent research has shown that gender biases can be reduced in letters of recommendation by requiring raters of such letters to elaborate and expand on interpretations of letters (Morgan et al., 2013). In particular, when participants were asked to read letters of recommendation and make ratings of the applicant, those who were asked to explain their ratings showed less gender bias against the applicants than those who were not asked to explain their ratings.

Another suggestion is that letters of recommendation should be structured in both their development and how they are used in the selection process. The low validity coefficients in Kuncel et al. (2014) were based on samples of letters that varied in how unstructured they were (some were structured and others were not). This relationship between structure and validity is found for interviews, particularly with structured interviews having greater validity than unstructured interviews. Thus, academic institutions and organizations can reduce gender bias in letters by being aware of the potential biases in letters of recommendation through formal organizational policies or diversity training (Hebl, Madera, & King, 2007), taking direct steps to deactivate the impact of these biases (Morgan et al., 2013), and adding structure to their development and use in their evaluations.

Limitations and Future Research

Although we used actual archival data and not hypothetical letters of recommendation in study 1, a potential limitation is that a variable that we did not include in our analyses caused some systematic differences in the extent of doubt raisers that were based on real gender differences. Since we controlled for number of years in graduate school, the number of total publications, the number of first author publications, the number of honors, the number of post-doc years, the position applied for, and the number of courses taught, however, we doubt the existence of other major differences. Furthermore, other research has shown that such differences still exist (Morgan et al., 2013), even when the quality of candidates is controlled (see Madera et al., 2009).

One fruitful area of future research is how the content of doubt raisers influences evaluations of applicants. In the current research, we manipulated different types of doubt raisers that were related to research but not to teaching (study 2). The doubt raisers did not affect the ratings of teaching skills, which used an academic sample, suggesting that faculty evaluate applicants based on the content of the doubt raiser (e.g., research) without generalizing to other domains (e.g., teaching). Future research might investigate, via standardized manipulations, how doubt raiser content potentially influences letters of recommendation for and appraisal of candidates.

In addition, the current studies did not examine the race of the applicants (in both study 1 and 2) nor of the letter writers in study 1. This is an area for future research to explore. In fact, qualitative research suggests that racial minority faculty face subtle forms of discrimination in academia (e.g., Kelly & McCann, 2014; Perry, Moore, Edwards, Acosta, & Frey, 2009; Peterson, Friedman, Ash, Franco, & Carr, 2004; Stanley, 2006); this body of literature has examined how discrimination manifests when one is already employed in academia. Very little research has examined how an academic racial minority applicant faces discrimination in the selection process or how this interacts with gender, particularly in letters of recommendation.

These data are from a single field, namely psychology. Specifically, the letters for study 1 were for eight assistant-level positions, but for one department (psychology) at one university. However, our results from study 1 are consistent with similar research that examined biases in letters of recommendation from non-psychology samples. In particular, past research using samples from the STEM fields has found similar gender effects in letters of recommendation (Isaac et al., 2011; Schmader et al., 2007; Trix & Psenka, 2003). In addition, the sample for study 2 included professors from various disciplines; only 43.3% were from psychology. These professors also worked in a variety of institutions, including four-year teaching schools. Therefore, the results from study 1 (i.e., how letters for women have more doubt raisers than letters for men) and study 2 (i.e., how doubt raisers influence applicant evaluations) can generalize to other academic fields and types of institutions. However, we do encourage future research to examine if these effects hold in other fields. Of particular importance are the STEM fields in which women are underrepresented in academia (U.S. Department of Commerce, 2011) and for occupations or positions outside of academia that are related to masculine schemas.

Relatedly, the current studies focused only on letters of recommendation. Another area for future research is to examine if other methods used early in the selection process (e.g., reference check phone calls) can also be biased by gender schemas, leading to gender differences in doubt raisers. For example, researchers have argued that many reference check phone calls are unstructured and therefore susceptible to biases (e.g., Hedricks, Robie, & Oswald, 2013; Taylor, Pajo, Cheung, & Stringfield, 2004). The unstructured nature of reference checks is an important feature in light of research that suggests that bias against women is less prevalent when structure reduces the uncertainty of whether a female applicant will be successful in a masculine-gendered work environment, role, or position (Heilman et al., 2004; Heilman, 2012).

Conclusion

The implications of the current research on letters of recommendations are particularly important because their use for academia is well established (Johnson et al., 1998; Landrum et al., 1994; Sheehan et al., 1998). Our studies show how bias in the letter-writing process can be propagated, even if evaluators do not necessarily display overt gender biases. The differences in word choice may seem negligible, but in fact, as our data show, doubt raisers have discernible penalties for women in academia (Eagly & Karau 2002; Eagly & Johannesen-Schmidt, 2001; Wood & Eagly, 2000). Awareness of and attention to these differences are critical areas of future research and application if we want to maximize fairness in occupations, such as academia, that rely on letters of recommendation.

Funding Information This paper was funded by an NIH Grant (1R01GM088530).

Appendix 1. Data Transparency Appendix

The data reported in this submitted manuscript (study 1 data only) have been previously published. Findings from the data collection have been reported in separate manuscripts. MS 1 (published) focuses on communal and agentic descriptions of applicants in letters of recommendations for academic positions as the dependent variables. MS 2 (the current submitted manuscript) focuses on doubt raiser descriptions of applicants in letters of recommendations as the dependent variables. The table below displays where each data variable appears in each study, as well as the current status of each study.

Variables in the complete dataset	MS 1 (status = pub)	MS 2 (status = current)
Communal adjectives	x	
Social-communal orientation	x	
Agentic adjectives	x	
Agentic orientation	x	
Doubt raisers: negatives		x
Doubt raisers: hedges		x
Doubt raisers: faint praises		x
Doubt raisers: irrelevancies		x
Applicant gender	x	x
Letter writer gender	x	x

Appendix 2. Letter Exemplar

Dear Search Committee,

It is with enthusiasm that I recommend AA for a tenure track faculty position (Assistant Professor) within the <DEPT> at WR99. I was AA's doctoral research advisor at WRNR and I know AA both professionally and personally. As a graduate student, AA also served as my teaching assistant for two undergraduate laboratory classes. AA was an impressive student who I have had the pleasure to work with at WRNR.

<MANIPULATION HERE> being successful in developing an independent research program at your institution. I have seen AA mature into a more careful scientist who demonstrates competence, leadership skills, and curiosity. I have kept in close contact with AA during <his/her> post doctoral training and know that <he/she> has matured scientifically and has expanded <his/her> knowledge base into other closely-related fields. AA has aptitude to continue developing in the field. In terms of research, AA has published two manuscripts based on <his/her> thesis work in my lab, and a third manuscript is pending submission. I know that AA detailed this work in <his/her> research statement so I will only state here that it is published in a solid journal and is theoretically strong and methodologically sound.

AA projects professionalism, whether it is in the lecture room and undergraduate laboratory, the research laboratory, or at conferences. AA is hardworking and also willing to take time to teach others. AA became a leader in my research lab, taking time to mentor undergraduate students and less senior

PhD students. AA has given a series of tutorial lectures on statistics in Psychology to the PhD students at WR99. AA is very willing to help others and I believe <he/she> demonstrates natural teaching abilities plus <he/she> greatly enjoys it. Both AA's skills and his vision are broad and fine-tuned.

In conclusion, I have come to regard AA with respect over the past several years. I hope you interview <him/her>. If you have any further questions about AA, please do not hesitate to phone me at [number removed].

Sincerely,
ZZ, PhD
Associate Professor of <DEP>

References

Aamodt, M. G., Nagy, M. S., & Thompson, N. (1998). *Employment references: Who are we talking about?*, Paper presented at the International Personnel Management Association Assessment Council, Chicago, IL.

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: Do metrics matter? *Nature News, 465*(7300), 860–862.

Adamo, S. A. (2013). Attrition of women in the biological sciences: Workload, motherhood, and other explanations revisited. *Bioscience, 63*(1), 43–48.

Aguirre Jr, A. (2000). *Women and minority faculty in the academic workplace: Recruitment, retention, and academic culture*. ASHE-ERIC Higher Education Report, Volume 27, Number 6. Jossey-Bass Higher and Adult Education Series. Jossey-Bass, 350 Sansome St., San Francisco, CA 94104–1342.

Aiston, S. J. (2014). Leading the academy or being led? Hong Kong women academics. *Higher Education Research & Development, 33*(1), 59–72. https://doi.org/10.1080/07294360.2013.864618.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411.

APPIC (2005). *Members survey: APPIC predoctoral internship members*. http://www.APPIC.Org

Applegate, B. K., Cable, C. R., & Sitren, A. H. (2009). Academia's most wanted: The characteristics of desirable academic job candidates in criminology and criminal justice. *Journal of Criminal Justice Education, 20*(1), 20–39.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*(1), 74–94. https://doi.org/10.1007/BF02723327.

Bailyn, L. (2003). Academic careers and gender equity: Lessons learned from MIT1. *Gender, Work & Organization, 10*(2), 137–153. https://doi.org/10.1111/1468-0432.00008.

Benson, T. A., & Buskist, W. (2005). Understanding "excellence in teaching" as assessed by psychology faculty search committees. *Teaching of Psychology, 32*(1), 47–49.

Broughton, W., & Conlogue, W. (2001). What search committees want. *Profession*, 39–51.

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, Public Policy, and Law, 5*, 665–692. https://doi.org/10.1037/1076-8971.5.3.665.

Carnegie Classification of Institutions of Higher Education (n.d.). *About Carnegie classification*. Retrieved from http://carnegieclassifications.iu.edu/.

Ceci, S. J., & Williams, W. M. (2015). Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men. *Frontiers in Psychology, 6*, 1532.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014a). Women in academic science: A changing landscape. *Psychological Science in the Public Interest, 15*(3), 75–141.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014b). Women in academic science: A changing landscape. *Psychological Science in the Public Interest, 15*(3), 75–141.

Cejka, M. A., & Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin, 25*(4), 413–423. https://doi.org/10.1177/0146167299025004002.

Crocker, J., Major, B., & Steele, C. (1998). Social stigma. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (Vol. 2, 4th ed., pp. 504–553). New York: McGraw-Hill.

Crockett, W. H. (1988). Schemas, affect, and communication. In L. Donohew, H. Sypher, & E. Higgins (Eds.), *Communication, social cognition, and affect*. Lawrence Erlbaum Association: Hillsdale, NJ.

Deo, M. E. (2014). Looking forward to diversity in legal academia. *Berkeley Journal of Gender, Law & Justice, 29*(2), 352.

Ding, W. W., Murray, F., & Stuart, T. E. (2013). From bench to board: Gender differences in university scientists' participation in corporate scientific advisory boards. *Academy of Management Journal, 56*(5), 1443–1464. https://doi.org/10.5465/amj.2011.0020.

Duehr, E. E., & Bono, J. E. (2006). Men, women, and managers: Are stereotypes finally changing? *Personnel Psychology, 59*(4), 815–846.

Dutt, K., Pfaff, D. L., Bernstein, A. F., Dillard, J. S., & Block, C. J. (2016). Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience, 9*(11), 805–808.

Eagly, A. H., & Johannesen-Schmidt, M. C. (2001). The leadership styles of women and men. *Journal of Social Issues, 57*, 781–797. https://doi.org/10.1111/0022-4537.00241.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*, 573–598. https://doi.org/10.1037/0033-295X.109.3.573.

Easterly, D. M., & Ricard, C. S. (2011). Conscious efforts to end unconscious bias: Why women leave academic research. *Journal of Research Administration, 42*(1), 61–73.

Ellemers, N., van den Heuvel, H., de Gilder, D., Maas, A., & Bovini, A. (2004). The underrepresentation of women in science: Differential commitment or the queen bee syndrome? *British Journal of Social Psychology, 43*, 1–24. https://doi.org/10.1348/0144666042037999.

Eveline, J. (2005). Woman in the ivory tower: Gendering feminised and masculinised identities. *Journal of Organizational Change Management, 18*(6), 641–658. https://doi.org/10.1108/09534810510628558.

Fiske, S. T., & Linville, P. W. (1980). What does the schema concept buy us? *Personality and Social Psychology Bulletin, 6*, 543–557. https://doi.org/10.1177/014616728064006.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312.

Fuerstman, D., & Lavertu, S. (2005). The academic hiring process: A survey of department chairs. *PS: Political Science & Politics, 38*(4), 731–736.

Gatewood, R., & Feild, H. (2001). *Human resource selection: Application forms, training and experience evaluations, and reference checks* (5th ed.). Mason, OH: Roche, M.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality.

Journal of Personality and Social Psychology, 101(1), 109–128. https://doi.org/10.1037/a0022530.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective*. New York, NY: Pearson.

Hebl, M. R., Madera, J. M., & King, E. B. (2007). Exclusion, avoidance, and social distancing. In K. M. Thomas (Ed.), *Diversity resistance: Manifestation and solutions* (pp. 127–150). Mahwah, NJ: Lawrence Erlbaum Associates.

Hebl, M. R., Tickle, J., & Heatherton, T. F. (2000). Awkward moments in interactions between nonstigmatized and stigmatized individuals. In T. Heatherton, R. Kleck, M. Hebl, & J. Hull's (Eds.), *The social psychology of stigma*. New York, NY: Guilford Press.

Hedricks, C. A., Robie, C., & Oswald, F. L. (2013). Web-based multisource reference checking: An investigation of psychometric integrity and applied benefits. *International Journal of Selection and Assessment, 21*(1), 99–110.

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. *Research in Organizational Behavior, 5*, 269–298.

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57*, 657–674. https://doi.org/10.1111/0022-4537.00234.

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior, 32*, 113–135. https://doi.org/10.1016/j.riob.2012.11.003.

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks? The implied communality deficit. *Journal of Applied Psychology, 92*, 81–92. https://doi.org/10.1037/0021-9010.92.1.81.

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male tasks. *Journal of Applied Psychology, 89*, 416–427. https://doi.org/10.1037/0021-9010.89.3.416.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1–2), 152–194.

Howe-Walsh, L., & Turnbull, S. (2016). Barriers to women leaders in academia: Tales from science and technology. *Studies in Higher Education, 41*(3), 415–428.

Isaac, C., Chertoff, J., Lee, B., & Carnes, M. (2011). Do students' and authors' genders affect evaluations? A linguistic analysis of medical student performance evaluations. *Academic Medicine, 86*(1), 59–66. https://doi.org/10.1097/ACM.0b013e318200561d.

Johnson, M., Elam, C., Edwards, J., Tayor, D., Heldberg, C., Hinkley, R., & Comeau, R. (1998). Medical school admission committee members' evaluations of and impressions from recommendation letters. *Academic Medicine, 73*, S41–S43. https://doi.org/10.1097/00001888-199810000-00040.

Kaminski, D., & Geisler, C. (2012). Survival analysis of faculty retention in science and engineering by gender. *Science, 335*(6070), 864–866.

Kelly, B. T., & McCann, K. I. (2014). Women faculty of color: Stories behind the statistics. *The Urban Review, 46*(4), 681–702.

Kervyn, N., Bergsieker, H. B., & Fiske, S. T. (2012). The innuendo effect: Hearing the positive but inferring the negative. *Journal of Experimental Social Psychology, 48*(1), 77–85. https://doi.org/10.1016/j.jesp.2011.08.001.

Knouse, S. B. (1983). The letter of recommendation: Specificity and favorability of information. *Personnel Psychology, 36*, 331–341. https://doi.org/10.1111/j.1744-6570.1983.tb01441.x.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin, 137*, 616–642.

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment, 22*, 101–107. https://doi.org/10.1111/ijsa.12060.

LaCroix, P. P. (1985). Sex in recs: gender bias in recommendation writing. Journal of College Admission, 109, 24–26.

Landrum, R. E., & Clump, M. A. (2004). Departmental search committees and the evaluation of faculty applicants. *Teaching of Psychology, 31*(1), 12–17.

Landrum, R., Jeglum, E., & Cashin, J. (1994). The decision-making process of graduate admissions committees in psychology. *Journal of Social Behavior and Personality, 9*, 239–248.

LeBreton, J. M., & Senter, J. L. (2007). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642.

Lee, Y. J., & Won, D. (2014). Trailblazing women in academia: Representation of women in senior faculty and the gender gap in junior faculty's salaries in higher educational institutions. *The Social Science Journal, 51*(3), 331–340.

Lerback, J., & Hanson, B. (2017). Journals invite too few women to referee. *Nature, 541*(7638), 455–457.

Levine, R. B., Lin, F., Kern, D. E., Wright, S. M., & Carrese, J. (2011). Stories from early-career women physicians who have left academic medicine: A qualitative study at a single institution. *Academic Medicine, 86*(6), 752–758.

Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2009). Using the standardized letters of recommendation in selection: Results from a multidimensional Rasch model. *Educational and Psychological Measurement, 69*, 475–492. https://doi.org/10.1177/0013164408322031.

Maass, A., & Arcuri, L. (1996). Language and stereotyping. In C. N. Macrae, C. Stangor, & M. Hewstone (Eds.), *Stereotypes and stereotyping* (pp. 193–226). New York, NY: Guilford Press.

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology, 94*(6), 1591–1599. https://doi.org/10.1037/a0016539.

McCarthy, J. M., & Goffin, R. D. (2001). Improving the validity of letters of recommendation: An investigation of three standardized reference forms. *Military Psychology, 13*, 199–222. https://doi.org/10.1207/S15327876MP1304_2.

Meizlish, D., & Kaplan, M. (2008). Valuing and evaluating teaching in academic hiring: A multidisciplinary, cross-institutional study. *The Journal of Higher Education, 79*(5), 489–512.

Mittenberg, W., Peterson, R. S., Cooper, J. T., Strauman, S., & Essig, S. M. (2000). Selection criteria for clinical neuropsychology internships. *The Clinical Neuropsychologist, 14*, 1–6.

Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation. *Journal of Applied Social Psychology, 43*(11), 2297–2306. https://doi.org/10.1111/jasp.12179.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering*. Washington, DC: The National Academies Press.

National Research Council (NRC). (2009). *Gender differences at critical transitions in the careers of science, engineering and mathematics faculty*. Washington, DC: National Academy Press.

National Science Foundation, Division of Science Resources Statistics (2004). *Gender differences in the careers of academic scientists* *and engineers, NSF 04-323, Project Officer, Alan I. Rapoport* (Arlington, VA).

Nicklin, M. J., & Roch, S. G. (2009). Letters of recommendation: Controversy and consensus from expert perspectives. *International Journal of Selection and Assessment, 17*, 76–91. https://doi.org/10.1111/j.1468-2389.2009.00453.x.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC 2001): A computerized text analysis program*. Mahwah, NJ: Erlbaum.

Perry, G., Moore, H., Edwards, C., Acosta, K., & Frey, C. (2009). Maintaining credibility and authority as an instructor of color in diversity-education classrooms: A qualitative inquiry. *The Journal of Higher Education, 80*(1), 230–244.

Peterson, N. B., Friedman, R. H., Ash, A. S., Franco, S., & Carr, P. L. (2004). Faculty self-reported experience with racial and ethnic discrimination in academic medicine. *Journal of General Internal Medicine, 19*(3), 259–265.

Pyke, J. (2013). Women, choice and promotion or why women are still a minority in the professoriate. *Journal of Higher Education Policy and Management, 35*(4), 444–454. https://doi.org/10.1080/1360080X.2013.812179.

Ragins, B. R., & Sundstrom, E. (1989). Gender and power in organizations. *Psychological Bulletin, 105*, 51–88. https://doi.org/10.1037/0033-2909.105.1.51.

Ragins, B. R., Townsend, B., & Mattis, M. (1998). Gender gap in the executive suite: CEOs and female executives report on breaking the glass ceiling. *Academy of Management Executive, 12*, 28–42 http://www.jstor.org/stable/4165439.

Ralston, S. M., & Thameling, C. A. (1988). Effects of vividness of language on information value of reference letters and job applicants' recommendation. *Psychological Reports, 62*, 867–870. https://doi.org/10.2466/pr0.1988.62.3.867.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical and nonlinear modeling [computer software]*. Lincolnwood, IL: Scientific Software International.

Rubini, M., & Menegatti, M. (2014). Hindering women's careers in academia gender linguistic bias in personnel selection. *Journal of Language and Social Psychology*, 0261927X14542436.

Rudman, L. A., & Glick, P. (2001). Perspective gender stereotypes and backlash toward agentic women. *Journal of Social Issues, 57*, 743–762. https://doi.org/10.1111/0022-4537.00239.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles, 57*(7–8), 509–514. https://doi.org/10.1007/s11199-007-9291-4.

Settles, I. H., Cortina, L. M., Malley, J., & Stewart, A. J. (2006). The climate for women in academic science: The good, the bad, and the changeable. *Psychology of Women Quarterly, 30*(1), 47–58.

Sheehan, E. P., McDevitt, T. M., & Ross, H. C. (1998). Looking for a job as a psychology professor? Factors affecting applicant success. *Teaching of Psychology, 25*, 8–11. https://doi.org/10.1207/s15328023top2501_3.

Shen, H. (2013). Mind the gender gap. *Nature, 495*(7439), 22–24.

Stanley, C. A. (2006). Coloring the academic landscape: Faculty of color breaking the silence in predominantly White colleges and universities. *American Educational Research Journal, 43*(4), 701–736.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135*(6), 859–884. https://doi.org/10.1037/a0017364.

Taylor, D. (2007). Employment preferences and salary expectations of students in science and engineering. *Bioscience, 57*, 175–185. https://doi.org/10.1641/B570212.

Taylor, P. J., Pajo, K., Cheung, G. W., & Stringfield, P. (2004). Dimensionality and validity of a structured telephone reference check procedure. *Personnel Psychology, 57*(3), 745–772.

Treviño, L. J., Gomez-Mejia, L. R., Balkin, D. B., & Mixon, F. G. (2015). Meritocracies or masculinities? The differential allocation of named professorships by gender in the academy. *Journal of Management., 44*, 972–1000. https://doi.org/10.1177/0149206315599216.

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse and Society, 14*, 191–220. https://doi.org/10.1177/0957926503014002277.

U.S. Department of Commerce (2011). Women in STEM: A gender gap to innovation. Executive summary. *Economics and Statistics Administration. ESA Issue Brief #04-11*. August Retrieved on 1/10/2015 at url: http://www.esa.doc.gov/sites/default/files/reports/documents/womeninstemagaptoinnovation8311.pdf.

U.S. Department of Education, National Center for Education Statistics (2015). *The condition of education 2016 (NCES 2016-144), characteristics of postsecondary faculty*. Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=61

Valian, V. (1998). *Why so slow? The advancement of women*. Cambridge: M.I.T. Press.

Van den Brink, M., & Benschop, Y. (2012). Slaying the seven-headed dragon: The quest for gender change in academia. *Gender, Work & Organization, 19*(1), 71–92. https://doi.org/10.1111/j.1468-0432.2011.00566.x.

Westring, A. F., Speck, M. R. M., Sammel, M. D., Scott, M. P., Tuton, L. W., Grisso, J. A., & Abbuhl, S. (2012). A culture conducive to women's academic success: Development of a measure. *Academic Medicine: Journal of the Association of American Medical Colleges, 87*(11), 1622–1631. https://doi.org/10.1097/ACM.0b013e31826dbfd1.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences, 112*(17), 5360–5365.

Wood, W., & Eagly, A. H. (2000). Once again, the origins of sex differences. *American Psychologist, 55*(9), 1062–1063. https://doi.org/10.1037/0003-066X.55.9.1062.

Yost, E., Winstead, V., Cotten, S. R., & Handley, D. M. (2013). The recruitment and retention of emerging women scholars in stem: Results from a national web-based survey of graduate students, postdoctoral fellows, and junior faculty. *Journal of Women and Minorities in Science and Engineering, 19*(2), 143–163. https://doi.org/10.1615/JWomenMinorScienEng.2013003021.