Impact of Gendered Topics in Letters of Recommendation on Perceived Importance for Making a Hiring Decision in Geosciences

Joshua Elmore

Submitted in partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy
Under the Executive Committee
Of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Impact of Gendered Topics in Letters of Recommendation on

Perceived Importance for Making a Hiring Decision in Geosciences

Joshua Elmore

Scientific fields are perceived of as more masculine than feminine and stereotypes of scientists are more closely associated with stereotypes of men than of women (Carli et al., 2016). Supporting this point, Elmore, Block, Bowers and Dutt (2019) uncovered 31 topics from letters of recommendation for post-doctoral applicants in the geosciences and found that these topics described male and female applicants differently and, in a manner consistent with gender stereotypes. As the number of women in the geosciences declines further up the academic ladder (National Science Foundation, 2017) and as letters of recommendation play an important role in academic hiring (Abbott et al. 2010), it is important to understand if the use of gendered topics in these letters may reduce the likelihood of female advancement. Thus, in the present study we gathered questionnaire data from 250 geoscience researchers and scientists asking them to rate and rank the topics uncovered by Elmore et al. (2019) in terms of their importance when making hiring decisions.

Results showed geoscientists valued research productivity and publishing over being a teacher, student, or department citizen. Topics expressed more in letters for male applicants in Elmore et al. (2019) were *rated* as significantly more important when making a hiring decision than topics expressed more in letters for female applicants. Further, the male topics identified in Elmore et al. (2019) were *ranked* more often as the most important topics and less often as least important topics when making a hiring decision compared to the female topics. Finally, 68% of

participants indicated they attribute 50% or more of their hiring decision to information found in letters of recommendation, underscoring the importance of letters of recommendation to career advancement in the geosciences.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ACKNOWLEDGEMENTS

I am humbled by the sheer number of people who have provided me support, encouragement, inspiration, guidance, mentorship, friendship, and partnership as I have navigated my academic journey and without whom I would not have earned this rare degree. First and foremost, I am grateful to my dissertation committee members, Caryn Block, Debra Noumair, Bryan Keller, Elissa Perry, and Kuheli Dutt. Your support, wisdom, and challenging feedback has made this work possible. Thank you to Bryan Keller for being an invaluable thought partner while I navigated the complex statistical analysis required of this project. Thank you to Debra Noumair for ensuring that I kept my eye on the big picture while down in the weeds. Thank you to Elissa Perry for sparking the need for this study in the first place; I distinctly recall sitting in your office describing the results of our topic model and your first reply being a challenge for a second study when you asked "do these topics matter to decision makers?" Thank you to Kuheli Dutt, for being our thought partner and resident geoscience expert; your perspective was invaluable to the design of the present study and your willingness to share letter data made this project possible. I would also like to thank Alex Bowers, for introducing us to topic modeling and guiding the early work that informed this dissertation. I would like to thank the American Geophysical Union for allowing us to recruit participants on their community forum. Finally, I would like to thank each and every person who took time to complete the questionnaire for this study; I could not have done it without you!

I am eternally grateful to my sponsor and advisor, Caryn Block. Caryn, since day one you have been there for me, as a mentor, guide, colleague, and friend. Your approach to research and workgroup has served as an example for how to do work with integrity and how to effectively collaborate. Your expectation for high quality research has continuously pushed me to do better

and your advocacy for myself and my fellow PhD students has ensured that our hard work leads to success in the program. Your lessons will last me a lifetime.

I am also grateful to the faculty of the Social-Organizational Psychology program. I am particularly grateful for your thoughtful design of this program. The degree requirements up to this point prepared me for taking on this dissertation and ensured that I would thrive under the challenge. Specifically, completing seminar proposals and qualifying papers provided highly relevant experiences that served me well throughout this process. To that end, I would like to thank Jim Westaby, Loriann Roberson, Caryn Block, Bill Pasmore, Elissa Perry, and Debra Noumair for your feedback on my writing and practical and theoretical reasoning. Thank you also to Warner Burke, Peter Coleman, Madhabi Chatterji, Sarah Brazaitiis, Gina Buontempo, Rebecca Stilwell, Yaron Prywes, Allan Church, and Marina Field for offering valuable schools of thought, upon which I will draw throughout my life.

I would like to thank the staff of the program, John Handal, Ambar Urena, and Lebab Fallin; without your support and guidance, I would probably be a year behind. Thank you so much for sharing your expertise and being willing to stop what you are doing to help a PhD student in need! Thank you also to Myra Luna-Lucero at the Teachers College Institutional Review Board for providing guide to performing research at TC and for making yourself available for my many questions. I am very thankful for my cohort, Allegra Chen-Carrell, Aimee Lace, Adam Parr, and Diego Ramos; I couldn't have picked four more wonderful people with whom to share this journey. I am also thankful for our wider PhD community for always being a source of support, guidance, inspiration, and friendship. I am grateful to the PhD students my senior who were always there for me and provided an example upon which I have drawn when interacting with PhD students my junior.

I would like to thank Caryn Block and Elissa Perry for welcoming me into their research workgroups; I owe much of my ability to complete this project to the work that was done each week in your offices. To my workgroup members and research associates, Dave Mendelsohn, Aitong Li, Jean Sohn, DaHee Shon, Julian McNeil, Lea Lynn Yen, Charlie Tantivivathanaphand, Camille Zubizarreta, and Mariana Garcia; I am a better researcher because of each of you.

There are several people and entities without whom I would not be here today. I am forever am grateful to C. Malik Boykin for introducing me to academic culture broadly and to the field of organizational psychology specifically; your openness, mentorship, and deep knowledge of academia has helped me successfully navigate academic life. To the Jack Kent Cooke Foundation for your enduring financial and emotional support for the last 6 years; my life will forever be better because of your willingness to believe in me and I will spend the rest of my life paying it forward. To my fellow scholars at the JKCF; I am so thankful to have you as a community. To Teachers College, Columbia University for offering a program that so uniquely fit my interests and for providing financial support, which enabled my attendance.

To my family, for always being there when I needed them. To my parents; you provided me a solid foundation, picked me up when I failed, and never stopped believing in me; you gave me the confidence to achieve hard things. To my brother; together we learned to face challenges, solve problems, and construct and deconstruct anything and everything; thank you for being my thought partner in life. To Natacha's family, for sharing with me their passion for art, food, wine, and science; your world views have provided a compass on my journey. And to my extended families, the Elmores, Janssens, Fischers, and Palmers for always expressing interest in my work for offering words of encouragement that sustained me during challenging times.

To my friends in California who started this journey with me and have ensured that life stays goofy. To my friends in New York who share the struggle of making it in the big city and always being up for an adventure.

To Natacha; you have stuck with me throughout this 8-year odyssey. I know it has not always been easy; we have endured long days, missed weekends, and stressful semester after stressful semester. I could not have made it without you, and I am so grateful for your patience, resolve, understanding, dedication, and ability to normalize challenge and make anything seem possible. Your passionate approach to life has instilled in me a drive to set lofty goals and achieve them through hard work. Thank you so much for being there for me. I love you.

J.J.G.E.

CHAPTER 1: INTRODUCTION

The number of women in the geosciences declines as they make their way up the academic ladder. In 2017, women earned 44.9% of all doctorate degrees in the geosciences, atmospheric sciences, and ocean sciences but received only 37.6% of all postdoctoral appointments and only 31% of all doctorate-holding non-faculty research positions (National Science Foundation, 2017). These statistics indicate that female doctorates graduate at almost equal rates as their male counterparts, but their numbers begin to decline in post-graduate university research positions. The gender gap becomes even wider as women enter faculty positions in the geosciences. Indeed, in 2017 women made up an average of only 20% of faculty positions across the geosciences (American Geosciences Institute, 2019).

This curious and troublesome trend of women out of academia in the geosciences is in need of further exploration. As the above statistics illustrate, the narrowing percent of women in academia begins at the post-doctorate level. This is critical as past research shows that those who hold post-doctorate positions are more likely than those who do not to be hired for an academic job (Lin & Chiu, 2016; Hanchane & Recotillet, 2003). Thus, to uncover the narrowing effect of women out of academia, it is important to understand what might constrain women's chances at being hired as a post-doctorate researcher. One area that has received some attention in the past regarding hiring for academic roles is selection materials, specifically academic letters of recommendation, the themes within them, and how they differ for male and female applicants.

Letters of recommendation are an important element to the selection process for academic positions (Abbott et al. 2010). Indeed, academic professionals use letters of recommendation and place more weight on their contents when making selection decisions than do applied professionals (Nicklin & Roch, 2009). Thus, these selection materials deserve critical

inspection. Indeed, there is reason to suspect that letters of recommendation may contribute to fewer female post-doctoral researchers in the geosciences. Dutt et al. (2016) used content coding to label letters of recommendation for post-doctoral fellowships in the geosciences as being of doubtful, good, or excellent quality. They found that female applicants were "half as likely to receive excellent letters versus good letters compared to male applicants," (p. 805). Empirical research and theory on traditional gender stereotypes helps explain why writers may provide male applicants with higher quality letters.

Evidence from organizational research explains that women are perceived to "lack fit" in (Heilman, 1983) or to be incongruent with (Eagly & Karau, 2002), traditionally male-typed roles. This phenomenon has been shown to exist for women in science. For instance, research by Carli et al. (2016) found that scientific fields such as biology, chemistry, and physics were perceived by participants as more masculine than feminine and that stereotypes of scientists were more closely associated with stereotypes of men than of women. Thus, men were perceived to be more congruent with the role of a scientist than were women.

Gender bias not only shows up in stereotype perceptions of scientists, it also shows up in the language used by writers of letters of recommendation. Many of the studies exploring gendered language in letters of recommendation have taken one of two methodological approaches; 1) developing *a priori* word dictionaries and then exploring frequencies of stereotype words within letters (e.g. Schmader, Whitehead, & Wysocki, 2007; Akos, & Kretchmar, 2017, Aggarwal et al., 2018; & Madera, Hebl, & Martin, 2009) and 2) by-hand analysis, which involves first exploring letters and then tabulating emergent categories (e.g. Trix & Psenka, 2003). The *a priori* method has found that female applicants compared to male applicants are described as empathetic (Aggarwal et al., 2018), communal (Hebl et al., 2009),

and receive fewer standout words (e.g. unique, magnificent) than male applicants. While used less frequently, the by-hand method has found that letters for women are shorter, contain more doubt, and are associated more with the role of teacher and student than of researcher and professional colleague, which is how men are portrayed (Trix and Psenka, 2003).

While insightful, past methods used for examining gender bias in the language of letters of recommendation have important limitations. First, using an *a priori* approach constrains the inquiry to researcher defined themes, which leaves open the possibility that gendered themes in the data may go unobserved as they were not considered prior to the analysis. Second, an ad hoc approach introduces the threat of selection bias where content that confirms the hypotheses receives greater attention by the researcher, which may bias their results. Thus, while valuable for identifying expected or emerging themes in the data, both approaches require the researcher to decide what content to focus on prior to, or during, data analysis, which serves to limit inquiries whereby some content will inevitably go unexamined. This is a problem because important differences may exist between male and female applicants that simply go unnoticed as they are not in the researcher's dictionary or part of their research question. Recent developments in computer-based modeling are well suited to address these limitations. One method in particular, Structural Topic Modeling (STM - Roberts, Stewart, & Tingley, in press), has been described as an "automated method for content analysis" (Schmiedel, Müller, & vom Brocke, 2018). STM takes a holistic view of a corpus (i.e. a group of documents) by evaluating a word frequency by document matrix and identifies the co-occurrences of words within and across documents. Based on their co-occurrences, STM then clusters the words together in topics which serve to describe various elements of the letters. STM allows researchers to approach their data with greater objectivity and also provides a more holistic representation of the many expressed

themes within and across a corpus. Finally, the STM model allows researchers to assess the prevalence of covariates (e.g. gender of applicant) in each topic, providing insight regarding which topics are used more frequently in some documents over others (e.g. topics for male applicants vs. topics for female applicants).

Exploring the efficacy of this method, Elmore, Block, Bowers and Dutt (2019) employed STM to uncover topics expressed for male and female applicants across 1,203 letters of recommendation for post-doctoral fellowships in geosciences. They identified 31 topics that described the applicants, which provided a more complete picture of how applicants were described. As with past studies, they found that these topics differed by applicant gender and that male and female applicants were described in a manner consistent with gender stereotypes. For instance, female applicants were described more frequently as *top students and teachers* and male applicants were described more frequently as *publishing frequently*. These findings were consistent with those of Trix and Psenka (2003) who found, across 300 letters of recommendation for medical faculty positions, that women were portrayed as teachers and students while men were portrayed as researchers and professional colleagues. Beyond supporting past findings, Elmore et al. (2019) were able to examine the nuanced differences between how male and female applicants were described. While some topics were expressed much more for male or female applicants, others were expressed at a similar frequency between genders. For instance, topics like *asks insightful question*s and *works well and quickly on tasks* were expressed in relatively equal proportion for male and female applicants. These nuances provided a more complete picture of how male and female applicants were described in letters and thus a more representative and realistic picture of what decision makers might see when making evaluations.

While findings on gender differences in letters are valuable, we have limited insight as to whether these differences play a role in evaluation decisions about applicants; two studies provide notable exceptions. Madera, Hebl, and Martin (2009) explored letters of recommendation for assistant professor positions to a psychology program and found that letters for female applicants were more likely to express communal (e.g. e.g. "affectionate," "helpful," "kind") words compared to men, who were more often described as agentic (e.g. "assertive," "confident," "aggressive"). In a follow-up study, six professors rated 100 unique letters each. More communality was negatively related to hireability. Using the same data as Madera et al. (2009), Madera et al. (2018) found that female applicants received more language associated with doubt (e.g. hedging, faint praise, etc.) than male applicants. In a follow-up study, they found that faculty rated letters containing doubt more negatively. Thus, female applicants are perceived of, and described in, communal and doubtful terms, which impact how letters are rated.

While valuable, these studies have some limitations. First, these studies only examined how *single themes* in letters impacted evaluations, which limits our understanding of how the many other themes used to represent applicants compound to produce ratings. Second, one might expect letters with expressed doubt to be rated more negatively than letters without doubt. Further, when coding 1,224 geoscience letters of recommendation into categories of doubtful, good, and excellent quality, Dutt et al. (2016) observed that only 2.5% of the letters were considered doubtful. Thus, doubt is not a highly prevalent theme across letters, making it less representative of applicants, in general. Third, the Madera et al. (2009; 2018) studies were performed on letters of recommendation from the field of psychology which is generally majority female (American Psychological Association, 2019) and compared to harder sciences often perceived of as more feminine (Carli et al. 2016). Thus, one might expect to observe larger

effects of gender bias in letters of recommendation for harder sciences. Finally, Madera et al. (2009) only asked 6 professors to evaluate letters, lowering the power of their results and potentially opening up their results to sample bias.

In their work to address the methodological limitations of past research on letters of recommendation, Elmore et al. (2019) provided a more comprehensive picture of how applicants are described and how those descriptions differ for male and female applicants. This more complete picture provides an opportunity to address the limitations posed by Madera et al. (2009; 2018). Thus, in the present study, we examine how the 31 topics identified by Elmore et al. (2019) are rated and ranked in terms of importance for making a hiring decision by researchers and scientists in the geosciences. To this end we examine 1) which topics are considered more or less important for making a hiring decision and 2) whether those topics considered more important are also expressed more frequently in letters written for male applicants compared to female applicants. The present study addresses the limitations posed by Madera et al. (2009; 2018), first, by taking a detailed look at which gendered topics serve to produce biased evaluations beyond single themes, second, by having decision makers evaluate a range of topics that are representative of letter content and include desirable applicant attributes, rather than just a single negative theme (e.g. doubt), third, by showing how topics impact evaluations in a majority male academic field, and forth, by gathering feedback from a more representative sample of participants (compared to Madera et al. 2009). More generally, as these letters came from a sample of applicants to post-doctoral fellowships in the geosciences, our inquiry will help uncover one potential reason for the narrowing population of women in more senior academic roles. Focusing on evaluation materials for post-doctoral positions may provide some clues as to

why women receive approximately one-third of all post-doctorate roles in the geosciences and offer specific examples of areas that can be corrected.

CHAPTER 2: LITERATURE REVIEW

Gender Bias in Hiring in Academic STEM Fields

Unequal outcomes persist for women academic STEM (Science, Technology, Engineering, and Mathematics - National Science Foundation, 2017) fields. Studies show that one likely culprit is gender bias in hiring. Moss-Racusin et al. (2012) found that when examining identical applications for a lab manager position, male and female faculty rated female applicants as significantly less competent than male applicants. Further, both male and female faculty participants offered female applicants less starting salary and fewer mentoring opportunities compared to male applicants. The researchers also found that these effects were mediated by the female applicants being perceived to lack competence and were moderated by pre-existing subtle biases against women. Eaton, Saunders, Jacobson and West (2019) asked 251 tenure-track faculty in physics and biology departments from eight large public universities in the U.S. to evaluate the formatting and design styles of one of eight fictious CVs from an ambiguously qualified post-doctoral applicant. However, the researchers manipulated the race and gender of these CVs, keeping all other information identical. The true purpose of the study was to learn if faculty would rate these hypothetical post-doctoral applicants differently based on their race and gender. They asked faculty to also rate the candidate in terms of likeability, hireability, and competence. They found that male candidates were rated as more competent and more hireable by physics faculty but not biology faculty and that women were rated as more likeable across both departments. However, women account for 41.9% of all post-doctoral appointees in biology and only 17.2% of all post-doctoral appointees in physics (National Science Foundation, 2017), lending evidence to the idea that the more male a field, the less likely women will be perceived of as fit for its ranks (Eagly, Wood, & Diekman, 2000). These studies illustrate that by simply

manipulating the gender of the applicant, researchers have been able to show bias in hiring preferences and competence perceptions in STEM.

While the above research provides evidence for bias at the hiring stage, other research extends to show how unequal outcomes persist beyond hiring. For example, when examining the employment behavior of 2,062 faculty from 39 biology (e.g. molecular, biochemistry, genetics, and/or cell) departments within 24 high-ranked institutions, Sheltzer & Smith (2014) found that male faculty employed significantly fewer female post-doctorates than did female faculty. The difference became even more pronounced for male faculty with high prestige (e.g. funding, awards, etc.). Further, prestigious laboratories of both male and female faculty were significantly more likely to feed the pipeline of assistant professors at top universities. These same laboratories were also significantly less likely to employ female post-doctorates, thereby increasing the likelihood of male over female advancement. Thus, not only are women discriminated against at the hiring stage (e.g. Moss-Racusin et al., 2012; Eaton, Saunders, Jacobson & West, 2019), but their lower numbers in the best laboratories widens the gender gap as they move up the academic ladder.

The findings by Sheltzer & Smith (2014) illustrates how faculty are largely the deciding factor for who advances in academic life and results showing their preference for males provides some insight into how hiring can act as a roadblock for professional female advancement. It is a fruitful avenue then, to examine *why* bias may exist for women at the hiring stage. Moss-Racusin et al. (2012) found their effects were mediated by female applicants being perceived as lacking competence. But what drives that perception? What about women leads to their being perceived as lacking competence in the academic domain? Two theories, role congruity theory (Eagly &

Karau, 2002) and the "lack-of-fit" model (Heilman, 1983) provide helpful frames to understand why this might be the case.

Theories of Gender Bias

Summarizing Eagly's (1987) social role theory, Eagly, Wood, and Diekman (2000, p. 126) explain that "the differences in the behavior of women and men that are observed in psychological studies of social behavior and personality originate in the contrasting distributions of men and women into social roles." That is, the behavior of men and women generally, is rooted in and thus consistent with behaviors suited to those roles in which their gender is the majority. Extending this, Eagly, Wood, and Diekman (2000) developed the social role theory framework, which explains that over time the divergent roles that men and women have held in society have produced accommodation in them, such that they enact the behaviors and personality traits associated with these roles. Thus, men and women have become associated with the divergent behaviors and personality traits common to the roles they have traditionally taken up (e.g. women as homemakers and men as breadwinners - Eagly, Wood, & Diekman, 2000). It is not surprising then, that expectations of success in male dominated roles are perceived to require male attributes, and success in female dominated roles are perceived to require female attributes (Cejka & Eagly, 1999). However, between 1948 and 2016 participation of working age women in the U.S. workforce increased from 32.7% to 56.8% (U.S. Bureau of Labor Statistics, 2016), thus increasing female entrance into a traditionally male domain. However, associations between gender roles and occupational roles have created challenges for women as they have traversed across traditional boundaries.

Organizational theorists have developed models to explain the challenges women face when entering traditionally male roles. Specifically, the "lack of fit" model (Heilman, 1983) and

role congruity theory (Eagly & Karau, 2002) provide useful frames for exploring phenomena in this area. Citing her foundational 1983 article positing the "lack-of-fit" model, Heilman (2012, p. 116) explains that descriptive stereotypes, or designations of what women are like "create problems for women when there is a perceived 'lack of fit' between a woman's attributes and the attributes believed to be required to succeed in traditionally male occupations and organizational positions." That is, female attributes are perceived to lack fit with jobs associated with male attributes. "Lack of fit" perceptions are a problem for women as they can reduce the likelihood of obtaining male-typed roles.

Similarly, role congruity theory (Eagly & Karau, 2002, p. 573), proposes that women face prejudice when taking on leadership roles, including being perceived as less favorable for leadership positions, and being held to a higher standard when in leadership positions. More generally, Eagly and Diekman (2005, p. 2) explain that "a member of a group whose stereotypical attributes are thought to facilitate performance in a role is ordinarily preferred over a member of a group whose stereotypical attributes are thought to impede performance." Thus, when attributes associated with women are perceived as impediments to success in a role, leadership for instance, chances of a woman being hired are reduced. For example, leadership research has shown that women are not considered good matches for leadership roles because of descriptive stereotypes about what women are like, such as compassionate and warm, do not match the qualities associated with what a leader is like, such as agentic and competent, which are traditional stereotypes of what men are like (Schein, 1973; Heilman et al. 1989; Eagly & Carli, 2007).

Role congruity theory and the "lack of fit" model help explain findings of gender inequity in academia. For example, Carli et al. (2016) found that scientific fields such as biology,

chemistry, and physics were perceived by participants as more masculine than feminine and that stereotypes of scientists were more closely associated with stereotypes of men than of women. Thus, male attributes were perceived as more congruent with the role of a scientist than were female attributes. Further, Rice & Barth (2017) explored gender bias in hiring decisions for professor positions in either stereotypically masculine (e.g. engineering, chemistry) or stereotypically feminine (e.g. education, art history) fields. Over multiple hiring decisions, the researchers presented participants with one male and one female applicant and a job description, then were asked which applicant they would hire. They found that male applicants were more likely to be hired for the masculine position, and female applicants were more likely to be hired to the feminine position, upholding the findings of Cejka and Eagly (1999) that male dominated roles are perceived to require male attributes and female dominated roles are perceived to require female attributes. Smyth & Nosek (2015) asked 100,000+ participants to take an implicit association test measuring reaction time to words that were stereotype-congruent (e.g. science-male and liberal arts-female) or stereotype-incongruent (e.g. science-female and liberal arts-male). They found a strong overall science-male stereotype bias. That is, participants were faster (in terms of milli-seconds) to associate male-type words (e.g. “man”, “father”, etc.) with science words (e.g. “geology”, “physics”, “math”, etc.) than female-type words (e.g. “woman”, “wife”, etc.) with science words.

Thus, over several studies using various methods, female targets are perceived to lack fit or be incongruent with the role of scientist (Rice & Barth, 2017), and attributes associated with women are perceived to lack fit or be incongruent with the attributes of a scientist (e.g. Carli et al., 2016; Smyth & Nosek, 2015).

Gender Differences in Letters of Recommendation

The research reviewed so far has focused on gender bias in academia at the evaluation, hiring, and placement stages. However, before these stages, applicants must first prepare materials for evaluation. Among these materials are letters of recommendation. Letters provide an opportunity for applicants to be described in greater detail, for evaluators to gain professional views of the applicant from previous advisors, and for advisors to share their experiences of the applicant. Letters of recommendation are an important element of hiring decisions in academia. Indeed, academic professionals use letters of recommendation and place more weight on their contents when making selection decisions than do applied professionals (Nicklin & Roch, 2009). Further, Potvin, Chari, and Hodapp (2017) surveyed 170 physics faculty from 149 U.S. institutions asking them to rate criteria such as GPA, GRE scores, personal statements, and quality of letters of recommendation among others, in terms of importance for admission to a doctoral program. They found that, out of the 20 different student criteria, quality of letters of recommendation was rated as second most important behind GPA. Thus, at least in physics, letters are highly important to admission decisions. However, as we have seen, past research shows that faculty evaluators are subject to the influence of gender stereotypes and bias when making evaluations and hiring decisions (e.g. Moss-Racusin et al., 2012). This influence is also present when recommenders write letters of recommendation. Past research shows that writers not only tend to describe male and female applicants in a manner consistent with their gender stereotypes, and in so doing send signals of female incongruence with the traditionally male-typed role of scientist, but also write poorer quality letters for female applicants compared to male applicants.

Dutt et al. (2016) used content coding to label letters of recommendation for post-doctoral fellowships in the geosciences as being of doubtful, good, or excellent quality. They found that female applicants were "half as likely to receive excellent letters versus good letters compared to male applicants," (p. 805). Hoffman et al. (2019) examined 311 letters of recommendation for an abdominal transplant surgery fellowship. The researchers content coded the letters and counted the frequency of communal (e.g. warm, thoughtful, kind) and agentic (e.g. leader, outstanding, assertive) terms. While they did not find a difference in use of communal terms, they did find that agentic terms were used significantly more often to describe male applicants compared to female applicants. In their examination of letters of recommendation for a medical faculty position, Trix & Psenka (2003) observed that letters for women were shorter, contained more doubt, and were associated more with the role of teacher and student than of researcher and professional colleague, which was how men were portrayed. Finally, across 2,523 letters of recommendation to an ophthalmology residency program Aggarwal et al. (2018) examined the presence of 22 words, including standout words, grindstone adjectives, and words that illustrated compassion, and ability. They found that female applicants compared to male applicants were more likely to be described as empathetic. Thus, we see that both men and women are described in a manner consistent with their gender stereotypes and further that female applicants receive poorer quality letters than do male applicants.

However, past research has found mixed results regarding the use of standout language (e.g. words like "unique" and "magnificent") for female and male applicants. For example, Schmader, Whitehead, and Wysocki (2007) investigated the difference in word frequencies between letters of recommendation for male and female applicants applying for chemistry and biochemistry jobs. They found that men received more standout words compared to women.

Alternatively, French et al. (in press) examined word use in letters of recommendation to a surgical residency program. They found that female applicants received more standout words than male applicants. While valuable, French et al.'s (in press) findings are anomalous among the otherwise consistent results that female applicants are described in letters of recommendation more negatively or in a manner more congruent with their gender stereotypes, which is incongruent with the stereotype of scientists (e.g. Carli et al., 2016; Smyth & Nosek, 2015). Indeed, in a study of letters of recommendation for assistant professor positions, Madera et al. (2018) trained three raters to examine letters and rate the extent to which they contained four types of doubt including, negativity, hedging, irrelevant information, and faint praise. They found that female applicants received more language associated with doubt than male applicants, and that certain types of doubt led to lower ratings of perceived competence in research by faculty. While each data set and field of study is different, there is far more evidence that female applicants are described in an undesirable manner compared to male applicants in letters of recommendation. Further, as the Madera et al. (2018) study shows, these differences impact evaluations.

Building on how gendered content impacts evaluations, Madera, Hebl, and Martin (2009) explored gendered language in 624 letters of recommendation to eight junior faculty positions in psychology. They found that female applicants were more likely to be described in communal terms (e.g. "affectionate," "helpful," "kind") and social-communal terms (e.g. "wife," "kids," "babies") than male applicants and that male applicants were more likely to be described in agentic terms (e.g. "assertive," "confident," "aggressive") compared to female applicants. In a follow-up study, the authors enlisted six psychology professors to read 100 unique letters and rate the applicant in terms of hireability. They found that the expression of communal

characteristics was negatively related to hireability and that communal attributes in letters mediated the relationship between applicant gender and ratings of hireability, controlling for other application details (e.g. publications, years teaching experience, etc.). Thus, female applicants were described by letter writers in terms congruent with their gender (e.g. communality - Eagly & Karau, 2002), and letters with communal terms led to lower ratings of hireability by faculty.

Thus, across a variety of studies, we see that female applicants receive fewer high-quality letters of recommendation compared to male applicants, are described in a manner that is congruent with female gender stereotypes and incongruent with scientist stereotypes, which signals a lack of fit and impacts subsequent hiring decisions. This work is valuable as it shows how details about applicants can operate to influence important hiring decisions. Further, these findings refute what Williams and Ceci (2015) conclude in their research; that gender is not a factor in academic hiring decisions. In their study, the researchers presented STEM faculty with identically qualified male and female applicants to an assistant professor position. Women were favored 2 to 1 over men. These findings refute those of Moss-Racusin et al. (2012) and Eaton, Saunders, Jacobson and West (2019), due, according to the authors, to the applicants appearing unambiguously strong. However, they present one important limitation; that applicants were identical in terms of attributes deemed important for science (e.g. competence). As we have seen, men and women are not portrayed as unambiguously strong in evaluation materials. Indeed, as the literature on gender differences in letters of recommendation show, women are described in a way that may contradict perceptions of competence (e.g. expressions of doubt found by Madera et al., (2018)). While, Madera, Hebl, and Martin (2009) found that men and women were rated the same in terms of hireability when subject matter experts evaluated their letters of

recommendation (similar to Williams and Ceci, (2015)), letters expressing communality were rated as less desirable and women were described in more communal terms, which is incongruent with science (Carli et al., 2016). Thus, it is not realistic to hold competence constant when language in letters of recommendation place competence in question more often for female applicants than male applicants (e.g. through doubt or incongruent signals). As letters of recommendation are a vital component to any strong application (Abbott et al., 2010) and serve to provide nuanced details about an applicant, such as their work ethic, interpersonal skills, research focus, and career potential, among other characteristics, none of which is held constant, it is important to examine the nuanced manner in which female applicants may be described compared than male applicants, and how those differences impact hiring decisions. A task past research has made progress in, but where limitations remain.

Limitations of Past Methods Studying Gender Differences in Letters of Recommendation

Previous methods used to explore gender bias in letters of recommendation vary, from by-hand analysis, which involves first exploring letters and then tabulating emergent categories (e.g. Trix & Psenka, 2003) to developing *a priori* word dictionaries and then exploring frequencies of stereotype words within letters (e.g. Schmader, Whitehead, & Wysocki, 2007; Akos, & Kretchmar, 2017). One popular dictionary method, Linguistic Inquiry Word Count (LIWC - Pennebaker, Francis, & Booth, 2001), allows researchers to compare dictionaries that illustrate gender stereotypes to word frequencies in their own data, providing an indication of the prevalence of stereotypic portrayals. While valuable, these methods have limitations.

For example, Trix and Psenka's (2003) by-hand analysis involved evaluating letters and then expanding their semantic categories throughout the coding process, thus allowing them to develop themes directly from their data. However, as they developed themes as they went along,

their inquiry was ad hoc. This is a challenge as this approach can introduce the threat of bias due to researcher interests shaping the results, and further is not easily reproducible (Moretti et al., 2011). Also, by-hand content coding is time consuming (Marks and Yardley, 2004).

Further, dictionary methods like LIWC are constrained by the finite word dictionaries developed by researchers (Grimmer & Stewart, 2013), where results are limited to expected content (e.g. words in the dictionary) rather than emergent content. This limited view poses a threat to picking up actual differences across letters. Indeed, while most studies using the LIWC have uncovered gender bias in letters, some have not. For example, French et al. (in press) compiled approximately 400 terms which were comprised of 24 categories using past literature and by gathering words in letters from a previous application year which they thought "might influence a reader." They then used these word categories to examine the frequency with which words were expressed in a more recent round of applications. Counter to the findings of Schmader et al. (2007), French et al. (in press) found that female applicants received more standout words (e.g. amazing, outstanding, etc.) compared to male applicants. While interesting, like all other studies using the dictionary method, their findings are limited by the fact that they did not evaluate their data directly. That is, they evaluated their corpus using a word list that was created independent of their data. While their corpus contained gender differences on one of the categories (e.g. standout words), they found no other differences on any of the other 23 categories. It may be that no differences existed, however, they could not know that because they did not examine the actual content of the letters as is done in content analysis (e.g. Trix and Psenka, 2003). Thus, by using the dictionary method, French et al. (in press) may have missed other gender differences that existed in the study data but were not accounted for by their dictionaries.

Thus, while providing valuable insights, past investigations of gender bias in letters of recommendation have been limited to inquiries that require significant researcher intervention before (e.g. dictionary methods) or while (e.g. by-hand tabulation) discovery occurs. Techniques in computer-based modeling are helping to address these limitations. Pennebaker and Tausczik (2010, p. 38) describe the emergence of new computational techniques as "transforming language analysis and modern social science," and that LIWC "represents only a transitional text analysis program in the shift from traditional language analysis to a new era of language analysis." That new era has arrived in the form of Natural Language Processing and one prominent approach under this suite of techniques is topic modeling (Blei & Jordan, 2003).

Topic Modeling

Schmiedel, Müller, and vom Brocke (2018, p. 3) explain that topic modeling "can be understood as an automated method for content analysis." This automation is the innovation of topic modeling as it allows for topics to be generated in an "unsupervised" fashion, where topics are inferred based on word frequencies occurring both within and across documents. This contrasts with a "supervised" content analysis where researchers must assume the topics of interest (e.g. Trix & Psenka, 2003 – Roberts et al., 2014a). Topic modeling is an algorithm-based approach that reveals latent topics across a collection of documents (corpus) by identifying the words that co-occur within and across documents and placing them in lists (topics). While there are a variety of topic model techniques available, Structural Topic Modeling (STM – Roberts et al., 2014a) provides an appropriate framework for exploring research questions in the social sciences.

Structural Topic Modeling is defined as mixed-membership model, meaning documents can be represented by a mixture of topics (Roberts et al., 2014a). That is, each document is

assigned a probability of association with all of the topics that are generated, summing to 1, with higher probabilities assigned to topics that are most representative of a given document. Thus, one can observe a cluster of documents highly associated with a given topic and surmise that those documents share a theme illustrated by words in that topic. The advantage of STM over other methods of topic modeling is that it allows the researcher to incorporate covariates into model estimation, thereby accounting for the influence those covariates may have on the prevalence and word use of topics in documents. As Roberts et al. (2014a, p. 1067) put it, "rather than assume that topical prevalence (i.e., the frequency with which a topic is discussed) and topical content (i.e., the words used to discuss a topic) are constant across all participants, the analyst can incorporate covariates over which we might expect to see variance." This is an advantage over earlier topic models such as Latent Dirichlet Allocation (LDA – Blei, Ng, & Jordan, 2003), which does not allow for the inclusion of covariates and thus requires the researcher to perform a "two-stage" approach, whereby (1) topics are estimated and then (2) covariate effects on topic proportions are examined (Roberts et al., 2014b). Roberts et al. (2014b) explains that accounting for covariates in the analysis rather than after provides a more "accurate estimation of quantities of interest" and that compared to STM, the two-stage process required by LDA tends to "attenuate continuous covariate relationships on topical prevalence." Therefore, STM provides a more appropriate test for how topics vary by covariates.

Schmiedel, Müller, and vom Brocke (2018) employ STM and provide an example of how covariates vary among topics. The researchers entered 428,492 employee reviews from Glassdoor.com into the STM model to learn how ratings of organizational culture vary by topics discussed in reviews. They identified a model with 70 topics to describe their data. As topic modeling provides a holistic representation of the topics discussed in corpus, there are often

topics considered outside the scope of a given study as they do not relate to the researcher's questions. Accordingly, Schmiedel et al. (2018) examined their 70 topics to identify those that related to perceptions of organizational culture; they found 45 topics. Among the 45 topics were those that covered social aspects of work (e.g. lying, helping, etc.) and career opportunities (e.g. development, advancement, etc.). They then evaluated how the topics varied by a covariate; company culture ratings associated with each review. They found that employees who gave their company high culture ratings emphasized career opportunities in their reviews, while employees who gave their company low culture ratings emphasized social aspects at work. Thus, without prior examination, the researchers identified concepts that mattered to employees and showed how they negatively or positively related to organizational culture. Thus, STM allows researchers to mix qualitative and quantitative methods in order to understand a collection of documents.

STM addresses limitations of past methods used to explore bias in letters of recommendation as it is an inductive approach that requires neither constraining exploration to *a priori* word lists, as is done with LIWC, nor manually coding letters by-hand to identify emergent themes; both of which may not totally represent aggregate content and may only represent interests of the researcher (DiMaggio, Nag, & Blei, 2013). That is, with STM, the researcher lets the algorithm examine aggregate content and generate associations within the data rather than telling it what to look for, as is done with the dictionary method. When there are large sums of data to be analyzed, STM cuts down on the time necessary to code but still accounts for representative themes in the data, similar to by-hand content coding but found lacking in the dictionary method. Finally, STM is also advantageous when seeking to remain agnostic about the themes expressed in a data set, which is found lacking in by-hand content analysis. Thus, topic

modeling generally, and STM in particular, provide advantages when evaluating letters of recommendation, which we illustrate below.

Topic Modeling – Exploring Gender Differences in Letters of Recommendation

Turrentine et al. (2019) provide an example of how LDA can be used to explore gender differences in letters or recommendation. The researchers examined 332 letters of recommendation for applicants to a surgical residency program who were invited to interview. They found that letters written for female applicants included topics describing work ethic, caregiving, and support, and that letters written for male applicants included topics describing performance and technical skill. While, this research corroborates past findings and provides one example of how topic modeling can be used to understand gendered themes in letters of recommendation, their approach presented some limitations. The researchers did not make it clear how they evaluated gender differences by topic (e.g. chi-square, multinomial logistic regression, etc.), which leaves the reader wondering the extent to which topics actually differed. Further, as they used LDA to evaluate their data, whatever differences they did calculate may been less reliable given the concerns of Roberts et al. (2014b), who show that results from LDA compared to STM exhibit higher variance when examining relationships with covariates.

Elmore et al. (2019) address this limitation by using STM to explore gender bias in 1,203 letters of recommendation to post-doctoral fellowships in the geosciences. As we employ their findings in the present study, we review their work in detail here. The data included letters for 452 applicants (averaging 2.71 letters per applicant), 137 female applicants (averaging 2.64 letters per applicant) and 315 male applicants (averaging 2.74 letters per applicant). Using the statistical package R (R Core Team, 2017), the researchers performed four detailed steps which included pre-processing, developing, estimating and evaluating, and interpreting their model.

Their first step was to pre-process the data, which included making the data machine readable, enhancing uniformity (e.g. reducing words to their stems by removing endings such as *ing* and *ed*), and removing irrelevant words (e.g. *the, it, and*, etc.).

Their second step involved developing their model by determining the number of words to be included in the analysis. Words that appeared in 10 percent of the letters or more were kept in the analysis as this enhanced the representativeness of the data, cut down on the idiosyncratic content related to the applicant's field of study (e.g. *volcanism, glaciology,* etc.), and enhanced the resolution on common words used to describe applicants across letters (e.g. *studious, achievement,* etc.). The analysis included 400 unique words. These words occurred 107,143 times, accounting for approximately 55% of the 192,913 words in the letters.

Their third step involved estimating and evaluating their model. The STM algorithm was performed on the 1,203 letters to estimate the correct number of topics given the words in the analysis (Lee & Mimno, 2004) and produced a model with 53 topics. To learn how well this model represented the data in the letters, the researchers quantitatively evaluated model fit using STM metrics and a qualitatively confirmed these metrics using a visual inspection of topics. This quantitative process began with two metrics, which provided insight into their model's semantic coherence (Mimno et al., 2011), which verifies that "high-probability words for [a] topic tend to co-occur within documents," (Roberts et al., 2014a, p. 1069) and exclusivity (Bischof & Airoldi 2012), which verifies that "top words for topic[s] are unlikely to appear within top words of other topics," (Roberts et al., 2014a, p. 1069). Optimizing semantic coherence and exclusivity produces a more interpretable set of topics, which aids qualitative inquiry (Chang, et al., 2009). Comparing models of between 10 and 100 topics the researchers observed the semantic coherence and exclusivity statistics to be optimized at the 53-topic model. To ensure topics from

this particular model were indeed semantically coherent and exclusive, they performed a visual inspection of the topics, which confirmed the model statistics. Finally, they performed a test of model validity using the held-out likelihood. The held-out likelihood determines model fit by generating topics from only part of the documents and then uses those topics to predict topics in the held-out parts of the documents, the number of topics that best predict the held-out document parts is deemed the best model. Based on this analysis they once again confirmed and were able to conclude that the 53-topic model specified by the STM algorithm was indeed the most appropriate model to represent the data.

Finally, in their fourth step, they interpreted and labeled the topics from the STM model using a procedure outlined by Schmiedel, Müller, and vom Brocke (2018). As they were interested in the way in which applicants were described rather than the science they were engaged in (e.g. researching *seismology*, *water, sediment,* etc.), the researchers narrowed their focus to 31 topics which described the applicants. To understand the meaning of each topic, they examined how the words from each topic were used in letters most associated with that topic. To identify those letters, they used topic-document proportions provided by the STM package. Following methods outlined by Schmiedel et al. (2018), the researchers identified the top five letters most associated with each topic and individually coded them using the top words from the topic as their guide. As Roberts et al. (2014, p. 1069) explains "'exemplar' documents and can be used to validate that [a] topic has the meaning the analyst assigns to it." Thus, their fourth step included a final check of model validity to bolster step 3. To date, there are no guidelines for the appropriate number of associated documents to investigate when validating a topic. Thus, the researchers surmised that a content analysis of five documents for each topic was sufficient to

ensure that 1) words from topics appeared in letters and that 2) topics were capturing meaningful themes.[1]

To label each topic, two independent raters keyword searched topic words within each of the top 5 letters, aggregated quotes that used topic words, then interpreted the collection of quotes. After norming on coding criteria, raters independently labeled the topics. To ensure coding reliability, coders normed several times throughout the coding process. Once labels were generated, coders compared labels and discussed any disagreements until agreement was reached (Schmiedel, Müller, & vom Brocke, 2018). There was high agreement between coders (81.6%) and high inter-coder reliability (Cohen's kappa = 0.81).[2] Table 1 provides a summary of topic labels and the most probable words for each topic from letters.

[1] Elmore et al. (2019) made this decision after consulting with their co-author who had published several times using the topic model methodology.

[2] Using the criteria of "same" vs. "different," where topics were considered the same if coder labels converged prior to discussion and were different if they converged after discussion. Comparing these frequencies Elmore et al. (2019) calculated a Cohen's Kappa coefficient. The aim of this procedure was to correct for if judges were coding randomly.

Table 1

Topic Labels and Most Probable Words for Each Topic

Topic Label	Highly Probable Words
Presenting and Publishing Nationally and Internationally	intern, research, present, univers, journal
Develops Models	model, numer, coupl, develop, result
Writer Served on Thesis Committee	committe, provid, program, member, serv
Top Student and Teacher	student, graduat, cours, undergradu, class, teach
Successful Thesis Defense	phd, thesi, supervis, innov, also
Strong Recommendation	doctor, post, posit, recommend, institut
Works Well and Quickly on Tasks	new, develop, task, well, time
Dedicated to Research	research, fellow, studi, associ, activ
Highly Capable Engineer	engin, degre, univers, complet, research
Skilled Experimentalist	experi, experiment, measur, laboratori, can
Achieves Scientific Results	result, obtain, scientif, phd, studi
Will Be Good Fit	good, found, sure, join, will
Engaged in Environmental Management Science	environment, manag, environ, particip, develop
Excels Academically and Professionally	academ, research, excel, profession, communic
Research Contributor Now and Future	state, project, month, futur, contribut
Outstanding Young Researcher	young, research, creativ, scientist, scientif
Making Progress and Maturing	part, made, progress, also, particular
Publishing Frequently	paper, publish, journal, author, two
Computer Programming Skills	geophys, comput, program, numer, background, skill
Writer Supports Application	postdoctor, applic, fellowship, strong, support
High Potential Academic and Researcher who Gets Funding	research, project, propos, fund, plan
Submitting Manuscripts	manuscript, will, result, posit, review
Hard Working	concern, may, phd, hard, enthusiast
Uses and Develops Methods	method, use, differ, test, develop
Capable Technician	knowledg, conduct, various, research, field
Tackles Research Problems	problem, clear, student, impress, abil

Understands Complex Systems	system, complex, design, project, skill
Accomplished Teacher and Department Citizen	depart, faculti, teach, student, assist
Pleasant Team-Member	candid, recommend, team, dear, postdoctor
Asks Insightful Questions	talk, meet, mani, question, exampl, insight, ask
Improving Their English	english, written, improv, write, well, time

Note. Topic labels were generated examining how highly probable words were used in exemplar letters. From Elmore et al. (2019).

In their analysis, each topic contributed some proportion to explain how female applicants were described overall, with proportions from the topics summing to 1 for female applicants, and similarly for male applicants. That is, one topic might express 3% of the topic proportions for 100% of female applicants, and that same topic might express only 2% of the topic proportions for 100% of male applicants. Thus, the researchers were able to examine the difference between how much a topic was discussed for male applicants overall versus female applicants overall. Figure 1 provides an illustration of the differences for topics which described applicants. Similar to prior research, Elmore et al. (2019) found that women and men were described differently among their collection of letters of recommendation and that those differences were in line with gender stereotypes.

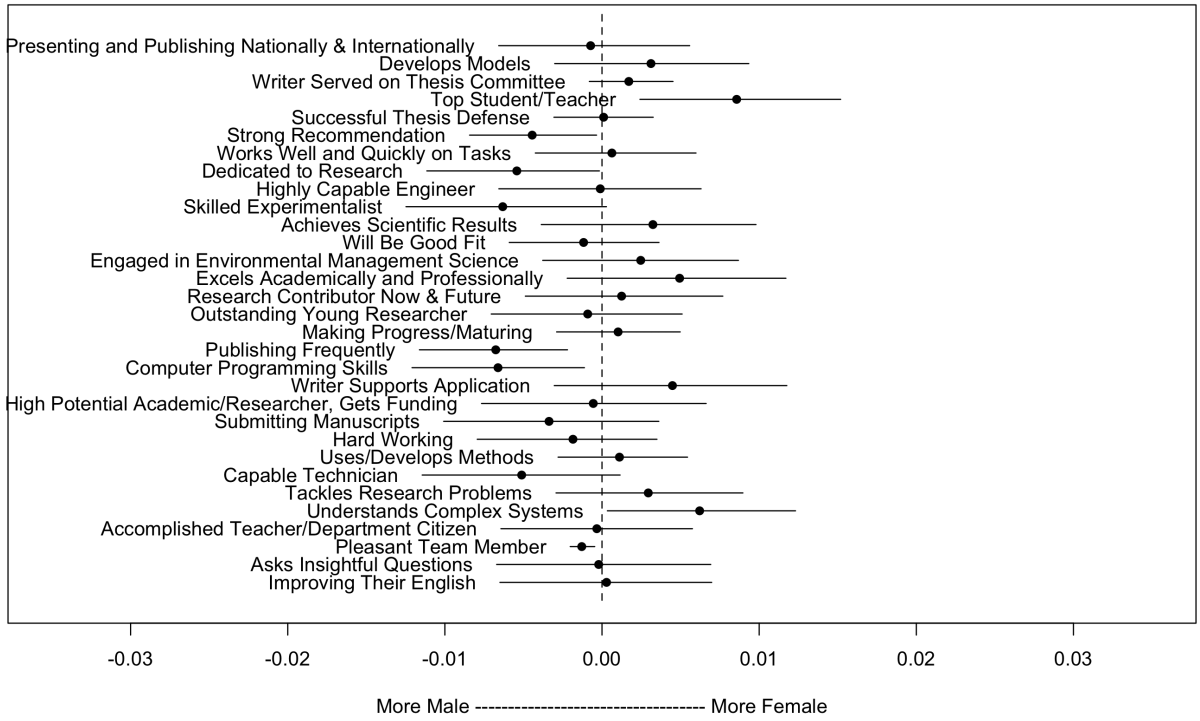

Figure 1. Difference in Topic Proportions by Applicant Gender. *Note.* Difference in average topic proportion for male vs. female applicants. Points to the right of zero indicate more topic use across letters written for female applicants and points to the left of zero indicate more topic use across letters written for male applicants. Lines indicate a 95% confidence interval indicating uncertainty in the model. From Elmore et al. (2019).

Among the topics which clearly describe one gender more than the other, they observed that female applicants were described more by the topics expressing success across roles, like *top student and teacher* and *excels academically and professionally.* Further, topics that contained expressions of support from the letter writers (e.g. *writer served on thesis committee* and *writer supports application*) were expressed more frequently for female applicants compared to male applicants. By contrast, compared to female applicants, male applicants were described by letter writers as technically skilled with topics like *computer programming skills* and *capable technician.* Further, male applicants were described more frequently as performing research and publishing, with topics like *dedicated to research, skilled experimentalist,* and *publishing frequently.* Finally, the researchers observed topics that were counter-stereotypic to applicant

gender. For instance, male applicants were described more frequently as a *pleasant team member*, which implies cooperation, which is a key feature of feminine leadership (Eagly & Karau, 2002), and female applicants were described more frequently as *understanding complex systems*, signaling competence, a contrast to lower perceptions of competence found in past research (Moss-Racusin et al., 2012). Thus, Elmore et al. (2019) found that similar to past research, topics expressed more frequently for female applicants included communal themes (e.g. student and teacher, writer involvement) and topics expressed more frequently for male applicants included competence themes (e.g. skills and research). However, the STM algorithm also picked up on counter-stereotypical topics (e.g. male-*pleasant team member* and female-*understands complex systems*), making the picture more complex than past research might have painted it. Topics used more frequently to describe male applicants were also more specific (e.g. *skilled experimentalist, computer programming skills*) than topics used to describe female applicants (e.g. *writer served on thesis committee, writer supports application*). This lack of detail may make the topics for female applicants less impactful in hiring decisions.

However, there are several likely criticisms of Elmore's et al. (2019) model: 1) whether topics have criterion-related validity; and 2) whether the fact that applicants were nested within letters inflated topic frequencies. Elmore et al. (2019) addressed these criticisms in their analysis. First, to address the doubt that topics were predictive of outcomes, Elmore et al. (2019) explored topic distribution on another covariate; letter quality. The letters used by Elmore et al. (2019) had been previously content coded in to doubtful, good, and excellent letter categories (e.g. Dutt et al., 2016). Using these categories as covariates, Elmore et al. (2019) were able to show that topics expressing attributes of an excellent applicant (e.g. *excels academically and professionally, outstanding young researcher*, etc.) were expressed more often in letters coded as

excellent and topics expressing attributes of a good candidate (e.g. *hard working, capable technician*, etc.) were expressed more frequently in letters coded as good. Thus, they were able to provide criterion-related validity for this novel approach to analyzing text. Second, one might argue that the topics observed by Elmore et al. (2019) were inflated due to the fact that each applicant was represented by more than one letter and thus, the same themes could have been expressed by multiple writers for a single applicant. To address the nested nature of the data and ensure that the topics were not inflated Elmore et al. (2019) randomly selected one letter from each applicant and ran their STM model again. The subset model used equivalent parameters in estimation as the full model described above. While they observed slightly different words in the topics, the semantic meanings were the same and described applicants in a manner consistent with topics from full model. Thus, they concluded that the nested nature of the data did not inflate their results.[3]

Thus, the model by Elmore et al. (2019) uncovered topics without being constrained by dictionaries developed *a priori*; the topics were derived using only the data set of interest. Further, they did not comb through the data making decisions on labels as they went along, they let the data speak for itself and interpreted what it told them. That is, they produced these topics in an inductive, rather than deductive manner, which provided a fuller, more representative view

[3] Elmore's et al. (2019) subset model used parameters equivalent to their full model estimation. They observed slightly different topics in the subset model, but in general, topics remained on theme with gender differences observed in the full model. For instance, topics with higher average frequencies across letters written for female applicants described applicants as students, teachers, thoughtful researchers, and department citizens, which were topic themes also used more frequently to describe female applicants in the full model. Conversely, topics with higher average frequencies across letters written for male applicants described them as successful, as researchers, and that they are publishing, which were topic themes also used more frequently to describe male applicants in the full model. Thus, trends across topics in the subset model were similar to those in the full model. Based on these results, the researchers concluded that each letter writer seemed to be contributing unique content to the corpus while remaining on theme with gender differences and that the differences observed between male and female applicants in the full model were not an artifact of inflated themes due to applicants having received more than one letter.

of the letter content. Beyond the topics reviewed above, which generally produced the largest differences between male and female applicants, there were other topics which provided a more detailed picture of how applicants were described and how those descriptions varied by applicant gender. This is important as much of the past work examining gender differences in letters of recommendation have focused on the big and obvious differences, casting aside other themes that may operate in an evaluators decision to hire an applicant.

Elmore's et al. (2019) comprehensive approach to exploring topics expressed for male and female applicants provides an opportunity to learn how these topics impact an evaluator's decision to hire an applicant. However, it is not yet clear whether the topics that vary by gender would be considered more or less important in hiring decisions and whether those differences vary systematically to favor male applicants, which past research on gender stereotypes would predict (e.g. Carli et al., 2016). Indeed, we do not know if these differences matter to those who actually hire post-doctoral fellows in the geosciences. As reviewed above, past studies investigating the impact of gender differences on evaluations indicate that differences do matter.

Impact of Letter Themes on Hiring Decisions

While limited in number, two studies illustrate how themes in letters of recommendation impact evaluations. First, in a follow-up to their finding that female applicants receive more doubt in letters compared to male applicants, Madera et al. (2018) operationalized the four types of doubt (e.g. negativity, hedging, irrelevant information, and faint praise) within an identical letter of recommendation. They recruited 305 faculty participants from various disciplines (43% psychology) and asked them to rate the applicant in terms of perceived teaching and research competence. They found that expressions of negativity and hedging in letters led to lower evaluations of research competence compared to the faint praise, irrelevant information, and

control conditions. Thus, negative content and hedging, which the researchers found occurring more frequently in letters for women, was also evaluated more negatively by likely decision-makers in hiring. Second, in a follow-up to their finding that female applicants received more words associated with communality compared to male applicants, Madera et al. (2009) recruited six psychology professors to read 100 unique letters and rate the applicant in terms of hireability. They found that the expression of communal characteristics was negatively related to hireability. Further, communal attributes in letters mediated the relationship between applicant gender and ratings of hireability, controlling for other application details (e.g. publications, years teaching experience, etc.). This study illustrates how one attribute found more frequently to describe female applicants also serves to undermine the likelihood of applicant advancement.

While the studies by Madera and colleagues (2009; 2018) are interesting, they have several limitations. First, both studies highlight only one way in which women are portrayed in letters of recommendation (e.g. with doubt or as communal), and the subsequent impact of that single difference on evaluations. This necessarily limits our understanding of how themes expressed in letters of recommendation impact hiring decisions. These controlled experiments are helpful to learn how a single construct operates in letters, but they do not help us understand how the many topics in letters compound to produce a decision. When evaluators read recommendation letters, they do not read them in a vacuum; many topics are present. Indeed, in evaluations, applicants are evaluated against one another, against *multiple* topics describing applicants, which impacts decisions to hire. Second, one might expect letters with expressed doubt to be rated more negatively than letters without doubt. Thus, Madera et al. (2018) failed to account for the impact that more positively valanced themes have on evaluations. Further, Dutt et al. (2016) observed that only 2.5% of the letters for applicants for a geoscience position raised

doubts, which indicates that doubt is not a highly prevalent theme across letters, making it less representative of applicants, in general. Third, while Madera et al. (2009) *did* explore evaluations based on how applicants were described, they investigated gender differences in the field of psychology which is generally majority female (American Psychological Association, 2019) and compared to harder sciences often perceived of as more feminine (Carli et al. 2016). Thus, the impact of female stereotypes in letters on evaluations, while evident, may be weaker in the psychology domain than in the harder sciences. Last, Madera et al. (2009) only asked six professors to evaluate letters, lowering the statistical power of their findings and potentially opening their results up to sample bias.

The Present Study

The more complete picture of how applicants are described across letters captured by Elmore's et al. (2019) STM model provides an opportunity to address the limitations of the work by Madera et al. (2009; 2018) by 1) taking a detailed look at which gendered topics serve to produce biased evaluations beyond single themes, 2) by having decision makers evaluate a range of topics that are representative of letter content and include desirable applicant attributes, rather than just a single negative theme (e.g. doubt), 3) by showing how topics impact evaluations in a majority male academic field, and 4), by gathering feedback from a more representative sample of participants (compared to Madera et al. 2009). Given that evaluators provide different ratings based on the themes expressed within letters (e.g. Madera et al., 2009; 2018) we are interested in learning whether topics will be rated differently in terms of importance in making a hiring decision and which topics evaluators perceive of as most and least important, in general. Thus, in the present study we use the topics from the Elmore et al. (2019) study to investigate several research questions and hypotheses:

Research Question 1: What topics used to describe post-doctoral applicants in the geosciences are *rated* as most important by geoscience faculty and researchers?

Research Question 2: What topics used to describe post-doctoral applicants in the geosciences are *ranked* as *most important* by geoscience faculty and researchers?

Research Question 3: What topics used to describe post-doctoral applicants in the geosciences are *ranked* as *least important* by the geoscience faculty and researchers?

As stereotypes of women are perceived as incongruent with the role of a scientist compared to stereotypes of men (Carli et al., 2016; & Smyth & Nosek, 2015) and that Elmore et al. (2019) found that topics expressed in letters produced gender differences consistent with gender stereotypes we predict that,

Hypothesis 1: Topics expressed more frequently for male applicants will be *rated* as more important for making a hiring decision compared to topics expressed more frequently for female applicants

Hypothesis 2: Topics expressed more frequently for male applicants will be *ranked* more frequently as most important and less frequently as least important compared to topics expressed more frequently for female applicants

These research questions and hypotheses will provide insight in to the subtle and complex ways that applicants are described that compound to produce perceptions of the most suitable applicant. If those attributes that are rated and ranked as more important are also those used more frequently to describe male applicants, then we will learn one way that women in the geosciences are disadvantaged at the evaluation stage and one reason why their proportion decreases as they make their way up the academic ladder.

CHAPTER 3: METHODS

Participants

To accurately capture results approximating the population, we estimated that we needed a sample size of 245 participants. This estimate was arrived at by setting a narrow 95% confidence interval and estimating a likely standard deviation, which allowed us to solve for the sample size that would achieve these parameters. As the items on our rating scale ranged between 1-9, we simulated a sample of 1000 observations that randomly chose either 1 or 9 on a single item. This sample produced a standard deviation of 4, the most extreme case possible. As we did not expect our participants to choose only 1 or 9 (e.g. 50% choosing 1 and 50% choosing 9) on any given item, we considered a less extreme standard deviation, 2 (e.g. 50% choosing 3 and 50% choosing 7), as something approaching what we would likely observe in our data. The width of the 95% confidence interval represents how closely a sample approximates the true population mean, where a narrower interval indicates a closer approximation of the population mean. Thus, to approximate the population mean accurately, we chose a narrow confidence interval; a width of 0.5. With the width of our confidence interval and estimated standard deviation, we were able to estimate the sample size that would capture enough data to approximate the population mean; 245.

We recruited participants from the online community of the American Geophysical Union (AGU) and directly from geoscience departments in American universities. To recruit through AGU, we posted our invitation on the main community forum on their site. Participants accessed the study through a link within the invitation on the forum. To recruit by email, we first developed a list of earth and geoscience programs at American universities broadly. We then visited each program's website, identified faculty members within the program, and then sent

them an email invitation to participate and encouraged them to share the invitation with members of their department. Participants recruited via email accessed the study through a link in the invitation. To take part in the study, participants had to be geoscience researchers (e.g. faculty, scientists, etc.) who had evaluated letters of recommendation for post-doctorate applicants in the past. Further, they had to be able to read and write in English. Sampling from AGU and geoscience departments is relevant as the topics used in the current study were derived from letters written for applicants to geoscience post-doctoral fellowships.

Two-hundred and sixty participants completed the questionnaire. Of those who finished, 17 had never used letters of recommendation for making a hiring decision. Of the 17, thirteen had either employed post-docs, made a hiring decision in the past, or used letters of recommendation in student admission decisions. The remaining four of the 17 participants had not done any of the above and were thus excluded from the sample for lacking relevant experience with hiring and/or evaluating letters of recommendation.

To ensure multiple questionnaires were not submitted by the same individual, we collected IP addresses associated with the device that each participant used to complete the questionnaire. We found 6 pairs of observations with the same IP address. One explanation for this is that more than one participant used the same device to take the questionnaire. For instance, it is possible that the questionnaire was taken on a lab computer, upon which multiple individuals participated. However, out of an abundance of caution we removed the second observation submitted by any IP address that appeared more than once in the data. If someone had completed the questionnaire twice, the first observation is likely more valid as the participant would have had less prior knowledge of the study and in turn would have provided more honest responses.

The final sample size was 250 participants. The demographic composition (see Table 2) included 166 (66%) men, 79 (32%) women, 1 (0.4%) transgender person (4 or 2% did not specify). The sample included 218 (87%) White or European Americans, 16 (6%) Asian or Asian Americans, 4 (2%) Hispanic or Latinos, 1 (0.4%) Black or African American, 1 (0.4%) Native American or American Indian, and 10 (4%) of another ethnicity. The sample ranged in age with 1 (0.4%) participant between 18-28 years old, 51 (20%) between 29-39, 62 (25%) between 40-49, 62 (25%) between 50-59, 52 (21%) 60-69, and 22 (9%) over 70 years old. The sample was comprised of 117 (47%) full professors, 55 (22%) assistant professors, 38 (15%) associate professors, 18 (7%) researchers, 11 (4%) professor emeriti, 4 (2%) adjunct professors, 5 (2%) in other roles, and 1 (0.4%) participant in a non-teaching position (1 or 0.4% did not specify). Two hundred and forty-three (97%) of the participants held a doctorate degree, 5 (2%) held a professional degree, 1 (0.4%) held a master's, and 1 (0.4%) held some other degree. In terms of time working in an academic institution, 15 (6%) participants had worked between 0-2 years, 31 (12%) had worked between 3-5 years, 14 (6%) worked between 6-8 years, 25 (10%) had worked between 9-11 years, and 164 (66%) had worked in an academic institution for 12 years or more (1 or 0.4% did not specify). Participants came from a range of fields in the earth sciences, including 85 (34%) from geology, 45 (18%) from geophysics, 33 (13%) from environmental science, 24 (10%) from atmospheric sciences, 23 (9%) from hydrology, 11 (4%) from space sciences, 10 (4%) from geography, 8 (3%) from glaciology, 7 (3%) from oceanography, and 4 (2%) who did not specify. Five (2%) worked in an academic department with 4-6 faculty members, 9 (4%) worked in a department with 7-9 faculty members, 27 (11%) worked in a department with 10-12 faculty members, 208 (83%) worked in an academic department with 13 faculty members or more, and 1 (0.4%) did not specify.

The question regarding a participant's country of employment was added mid-way through data collection. Thus, we were only able to collect partial data for this question. We observed 63 (25%) missing observations. Of those who reported their country of employment, 184 (74%) worked in the United States of America, 1 (0.4%) worked in Canada, 1 (0.4%) worked in the United Kingdom, and 1 (0.4%) worked in Australia. In terms of how they were recruited, 212 (85%) participants responded to an invitation by direct email recruitment or department announcement (which originated from direct email recruitment) and 38 (15%) responded to one of several announcements made on the American Geophysical Union's online community forum. Given that those who responded via direct email (i.e. 212 participants) came only from American universities, we have some confidence that at least 85% of the sample worked in the United States.

Table 2

Demographic Characteristics of Study Sample

Characteristic	Frequency	Percent
Gender		
Men	166	66.4%
Women	79	31.6%
Transgender	1	0.4%
N/A	4	1.6%
Ethnicity		
White or European American	218	87.2%
Asian or Asian American	16	6.4%
Hispanic or Latino	14	1.6%
Black or African American	1	0.4%
Native American or American Indian	1	0.4%
Age Range		
18-28	1	0.4%
29-39	51	20.4%
40-49	62	24.8%
50-59	62	24.8%
60-69	52	20.8%
70+	22	8.8%
Job Title		
Full Professor	117	46.8%
Assistant Professor	55	22.0%
Associate Professor	38	15.2%
Researcher	18	7.2%
Professor Emeriti	11	4.4%
Adjunct Professor	4	1.6%
Other	5	2.0%
Non-teaching role	1	2.0%
N/A	1	0.4%
Education		
Doctorate	243	97.2%
Professional	5	2.0%
Master's	1	0.4%
Other	1	0.4%
Time Working in Academia		
0-2 years	15	6.0%
3-5 years	31	12.4%
6-8 years	14	5.6%
9-11 years	25	10.0%
12+ years	164	65.6%
N/A	1	0.4%
Academic Field		

Geology	85	34.0%
Geophysics	45	18.0%
Environmental Science	33	13.2%
Atmospheric Sciences	24	9.6%
Hydrology	23	9.2%
Space Sciences	11	4.4%
Geography	10	4.0%
Glaciology	8	3.2%
Oceanography	7	2.8%
N/A	4	1.6%
Number Faculty in Dept		
4-6	5	2.0%
7-9	9	3.6%
10-12	27	10.8%
13+	208	83.2%
Country of Employment		
USA	184	73.6%
N/A	63	25.2%
Canada	1	0.4%
UK	1	0.4%
Australia	1	0.4%
Response Type		
Direct email/department announcement	212	84.8%
AGU board	38	15.0%

Note. $n = 250$.

Procedure

After clicking the email or website link, participants were provided information about the study and asked to give their informed consent. Once informed consent was given, participants were asked to read instructions which explained the purpose of the study (Appendix A). The instructions explained that we were interested in learning which topics used to describe applicants in letters of recommendation are important for making hiring decisions for post-doctoral fellowships in the geosciences. Further, we explained that we had compiled a representative list of topics used to describe applicants from a sample of letters of recommendation for applicants to post-doctoral fellowships in the geosciences. We then explained that as experts in the geosciences, we wanted them to examine the topics and rate and rank them in terms of importance in their decision to hire an applicant.

To give the participants an idea of the type of fellowships to which applicants applied, we asked them to first read a relevant job description for this position (Appendix B) before examining the topics. The job description was an adapted version of the announcement made by the program for which the recommendation letters were written. The reason for including this job description is that, we wanted to provide participants with the criteria that writers likely used to draft their letters, which made the topics relevant to this job description in particular. The announcement was relatively general as it invited applications from a variety of Earth science fields. The principal criteria for selection described in the announcement was scientific excellence and for Fellows to have a "clearly expressed plan to investigate problems at the forefront of Earth science."

Once participants read the job description, we asked them to review each of the 31 topics expressed across the letters of recommendation. With this step we sought to ensure that, prior to evaluating the topics, participants had become familiar with them. To this end, we asked participants to select a radio button next to each topic to indicate that they had reviewed each one (Appendix C).

We then asked them to complete a questionnaire where they rated each of the 31 topics in terms of importance for making a hiring decision (Appendix D). Next, we asked participants to choose and rank their top five most important and bottom five least important topics when making a hiring decision (Appendix E). To control for the order of topics impacting the way participants made ratings and rankings, the order was randomized between participants, but consistent within. That is, participants saw the same order of topics throughout the rating and ranking sections, but each participant was presented a unique random ordering of topics. This ensured that topic order would not impact our results. Following the rating and ranking tasks, we

asked participants how important letters of recommendation are in their overall decisions to hire (Appendix F). Last, we asked for demographic information (Appendix G). Following completion of the questionnaire, we provided participants with a debriefing statement where we explained the full purpose of the research and provided them information on how to get in touch with the lead researcher (Appendix H).

Materials

Formatting topics. To allow participants to rate and rank the topics given the context of the study, we formatted the topics to fit the phrasing of our questions. That is, when prompted to rate each topic, we asked participants, "in your decision to hire a post-doctorate fellow in the geosciences, how important would it be that...," followed by the list of topics and rating scales. To fit the format of this question we began each topic with the phrase "an applicant," sometimes followed by a verb to fit the topic (e.g. is, has, had, etc.). For example, the topic *hard working* became *an applicant is hard working*. Thus, to provide coherence between topic and inquiry the question for participants became "in your decision to hire a post-doctorate fellow in the geosciences, how important would it be that *an applicant is hard working*," and so on for each topic.

Addressing double-barreled nature of topics. Among the topics, we recognized that some of the topic labels from Elmore et al. (2019) were double-barreled (i.e. they contained more than one theme). For instance, *top student and teacher* captures two roles; student and teacher. Given this, it would be difficult for participants to accurately rate or rank the topic and in turn for us to know to which part of the topic participants were referring. While it would be ideal to separate these themes, the fact is roles sometimes co-occurred within letters for applicants. That is, Elmore et al. (2019) found that letter writers described applicants as both great teachers *and*

students. Therefore, while double-barreled, those evaluating applicants would, in general, be interpreting the content in combination as it was written in the letter. Thus, topics with multiple themes were realistic examples of content intended for evaluation. Further, it is a common feature of topic modeling to observe topics that express two or more tightly linked themes (Lucas, et al., 2015; Tvinnereim & Fløttum, 2015; Wang, Bowers, & Fikis, 2017). Thus, we kept the double-barreled topics with the understanding that this may have introduced some error in responses.

Measures

Ratings. When rating topics, participants were asked to rate the extent to which each topic would be important in their decision to hire an applicant. Topics were rated using a 9-point Likert scale with anchor words on each end, where selections closer to 1 indicated the topic is "not at all important" and selections closer to 9 indicated that the topic is "extremely important," (Appendix D).

Rankings. We asked each participant to choose five topics they considered the most important and five topics they considered the least important for making a decision to hire a post-doctoral applicant. The 31 topic options were presented in a list for each task (Appendix E).

Importance of letters of recommendation. Participants were asked to indicate the percent of their total decision to hire that they attribute to information provided in letters of recommendation. The rating scale included 11 radio buttons ranging from 0% to 100%. This question extended our research questions and hypothesis by putting into context the amount of weight that letters have in the overall evaluation process. In turn, this question helped us understand the implications of our findings (Appendix F).

Demographic information. We requested background information from participants including their ethnicity, gender, age, work status, how long they have worked in an academic institution, education, years since highest level of education, the frequency with which they make hiring decisions, whether they employ post-doctorates, the frequency with which they use letters of recommendation and for what purpose, the number of faculty in their program, their field, and the country in which they are employed (Appendix G).

CHAPTER 4: RESULTS

Attention Checks

To ensure participants were paying attention, we examined their pattern of ranking topics as well as their time to complete the questionnaire. Choosing five topics as the most important and five topics as the least important required participants to attend to opposite ends of the spectrum regarding criteria. As the topics were presented in the same order for both tasks it would be easy for someone not paying attention to simply choose the first five topics in the list. However, the nature of the task requires that choices from both lists be mutually exclusive. Thus, if we observed any overlap between the most and least important topic choices, then we could conclude that a participant was likely not paying attention. No participant picked the same topics for both ranking tasks, an indication that those in the sample were indeed paying attention. We also examined the time it took each participant to complete the study, imagining that an inordinately quick response time would signal a lack of thoughtful consideration toward the questionnaire. Prior to conducting the study, we estimated the time to complete the questionnaire would be between 10-15 minutes. After correcting for those who started and came back later, the average response time was 11 minutes. Further, no response was recorded in less than 5 minutes. These findings indicate that all participants spent a meaningful amount of time completing the questionnaire. Thus, we concluded that all participants in our sample were paying attention when completing the study.

Data Analysis

Research Questions

Research question 1 asked, what topics used to describe post-doctoral applicants in the geosciences are *rated* as most important by geoscience faculty and researchers? To explore our

first research question, we calculated the average rating for each topic across participants and examined the distribution of means. Table 3 provides an overview of the average ratings and standard deviations for each topic. Average topics ratings ranged between 3.13 and 7.72. The average rating across topics was 6.17 and the standard deviation across topics was 2.01. Thus, in our estimation of the required sample size, the use of 2 as our estimated standard deviation was found to be justified. The three topics rated highest on average in terms of importance for making a hiring decision were *presenting and publishing nationally & internationally*, *excels academically and professionally*, and *high potential academic and researcher who gets funding.* In general, among those topics rated 7 or higher, on average, we can see that research productivity, getting funding, getting results, and generally being active as a researcher are themes rated highest in terms of importance for making a hiring decision. Alternatively, we can see that the three topics rated lowest in terms of importance for making a hiring decision were *writer served on thesis committee*, *highly capable engineer*, and *engaged in environmental management science*. In general, among those topics rated below 6, on average, we can see that being technically skilled, showing improvement, being involved both interpersonally or with their department, and general interests or success were themes rated lowest in terms of importance for making a hiring decision.

Table 3

Average Ratings of Topic Importance for Making a Hiring Decision

Topic Label	Average Rating
	Mean (SD)
Presenting and Publishing Nationally & Internationally	7.72 (1.49)
Excels Academically and Professionally	7.58 (1.53)
High Potential Academic and Researcher who Gets Funding	7.38 (1.65)
Submitting Manuscripts	7.32 (1.70)
Publishing Frequently	7.30 (1.53)
Asks Insightful Questions	7.29 (1.58)
Research Contributor Now & Future	7.26 (1.89)
Achieves Scientific Results	7.25 (2.07)
Outstanding Young Researcher	7.22 (1.90)
Strong Recommendation	6.98 (2.11)
Tackles Research Problems	6.95 (2.19)
Works Well and Quickly on Tasks	6.94 (1.66)
Hard Working	6.89 (1.94)
Dedicated to Research	6.83 (1.92)
Understands Complex Systems	6.68 (1.81)
Will Be Good Fit	6.44 (2.21)
Pleasant Team Member	6.35 (1.93)
Computer Programming Skills	6.19 (1.96)
Uses and Develops Methods	6.14 (2.11)
Skilled Experimentalist	5.86 (2.19)
Top Student and Teacher	5.79 (2.00)
Successful Thesis Defense	5.55 (2.75)
Making Progress and Maturing	5.44 (2.34)
Develops Models	5.35 (1.95)
Writer Supports Application	5.32 (2.83)
Accomplished Teacher and Department Citizen	5.27 (1.97)
Improving Their English	4.69 (2.17)
Capable Technician	4.41 (2.17)
Writer Served on Thesis Committee	4.00 (2.46)
Highly Capable Engineer	3.74 (2.17)
Engaged in Environmental Management Science	3.13 (2.17)

Note. $n = 250$. Higher ratings indicate that the topic was, on average, considered more important for making a hiring decision.

Research question 2 asked, what topics used to describe post-doctoral applicants in the geosciences are *ranked* as *most important* by geoscience faculty and researchers? To explore our second research question, we examined the frequency with which participants chose each topic as one of their five most important for making a hiring decision. Table 4 provides an overview of

the frequency with which topics were chosen and the percent of the sample who chose each topic as most important. Topics participants chose as one of their five most important for making a hiring decision ranged in frequency from 0 to 129. Interestingly, the topics ranked most frequently as most important were the same as those that received the highest average ratings of importance. Again, they were topics explaining that an applicant was *presenting and publishing nationally & internationally*, that the applicant *excels academically and professionally*, or that they are a *high potential academic and researcher who gets funding*. In general, topics that described applicants as publishing, being productive, or as promising researchers were ranked as most important for making a hiring decision.

Table 4

Ranking Frequency of Topics Considered Most Important

Topic Label	Ranked Most Important Freq (% of sample who chose topic)
Presenting and Publishing Nationally & Internationally	129 (51.6%)
High Potential Academic and Researcher who Gets Funding	98 (39.2%)
Excels Academically and Professionally	90 (36.0%)
Outstanding Young Researcher	86 (34.4%)
Asks Insightful Questions	79 (31.6%)
Hard Working	66 (26.4%)
Strong Recommendation	65 (26.0%)
Achieves Scientific Results	64 (25.6%)
Publishing Frequently	64 (25.6%)
Will Be Good Fit	58 (23.2%)
Pleasant Team Member	56 (22.4%)
Submitting Manuscripts	52 (20.8%)
Works Well and Quickly on Tasks	51 (20.4%)
Understands Complex Systems	36 (14.4%)
Dedicated to Research	34 (13.6%)
Computer Programming Skills	34 (13.6%)
Research Contributor Now & Future	32 (12.8%)
Skilled Experimentalist	28 (11.2%)
Tackles Research Problems	27 (10.8%)
Uses and Develops Methods	20 (8.0%)
Accomplished Teacher and Department Citizen	20 (8.0%)
Writer Supports Application	13 (5.2%)
Top Student and Teacher	11 (4.4%)
Making Progress and Maturing	11 (4.4%)
Successful Thesis Defense	9 (3.6%)
Develops Models	8 (3.2%)
Writer Served on Thesis Committee	4 (1.6%)
Improving Their English	3 (1.2%)
Capable Technician	1 (0.4%)
Highly Capable Engineer	1 (0.4%)
Engaged in Environmental Management Science	0 (0.0%)

Note. $n = 250$. Percentages in parentheses indicate the percent of the sample who chose to rank that particular topic as most important. Each participant chose 5 topics.

Research question 3 asked what topics used to describe post-doctoral applicants in the geosciences are *ranked* as *least important* by the geoscience faculty and researchers? To explore this question, we examined the frequency with which participants chose each topic as one of

their five least important for making a hiring decision. Table 5 provides an overview of the frequency with which topics were chosen and the percent of the sample who chose each topic as least important. Topics participants chose as one of the five least important for making a hiring decision ranged in frequency from 2 to 127. The topics chosen most frequently as least important were the same as those that received the lowest average ratings of importance. Again, they were topics explaining that an applicant's *writer served on thesis committee*, that they were a *highly capable engineer*, or that they were *engaged in environmental management science*. In general, topics that described applicants as being engaged in their field or department, having technical skills, improving their skills, or referenced a relationship with the letter writer were ranked as least important for making a hiring decision.

Table 5

Ranking Frequency of Topics Considered Least Important

Topic Label	Ranked Least Important
	Freq (% of sample who chose topic)
Engaged in Environmental Management Science	127 (50.8%)
Highly Capable Engineer	115 (46.0%)
Writer Served on Thesis Committee	112 (44.8%)
Improving Their English	108 (43.2%)
Capable Technician	107 (42.8%)
Successful Thesis Defense	76 (30.4%)
Writer Supports Application	63 (25.2%)
Develops Models	60 (24.0%)
Making Progress and Maturing	59 (23.6%)
Accomplished Teacher and Department Citizen	55 (22.0%)
Top Student and Teacher	51 (20.4%)
Skilled Experimentalist	43 (17.2%)
Pleasant Team Member	38 (15.2%)
Will Be Good Fit	37 (14.8%)
Uses and Develops Methods	29 (11.6%)
Computer Programming Skills	25 (10.0%)
Dedicated to Research	16 (6.4%)
Tackles Research Problems	15 (6.0%)
Hard Working	15 (6.0%)
Publishing Frequently	13 (5.2%)
Strong Recommendation	13 (5.2%)
Understands Complex Systems	12 (4.8%)
Research Contributor Now & Future	12 (4.8%)
Achieves Scientific Results	9 (3.6%)
Works Well and Quickly on Tasks	8 (3.2%)
Asks Insightful Questions	7 (2.8%)
High Potential Academic and Researcher who Gets Funding	7 (2.8%)
Submitting Manuscripts	7 (2.8%)
Outstanding Young Researcher	5 (2.0%)
Presenting and Publishing Nationally & Internationally	4 (1.6%)
Excels Academically and Professionally	2 (0.8%)

Note. $n = 250$. Percentages in parentheses indicate the percent of the sample who chose to rank that particular topic as least important. Each participant chose 5 topics.

The consistency in findings across Research Questions 1-3 allowed us to extract general themes from the pattern of topic ratings and rankings. In general, study participants considered research productivity (e.g. publishing, making contributions, ideation, etc.) of increasing

importance, practical experience and motivation (e.g. task work, dedication, etc.) of medium importance and themes of capabilities, development, and relationship with the letter writer (e.g. application support, personal development, etc.) of lesser importance for making a hiring decision.

Operationalizing Independent and Dependent Variables

To evaluate our hypotheses, we incorporated the findings from Elmore et al. (2019) which specified the extent to which each topic was expressed for male and female applicants. These data informed out independent variable, *gendered topics*. That is, for each hypothesis test, we relied upon the findings from Elmore et al. (2019) to specify how much topics were expressed for one applicant gender over the other. This allowed us to examine if our dependent variable, topic ratings and rankings, were related to the extent to which topics were gendered in Elmore et al. (2019). Thus, in the present study, Elmore's et al. (2019) model provided the basis for our independent variables and the findings from our questionnaire provide the basis for our dependent variables. To explore the relationship between *gendered topics* and participant topic ratings (hypothesis 1) and rankings (hypothesis 2), we tested several variations of our independent and dependent variables. Specifically, we created three different operationalizations of *gendered topics*; 1) all topics split into either male or female categories, 2) highly gendered topics split into either male or female categories, and 3) the relative frequency that all topics were expressed in letters for both male and female applicants. Each of these variations were specified in sub-hypotheses of hypothesis 1 with ratings as the dependent variable (1a-c) and hypotheses 2 with rankings as the dependent variable (2a-d). Each hypothesis will be reviewed after each section where we operationalize the variations of our independent variables.

Independent variable operationalization 1: all gendered topics. In Table 6, one can observe topics that appeared more frequently in female letters and topics that appeared more frequently in male letters from Elmore et al. (2019). Positive percentages indicate topics appeared that much more in letters for female applicants and negative percentages indicate topics appeared that much more in letters for male applicants. For example, among the 15 topics expressed more frequently in letters for female applicants, the topic *understands complex systems* was expressed 18.18% more often in female letters compared to male letters. Alternatively, among the 16 topics expressed more frequently in letters for male applicants, the topic *computer programming skills* was expressed 29.69% more often in male letters compared to female letters.[4]

[4] In this operationalization, we calculated percentages using the proportion with which each topic was expressed among the 53 topics generated by Elmore et al. (2019). Percentages were calculated by dividing the proportion with which each topic was expressed in letters for female and male applicants, respectively, by the total proportion with which the topic was expressed in letters for both male and female applicants, then taking the difference of the two. For example, the proportion with which the topic *understands complex systems* was 0.020 for female letters and 0.014 for male letters. We summed these proportions to get a total of 0.034. We then divided each gender proportion by the total arriving at 0.020/0.034 = 0.59 for female applicants and 0.014/0.034 = 0.41 for male applicants. Last, we took the difference of these two numbers, which revealed that for the topic *understands complex systems* was expressed 0.59-0.41 = 0.18 or 18% more frequently in letters for female applicants.

Table 6

Topic Expression in Letters for Female vs. Male Applicants

Topic	Percent Expression Female vs. Male	Topic Expressed More For
Understands Complex Systems	18.18%	Female Applicants
Top Student and Teacher	13.75%	Female Applicants
Excels Academically and Professionally	8.69%	Female Applicants
Writer Supports Application	7.76%	Female Applicants
Develops Models	7.51%	Female Applicants
Achieves Scientific Results	7.25%	Female Applicants
Tackles Research Problems	6.57%	Female Applicants
Engaged in Environmental Management Science	6.23%	Female Applicants
Writer Served on Thesis Committee	5.79%	Female Applicants
Research Contributor Now & Future	5.13%	Female Applicants
Uses and Develops Methods	3.64%	Female Applicants
Making Progress and Maturing	3.57%	Female Applicants
Works Well and Quickly on Tasks	1.33%	Female Applicants
Improving Their English	0.58%	Female Applicants
Successful Thesis Defense	0.30%	Female Applicants
Highly Capable Engineer	-0.22%	Male Applicants
Asks Insightful Questions	-0.36%	Male Applicants
Accomplished Teacher and Department Citizen	-0.99%	Male Applicants
High Potential Academic and Researcher who Gets Funding	-1.32%	Male Applicants
Presenting and Publishing Nationally & Internationally	-1.93%	Male Applicants
Outstanding Young Researcher	-1.93%	Male Applicants
Will Be Good Fit	-2.70%	Male Applicants
Hard Working	-3.24%	Male Applicants
Pleasant Team Member	-9.78%	Male Applicants
Submitting Manuscripts	-10.86%	Male Applicants
Capable Technician	-11.73%	Male Applicants
Dedicated to Research	-12.34%	Male Applicants
Skilled Experimentalist	-15.31%	Male Applicants
Publishing Frequently	-17.42%	Male Applicants
Strong Recommendation	-17.54%	Male Applicants
Computer Programming Skills	-29.69%	Male Applicants

Note. The extent to which a topic was expressed in letters for female applicants versus male applicants. Topics with positive percentages were that much more likely to show up in letters for female applicants compared to letters for male applicants. Topics with negative percentages were that much more likely show up in letters for male applicants compared to female applicants. Based on topic proportions from Elmore et al. (2019).

This split between positive and negative percentages allowed us to operationalize our first independent variable, which had two levels, 1) *female topics*, which included the 15 topics that were expressed more in letters for female applicants (compared to male applicants), and 2) *male topics*, which included the 16 topics that were expressed more in letters for male applicants (compared to female applicants). Thus, using this operationalization of our independent variable, we made two predictions based on hypothesis 1 and 2,

1a. *Male topics* will be rated as more important in letters of recommendation than *female topics*

2a. *Male topics* will be ranked more as most important and less as least important in letters of recommendation compared to the *female topics*

Calculating the dependent variable for hypothesis 1a. To calculate the dependent variable for hypothesis 1a, we took the average of participant ratings across the *female topics* and the average of participant ratings across the *male topics*, creating two composite scores; the *female topics composite* and the *male topics composite*.

Calculating the dependent variable for hypothesis 2a. To calculate the dependent variable for hypothesis 2a, we calculated the average frequency with which the *male topics* and *female topics* were ranked as one of a participant's five most or five least important topics. Accordingly, we took four averages, 1) the average frequency that the *male topics* were ranked most important, 2) the average frequency that the *male topics* were ranked least important, 3) the average frequency that the *female topics* were ranked most important, and 4) the average frequency that the *female topics* were ranked least important.

Independent variable operationalization 2: highly gendered topics. It could be argued that topics split at the center of the distribution in Table 6, which have very small percent

differences, should not be included in a gender category as they are not very gendered. To address this concern, we developed a second operationalization of our independent variable where we specifically grouped the five topics that were expressed most in letters for either gender. Table 6 shows the topics *understands complex systems, top student and teacher, excels academically and professionally, writer supports applicants,* and *develops models,* were expressed most in letters for female applicants compared to male applicants and the topics *dedicated to research, skilled experimentalist, publishing frequently, strong recommendation,* and *computer programming skills* were expressed most in letters for male applicants compared to female applicants. We used these highly gendered topics to create the second operationalization of our independent variable, which had two levels, 1) the five topics expressed most in letters for female applicants, which we called the *top 5 female topics* and 2) the five topics expressed most in letters for male applicants, which we called the *top 5 male topics*. Thus, using this operationalization of our independent variable, we made two predictions based on hypothesis 1 and 2,

1b. The *top 5 male topics* will be rated as more important in letters of recommendation than *top 5 female topics*

2b. The *top 5 male topics* will be ranked more as most important and less as least important in letters of recommendation compared to *top 5 female topics*

Calculating the dependent variable for hypothesis 1b. To calculate the dependent variable for hypothesis 1b, we took the average of the ratings across the *top 5 female topics* and an average of the ratings across the *top 5 male topics*, creating two composite scores; the *top 5 female composite* and the *top 5 male composite*.

Calculating the dependent variable for hypothesis 2b. To create the dependent variable for hypothesis 2b, we calculated the average frequency with which the *top 5 female topics* and *top 5 male topics* were ranked as one of a participant's five most or five least important topics. Accordingly, we took four averages, 1) the average frequency that the *top 5 male topics* were ranked most important, 2) the average frequency that the *top 5 male topics* were ranked least important, 3) the average frequency that the *top 5 female topics* were ranked most important, and 4) the average frequency that the *top 5 female topics* were ranked least important.

Independent variable operationalization 3: relative frequency of gendered topics. In our final operationalization of *gendered topics*, we considered topic gender proportions directly.[5] For instance, rather than the topic *top student and teacher* being considered as only a female topic, as was done in our first two operationalizations, in this operationalization we accounted for the fact that it was expressed 5.67% of the time among the 31 topics expressed in letters written for female applicants and 4.27% of the time among the 31 topics expressed in letters written for male applicants (see Table 7).[6] This allowed us to account for the fact that each topic was expressed in some proportion for *both genders,* not just for one gender.

[5] To properly operationalize our independent variable, we had to modify the gender proportions from Elmore et al. (2019) to accommodate our research inquiry. The model developed by Elmore et al. (2019) accounted for all of the data in the letters, which generated 53 topics. However, in our study, we were only interested in the 31 topics that described applicants directly. As the 53 topics from Elmore et al. (2019) accounted for all the data in letters for male and female applicants, the proportions of these topics summed to one for both genders. That is, the 53 topics they generated in their model accounted for 100% of the data across female letters and 100% of the data across male letters, in varying proportions for either gender. As we only asked participants to rate and rank 31 topics in the present study, the proportional estimates for these topics did not sum to one for either gender. Indeed, the sum of the average estimated topic proportions for these 31 topics accounted for 61.9% of the content from female letters and 62.1% of the content for male letters. To allow these 31 topics to account for 100% of the letter content for the purposes of our study, we scaled them to 1 by dividing each topic proportion by the sum of the 31 topic proportions. For example, across the 53 topics *top student and teacher* accounted for 2.6% of the corpus for male applicants. When scaled, this number became 2.6/62.1 = 4.2%, which represented a proportion of .042 of the 1 that made up all 31 topic proportions for male applicants. This ensured that we accounted for the relative frequency of each topic among the other 31 topics and allowed us to accurately specify the nuance in relative gender proportions in our hypothesis testing.

[6] Because Table 7 reflects scaled proportions, the gender ratios presented in Table 6 do not exactly correspond to the percentages in Table 7. This is because operationalization 1 and 2 consider gender proportions differently. In Table 7, we use scaled proportions, which account for gender differences among the 31 topics, and in Table 6 we use non-

Table 7

Topic Proportions by Applicant Gender

Topic	Percent in Female Letters	Percent in Male Letters
Understands Complex Systems	3.27%	2.25%
Top Student and Teacher	5.67%	4.27%
Excels Academically and Professionally	5.03%	4.20%
Writer Supports Application	4.91%	4.18%
Develops Models	3.57%	3.05%
Achieves Scientific Results	4.00%	3.43%
Tackles Research Problems	3.98%	3.47%
Engaged in Environmental Management Science	3.33%	2.93%
Writer Served on Thesis Committee	2.47%	2.18%
Research Contributor Now & Future	2.11%	1.89%
Uses and Develops Methods	2.57%	2.37%
Making Progress and Maturing	2.45%	2.27%
Works Well and Quickly on Tasks	3.59%	3.47%
Improving Their English	3.19%	3.14%
Successful Thesis Defense	3.22%	3.18%
Highly Capable Engineer	3.67%	3.66%
Asks Insightful Questions	3.37%	3.38%
Accomplished Teacher and Department Citizen	2.89%	2.93%
High Potential Academic and Researcher who Gets Funding	3.54%	3.62%
Outstanding Young Researcher	3.77%	3.89%
Presenting and Publishing Nationally & Internationally	3.34%	3.45%
Will Be Good Fit	3.47%	3.64%
Hard Working	4.59%	4.87%
Pleasant Team Member	0.97%	1.17%
Submitting Manuscripts	2.26%	2.79%
Capable Technician	3.20%	4.03%
Dedicated to Research	3.11%	3.96%
Skilled Experimentalist	2.82%	3.81%
Publishing Frequently	2.67%	3.77%
Strong Recommendation	1.70%	2.41%
Computer Programming Skills	1.28%	2.34%

Note. Percentages represent the proportion with which each topic was expressed in letters for female applicants and male applicants. Higher percentages indicate that a topic was expressed more frequently across letters for either gender. Based on topic proportions from Elmore et al. (2019).

scaled proportions to calculate the ratios. We used non-scaled proportions in Table 6 because they provided an exact accounting of whether each particular topic was either male or female in Elmore et al. (2019) (i.e. IV operationalization 1). We used scaled proportions in Table 7 as they indicate the relative frequency that topics were expressed within applicant gender among the 31 topics (i.e. IV operationalization 2).

This operationalization had two levels, 1) *female topic proportions*, which accounted for the proportion with which each topic was expressed in letters for female applicants, and 2) *male topic proportions*, which accounted for the proportion with which each topic was expressed in letters for male applicants. As this variable incorporates gender proportions for all topics for both male and female applicants rather than splitting them into two groups, it was the most conservative test of our hypotheses. Thus, using this operationalization of our independent variable, we made three predictions based on hypothesis 1 and 2,

1c. Topics with higher *male topic proportions* will be rated as more important in letters of recommendation than topics with higher *female topic proportions*

2c. Topics with higher *male topic proportions* will be *ranked more* as *most important* in letters of recommendation compared to topics with higher *female topic proportions*

2d. Topics with higher *female topic proportions* will be *ranked more* as *least important* in letters of recommendation compared to topics with higher *male topic proportions*

Calculating the dependent variable for hypothesis 1c. To create the dependent variable for hypothesis 1c, we weighted each participant's topic ratings by *female topic proportions* and *male topic proportions*. Consider a hypothetical participant who rated the topic *top student and teacher* 6 on the 1-9 scale in terms of importance for making a hiring decision. The *female proportion* for this topic was 5.6% and the *male proportion* was 4.2%. Using these gender proportions, we weighted this participant's rating, thereby accounting for gender differences in topic expression. Thus, for this hypothetical participant, the rating of 6 became 6*.057 = 0.34 for female applicants and 6*.043 = 0.26 for male applicants. From this weighting procedure, one can observe that the rating of 6 means more for female applicants because *top student and teacher* was expressed more in their letters compared to male letters. Alternatively, consider that our

hypothetical participant rated the topic *publishing frequently* with a score of 8 in terms of importance in making a hiring decision. The *female proportion* for this topic was 2.7% and the *male proportion* was 3.8%. When we weighted this rating by these frequencies, the participant's rating became 8*.027 = 0.22 for female applicants and 8*.038 = 0.30 for male applicants. Here, one can observe that the rating of 8 means more for male applicants because *publishing frequently* was expressed more in their letters compared to female letters. Continuing this procedure, we weighted each participant's ratings by the gender topic proportions for each topic. As we observed with our hypothetical participant, this created two scores for each topic, one based on *female topic proportions* and one based on the *male topic proportion*. We then took the sum of ratings weighted by *male* and *female topic proportions* to calculate our dependent variable, which included two scores for each participant, 1) the *weighted female* score, which was the sum of ratings weighted by *female topic proportions*, and 2) the *weighted male* score, which was the sum of ratings weighted by *male topic proportions*. With this calculation, these two sum scores captured both the importance that participants placed on each topic as well as the extent to which the topics were expressed in letters for male and female applicants.[7]

Calculating the dependent variable for hypothesis 2c. To create the dependent variable hypothesis 2c, we examined the five topics each participant ranked as most important and calculated the sum of *female topic proportions* and *male topic proportions* for those five topics. Consider a hypothetical participant who chose as their most important topics *writer served on thesis committee, works well and quickly on tasks, excels academically and professionally, publishing frequently,* and *capable technician*. The *female topic proportions* based on Elmore et al. (2019) for these 5 topics were .025, .036, .050, .027, .032 out of 1, respectively, and the *male*

[7] As differences between gender topic proportions existed in the hundredths and thousandths place, we report to the thousandths place on tests that used this operationalization as the independent variable.

topic proportions for these topics were .022, .035, .042, .038, .040 out of 1, respectively. When summed together, these five topics occurred 17% of the time in letters for female applicants and 17.7% of the time in letters for male applicants. Thus, for this particular participant, the five topics they chose as most important were expressed more frequently (difference = 0.7%) in letters for male applicants compared to female applicants. Following this methodology, we calculated our dependent variable, which included two scores for each participant, 1) the *MI (Most Important) female* score, which was the sum of the *female topic proportions* for topics ranked as most important, and 2) the *MI male* score, which was the sum of the *male topic proportions* for topics ranked as most important.

Calculating the dependent variable for hypothesis 2d. To create the dependent variable hypothesis 2d, we performed the same procedure as was done for the dependent variable of hypothesis 2c. However, for this calculation, we used the five topics ranked by participants as *least* important and calculated the sum of *female topic proportions* and *male topic proportions* for those five topics. Following this methodology, we created two scores for each participant, 1) the *LI (Least Important) female* score, which was the sum of the *female topic proportions* for topics ranked as least important, and 2) the *LI male* score, which was the sum of the *male topic proportions* for topics ranked as least important.

Hypothesis Testing

Test of hypothesis 1a. To learn if the *female topics composites* and *male topics composites* were approximately normally distributed, we ran a Shapiro-Wilk test.[8] We found significant p-values for the *female topics composite* variable (p = .027), as well as the *male topics*

[8] To ensure our data was approximately normally distributed, we used Shapiro-Wilk tests for normality on all of our dependent variables prior to each hypothesis test. A significant p-value for the Shapiro-Wilk test indicates that assumptions are violated in terms of the data being normally distributed. We followed up on any significant Shapiro-Wilk p-value with z-tests of skewness and kurtosis and outlier analysis to verify results.

composite variable (p = .038). To verify that these composite variables were indeed not normally distributed, we followed up with a z-test using the skew and kurtosis of each variable's distribution.[9] For samples between 50-300 observations, Kim (2013) recommends rejecting the null hypothesis of normality over absolute z-scores of 3.29. The *female topics composite* had an absolute skewness z-score of 1.45 and an absolute kurtosis z-score of 1.56, whereas the *male topics composite* had an absolute skewness z-score of 2.37 and an absolute kurtosis z-score of 0.64, all of which was below the 3.29 threshold that Kim (2013) recommends. Given the inconclusive results regarding normality, we proceeded with our statistical test.

As ratings for *male topics* and *female topics* came from the same participant, the design was within-subjects, thus, to test hypothesis 1a, we ran a paired t-test comparing the means of the *female topics composite* and the *male topics composite.* Table 8 provides an overview of composite means as well as a test of the difference between these two means. The difference between the average *female topics composite* (M = 5.87, SD = 1.23) and the average *male topics composite* (M = 6.45, SD = 0.97) was significant $t(249) = -12.45$, $p < .000$, indicating that the topics expressed more in letters for male applicants were rated higher in terms of importance for making a hiring decision than the topics expressed more in letters for female applicants. Thus, hypothesis 1a was confirmed. Following this significant result, we examined the effect size of the difference score. For hypothesis 1a, we found a Cohen's D of 0.79, which indicates a large effect size.[10]

[9] Z-scores were calculated by dividing the skewness and kurtosis values by their respective standard errors.
[10] Effect size for a paired t-test is calculated by dividing the average difference between means by the standard deviation of that difference, where 0.2 indicates a small effect size, 0.4 indicates a medium effect size, and 0.8 indicates a large effect size (Cohen, 1988).

Table 8

Hypothesis 1a Results: Means, Standard Deviations, & Paired t-test for Gender Composite

	Independent Variable						
	Female Topics	Male Topics					
Dependent Variable	Mean (SD)	Mean (SD)	Diff	SD Diff	t-value	sig.	*d*
Gender Composite	5.87 (1.23)	6.45 (0.97)	-0.58	0.73	-12.45***	.000	0.79

Note. $n = 250$. Diff = difference between mean values for the female and male topic composites, SD Diff = the standard deviation of the difference scores, *d* = effect size the paired t-test using Cohen's D (1988). Average composite ratings on 9-point Likert scale, where the higher the mean, the more important topics were considered for making a hiring decision.
***$p < .001$.

Test of hypothesis 1b. To learn if the *top 5 female composites* and *top 5 male composites* were approximately normally distributed, we ran a Shapiro-Wilk test. We found significant p-values for the *top 5 female composite* variable (p = .010), and *top 5 male composite* variable (p = .000). To verify that these composite variables were indeed not normally distributed we followed up with a z-test using the skew and kurtosis of each variable's distribution. Again, for samples with 50-300 observations, Kim (2013) recommends rejecting the null hypothesis of normality over absolute z-scores of 3.29. The *top 5 female composite* had an absolute skewness z-score of 0.83 and an absolute kurtosis z-score of 2.31, whereas the *top 5 male composite* had an absolute skewness z-score of 4.05 and an absolute kurtosis z-score of 2.77. As the *top 5 male composite* was above the 3.29 threshold, we further examined this variable for potential outliers. The *top 5 male composite* included scores that ranged between 2-9 on the 1-9 Likert scale. However, the bottom quartile range had a minimum of 4. Thus, scores below 4 lay outside the bottom quartile, making them potential outliers. Three cases had composite scores below 4, and the most extreme of the three cases was a composite score of 2. To limit discretionary case omission, we focused on the single most extreme case. After omitting the composite score of 2 from our analysis, we

observed an absolute skewness z-score of 2.79 down from 4.05. Thus, the single case had caused the distribution of scores for the *top 5 male composite* to appear non-normally distributed. We therefore omitted the single case of 2 from our analysis of the *most gendered ratings*. Given the above results, we proceeded with our analyses with general confidence that normality assumptions were not violated.

As topic ratings for both groupings came from the same participant, the design was within-subjects, thus, to test hypothesis 1b, we ran a paired t-test comparing the means of the *top 5 female topics composite* and the *top 5 male topics composite.* Table 9 provides an overview of composite means as well as a test of the difference between these two means. The difference between the average *top 5 female composite* ($M = 6.16$, $SD = 1.24$) and the average *top 5 male composite* ($M = 6.65$, $SD = 1.09$) was significant $t(248) = -7.75$, $p < .000$, indicating that the five topics expressed most in letters for male applicants were rated higher in terms of importance for making a hiring decision than the five topics expressed most in letters for female applicants. Thus, hypothesis 1b was confirmed. Following this significant result, we examined the effect size of the difference score and found a Cohen's D of 0.49, which indicates a medium effect size.

Table 9

Hypothesis 1b Results: Means, Standard Deviations, & Paired t-test for Top Five Gender Composite

	Independent Variable						
	Top 5 Female Topics	Top 5 Male Topics					
Dependent Variable	Mean (SD)	Mean (SD)	Diff	SD Diff	t-value	sig.	*d*
Top 5 Gender Composite	6.16 (1.24)	6.65 (1.09)	-0.50	1.01	-7.75***	.000	0.49

Note. As one outlying observation for the top 5 male composite was excluded from the analysis, $n = 249$. Diff = difference between mean values for the top 5 female and top 5 male composites, SD Diff = the standard deviation of the difference scores, *d* = effect size the paired t-test using Cohen's D (1988). Average composite ratings on 9-point Likert scale, where the higher the mean, the more important topics in that grouping were considered for making a hiring decision. ***$p < .001$.

Test of hypothesis 1c. To learn if the *weighted female* and *weighted male* scores were approximately normally distributed, we ran a Shapiro-Wilk test. The Shapiro-Wilk test for normality was not significant for the *weighted female* variable (p = .104) or *weighted male* variable (p = .113). Thus, we concluded that these variables were normally distributed and proceeded with our analysis.

As both sums of scores were derived from the same participant, the design was within-subjects, thus, we ran a paired t-test comparing the means of the *weighted female* and the *weighted male* ratings. Table 10 provides an overview of weighted means as well as a test of the difference between these two means. The difference between the average *weighted female* ($M = 6.177$, $SD = 1.046$) and the average *weighted male* ratings ($M = 6.202$, $SD = 1.033$) was significant $t(249) = -6.95$, $p < .000$, indicating that topics expressed in higher proportions in male letters were also rated as more important for making a hiring decision compared to topics expressed in higher proportions in letters for female applicants. Thus, hypothesis 1c was

confirmed. Following this significant result, we examined the effect size of the difference score and found a Cohen's D of 0.44, which indicates a medium effect size.

Table 10

Hypothesis 1c Results: Means, Standard Deviations, & Paired t-test for Weighted Ratings

	Independent Variable						
	Female Topic Proportion	Male Topic Proportion					
Dependent Variable	Mean (SD)	Mean (SD)	Diff	SD Diff	t-value	sig.	*d*
Weighted Ratings	6.177 (1.046)	6.202 (1.033)	-0.025	0.058	-6.95***	.000	0.44

Note. $n = 250$. Diff = difference between mean values for the weighted female and weighted male scores, SD Diff = the standard deviation of the difference, d = effect size the paired t-test using Cohen's D (1988). Higher average means indicate that topics rated higher on the 1-9 Likert scale had larger relative gender proportions in letters.
***$p < .001$.

Examination of hypothesis 2a. As we asked participants to rank five topics as most important and five topics as least important, it is possible that participants chose topics within several categories, which would threaten the independence of observations between categories. Thus, we did not perform an inferential test on hypothesis 2a. Instead, we provide a descriptive overview of the differences in average frequencies that *gendered topics* were ranked as most and least important. Also, as the dependent variable for hypothesis 2a focused specifically on ranking frequencies, the variable was categorical and did not have a distribution. Thus, we did not perform tests of normality on this variable. Figure 2 provides an illustration of the average frequency with which *female* and *male topics* were ranked as least and most important. *Female topics* were ranked most important 25 times, on average, whereas the *male topics* were ranked most important 54 times, on average. Alternatively, *female topics* were ranked least important 50 times, on average, whereas the *male topics* were ranked least important 32 times, on average.

Thus, one can observe that, in general, the *male topics* were ranked more as most important and less as least important compared to the *female topics*.

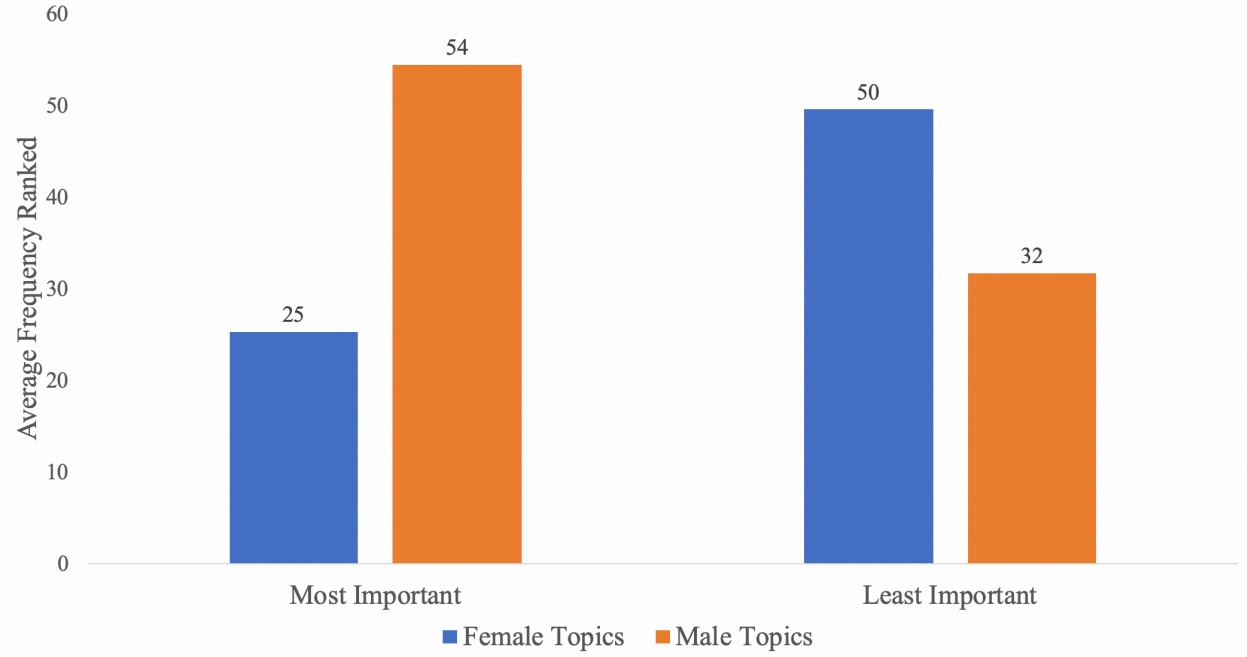

Figure 2. Average Frequency Female and Male Topics Ranked Most and Least Important. *Note.* Higher bars indicate higher average endorsement of topics as most or least important.

Examination of hypothesis 2b. Again, as we asked participants to rank five topics as most important and five topics as least important, it is possible that participants chose topics within several categories, which would threaten the independence of observations between categories. Thus, we did not perform an inferential test on hypothesis 2b. Instead, we provide a descriptive overview of the differences in average frequencies that the *top 5 female* and *top 5 male topics* were ranked as most and least important. Also, as the dependent variable for hypothesis 2b focused specifically on ranking frequencies, the variable was categorical and did not have a distribution. Thus, we did not perform tests of normality on this variable. Figure 3 provides an illustration of the frequency with which the *top 5 female* and *top 5 male topics* were ranked as least and most important. The *top 5 female topics* were ranked most important 32 times, on average, whereas the *top 5 male topics* were ranked most important 45 times, on average.

Alternatively, the *top 5 female topics* were ranked least important 38 times, on average, whereas the *top 5 male topics* were ranked least important 22 times, on average. Thus, one can observe that, in general, the *top 5 male topics* were ranked more as most important and less as least important compared to the *top 5 female topics*.

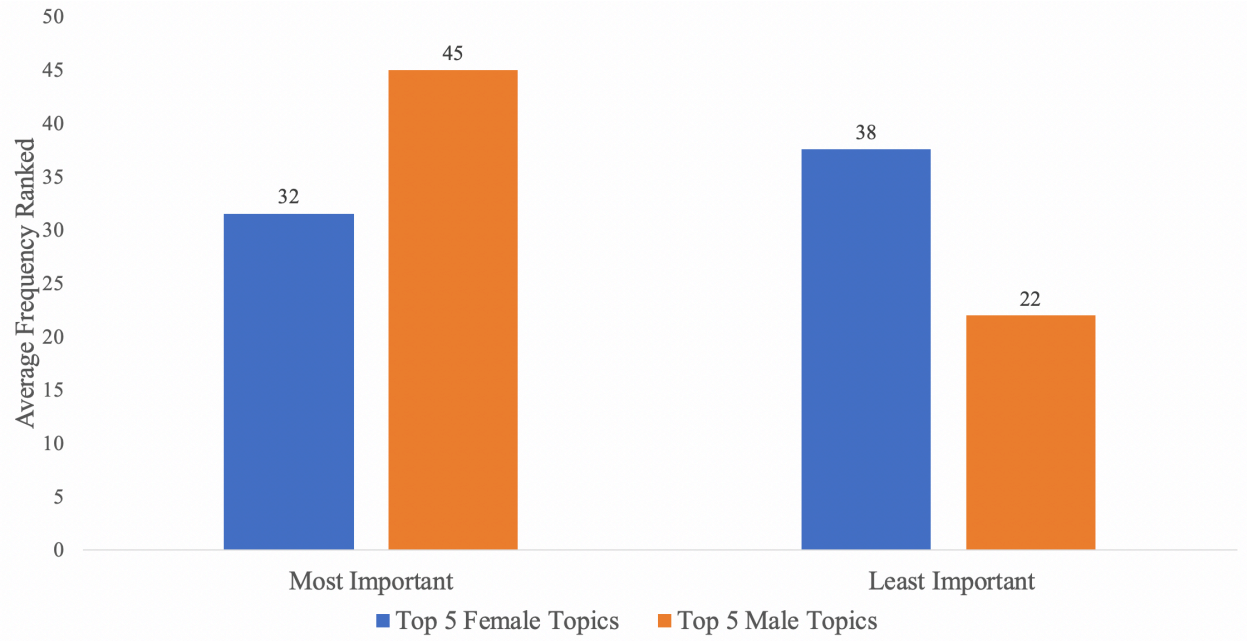

Figure 3. Average Frequency Top 5 Female and Top 5 Male Topics Ranked Most and Least Important. *Note.* Higher bars indicate higher average endorsement of topics as most or least important.

Test of hypothesis 2c. To learn if the *MI female* and *MI male* scores were approximately normally distributed, we ran a Shapiro-Wilk test. The Shapiro-Wilk for normality was not significant for the *MI female* variable (p = .614), *MI male* variable (p = .355). Thus, we concluded that these variables were normally distributed and proceeded with our analysis.

As these two scores came from the same participant's ranking, the design was within-subjects. Thus, we tested the difference between the *MI female* and the *MI male* scores using a paired t-test. Table 11 provides an overview of means for *MI female* and *MI male* scores as well as a test of the difference between these two means. The difference between the *MI female* (M = 0.163, SD = 0.021) and *MI male* scores (M = 0.167, SD = 0.017) was significant $t(249)$ = -

5.459, $p < .000$, indicating that topics expressed in higher proportions in letters for male applicants were ranked more frequently as most important for making a hiring decision compared to topics expressed in higher proportions in letters for female applicants. Thus, hypothesis 2c was confirmed. Following this significant result, we examined the effect size of the difference score and found a Cohen's D of 0.35, which indicates a small-medium effect size.

Table 11

Hypothesis 2c Results: Means, Standard Deviations, & Paired t-test for Most Important Topic Gender Proportions

	Independent Variable						
	Female Topic Proportion	Male Topic Proportion					
Dependent Variable	Mean (SD)	Mean (SD)	Diff	SD Diff	t-value	sig.	*d*
MI Gender	0.163 (0.021)	0.167 (0.017)	-0.004	0.012	-5.46***	.000	0.35

Note. $n = 250$, Diff = difference between mean values for MI male and MI female scores, SD Diff = the standard deviation of the difference, d = effect size for each test using Cohen's D (1988). Average summed proportions indicate the extent to which topics ranked as most important were expressed in letters for male or female applicants. Higher means indicate that topics chosen were expressed more frequently in letters of recommendation for male or female applicants. ***$p < .001$.

Test of hypothesis 2d. To learn if the *LI female* and *LI male* scores were approximately normally distributed, we ran a Shapiro-Wilk test. The Shapiro-Wilk for normality was not significant for the *LI female* variable (p = .667), or *LI male* variable (p = .679). Thus, we concluded that these variables were normally distributed and proceeded with our analysis. As these two scores came from the same participant's ranking, the design was within-subjects, thus, we tested the difference between the *LI female* and the *LI male* scores using a paired t-test. Table 12 provides an overview of means for *LI female* and *LI male* scores as well as a test of the difference between these two means. The difference between the *LI female* ($M = 0.161$, $SD =$

0.019) and *LI male* scores ($M = 0.159$, $SD = 0.015$) was significant $t(249) = 3.62$, $p < .000$, indicating that topics expressed in higher proportions in letters for female applicants were ranked more frequently as least important for making a hiring decision compared to topics expressed in higher proportions in letters for male applicants. Thus, hypothesis 2d was confirmed. Following this significant result, we examined the effect size of the difference score and found a Cohen's D of 0.23, which indicates a small effect size.

Table 12

Hypothesis 2d Results: Means, Standard Deviations, & Paired t-test for Least Important Topic Gender Proportions

	Independent Variable						
	Female Topic Proportion	Male Topic Proportion					
Dependent Variable	Mean (SD)	Mean (SD)	Diff	SD Diff	t-value	sig.	*d*
LI Gender	0.161 (0.019)	0.159 (0.015)	0.003	0.012	3.62***	.000	0.23

Note. $n = 250$, Diff = difference between mean values for LI male and LI female scores, SD Diff = the standard deviation of the difference, d = effect size for each test using Cohen's D (1988). Average summed proportions indicate the extent to which topics ranked as least important were expressed in letters for male or female applicants. Higher means indicate that topics chosen were expressed more frequently in letters of recommendation for male or female applicants.
***$p < .001$.

Summary of results. Across all three tests of hypothesis 1, we found consistent evidence that topics expressed more in letters for male applicants were rated as significantly more important for making a hiring decision than topics expressed more in letters for female applicants. In addition, across all four examinations of the data for hypothesis 2, we found consistent evidence that topics expressed more in letters for male applicants were ranked more as most important and topics expressed more in letters for female applicants were ranked more as least important for making a hiring decision.

Importance of information in letters

To learn how much weight participants ordinarily place on letters as evaluation tools for hiring decisions, we asked them to indicate the percent of their total decision to hire that they attribute to information provided in letters of recommendation (see Figure 4). One hundred and sixty-nine (68%) participants attributed 50% or more of their total hiring decision to information provided in letters of recommendation. While 64 (25.6%) participants attributed 30-40% of their total decision hiring decision to information in letters of recommendation, only 17 (7%) participants chose an option lower than 30%. These results indicate that letters of recommendation play an outsized role in influencing hiring decisions.

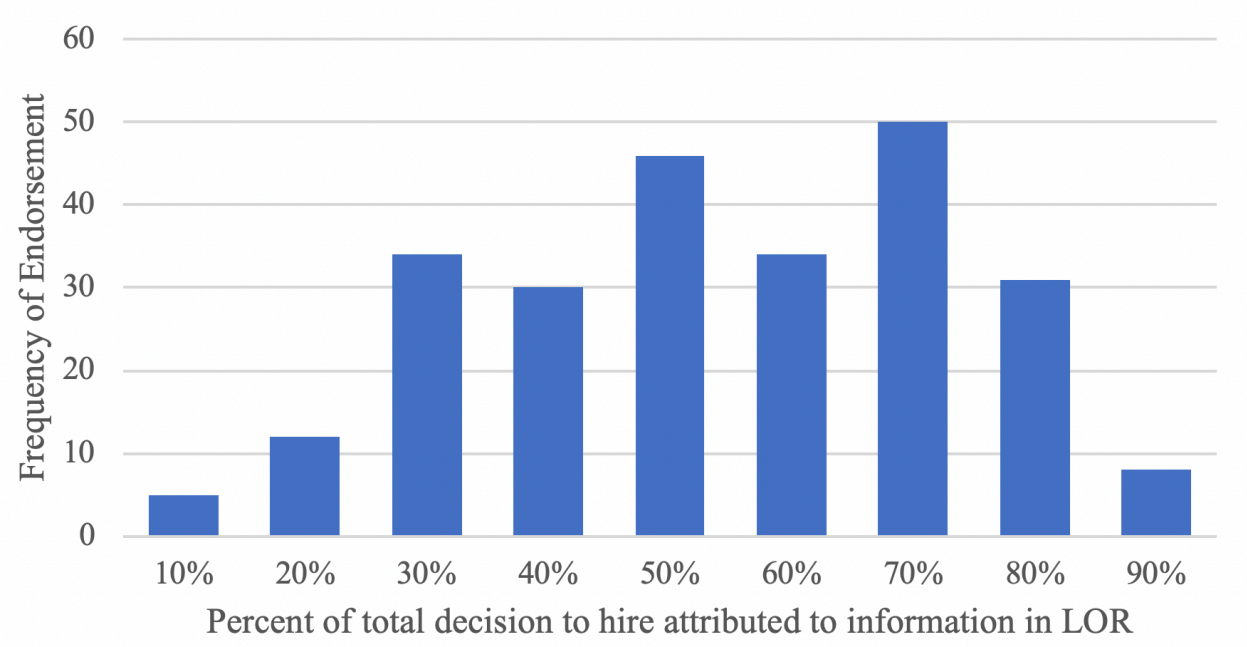

Figure 4. Importance of Letters of Recommendation (LOR). *Note.* Higher bars indicate more frequent endorsement for a given percent.

CHAPTER 5: DISCUSSION

Overview of Findings

The present study investigated how geoscientists evaluated topics expressed in letters of recommendation for applicants to a post-doctoral fellowship in the geosciences and whether topics expressed more in letters for male applicants were evaluated more favorably than topics expressed more in letters for female applicants. We observed from our research questions that participants valued certain topics over others; that being a productive researcher and publishing was considered more important for making a hiring decision than applicants being generally engaged in science, being a teacher, student, or department citizen, and/or showing improvement. This variation in preference for some topics over others had implications for applicants, depending on their gender.

The results for hypothesis 1a showed that the 15 topics expressed more in letters for female applicants were rated as significantly lower in importance for making a hiring decision, compared to the 16 topics expressed more in letters for male applicants. The results for hypothesis 1b found that the five topics expressed most in letters for male applicants were rated as significantly more important for making a hiring decision than the five topics expressed most in letters for female applicants. Finally, lending further evidence to the perceived importance of the topics expressed in letters of recommendation for male applicants, results from hypothesis 1c revealed that, across all topics, those expressed in higher proportions in male letters were rated as significantly more important for making a hiring decision than topics expressed in higher proportions in female letters. Thus, we observed that, compared to topics expressed in letters of recommendation for female applicants, the topics expressed in letters of recommendation for male applicants were consistently rated as more important across several different

operationalizations of our variables. The effect size for these tests were medium to large indicating that these differences are meaningful.

The results of hypothesis 2a descriptively illustrated a pattern in participant ranking consistent with the pattern of participant ratings in hypothesis 1a-c. That is, when we examined the average frequencies with which participants ranked the 15 female topics and the 16 male topics as most and least important for making a hiring decision, we observed that the male topics were ranked more frequently as most important and less frequently as less important compared to the female topics. Examining these frequencies for the five most female and five most male topics for hypothesis 2b, we observed participants ranked the most male topics more frequently as most important for making a hiring decision compared to the most female topics. In addition, participants ranked the five most female topics more frequently as least important for making a hiring decision compared to the five most male topics. The test of hypothesis 2c examined these trends inferentially and showed that topics expressed in higher proportions in male letters were ranked more by participants as most important compared to topics expressed in higher proportions in female letters. In addition, results from hypothesis 2d showed that topics expressed in higher proportions in female letters were ranked more by participants as least important compared to topics expressed in higher proportions in male letters. The effect size for these latter tests were small to medium, indicating that the differences were slight, but meaningful. Thus, we observed that, compared to female topics, male topics were consistently ranked as most important and, compared to male topics, female topics were consistently ranked as least important across several different operationalizations of our variables.

Finally, we observed that participants attribute an outsized amount of their decision to hire applicants to information provided in letters of recommendation. Indeed, a majority

attributed 50% or more of their decision to hire to information found in letters of recommendation.

Theoretical Implications

Our study extends previous research on gender bias in letters of recommendation in a number of ways. First, results from the present study were based on a comprehensive and holistic method of examining gender bias in letters of recommendation. Indeed, the topics from Elmore et al. (2019) used in the present study represent a detailed picture of the actual content within letters of recommendation and how this content differs for male and female applicants. Thus, in this study by using topics that emerged from 1,203 actual letters of recommendation, we were able to closely approximate the psychological circumstances under which evaluators find themselves while evaluating multiple letters and multiple applicants. Second, this approach revealed which representative themes in letters are considered most important and accounted for how these topics were differentially used to describe male and female applicants. Specifically, our study was able to account for the exact amount that a given topic was gendered and show how topic evaluations can compound to tip the scales in favor of hiring male applicants. These features helped illustrate two important points that were previously unknown until now; 1) there is a clear hierarchy of hiring criteria for post-doctoral fellows in the geosciences, and 2) this hierarchy serves to systematically disadvantage female applicants in hiring decisions. These insights illustrate the value of our study and carry important theoretical implications that build on past research examining gender differences in letters of recommendation.

The present study extends our understanding of how gender differences in letters of recommendation impact evaluations. Trix and Psenka (2003) found that female applicants were described more frequently as teachers and students and male applicants were described more

frequently as researchers and professional colleagues. We observed that the topic *top student and teacher,* which was expressed more in letters for female applicants (Elmore et al., 2019), was rated and ranked as less important for hiring decisions than the topic *dedicated to research*, which was expressed more frequently in letters for male applicants (Elmore et al., 2019). Thus, the present study provides evidence that being described as students and teachers is a detrimental and being described as researchers is beneficial for hiring decisions. As these two topics contained gender differences in their distribution among letters, these evaluations served to benefit male applicants and disadvantage female applicants.

Schmader, Whitehead, and Wysocki (2007) found that letters for male applicants contained more standout words (e.g. excellent, outstanding, etc.) compared to letters for female applicants. In the present study, we found that topics containing standout words are evaluated more favorably. Indeed, the topics *excels academically and professionally* and *outstanding young researcher*, were among the highest rated and ranked topics. Thus, it would seem from the above findings that standout words are important and that male applicants likely have an advantage. However, Elmore et al. (2019) found that the topic *excels academically and professionally* was expressed more frequently in letters for female applicants, whereas *outstanding young researcher* was expressed only slightly more frequently in letters for male applicants. This is similar to what French et al. (in press) observed in letters of recommendation to a surgical residency program; that female applicants received more standout words (e.g., amazing, outstanding, etc.) than male applicants. Thus, in some cases female applicants are more likely to receive standout words, and in the present study, topics with those words benefited female ratings and rankings. Indeed, the topic *excels academically and professionally* was the second highest rated topic and the third most female topic. However, it is also clear from our

findings that one standout word among the topics expressed more for female applicants was not enough to counter the cumulatively low ratings of other female topics. Thus, while standout words may signal positivity, they may not be enough if other topics negate their impact.

In addition, the present study extends past research on standout words by illustrating the importance of context with regard to how standout words are used. For instance, the standout words in the topics *excels academically and professionally* and *outstanding young researcher* described applicants as being academics and professionals as well as young researchers, respectively. While both topics were rated highly and ranked as most important at almost the same frequency, *outstanding young researcher* was rated slightly lower. Thus, being academic and professional was considered slightly more important than being a young researcher. One might imagine further variability in ratings if a standout word were used to describe an applicant as a teacher, for instance. Indeed, we observed much lower ratings for the topic *top student and teacher*. However, the descriptor in this topic, *top*, connotes some standout value; that an applicant is ranked above others. However, it seems that the value of this word is less powerful when followed by *student and teacher*. Thus, we can see that, while topics with standout words do indeed seem to be rated highly, the content being emphasized by standout words seems to play a larger role. Therefore, a focus on standout words in letters may miss the larger picture; the substantive attributes used to describe applicants.

In addition to extending past research findings that have explored gendered content in letters of recommendation, the present study also addressed limitations posed by past research on how differences in letter content influence evaluations. Two studies have been done that directly examined both the content of gender differences in letters of recommendation, and the subsequent evaluations of letters containing this content. Madera et al., (2009) showed that

letters for female applicants were more likely to express communal attributes (e.g. "affectionate," "helpful," "kind") and that applicants with letters containing more communal attributes received lower ratings of hireability. Further, Madera et al. (2018) showed that letters for female applicants were more likely to express doubt (e.g. negativity, hedging, irrelevant information, and faint praise) and that applicants with letters that contain doubt were rated lower in terms of research competence. Thus, attributes expressed more frequently in letters for female applicants led to lower subsequent evaluations. Although both of these studies did examine subsequent evaluations of the content used to differentially describe male and female applicants in letters of recommendation, they have limitations, which the present study addressed.

First, these studies only examined how *single themes* in letters impacted evaluations. This approach limits our understanding of how the many other themes used to represent applicants compound to produce ratings. The present study provides a detailed look at which gendered topics serve to produce biased evaluations beyond single themes. Indeed, the present study provided a deeper understanding for how a whole corpus of letters gets evaluated. For instance, when evaluators read recommendation letters, they do not read them in a vacuum; many topics are present. Indeed, in real-life evaluations, applicants are evaluated against one another, using *multiple* topics, all of which impact decisions to hire. In the present study we included 31 topics used to describe applicants. Not all topics appeared in all letters and not all topics appeared in the same proportion in letters for female and male applicants. By accounting for this complexity across letters we more closely approximated the conditions associated with letter evaluation. Through this design, we were able to show how topics are evaluated in relation to one another and illustrate how topics compound to advantage male applicants and disadvantage female applicants.

Second, one might expect letters with expressed doubt to be rated more negatively than letters without doubt. Thus, Madera et al. (2018) failed to account for the impact that more positively valanced themes have on evaluations. Further, Dutt et al. (2016) observed that only 2.5% of the letters for applicants for a geoscience position raised doubts, which indicates that doubt is not a highly prevalent theme across letters, making it less representative of applicants, in general. The present study addressed this limitation by asking decision makers to evaluate a wide range of topics that better represented the data within letters and included desirable applicant attributes, rather than just a single negative theme (e.g. doubt). Thus, by incorporating a range of applicant attributes, findings from the present study illustrated not just which gendered topics lead to applicants being rejected, but which gendered topics lead to their being accepted.

Third, the Madera et al. (2009; 2018) studies were performed on letters of recommendation from the field of psychology which is generally majority female (American Psychological Association, 2019) and compared to harder sciences often perceived of as more feminine (Carli et al. 2016). Our research was based on a sample of letters from the geosciences, which is a majority male field (National Science Foundation, 2017). Social role theory (Eagly, Wood, & Diekman, 2000) explains that the behavior of men and women is generally rooted in and thus consistent with behaviors suited to those roles in which their gender is the majority. Thus, one might expect feminine attributes to be evaluated more favorably in a field where women are the majority. The present study provides a look at how favorably stereotypically female attributes (e.g. as teachers) are evaluated in a field where men are the majority.

Finally, Madera et al. (2009) only asked 6 professors to evaluate letters, lowering the power of their results and potentially opening up their results to sample bias. Our findings are based on evaluations from 250 participants with relevant experience in the geosciences. Thus,

our study captured results that are likely to better represent evaluator perceptions in geosciences overall.

In summary, our study addressed limitations by past research in four important ways; 1) we uncovered how *multiple* topics expressed at different rates in letters for male and female applicants compound to produce biased evaluations, 2) we explored evaluations of a range of topics that were representative of letter content and included desirable applicant attributes, not just negative themes (e.g. doubt), 3) we explored this effect in a majority male academic field, and 4) we collected data from a large sample of participants, helping to increase confidence in our results.

The findings of the present study also provide an illustration for how role congruity theory operates in letters of recommendation. Role congruity theory generally stipulates that "a member of a group whose stereotypical attributes are thought to facilitate performance in a role is ordinarily preferred over a member of a group whose stereotypical attributes are thought to impede performance," (Eagly & Diekman, 2005, p. 2). Elmore et al. (2019) found that student and teacher attributes, such as *top student and teacher*, were expressed more frequently in letters for female applicants and researcher attributes, such as *publishing frequently* and *dedicated to research*, were expressed more frequently in letters for male applicants. In the present study, we observed that researcher attributes were rated and ranked higher for making a hiring decision than student and teacher attributes. Thus, the findings of the present study indicate that, because participants preferred researcher attributes, which were expressed more in letters for male applicants, male applicants became more congruent with the role of geoscience post-doctoral fellow, and because participants showed less preference for student and teacher attributes, which were expressed more in letters for female applicants, female applicants became incongruent with

the role of geoscience post-doctoral fellow. This process illustrates how content from letters can serve to convey information that would lead to higher role congruence for one gender over the other.

Finally, student and teacher attributes may operate as signals writers use to indicate that an applicant is not a good fit. In the present study we observed that student and teacher attributes were evaluated less favorably than researcher attributes. However, participants preferences indicate a privileging of research over being a student and teaching as if they are not related. The process of researching is inherently a learning and teaching exercise. That is, by researching, a person remains a student and by publishing they become a teacher. Thus, the attributes of a student, teacher, and researcher are all required to for a person to be any of the three. However, favoring one (performing research) much more than others (being a student and teacher) illustrates a misunderstanding of their relatedness. Thus, one might conclude from our study that student and teacher attributes simply serve as shorthand for a poor applicant rather than a desired feature which is integral to the abilities of a researcher.

Practical Implications

We observed that a majority of participants in the present study attributed 50% or more of their decision to hire to information found in letters of recommendation. This finding lends support to past research, which has found that academic professionals place more weight on the content from letters of recommendation when making selection decisions than do applied professionals (Nicklin & Roch, 2009) and that academic faculty consider the quality of letters of recommendation as second most important behind GPA for making student admission decisions (Potvin, Chari, & Hodapp, 2017). These past results, along with our own, lend support the idea

that letters of recommendation are critically important for making hiring decisions. Thus, our findings carry important practical value.

Our results indicate that faculty and researchers in the geosciences clearly favor getting funding, being productive in research, and presenting and publishing findings. These results provide support for a notion expressed for some time; that the geosciences have a publish or perish culture (Ward, 1989; Castleford, 1998; Laird, Bell, & Pfirman, 2007). De Rond and Miller (2005, p. 322) explain that a publish or perish culture "signifies the principle according to which a faculty member's tenure is primarily a function of his or her success in publishing," In their definition, the criterion of publishing is applied only to those seeking tenure. However, based on the findings of the present work, it seems that this criterion is also applied to post-doctorate roles. This highly specific role expectation has implications for the attributes people assign to the role of scientist. Lane, Hardison, Simon, and Andrews (2019) interviewed 33 doctoral students in the life sciences at one research-university to investigate the factors that led to students developing a teaching identity. At least in this single institution, students reported on a culture that placed much more value on research over teaching. Students generally explained that their mentors saw teaching as a secondary task. Capturing this, one student explained that the sentiment they experienced in the department was that "a real scientist only cares about their research and they teach because they have to," (p. 150). Further, for some students there was a struggle to marry the identities of being a teacher and scientist. When discussing their desire to teach, another student wrestled with the two identities and reasoned "I'm not going to stop being a scientist just because I'm not conducting publishable research," (p. 152). This struggle illustrates the intense requirement that there is only room for one type of behavior to be considered a scientist; performing research and publishing. Our findings provide support for the publish or perish

culture in the geosciences as well as validates the associated anxieties that graduate students experience when planning their career trajectory. Given the strong evidence for a publish or perish culture in the geosciences and the tendency for male applicants to be described of as fitting into this culture, as was observed in the present study, it seems worthwhile for the field to reflect on their values and subsequently how they might improve their hiring processes in the future. We provide some guidance in this regard below.

There are several important elements of hiring processes upon which geoscience departments should reflect. First, what makes a successful post-doctoral fellow? In his examination of talent management in corporate settings, Church (2018) challenges hiring managers and decision makers to consider future potential and, more specifically, potential for what is required for success in the candidate's future role. It is clear from the topics rated and ranked highest in importance that the "potential for what?" question is answered as "potential for tenure-track academic roles in a research university." Thus, by this standard, one can observe that, at least in the geosciences, post-doctoral fellowships are being treated as a means for succession planning for tenure-track roles. Geoscience departments should consider whether research productivity is the sole purpose of a post-doctoral fellowship and if not, what other elements should be considered when evaluating what makes a successful post-doctoral fellow. For instance, while considered secondary in the current academic climate, tenure-track faculty still must teach.

Second, does the content of letters accurately represent the attributes of male and female applicants without introducing bias? Topics describing research productivity were emphasized more in letters for male applicants (Elmore et al., 2019) and it is possible that in the sample of letters used to create the topics, male applicants were actually publishing more. Indeed, in an

examination of the gender of first authors submitting and publishing articles in 20 journals published by the American Geophysical Union between 2012-2015, Lerback and Hanson (2017) found that female authors submitted 0.79 fewer papers compared to male authors. However, manuscripts from female authors were more likely to be accepted (61%) compared to male authors (57%). Thus, while there is evidence that indeed male applicants may have been *submitting manuscripts* more frequently, a topic that occurred 20% more often in letters for male applicants compared to female applicants (Elmore et al., 2019), female applicants are actually *publishing frequently*, a topic that still occurred much more in letters for male applicants (30% – Elmore et al., 2019). Geoscience departments should take steps to identify when letters emphasize critical hiring criteria disproportionately in letters for male applicants. Employing techniques such as Structural Topic Modeling (Roberts et al., in press) can aid as a first line defense for surfacing gender disparities in letters.

The discussion above provides guidance for an informed three-step process whereby geoscience departments can mitigate gender disparities in hiring. Departments should 1) define what it means to be a post-doctoral fellow in your specific department, 2) determine what criteria would make candidates successful given your definition, and 3) evaluate letters against these criteria. This final step should be preceded by an investigation of systematic gender differences across applicant letters of recommendation, which could be achieved by instituting preliminary evaluations of letters using methods such as Structural Topic Modeling. This proactive approach may serve to limit existing gender disparities and aid in the advancement of female geoscientists.

Findings from the present study also offer a unique illustration of how letter writers' gender bias can have an impact on subsequent evaluations of applicant fit. Participants in the present study were not aware of the gender of applicants for which letters were written. That is,

participants were blind to applicant gender. However, those who originally wrote these letters were not blind to the applicant's gender. Thus, whereas writers may have relied on gender stereotypes when writing about applicants, participants in the present study could not have, as they did not know the gender of applicants. Indeed, participants in the present study only provided feedback on the criteria they found most important for making a hiring decision. Given the arrangement of writer as aware of applicant gender and evaluator as blind to applicant gender, the present study illustrates a novel two-stage process whereby role congruity theory (Eagly & Karau, 2002) played out. In the first stage, writers wrote letters which expressed attributes important for success in science (e.g. researching and publishing) and did so more for male applicants and expressed attributes less important for success in science (e.g. student and teacher) and did so more for female applicants. In the second stage, evaluators from our study examined these attributes and indicated their preference for researchers and publishers over students and teachers. By confirming that content, which was disproportionality expressed for male applicants, was more important for making a hiring decision, evaluators were also confirming that male attributes were more congruent with the role of post-doctoral fellow in the geosciences. Alternatively, by confirming that content, which was disproportionality expressed for female applicants, was less important for making a hiring decision, evaluators were confirming that female attributes were less congruent with the role of post-doctoral fellow in the geosciences. This process illustrates the relationship between content creators and content interpreters. That is, writer biases get sown into letters in the first stage and inadvertently influence decision making in the second stage. Indeed, in this process evaluators need not know the gender of the applicant to assess congruence as the writer has already performed that task.

This process calls into question whether redacting gender from evaluation materials is enough to stop bias, which has implications for blind review processes.

Blind review processes have been shown to reduce gender bias. Indeed, Goldin and Rouse (2000) examined the impact of blind auditions on female advancement across 14,133 individuals in 592 audition segments for eight symphony orchestras between 1970-1995. Blind auditions were carried out using a screen to disguise musicians while they performed. The researchers found that blind auditions increased female advancement from early rounds by 50% and female hiring by 7.5-13.7%. However, these findings were based on a situation where evaluators were judging candidates directly (e.g. while they played on the stage). In the case of letters of recommendation, content creators are the primary evaluator and content interpreters are secondary evaluators. Therefore, as was observed in the present study, biases of the primary evaluator can seep in and influence how the secondary evaluator interprets an applicant, even if they are blind to applicant gender. Thus, while the idea of redacting gender from evaluations seems appealing, findings from the present study indicate that care should be taken to ensure that evaluation materials themselves are free of gender bias.

Given the observation that bias starts with writers, there are several steps departments can take to mitigate the risk of gender bias influencing hiring. First, geoscience departments should invest in training letter writers on techniques for avoiding gender stereotypes when describing applicants. This is especially important because, as we observed in the present study, evaluators being blind to applicant gender did not decrease the preference for content expressed more in letters for male applicants. While a variety of research has found that training faculty decision makers leads to greater awareness of gender bias and gains in female placement in academic departments (Fine et al., 2014; Carnes et al., 2015; Smith, Handley, Zale, Rushing, and Potvin

2015; Devine et al., 2017), this same work should incorporate how best to recommend applicants to other departments. Kuncel et al., (2014) found that letters of recommendation are weakly but positively associated with performance in degree programs. They recommend introducing greater structure into the process of writing letters of recommendation as they are currently a completely unstructured evaluation tool. Training could emphasize how to properly structure letters as well as how to maintain focus on attributes of applicants that are relevant to the role for which they are applying.

Second, departments should consider building more structure into the data they receive from recommenders. For instance, standardized letters of recommendation, which use a more structured methodology such as rating scales along with narrative evaluation, have been adopted in some fields. However, even these tools have seen mixed results in terms of validity and reduction in gender biases (Kaffenberger et al., 2016; Kominsky, Bryson, Benninger, & Tierney, 2016; Li, et al., 2017; Samade, Samora, Scharschmidt, & Goyal, 2020). Thus, it is clear that more work needs to be done to identify the right tools for reducing gender bias.

Finally, departments should consider how much weight evaluators place on content within letters for making a hiring decision. As we observed in the present study letter content plays an outsized role in influencing hiring decisions, but letters favor content that is expressed more in letters for male applicants. This necessarily increases the likelihood that gender bias will operate in hiring decisions leading to unfair advantages for male applicants. Thus, departments should also reduce the amount of weight that is placed on letters of recommendation in favor of more objective measures of performance (e.g. GPA, research products, etc.). Departments should create guidelines which specify how much weight should be applied to each type of performance

indicator within applications. In turn, evaluators could use these guidelines when making hiring decisions.

The present study offers an illustration of how findings from machine learning research can be applied to understanding social science problems. The results from Elmore's et al. (2019) structural topic model provided an in-depth, accurate, and highly nuanced look at how content in letters varied by applicant gender. This enabled us to explore the impact of letter content on hiring decisions. Data from letters of recommendation are messy and highly unstructured. Novel machine learning techniques are allowing social scientists to examine the impact of highly relevant but previously unwieldy information on outcomes. These new approaches provide the opportunity for researchers to achieve higher realism without sacrificing too much control. That is, machine learning allows one to account for the complexity of the real world, while offering confidence in the validity of relationships between variables. While the learning curve may be steep in the early days of these techniques, the rewards of labor will pay off with deeper insight into the human experience.

Limitations and Future Directions

The present study has several limitations that should be considered when interpreting the results. First, while we recruited participants from highly relevant domains, those who participated had to make the choice to take the questionnaire. That is, they clicked the link and voluntarily participated. However, not everyone who started the questionnaire finished the questionnaire. Four hundred and thirty-seven individuals started the questionnaire and 260 (54%) completed. Thus, slightly more than half of those who selected-in finished the task. There are a variety of explanations for why participants stopped while taking the questionnaire, including fatigue, boredom, lack of time, or that they determined they did not actually have relevant

experience to provide helpful data. However, given these possible explanations, our sample may be unique in that it included participants who were more persistent, more interested, had time, or had more relevant experience regarding the purpose of the study. Second, our sample may be considered somewhat unique given the number of people who were solicited by email and responded. Indeed, of those who were emailed an invitation, only 5.3% completed the study. Thus, this sample may contain individuals who are a particular subset within the geoscience community. However, anecdotally, within the geoscience departments from which we sampled, we observed that not all faculty and researchers had post-doctoral fellows in their labs. Therefore, it is possible that those with relevant experience related to the purpose of our study were a select group within the geosciences. Further, the rates of interest we observed in the present study are high compared to other email campaigns using similar cold email methodologies. Indeed, mailchimp.com provides distribution services for companies running email marketing campaigns by sending out cold emails to potential customers. In an examination across thousands of campaigns, mailchimp.com observed that the average rate that respondents clicked links within emails was 2.62% (Mailchimp.com, 2019). Thus, compared to other online recruitment attempts, our study would be considered successful.

In addition, our sample is similar to the population of geoscientists in terms of gender distribution. In 2018, women made up 29% of the membership of the American Geophysical Union, which has 60,000 members globally (American Geophysical Union, 2018). Our sample was 31.6% female. Thus, in terms of gender distribution, our sample might be considered representative of the larger geoscience population.

Another potential limitation of the study is that the topics we asked participants to evaluate were double-barreled in some cases. For instance, the topic *top student and teacher*

contained two roles; student and teacher. This feature of the topics reduced our ability to know whether participants were evaluating whether participants were top students or top teachers and therefore introduced error into our results. However, as Elmore et al. (2019) observed, topics with multiple themes were coded as such because those themes appeared in combination in single letters. Thus, while this feature of the topics reduced our ability to specify which criterion was being evaluated by participants, it enabled us to examine how evaluators interpret content that more closely approximates what would be observed in letters. That is, because the roles of student and teacher were closely intertwined in the letters, evaluators reading those letters would likely consider them in combination as we asked them to do in the present study. This trade-off between experimental control and ecological validity illustrates an inescapable feature of research methods and in this case, we sacrifice control for higher realism.

Our study provides a detailed look at how one element of letters of recommendation, the content, impacts evaluations. That is, we focused on how applicants were described in letters. Thus, we do not know how other factors such as letter length, gender of letter writer, or the reputation of the letter writer, for instance, impact evaluations. Future research should examine the impact of these other letter features on evaluations. Further, letters of recommendation are only one element of applications. While participants in the present study indicated that information in letters plays an outsized role in their decision to hire, they were first exposed to 31 succinct applicant attributes prior to answering that question. These attributes may have served to illustrate the usefulness of information in letters of recommendation. Thus, our study procedure itself could have influenced the amount of importance participants placed on information in letters. However, other studies have illustrated the important role that letters of recommendation play in academic admissions decisions (Potvin et al., 2017) and academic hiring decisions

(Abbott et al., 2010). While there is evidence that letters play a large role, it is important for future research to consider the impact of letter in the context of other hiring materials.

The present study illustrated that participants considered publishing, getting funding, and performing research as most important for making a hiring decision. However, participants could have simply been evaluating topics based on the criteria expressed in the job announcement they read prior to evaluating the topics. The job announcement explains that "the principal selection criteria for Fellows are scientific excellence and a clearly expressed plan to investigate problems at the forefront of Earth science." This selection criteria could have influenced participant's favoring of publishing and researching over other attributes. However, the announcement used in the present study was the same as that which was used to advertise the actual post-doctoral role for which letters were written. That is, it conveys the criteria by which these letters were judged in an actual hiring scenario. Thus, while it is possible that the announcement led participants to favor researcher attributes, the present study sought to approximate the psychological circumstances that were likely at play when actual hiring decisions were made based on these letters, thereby producing more valid results. However, to get a sense of the general values of a field, future research could examine the amount of importance that faculty and researchers place on topics absent predefined hiring criteria.

Finally, while gender bias in letters of recommendation and lower evaluations of themes expressed for female applicants has been observed in other fields (e.g. psychology – Madera et al. 2009; 2018), one should consider our findings limited to context in which the letters were written and evaluated; the geosciences. That is, the present findings hold strong relevance for the geosciences but should be explored more broadly by future researchers in other fields and organizational contexts.

Conclusion

Letters of recommendation provide academic programs the opportunity to learn about an applicant from the perspective of a previous supervisor. It is clear from the findings of the present study that these opinions matter a great deal to those making decisions. Thus, letter writers become gatekeepers to advancement in the geosciences. The fact that these gatekeepers write letters that depict female and male applicants differently is one issue, but that those differences serve to advantage male applicants and disadvantage female applicants with regard to hiring criteria, illustrates how the findings of the present study play a role in the loss of female talent from the geoscience career pipeline. Female applicants are still subject to stereotyping and in turn incongruence or "lack-of-fit" perceptions. To fix the leak, it behooves the geosciences to actively engage the issue and to critically evaluate their methodology for evaluating applicants.

REFERENCES

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

Aggarwal, S., Grob, S., Banerjee, D., Putzel, P. J., and Tao J. (2018). Key Word Use in Letters of Recommendation for Ophthalmology Residency Applicants According to Race, Gender, and Achievements. *Journal of Academic Ophthalmology*, *10*(1), 163-171.

Akos, P., & Kretchmar, J. (2017). Gender and Ethnics Bias in Letter of Recommendation: Considerations for School Counselors. *Professional School Counseling, 20*(1), 102-113.

American Geosciences Institute, (2019). Percentage of Female Faculty Working within Geoscience Research Fields. *Geoscience Workforce.* Retrieved from

https://www.americangeosciences.org/sites/default/files/currents/Currents-136-WomenResearchFields.pdf

American Geophysical Union, (2018). 2018 AGU Honors Demographic Report. *Diversity, Resources and Demographics.* Retrieved from https://honors.agu.org/files/2018/12/2018_Honors_Cycle_Demographics.pdf

American Psychological Association, (2019). *Psychology faculty salaries for the 2018-2019 academic year: Results from the 2019 CUPA-HR survey for four-year colleges and universities.* Washington, DC: Author.

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human relations*, *61*(8), 1139-1160.

Bischof, J. and Airoldi, E. (2012). Summarizing topical content with word frequency and exclusivity. *arXiv preprint arXiv:1206.*4631.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems, 988*, 601–608.

Carli, L. L., Alawa, L., Lee, Y., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women scientists. *Psychology of Women Quarterly, 40*(2), 244.

Carnes, M, Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., Forscher, P., Issac, C., Kaatz, A., Magua, W., Palta, M., & Sheridan, J., (2015). The Effect of an Intervention to Break the Gender Bias Habit for Faculty at One Institution: A Cluster Randomized, Controlled Trial. *Academic Medicine: Journal of the Association of American Medical Colleges 90*(2), 221–30.

Castleford, J. (1998). Links, lecturing and learning: some issues for geoscience education. *Computers & Geosciences*, *24*(7), 673-677.

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,* 413–423.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, 288–296.

Church, A. H., (2018). Think Outside the 9 Box. *Talent Quarterly, 19*, 39-43.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

De Rond, M., & Miller, A. N. (2005). Publish or perish: bane or boon of academic life? *Journal of management inquiry*, *14*(4), 321-329.

Devine, P. G., Forscher, P. S., Cox, W. T. L., Kaatz, A., Sheridan, J., & Carnes, M., (2017). A Gender Bias Habit- Breaking Intervention Led to Increased Hiring of Female Faculty in STEMM Departments. *Journal of Experimental Social Psychology* 73: 211–15.

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.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics, 41*(6), 570-606.

Eagly, A., & Karau, S. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*, 573–598.

Eagly, A. H., & Diekman, A. B. (2005). What is the problem? Prejudice as an attitude-in-context. In J. F. Dovidio, P. Glick, & L. A. Rudman (Eds.), *On the nature of prejudice*: *Fifty years after Allport* (pp. 19–35). Malden, MA: Blackwell.

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (pp. 123–174). Mahwah, NJ: Erlbaum.

Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2019). How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, 1-15.

Elmore, J. J. G., Block, C. J., Bowers, A., Dutt, K., (2019). A Structural Topic Model Approach to Exploring Gender Bias in Letters of Recommendation, Society for Industrial Organizational Psychology Conference, Washington D.C., April 4-6.

Fine, E., Sheridan, J., Carnes, M., Handelsman, J., Pribbenow, C., Savoy, J., & Wendt, A., (2014). Minimizing the Influence of Gender Bias on the Faculty Search Process. In *Gender Transformation in the Academy* (pp. 267–89). UK: Emerald Group.

French, J. C., Zolin, S. J., Lampert, E., Aiello, A., Bencsath, K. P., Ritter, K. A., Strong, A. T., Lipman, J. M., Valente, M. A., and Prabhu, A. S. (In press). Gender and Letters of Recommendation: A Linguistic Comparison of the Impact of Gender on General Surgery Residency Applicants. *Journal of Surgical Education*.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, *90*(4), 715-741.

Grimmer, J., & Stewart, B. M, (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21, 267–297.

Hanchane, S., & Recotillet, I. (2003). Academic careers: the effect of participation to post-doctoral program. *SASE 2003 Knowledge, Education, and Future Societies, LEST,* 26-28, 16.

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

Heilman, M. E. (2012). Gender stereotypes and workplace bias. *Research in Organizational Behavior, 32,* 113-135.

Heilman, M. E., Block, C. J., Martell, R. F., & Simon, M. C. (1989). Has anything changed? Current characterizations of men, women, and managers. *Journal of Applied Psychology*, *74*, 935-942.

Hoffman A., Grant W., Mccormick M., Jezewski E., Matemavi P., Langnas A.(2019). Gendered differences in letters of recommendation for transplant surgery fellowship applicants. *Journal of Surgical Education. 76,* 427–432.

Kaffenberger, J. A., Mosser, J., Lee, G., Pootrakul, L., Harfmann, K., Fabbro, S., Fernandex-Faith, E., Carr, D., Plotner, A., Zirwas, M., & Kaffenberger, B. H. (2016). A retrospective analysis comparing the new standardized letter of recommendation in dermatology with the classic narrative letter of recommendation. *The Journal of clinical and aesthetic dermatology*, *9*(9), 36.

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics*, *38*(1), 52-54.

Kominsky, A. H., Bryson, P. C., Benninger, M. S., & Tierney, W. S. (2016). Variability of ratings in the otolaryngology standardized letter of recommendation. *Otolaryngology–Head and Neck Surgery*, *154*(2), 287-293.

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*(1), 101-107.

Laird, J. D., Bell, R. E., & Pfirman, S. (2007). Assessing the publication productivity and impact of eminent geoscientists. *Eos, Transactions American Geophysical Union*, *88*(38), 370-371.

Lane, A. K., Hardison, C., Simon, A., & Andrews, T. C. (2019). A model of the factors influencing teaching identity among life sciences doctoral students. *Journal of Research in Science Teaching*, *56*(2), 141-162.

Lee M., & Mimno D. (2014). Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, *Association for Computational Linguistics*, 1319–1328.

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

Li, S., Fant, A. L., McCarthy, D. M., Miller, D., Craig, J., & Kontrick, A. (2017). Gender differences in language of standardized letter of evaluation narratives for emergency medicine residency applicants. *AEM education and training*, *1*(4), 334-339.

Lin, E. S., & Chiu, S. (2016). Does holding a postdoctoral position bring benefits for advancing to academia? *Research in Higher Education, 57*(3), 335-362.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254-277.

Mailchimp.com (2019). Email Marketing Benchmarks by Industry. *Mailchimp.com*. Retrieved from https://mailchimp.com/resources/email-marketing-benchmarks/

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2018). Raising doubt in letters of recommendation for academia: Gender differences and their impact. *Journal of Business and Psychology,* 1-17.

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,* 1591-1599.

Milkman, K. L., Akinola, M., & Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *Journal of Applied Psychology, 100*(6), 1678.

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. (2011). Optimizing Semantic Coherence in Topic Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011. *Association for Computational Linguistics*, 262–272.

Moss-Racusin, C., 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 of the United States of America, 109*(41), 16474.

National Science Foundation (2017), Survey of Graduate Students and Postdoctorates in Science and Engineering, *National Center for Science and Engineering Statistics.* Retrieved from https://ncsesdata.nsf.gov/gradpostdoc/2017/html/gss17-dt-tab004-2.html

Nicklin, J. M., & Roch, S. G. (2009). Letters of recommendation: Controversy and consensus from expert perspectives. *International Journal of Selection and Assessment, 17*(1), 76-91.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria.

Pennebaker, J. W. & Tausczik, Y. R. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Potvin, G., Chari, D., & Hodapp, T. (2017). Investigating approaches to diversity in a national survey of physics doctoral degree programs: The graduate admissions landscape. *Physical Review Physics Education Research, 13,* 020142.

Rice, L., & Barth, J. M. (2017). A tale of two gender roles: The effects of implicit and explicit gender role traditionalism and occupational stereotype on hiring decisions. *Gender Issues, 34*(1), 86-102.

Roberts, M. E., Stewart, B. M., & Airoldi E. M., (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association, 111(515),* 988-1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (In press). STM: R package for structural topic models. *Journal of Statistical Software.*

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014a). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064-1082.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014b). Structural topic models for open-ended survey responses; Online Appendix. *American Journal of Political Science, 58*(4), 1064-1082.

Samade, R., Samora, J. B., Scharschmidt, T. J., & Goyal, K. S. (2020). Use of Standardized Letters of Recommendation for Orthopaedic Surgery Residency Applications: A Single-Institution Retrospective Review. *Journal of Bone and Joint Surgery*, *102*(4), 14.

Schein, V.E. (1973). The relationship between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology. 60,* 340-344.

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*, 509-514.

Schmiedel, T., Müller, O., vom Brocke, J. (2018). Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial with an Application Example on Organizational Culture. *Organizational Research Methods*, 1-28.

Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences of the United States of America, 111*(28), 10107.

Smith, J. L., Handley, I. M., Zale, A. V., Rushing, S., & Potvin M. A., (2015). Now Hiring! Empirically Testing a Three-Step Intervention to Increase Faculty Gender Diversity in STEM. *BioScience, 65*(11),1084–87.

Smyth, F. L., & Nosek, B. A. (2015). On the gender-science stereotypes held by scientists: Explicit accord with gender-ratios, implicit accord with scientific identity. *Frontiers in Psychology,* 6, 415.

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.

Turrentine, F. E., Dreisbach, C. N., Ivany, A. R. S., Hanks, J. B., Schroen A. T. (2019). Influence of Gender on Surgical Residency Applicants Recommendation Letters. *Journal of the American College of Surgeons 4*(228), 356-365.

Tvinnereim, E., & Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change*, *5*(8), 744-747.

U.S. Bureau of Labor Statistics (2016). 1948-2016 annual averages, Current Population Survey. *Facts Over Time: Women in the Labor Force.* Retrieved from: https://www.dol.gov/wb/stats/NEWSTATS/facts.htm#WomenLF

Wang, Y., Bowers, A. J., & Fikis, D. J. (2017). Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of EAQ articles from 1965 to 2014. *Educational administration quarterly*, *53*(2), 289-323.

Ward, D. (1989). Information-Seeking Behavior of Geoscientists. *How to Tame your Library*, 169.

Williams, W. M., and 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 of the United Stated of America. 112,* 5360–5365.

APPENDICES

Appendix A – Study Materials: Introduction and Instructions

Introduction

Thank you very much for agreeing to participate in our study. Your participation is invaluable to our work.

The study should take approximately 10-15 minutes to complete. Please read the instructions on the next page and navigate through each page, responding when prompted.

Instructions

In the present study, we are interested in learning which topics used to describe applicants in letters of recommendation are important for making hiring decisions for post-doctoral fellowships in the geosciences.

We have compiled a representative list of topics used to describe applicants from a sample of letters of recommendation for applicants to post-doctoral fellowships in the geosciences. As an expert in the geosciences, we ask that you,

1. Review a job description relevant to the fellowships for which the letters were written
2. Review the list of topics discussed across the letters of recommendation
3. Indicate how important topics would be in your decision to hire an applicant
4. Complete a demographic questionnaire

In the following pages, please read the job description, review the topics, and respond to the questions and statements.

Appendix B – Study Materials: Job Description

Postdoctoral Fellowships in the Earth and Environmental Sciences

An Earth science institution invites applications for Postdoctoral Fellowships in the fields of Earth and environmental sciences. Candidates should have recently completed their Ph.D. or should expect to complete their degree requirements by September 2020.

Researchers at the institution work to understand the dynamics of the Earth's chemical, physical, and biological systems, from the core to the upper atmosphere, including Earth's interactions with human society. Our scientists lead research in the fields of solid Earth dynamics; ocean, atmospheric, and climate systems; cryospheric dynamics; paleoclimate; and biogeoscience.

The principal selection criteria for Fellows are scientific excellence and a clearly expressed plan to investigate problems at the forefront of Earth science. Applications from all related fields are welcomed.

Fellowships are supported institutionally for 24 months, include a $7,500 research allowance, and carry an annual salary of $66,000. Successful candidates will be encouraged to apply for external funding and may be eligible for further internal awards and positions. Our institution is especially interested in qualified candidates whose record of achievement will contribute to the diversity of our scientific personnel.

We are committed to diversity and are an Equal Opportunity/Affirmative Action employer – Race/Gender/Disability/Veteran.

Appendix C – Study Materials: Topic Review

Below is a list of topics used to describe applicants in letters of recommendation for a fellowship similar to the job description you just read.

Please note, as this is a representative sample of the topics discussed, some topics may seem similar, however, we ask that you please consider each one in turn. Also, some topics may contain more than one theme, please do your best to consider the whole topic, not just one element.

Please review the topics used to describe applicants below and click the circle next to each topic to show that you have read it.

A variety of topics were used across individual applicant letters. For any given applicant, topics might explain that…

	Topic Reviewed
An applicant is engaged in environmental management science	○
An applicant understands complex systems	○
An applicant is a skilled experimentalist	○
An applicant is publishing frequently	○
An applicant will be a good fit	○
An applicant had a successful thesis defense	○
An applicant achieves scientific results	○
An applicant tackles research problems	○
An applicant has computer programming skills	○
An applicant is hard working	○
An applicant received a strong recommendation	○
An applicant asks insightful questions	○
An applicant uses and develops methods	○

An applicant is a high potential academic and researcher, gets funding	○
An applicant excels academically and professionally	○
An applicant received support for their application by letter writer	○
An applicant is a capable technician	○
An applicant is a research contributor now and future	○
An applicant is improving their English	○
An applicant is a top student and teacher	○
An applicant is dedicated to research	○
An applicant is a pleasant team member	○
An applicant is an outstanding young researcher	○
An applicant develops models	○
An applicant is submitting manuscripts	○
An applicant is an accomplished teacher and department citizen	○
An applicant is making progress and maturing	○
An applicant is presenting and publishing nationally and internationally	○

Appendix D – Study Materials: Topic Rating

Please now rate how important each topic would be in your decision to hire an applicant.

Smaller numbers indicate that the topic is less important and larger numbers indicate that the topic is more important. For example, as '1' is the smallest number, it indicates that the topic is not at all important in your decision to hire an applicant and as '9' is the largest number, it indicates that the topic is extremely important in your decision to hire an applicant.

In your decision to hire a post-doctorate fellow in the geosciences, how important would it be that...

	Not at all important							Extremely important	
	1	2	3	4	5	6	7	8	9
An applicant is engaged in environmental management science	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant understands complex systems	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a skilled experimentalist	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is publishing frequently	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant will be a good fit	○	○	○	○	○	○	○	○	○

Matrix continued on next page

	Not at all important								Extremely important
	1	2	3	4	5	6	7	8	9
An applicant had a successful thesis defense	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant achieves scientific results	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant tackles research problems	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant has computer programming skills	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is hard working	○	○	○	○	○	○	○	○	○

	Not at all important								Extremely important
	1	2	3	4	5	6	7	8	9
An applicant received a strong recommendation	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant asks insightful questions	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant uses and develops methods	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a high potential academic and researcher, gets funding	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant excels academically and professionally	○	○	○	○	○	○	○	○	○

Matrix continued on next page

	Not at all important								Extremely important
	1	2	3	4	5	6	7	8	9
An applicant received support for their application by letter writer	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a capable technician	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a research contributor now and future	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is improving their English	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a top student and teacher	○	○	○	○	○	○	○	○	○
	Not at all important								Extremely important
	1	2	3	4	5	6	7	8	9
An applicant is dedicated to research	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a pleasant team member	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is an outstanding young researcher	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant develops models	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is submitting manuscripts	○	○	○	○	○	○	○	○	○

Matrix continued on next page

	Not at all important							Extremely important	
	1	2	3	4	5	6	7	8	9
An applicant is an accomplished teacher and department citizen	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is making progress and maturing	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is presenting and publishing nationally and internationally	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant is a highly capable engineer	○	○	○	○	○	○	○	○	○
	1	2	3	4	5	6	7	8	9
An applicant works well and quickly on tasks	○	○	○	○	○	○	○	○	○
	Not at all important							Extremely important	
	1	2	3	4	5	6	7	8	9
An applicant had letter writer serve on their thesis committee	○	○	○	○	○	○	○	○	○

Appendix E – Study Materials: Topic Ranking

Choose Five Most Important Topics

Please now consider all topics in relation to one another. Among the topics from letters of recommendation, select the those that you would consider the top five most important when making a decision to hire a post-doctorate fellow in the geosciences. Please choose five different topics.

- ☐ An applicant is engaged in environmental management science
- ☐ An applicant understands complex systems
- ☐ An applicant is a skilled experimentalist
- ☐ An applicant is publishing frequently
- ☐ An applicant will be a good fit
- ☐ An applicant had a successful thesis defense
- ☐ An applicant achieves scientific results
- ☐ An applicant tackles research problems
- ☐ An applicant has computer programming skills
- ☐ An applicant is hard working
- ☐ An applicant received a strong recommendation
- ☐ An applicant asks insightful questions
- ☐ An applicant uses and develops methods
- ☐ An applicant is a high potential academic and researcher, gets funding
- ☐ An applicant excels academically and professionally
- ☐ An applicant received support for their application by letter writer
- ☐ An applicant is a capable technician
- ☐ An applicant is a research contributor now and future
- ☐ An applicant is improving their English
- ☐ An applicant is a top student and teacher
- ☐ An applicant is dedicated to research
- ☐ An applicant is a pleasant team member
- ☐ An applicant is an outstanding young researcher

- [ ] An applicant develops models
- [ ] An applicant is submitting manuscripts
- [ ] An applicant is an accomplished teacher and department citizen
- [ ] An applicant is making progress and maturing
- [ ] An applicant is presenting and publishing nationally and internationally
- [ ] An applicant is a highly capable engineer
- [ ] An applicant works well and quickly on tasks
- [ ] An applicant had letter writer serve on their thesis committee

Choose Five Least Important Topics

Again, please consider all topics in relation to one another. Among the topics from letters of recommendation, select the those that you would consider the **bottom five least important** when making a decision to hire a post-doctorate fellow in the geosciences. Please choose five different topics.

- [ ] An applicant is engaged in environmental management science
- [ ] An applicant understands complex systems
- [ ] An applicant is a skilled experimentalist
- [ ] An applicant is publishing frequently
- [ ] An applicant will be a good fit
- [ ] An applicant had a successful thesis defense
- [ ] An applicant achieves scientific results
- [ ] An applicant tackles research problems
- [ ] An applicant has computer programming skills
- [ ] An applicant is hard working
- [ ] An applicant received a strong recommendation
- [ ] An applicant asks insightful questions
- [ ] An applicant uses and develops methods
- [ ] An applicant is a high potential academic and researcher, gets funding
- [ ] An applicant excels academically and professionally
- [ ] An applicant received support for their application by letter writer
- [ ] An applicant is a capable technician
- [ ] An applicant is a research contributor now and future
- [ ] An applicant is improving their English
- [ ] An applicant is a top student and teacher
- [ ] An applicant is dedicated to research
- [ ] An applicant is a pleasant team member
- [ ] An applicant is an outstanding young researcher

- [ ] An applicant develops models
- [ ] An applicant is submitting manuscripts
- [ ] An applicant is an accomplished teacher and department citizen
- [ ] An applicant is making progress and maturing
- [ ] An applicant is presenting and publishing nationally and internationally
- [ ] An applicant is a highly capable engineer
- [ ] An applicant works well and quickly on tasks
- [ ] An applicant had letter writer serve on their thesis committee

Appendix F – Study Materials: Importance of Letters

What percent of your total decision to hire would you attribute to information provided in letters of recommendation?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
○	○	○	○	○	○	○	○	○	○	○

Appendix G – Study Materials: Demographic Questions

Instructions: Please provide some demographic information about yourself.

What is your ethnicity?
- o Hispanic or Latino
- o Black or African American
- o Native American or American Indian
- o Asian or Asian American
- o White or European American
- o Native Hawaiian or other Pacific Islander
- o Other, please specify

What is your gender?
- o Female
- o Male
- o Transgender
- o Gender Neutral
- o Non-Binary
- o Other, please specify

What is your work status?
- o Non-teaching position
- o Adjunct – non-tenure track
- o Instructor
- o Assistant Professor
- o Associate Professor
- o Full Professor
- o Professor Emeritus
- o Researcher
- o Other, please specify

For how long have you been working in an academic institution?
- o 0-2 years
- o 3-5 years
- o 6-8 years
- o 9-11 years
- o 12+ years

What is your highest level of education?
o Less than high school degree
o High school degree or equivalent (e.g., GED)
o Some college
o Associate's degree
o Bachelor's degree
o Master's degree
o Doctoral Degree
o Professional Degree
o Other, please specify

How many years has it been since you completed your highest level of education?
o 0-2 years
o 3-5 years
o 6-8 years
o 9-11 years
o 12+ years

What is your age?
o 18-28 years-old
o 29-39 years-old
o 40-49 years-old
o 50-59 years-old
o 60-69 years-old
o 70+ years-old

Do you now or have you in the past employed post-doctorates?
o Yes
o No
o Other, please specify

Have you made hiring decisions in your current or past academic department(s)?
o Yes
o No
o Other, please specify

In general, how many times have you made a hiring decision?
o I have never made a hiring decision
o 1-3 times
o 4-6 times
o 7-9 times
o 10-12 times
o 13+ times

How many times have you used letters of recommendation for making a hiring decision?
o 0 times
o 1-3 time(s)
o 4-6 times
o 7-9 times
o 10-12 times
o 13+ times

How many times have you used letters of recommendation for making a decision in student admission?
o I have never used letters of recommendation for student admissions
o 1-3 time(s)
o 4-6 times
o 7-9 times
o 10-12 times
o 13+ times

How many faculty members are in your program?
o 1-3 faculty member(s)
o 4-6 faculty members
o 7-9 faculty members
o 10-12 faculty members
o 13+ faculty members

Which area best represents your field of study?
o atmospheric science
o environmental science
o glaciology
o geography
o geology
o geophysics
o hydrology (including oceanography and limnology)
o soil science
o space sciences

In what country do you work?
o Drop down listing to choose one of 195 countries in the world

How did you hear about this survey?
o American Geophysical Union (AGU) announcement
o Department announcement

Thank you for your participation!

Please be sure to press >> to submit your responses.

Appendix H – Study Materials: Debriefing Statement

Thank you for your participation. This study is interested in geoscience researcher ratings of topics describing applicants expressed across a sample of letters of recommendation for post-doctoral fellowships in the geosciences. You were asked to read a job announcement, and review, rate, and rank topics describing applicants in terms of how important they are in your decision to hire.

In a previous study we found that these topics were expressed in different proportions for male and female applicants. That is, some topics were used more often to describe male or female applicants. Based on previous research and an extensive literature review, which has shown, for instance, that stereotypes of scientists are more closely associated with stereotypes of men than of women (Carli et al., 2016), and that letters expressing traits more frequently used to describe female applicants in letters of recommendation are also rated lower in terms of hireability (Madera et al., 2009), we hypothesized that the topics expressed more frequently for male applicants are also those that you would rate and rank as more important in your decision to hire.

Women receive 44.9% of all doctorate degrees in the geosciences, atmospheric sciences, and ocean sciences but only 37.6% of all postdoctoral appointments (National Science Foundations, 2017). If the expected results are found, the findings will contribute greatly to understanding one possible reason for this decline. The findings can bring awareness regarding subtle differences in applicant descriptions that compound to create unequal outcomes. The results may help inform the development of training for how evaluators and letter writers can most equitably read and write letters of recommendation, respectively. These findings will also contribute to learning in the fields of organizational psychology, geosciences, and gender studies.

Carli, L. L., Alawa, L., Lee, Y., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women scientists. *Psychology of Women Quarterly, 40*(2), 244.

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,* 1591-1599.

National Science Foundation (2017), Survey of Graduate Students and Postdoctorates in Science and Engineering, *National Center for Science and Engineering Statistics.* Retrieved from: https://ncsesdata.nsf.gov/gradpostdoc/2017/html/gss17-dt-tab004-2.html

Thank you for your cooperation. If you have any questions, comments, or suggestions, please contact Josh Elmore at je2467@tc.columbia.edu.