Is There Really a Pro-Woman Bias in Academic Hiring?

A Replication and Extension of Williams & Ceci (2015)

RUNNING HEAD: IS THERE REALLY A PRO-WOMAN BIAS IN ACADEMIC HIRING?

Abstract

Decades of social scientific research has found that women face discrimination in stereotypically masculine occupations and domains, such as leadership, the workplace, and academia. However, a recent series of large-scale hiring experiments by Williams and Ceci (2015A) challenged this conclusion, finding that not only were women *not* disadvantaged in academic hiring, they were actually favored at a rate of 2 to 1. These findings raise questions about whether gender bias may have declined – or perhaps even reversed – in the decades that have elapsed since most classic research on gender bias was conducted. In this work, we propose a replication and extension of Williams and Ceci (2015A) to provide additional insight into the questions of whether and when women may be advantaged in academic hiring. In four pilot studies (total N = 2,459), we identify two possible boundary conditions that may limit the generalizability of Williams and Ceci (2015A), suggesting that this pro-woman bias may be limited to 1) exceptionally qualified women and 2) subjective, non-zero-sum outcome measures (e.g., those measuring verbal praise rather than allocations of objective resources like salary and start-up funding). In our registered report proposal, we plan to extend these findings to a sample of tenure-track academics to provide a more ecologically valid test of these questions. In doing so, we aim both to provide a better understanding of this highly influential set of studies, as well as to shed greater light on the current state of gender bias in academia and beyond.

Keywords: Gender bias, Representation, Replication, STEM, Disparities

Is There Really a Pro-Woman Bias in Academic Hiring?

A Replication and Extension of Williams & Ceci (2015)

"These results suggest it is a propitious time for women launching careers in academic science."

– Wendy M. Williams and Stephen J. Ceci (2015)

Women are underrepresented in a variety of academic disciplines (Cheryan, Ziegler, Montoya, & Jiang, 2017). For instance, in the life and social sciences women now earn the majority of doctorates, yet make up a minority of assistant professors (Williams & Ceci, 2015A). In 1993-1995, women earned 41.6% of Ph.D.s but received only 28.4% of assistant professorships (Ceci, Ginther, Kahn, & Williams, 2014). By 2008-2010, this gap had actually widened: women received 53.2% of doctorates, yet only 31.6% of assistant professorships. This gender divide holds even after controlling for demographic factors, degree characteristics, and field (Williams & Ceci, 2015A). Further, this gap in achievement between women and men is not limited to academia, but is paralleled by comparable gender divides across numerous domains ranging from the workplace to politics (U.N. Women, 2016).

For decades, the dominant explanation for this gender gap has been discrimination against women in the workforce (Burgess & Borgida, 1999; Eagly & Karau, 2002; Heilman, 2012). This perspective is bolstered by a large body of experimental research that shows that women tend to experience discrimination in stereotypically "male-typed" domains such as academia, politics, and the workplace (e.g., Johnson, Murphy, Zewdie, & Reichard, 2008; Knobloch-Westerwick, Glynn, & Huge, 2013; Okimoto & Brescoll, 2010). This research has demonstrated that there is a "lack of fit" (Heilman, 1983, 2012) between the traits and

characteristics that women are stereotypically believed to possess (communality but not agency) and the traits and characteristics that are seen as necessary for success in these male-typed domains (agency but not communality; Heilman, 1983; Kite, Deaux & Haines, 2008; Wood & Eagly, 2010). Because of this perceived lack of fit, women are believed to be ill-equipped to succeed in these domains, and as a result they are less likely to be hired for or promoted in these positions (Gaucher, Friesen, & Kay, 2011; Hoobler, Wayne, & Lemmon, 2009; Lyness & Heilman, 2006; Schmader, Whitehead & Wysocki, 2007), and when in these roles tend to receive fewer resources, lower salaries, and suffer other negative outcomes (Institute for Women's Policy Research, 2017).

Given the persistent gender gap in academia and other domains, as well as the extensive history of empirical evidence documenting bias against women in male-typed roles and occupations, it was somewhat surprising to see that a recent series of studies found that women were actually favored over men in tenure-track faculty hiring decisions at a rate of 2 to 1 (Williams and Ceci, 2015A; hereafter W & C) . This work was highly publicized – already ranking among the most widely discussed articles ever published by the *Proceedings of the National Academy of Sciences* (Altmetric, 2019) – and elicited a great deal of heated debate (e.g., Francis, 2015; W. M. Williams & Ceci, 2015B; J. C. Williams & Smith, 2015). Although some researchers questioned W & C's results and methods (e.g., Blau & Kahn, 2016; Francis, 2015; Haynes & Sweedler, 2015; J. C. Williams & Smith, 2015), others embraced the findings (e.g., Boynton et al., 2018; Mulligan, 2017; Stewart-Williams & Halsey, 2018), declaring the end of gender discrimination and concluding – as did the authors themselves – that "it is a propitious time for women launching careers in academic science" (pg. 5360).

But what explains the divergence between past social scientific research, which has tended to find bias *against* women (for a meta-analysis, see Koch, D'Mello, & Sackett, 2015), and W & C's studies, which documented bias *in favor* of women? Were W & C's results (a) simply due to chance or (b) to the particular design and methodology of their studies? Or might they suggest (c) that the anti-woman gender bias that has been documented elsewhere does not extend to this particular domain? Or, alternatively, is it possible that (d) the landscape of bias has truly changed, and that women no longer experience discrimination – and may now even be advantaged – in academic hiring?

At first blush, the evidence seems to support the existence of continued gender-based discrimination: as noted, real-world gender gaps in male-typed domains persist, and there is a large body of experimental research documenting gender bias against women in these domains. However, there are also reasons to believe that W & C's findings may indicate that gender bias has truly diminished. Although gender gaps in academia and other male-typed domains clearly exist, factors other than discriminatory hiring have been argued to explain this divide – such as "leaky pipeline" explanations, which contend that women are underrepresented in male-typed roles and occupations not because they experience discrimination in being hired or promoted in these roles, but because they choose different career paths, or elect to leave their careers to focus on raising a family (Ferriman, Lubinski, & Benbow, 2009; Hakim, 2006; though see also Dennehy & Dasgupta, 2017 and Dasgupta, Scircle, & Hunsinger, 2015 for evidence against this explanation).[1] Similarly, although there is extensive experimental evidence documenting discrimination against women in male-typed roles and occupations, the great majority of this

[1] Important to note is that these explanations do not necessarily posit that bias and discrimination do not contribute to women's choice to leave these male-typed domains – they simply argue that women's underrepresentation is due to them not applying for these positions, rather than applying and not being selected.

research was conducted years – or even decades – ago (for a review, see Koch et al., 2015), and it is possible that gender-based discrimination may have diminished in the intervening years, in line with documented decreases in other social biases (Charlesworth & Banaji, in press). Supporting this possibility, there is research suggesting that stereotypes may have shifted over the course of the last several decades, such that women are no longer seen to be as ill-equipped to succeed in some male-typed domains (Koenig, Eagly, Mitchell & Ristikari, 2011; Lewis & Michalak, 2018; Sczesny, Bosak, Neff, & Schyns, 2004).

Thus, W & C's findings raise many questions regarding the current state of gender bias in academia and beyond. In this proposed research, which constitutes a replication and extension of W & C, we seek to further examine whether anti-woman discrimination in academic hiring has come to an end – or whether there may even now be a pro-woman bias – or whether there might be particularities about W & C's research design that explain these divergent results. After a closer examination of the literature, we identified two key factors that differed between W & C's design and most past research, which we suspect may explain the pro-woman bias found in W & C's studies: (1) the candidates in W & C's studies were exceptionally qualified, which may have eliminated the ambiguity that typically gives rise to gender bias against women, and (2) W & C's outcome variable was more subjective than those used in previous research, and as a result may have been be more susceptible to social desirability bias and shifting standards.

In this research, we test these factors to examine whether there is truly a 2:1 hiring preference for women in academic science, or whether that finding might be constrained to particular kinds of women (specifically, exceptionally qualified women), and only emerge when these women are evaluated using W & C's specific dependent measure. Answering this question requires both (a) a direct replication of W & C's findings, and (b) extending those findings by

testing their robustness to different operationalizations (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018) – in this case, different levels of candidate qualification strength and a different dependent variable. Conducting this research will provide a better understanding of the factors that can give rise to gender bias (whether for or against women) in hiring decisions in academic science.

Qualification Strength:

In W & C's original studies, they contacted faculty members from across a range of different universities and four different STEM disciplines and asked them to evaluate hypothetical candidates for an assistant professorship. They created different sets of application materials, and randomized whether the candidate described was a man or a woman (indicated by gender pronoun only). They found that the woman candidate was overwhelmingly favored in these studies, at a rate of 2 to 1.

Notably, however, the candidate described in W & C's studies was no average applicant; rather, s/he was unambiguously extraordinary, being described as nearly perfect on every dimension. For example, in one set of materials the candidate is said to have scored a 9.5 out of 10 on the job talk and interview, and to have award-winning teaching skills. S/he is also said to have worked with an "eminent advisor" in a "hot" research area, and to be "poised to break new ground" with an exceptional research program. The candidate also excelled interpersonally, impressing the faculty and being described as "very likable, kind, and socially skilled." The description also noted that the faculty were unanimous in their agreement regarding the exceptional nature of the candidate's qualifications, and that s/he would be "a great potential hire."

Although it is interesting that W & C found an advantage for this exceptionally qualified woman candidate, there is reason to question whether these effects necessarily represent a general hiring advantage for women. That is, the candidate presented in W & C's materials, as described above, is clearly extraordinary – and past research has shown that in cases such as these, when a woman's qualifications are wholly and unambiguously exceptional, she may not suffer discrimination (Koch et al., 2015) and under some conditions may even be advantaged (Biernat & Fuegen, 2001). However, the average candidate is, by definition, not extraordinary. Even among highly qualified people such as Ph.D. candidates, many individuals have at least one area in which they are less than perfect. Further, it is unlikely that groups of 20+ academics typically agree unanimously regarding the qualifications and fit of a candidate (indeed, faculty searches often fail, even when a department has interviewed multiple candidates). Put another way, qualification information in the real world almost invariably provides some source of ambiguity. And ambiguity, as much past research has shown, increases the likelihood that gender bias will emerge (for a review, see Heilman, 2012).[2]

Past research, then, would suggest that if W & C's candidates were presented as somewhat less extraordinary, then this pro-woman bias might disappear. This is an important question for a number of reasons. From a theoretical perspective, the question of whether women suffer bias in traditionally male-typed domains like academia and the workplace has implications for a number of theories of gender bias (e.g., lack of fit model, Heilman, 1982, 2012; role congruity theory, Eagly & Karau, 2002). However, beyond its theoretical implications, this

[2] W & C have mentioned the question of qualification strength in other discussions of their work (W & C, 2015C). However, they argue, based on their personal experience at an ivy league university, that the materials they designed are representative of the typical job candidate at a top-tier research university. However, the two authors of this paper who have participated in search committees (for job searches that were conducted at the very same university as W & C) have found this to be quite an uncommon occurrence: it is not often that a candidate excels on every possible dimension, nor that faculty unanimously agree on the superiority of a given prospective hire.

question is also important from a practical perspective. W & C's findings garnered a great deal of press (Altmetric, 2019) and, as the authors argue, suggest that gender bias against women in academia has not only disappeared, but that women are now advantaged in academic hiring. If this is true, this has important implications for real-world policies aimed at creating gender diversity – both in academia and beyond – and suggests that targeted-placement and other affirmative-action-style policies that ensure women's representation in these domains are either no longer necessary, or may need to be altered in order to address other possible (non-bias-related) explanations for the gender gap in academia (cf. Dennehy & Dasgupta, 2017)

Conversely, if the pro-woman bias found by W & C is limited to situations in which candidates have flawless and exceptional qualifications, then these findings, though interesting from a theoretical perspective, may have more limited practical import, as the majority of female candidates may still face gender bias in academic hiring. A closer examination of this question is therefore needed to examine the constraints on the generalizability (Simons, Shoda, & Lindsay, 2017) of W & C's findings in order to ensure that we do not prematurely conclude that anti-woman gender bias is at an end, or that women enjoy a hiring advantage that they do not actually have.

For the reasons outlined above, we wished to revisit W & C's original research and to determine whether the pro-woman bias that they observed would be limited to situations in which the target candidate is exceptionally qualified. So far we have tested this question in four pilot studies (total N = 2,459). Below, we describe the results of these preliminary studies. We then outline additional proposed field research that can more conclusively test this question and provide further insight into the current state of gender bias. All materials, data, syntax, and

preregistration documentation can be found on the Open Science Framework at https://osf.io/j8yz6/?view_only=13c23f6afb1444a0aafaf3b2fd55d730.

Pilot Study 1: Does Ambiguity Moderate the Effect of Candidate Gender on Hiring Preferences?

In our first study, we tested the hypothesis that the hiring advantage for women observed by W & C would be limited to exceptionally qualified candidates. We hypothesized that moderating or "toning down" some of the positive information about the candidates – and thus adding a degree of ambiguity to their qualifications – would attenuate, or perhaps even reverse, the pro-female bias documented by W & C.

In their paper, W & C used multiple variations of their candidate-evaluation paradigm, raising the question of which specific design would be most appropriate for this replication. However, some of their paradigms have been criticized on the grounds that they (1) were overly methodologically complex and offered poor experimental control (e.g., involving unmatched stimuli and more than 20 different sets of experimental materials; J. C. Williams & Smith, 2015) and (2) that the results were likely to have been skewed by socially desirable responding (i.e., a desire to present oneself as egalitarian and unbiased; Haynes & Sweedler, 2015; J. C. Williams & Smith, 2015). Specifically, because their within-subjects design required people to explicitly choose between a male and female candidate, this may have led some participants to suspect that the study dealt with gender bias. Based on these critiques, we chose to replicate W & C's Experiment 5, which (1) used identical materials for both the male and female candidate, providing good experimental control and (2) asked participants to judge only a single (male or

female) candidate, thereby likely reducing suspicion regarding the purpose of the study and minimizing socially desirable responding.

Participants

For our initial pilot studies, we collected participants from Amazon's Mechanical Turk (for a discussion of Mechanical Turk as a research tool, see Buhrmester, Kwang, & Gosling, 2011).

Design

As in W & C's original paradigm, participants in this study (N = 394; target sample size of n = 50 per cell; 43% women, median age = 32.5) reviewed information about a male or female candidate for a tenure-track assistant professor position. They were asked to imagine that they were in the role of a university professor and that their department was searching for a new faculty member. The information that participants received was said to be the search committee chair's summary of the committee's general perceptions and impressions of the candidate. This summary discussed several relevant aspects of the candidate's qualifications, such as his/her past research performance and teaching skills. After reviewing the information, participants rated the candidate on a 10-point scale designed by W & C, which assessed the degree to which they felt that the s/he was qualified for the position and should be hired.

In addition to manipulating the gender of the candidate (man vs. woman) we also manipulated the strength of the candidate's qualifications. Participants were randomly assigned to receive one of four different sets of materials – either the original W & C vignette, or one of three novel, more ambiguous sets of materials that we created by slightly altering the wording of the original vignette. In the original W & C materials, the candidate is described as exceptionally qualified on all four of the dimensions that are discussed: research track record, teaching skills,

job talk/interview, and interpersonal skills. In the second set of materials (the "Warm Materials" condition), the candidate was described in positive – but less hyperbolic – terms. The remaining two conditions depicted the candidate as having "mixed" qualifications, such that s/he was described as exceptional on two dimensions (e.g., teaching and research skills) but more negative on the other two dimensions (e.g., job talk/interview and interpersonal skills). For example, in one mixed-qualifications condition, the candidate is said to have an exceptional teaching record, being described as "an effective and supportive mentor" and having won a teaching award in graduate school. However, this same candidate is described as having sub-par interpersonal skills, having "failed to impress" at dinner with the faculty and "not coming off as particularly likable or kind." Conversely, the second set of mixed-qualifications materials reversed the specific dimensions on which the candidate was said to be good and bad, describing the candidate as having poor teaching skills (e.g., being a "difficult and distant" mentor) but good interpersonal skills (e.g., being "very likable and kind").[3] In sum, then, while the former two conditions described a consistent (excellent or positive) candidate, the latter two conditions described a candidate with both strengths and weaknesses (for exact wording of all materials, see Supplementary Information; SI).

Following the candidate rating task, participants completed a few exploratory measures (discussed below) and provided demographic information. They also answered an attention-check question in which they were asked to indicate the gender of the candidate in the vignette

[3] Past research has shown that domains perceived as requiring more communal (vs. agentic) traits tend to be seen as "female-typed," while domains perceived as requiring more agentic (vs. communal) traits tend to be seen as "male-typed" (Cejka & Eagly, 1999). Based on this work, we identified two of these domains (teaching and interpersonal skills) as being female-typed, and two domains (research and job talk) as being male-typed. To balance perceived gender (in)congruity in these candidate descriptions, we designed our mixed materials so that the candidate was always said to excel in one male- and one female-typed domain, and to be less exceptional in one male- and one female-typed domain.

that they read. (In keeping with our preregistered analysis plans, in all studies we excluded participants who failed the attention check by not accurately identifying the candidate's gender.)

Results

Manipulation Check:

Before testing our primary hypotheses, we first examined the mean ratings of the candidates in the four different materials conditions, collapsing across gender condition, to ensure that our new materials were in fact judged to be less positive than those designed by W & C. We found a significant effect of condition on candidate rating ($F(3, 316) = 74.70$, $p < .001$). As expected, all three of our novel materials conditions were rated as significantly less positive than the original W & C materials (all *p*s < .001, Bonferroni correction for multiple comparisons). The original W & C materials were rated most positively (M = 8.3 out of 10), followed by the warm materials (M = 7.1), then the first set of mixed-qualifications materials (exceptional research and interpersonal skills, M = 5.4), and then the second set of mixed-qualifications materials (exceptional teaching skills and job talk, M = 4.4). Ratings of all conditions were significantly different from one another at $p <= .003$ (Bonferroni correction for multiple comparisons).

When interpreting these mean ratings, it is important to note that W & C's rating scale (a novel measure they designed for their studies) was intentionally structured in a way that allowed for fine-grained distinctions between exceptional candidates, with only the lowest scale points describing below-average candidates (scale points 1 and 2) and the majority of scale points describing good-to-exceptional candidates (scale points 4-10). Thus, the mean ratings of these four materials conditions do not correspond to the evaluations that a typical "balanced" scale might be expected to (e.g., a mean rating near the midpoint of the scale does *not* indicate that

participants thought that the candidate was average, but rather, quite good). Therefore, although these novel conditions were rated as less positive than the original W & C materials, they were not rated negatively. The mean rating of the warm materials (7.1) corresponds to "Extremely impressive candidate/offer all typical recruitment incentives," and the mean rating of the first set of mixed-qualifications materials (5.4) corresponds to "Very good candidate/I am enthusiastic about hiring this person." Even the lowest rated of the four conditions still had a mean rating (4.4) that corresponded to "Good candidate, pursue if resources allow." On the whole, then, even the candidates with more ambiguous qualifications were still judged to be quite good – which is likely to provide a particularly stringent test of our predictions (Biernat & Fuegen, 2001; Heilman, 2012; Koch et al., 2015).

Main Effect of Gender: If, as W & C suggest, women now enjoy a general hiring advantage, then we should expect to see a significant main effect of gender on ratings, such that the woman would be rated more highly in general. However, examining the mean ratings of the man and woman candidate revealed no main effect of gender (p = .89), with both candidates being rated almost identically (man mean = 6.32; woman mean = 6.35). These results speak against the existence of a general hiring advantage for women, at least in this sample.

Focal Hypothesis: Gender x Qualification-Strength Interaction: Our original prediction was that the pro-woman bias found by W & C would disappear when the candidates' qualifications were made somewhat less exceptional. Unexpectedly, however, we did not replicate W & C's results using their original materials, instead finding that ratings of the exceptional woman candidate were not significantly different from those of the exceptional man candidate (p = .23).

However, there was a weak directional (though non-significant) interaction between gender condition (male vs. female) and materials condition (excellent vs. ambiguous) that provided some tentative support for W & C's original findings, as well as our own predictions ($F(1,316) = 2.44$, $p = .119$, Fig. 1). When W & C's original materials were used, we directionally replicated the general pattern that they observed (albeit with a smaller, and statistically non-significant, effect size: $\eta p^2 = .025$ vs. W & C $\eta p^2 = .118$), such that the exceptionally qualified woman was rated somewhat more highly than the exceptionally qualified man (exceptional man M = 7.97; exceptional woman M = 8.52; $t(316) = 1.20$, $p = .23$)

Conversely, when the woman's qualifications were more ambiguous, she was rated somewhat (though non-significantly) more *negatively* than the man (ambiguous man M = 5.84; ambiguous woman M = 5.57; $t(316) = 1.05$, $p = .29$). This trend towards lower ratings for the more ambiguously qualified woman was true in all three of our novel materials conditions, with the woman being rated lower than the man in all cases (warm materials: man M = 7.26, woman M = 6.92, $p = .37$; mixed materials 1: man M = 5.67, woman M = 5.21, $t(316) = 1.20$, $p = .23$; mixed materials 2: man M = 4.67, woman M = 4.06, $p = .13$).

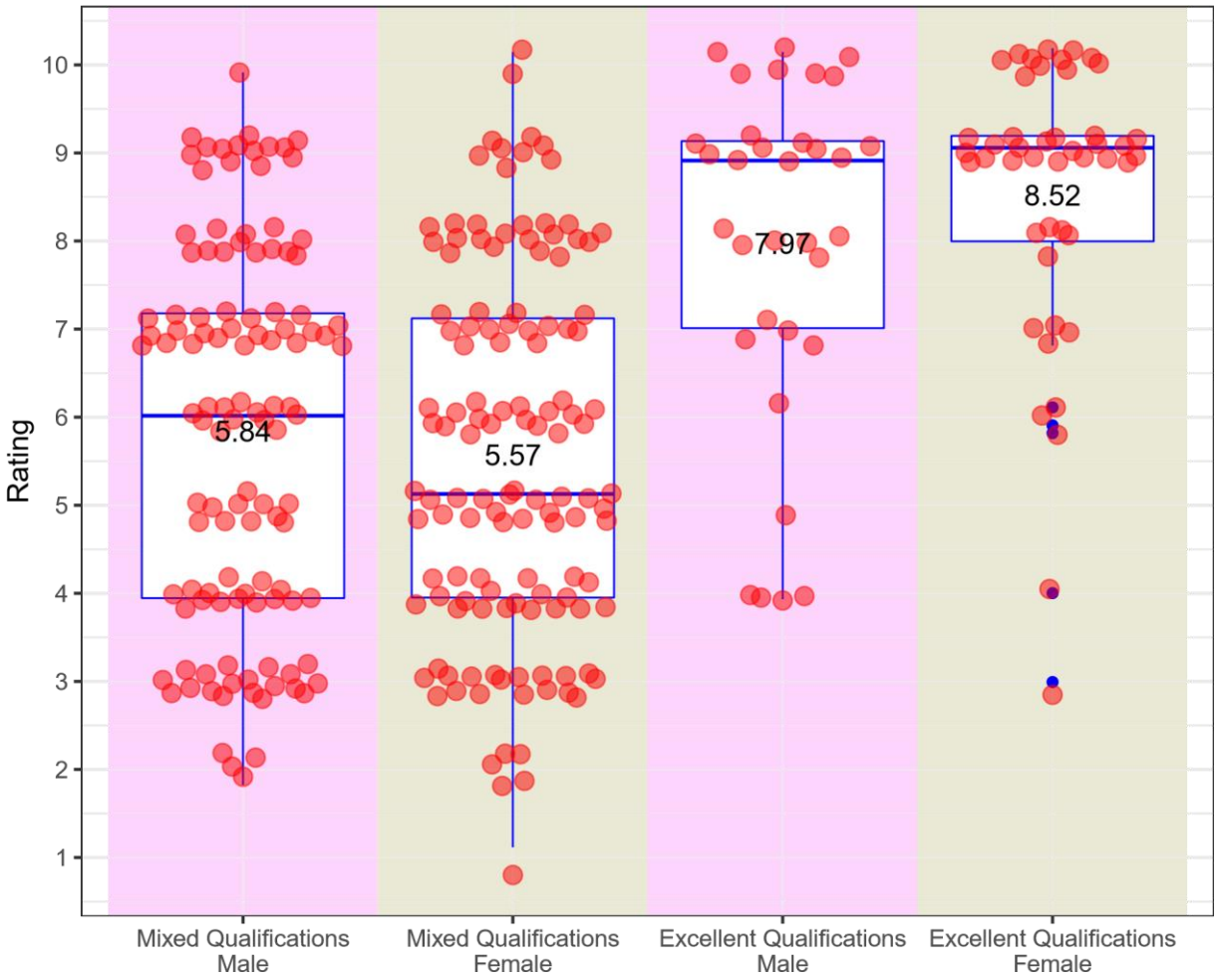

Fig 1 | Gender (man vs. woman) by qualification-level (mixed vs. excellent) interaction on 10-point ratings, Pilot 1.

Discussion

For this first pilot, we predicted that the pro-woman bias found by W & C would be attenuated if the candidates' qualifications were less unambiguously positive. Unexpectedly, however, we failed to replicate W & C's original effect, making the interpretation of our own prediction somewhat more difficult. Nevertheless, there was a weak directional trend that was consistent with our predictions, such that the (non-significant) preference for the exceptional woman (vs. exceptional man) was somewhat attenuated when the candidates' qualifications were less positive. In our second pilot, we increased our sample size (particularly in the exceptional-

materials condition) in order to determine whether additional statistical power would allow us to detect W & C's pro-woman bias and to more decisively test our own prediction that this pro-woman bias will be attenuated when the candidates' qualifications are less exceptional.

Pilot Study 2

Design

The design of this study (N = 479; target sample size of n = 60 per cell; 45% women, median age = 33) was very similar to that of our first pilot. As before, we randomly assigned participants to review either a male or female candidate, indicated by gender pronoun only. We also randomly assigned people to one of two materials conditions: either the original, exceptional candidate materials or to one of the "mixed qualifications" conditions from our first pilot (specifically, Mixed Qualifications 2, in which the candidate is described as having exceptional teaching skills and being rated near perfect on the job/talk interview, but as being a mediocre researcher and having poor social skills). After reviewing the materials, participants rated the candidate on the same 10-point scale used previously. They then completed some exploratory dependent measures (discussed in detail below), answered the attention-check question, and provided demographic information. As an additional exploratory manipulation, we also assigned half of participants to a novel accuracy motivation condition, in order to determine whether instructions that encouraged accurate responses in the task (see SI) would attenuate the pro-female bias observed by W & C. However, this manipulation had no effect and is therefore not discussed further in the main text. (Further information is available on page 20 of the Supplementary Information, for interested readers.)

Results

As predicted, in this study we found a significant interaction between gender condition (male vs. female) and materials type (excellent vs. mixed; $F(1,357) = 6.55$, $p = .011$). With the original materials, we replicated W & C's effect (albeit with a substantially smaller effect size: original W & C study $\eta p^2 = .118$; this study $\eta p^2 = .025$), finding that the exceptional woman candidate was rated more positively than the exceptional man (man M = 7.40, woman M = 8.41, $p < .001$). Conversely – and as predicted – we found that this pro-woman bias was significantly attenuated when the candidates' qualifications were less exceptional. In fact, in this mixed materials condition, there was no advantage for the female candidate, with ratings of the man and woman being virtually identical (man M = 4.21, woman M = 4.22, $p = .97$, Fig. 2).

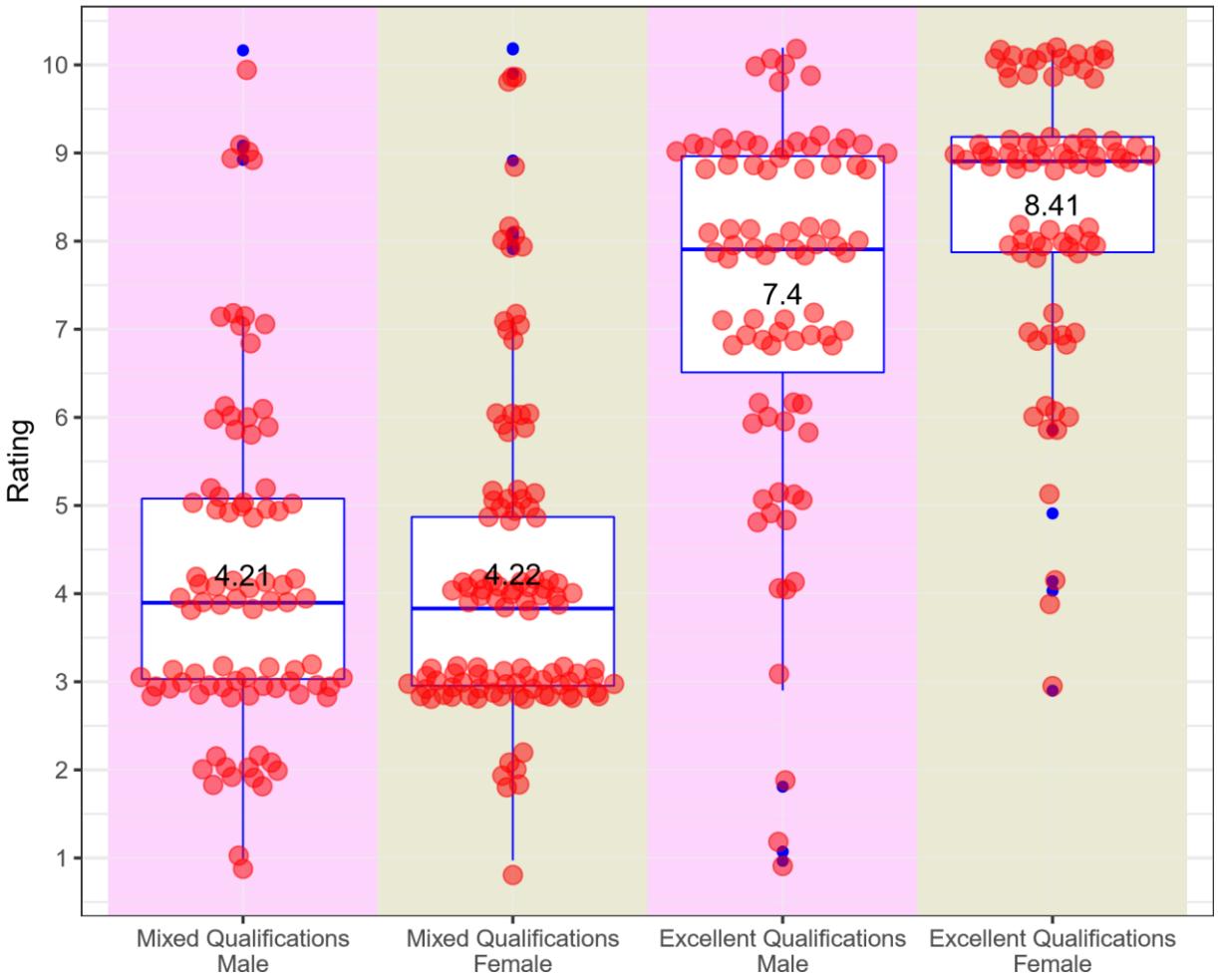

Fig 2 | Gender (man vs. woman) by Qualification-Level (Mixed vs. Excellent) Interaction on 10-Point Ratings, Pilot 2.

Discussion

Taken together, the results of our first two pilots provided tentative support for our predictions. Specifically, they show that the pro-woman bias observed by W & C (which was directionally supported in Pilot 1 and significant in Pilot 2) was not only reduced, but in fact disappeared entirely when the candidates' qualifications were less exceptional. These results suggest that the hiring advantage observed by W & C does not extend to all women, but seems to be limited to situations in which the woman is judged as extraordinarily qualified for the

position. In Pilot 3, we sought to provide an additional replication of this effect with 1) a larger sample and 2) a different set of ambiguous materials.

Pilot 3

Design

For this study, we set a target sample size of 787 participants, based on 80% power to detect an effect size of $d = .2$ / $f = .1$ (our smallest effect size of interest; Lakens & Evers, 2014). The design of this study was very similar to the previous two pilots. Participants ($N = 783$; 52% women, median age = 34) were randomly assigned to evaluate either a male or female candidate, and were randomly assigned to one of two materials conditions (excellent vs. mixed qualifications). In this study, we used the other set of mixed qualification materials from our first pilot (specifically, "Mixed Materials 1," in which the candidate is described as being an exceptional researcher and having great social skills, but as being a below average teacher and as having been rated poorly on the job talk/interview). After rating the candidate, participants answered the attention-check question and provided demographic information.

Results

As in our previous two studies, our prediction in this pilot was that the pro-woman bias observed by W & C would be attenuated when the candidate's qualifications were less exceptional (preregistration at http://aspredicted.org/blind.php?x=kn792d). Unexpectedly, however, in this study we once again failed to replicate the pro-woman bias observed by W & C. Instead, we found no preference for the woman candidate whatsoever ($p = .94$). These ratings of

the female (vs. male) candidate also did not differ as a function of the strength of their qualifications (p = .84): gender had no effect on ratings for either the exceptionally qualified candidate (p = .3) or for the more ambiguously qualified candidate (p = .59, Fig. 3).

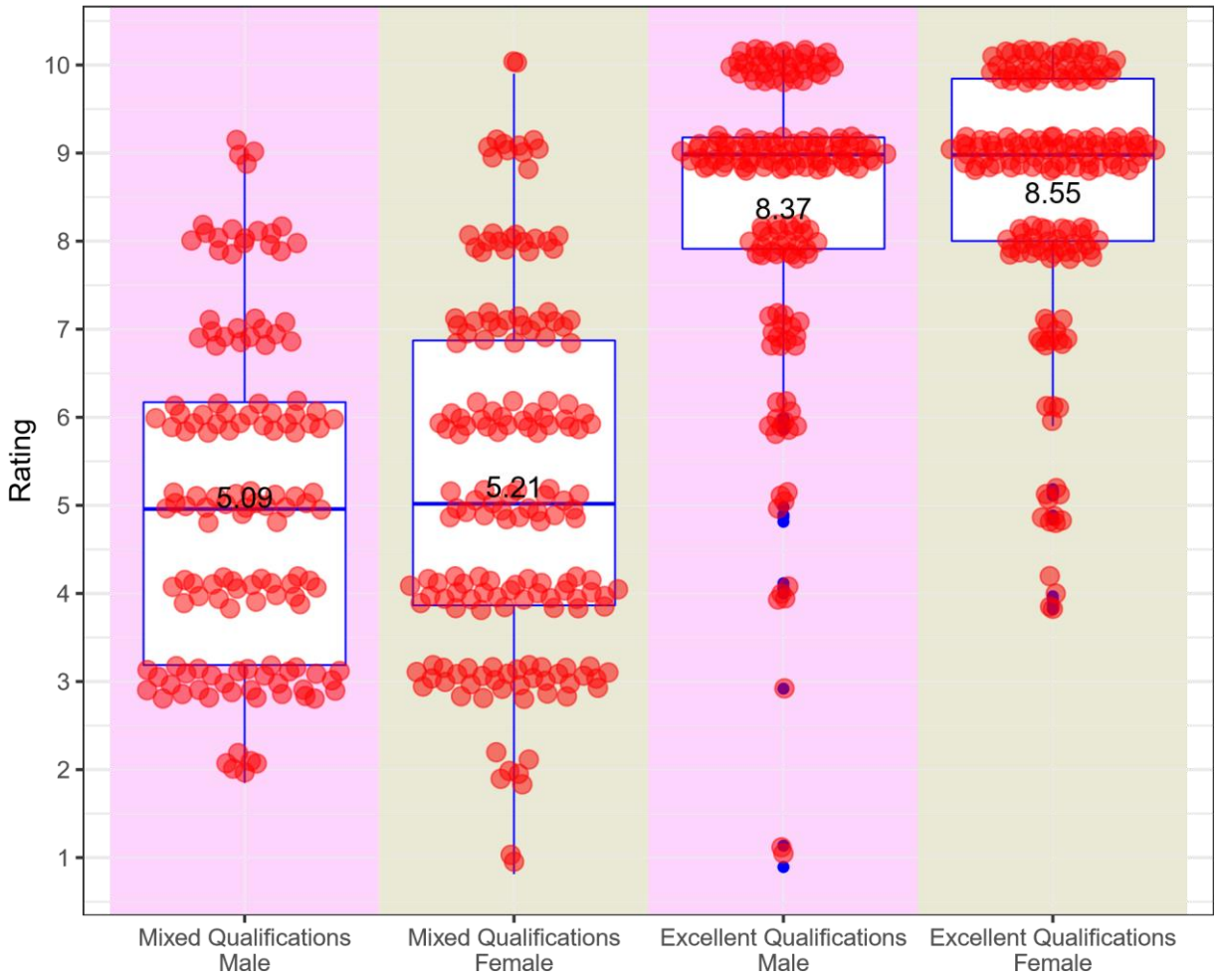

Fig 3 | Gender (man vs. woman) by Qualification-Level (Mixed vs. Excellent) Interaction on 10-Point Ratings, Pilot 3.

Discussion

In this study, we again failed to replicate the pro-woman bias observed by W & C, making it difficult to evaluate our hypothesis. From one perspective, the strength of the candidate's qualifications, contrary to our predictions, did not affect relative ratings of the female (vs. male) candidate. However, we had specifically predicted that the pro-woman bias observed

by W & C would be *attenuated or eliminated* for a less extraordinary woman candidate (as was the case in Pilot 2 and, directionally, in Pilot 1). From this perspective, the results of Pilot 3 could be interpreted as supporting our hypotheses. Indeed, in keeping with our predictions, we found no preference for the woman candidate whatsoever – it was simply the case that this lack of preference for the woman was more widespread than we predicted, holding true not only when the candidates had mixed qualifications, but also when they were exceptionally qualified for the position.

Why did we fail to replicate W & C's pro-woman bias in this study? One possibility is that there may have been some difference in the samples that explains the different pattern of results. Indeed, past work has shown that the characteristics of Mechanical Turk samples can differ as a function of many factors, including the specific time of day and day of the week that the data were collected (Casey, Chandler, Levine, Proctor, & Strolovitch, 2017). Although we were unable to detect any differences (demographic or otherwise) that explained the differing effects observed in these studies, it is possible that some unidentified sample difference(s) might be the cause.

Conversely, given that this sample was substantially larger than those collected in our previous studies, it is also possible that the results of Pilot 2 (and the directional but non-significant results of Pilot 1) may have been false positives. In this pilot, we had 99.99% power to detect an effect of the size obtained by W & C ($\eta p^2 = .118$) which would make a failure to replicate extremely unlikely. Even for an effect half the size of W & C's ($\eta p^2 = .059$), our power in this study would have been 99.57%. The high statistical power in this study therefore makes it seem relatively unlikely that this failure to replicate is due to chance alone, and, at the very least, suggest that the true effect in this sample is considerably smaller than that obtained by W & C.

In sum, then, the results of these first 3 pilots present a mixed picture. Although we do find some evidence of W & C's pro-woman bias among exceptional candidates, this effect appears to be less consistent and robust – at least among this particular sample – than would be suggested by W & C's results. In our fourth and final pilot, we sought to provide an additional replication of this effect in order to provide additional insight into this question.

In Pilot 4, we also manipulated the degree of ambiguity present in the candidates' qualifications in order to determine whether a high degree of ambiguity is necessary for these effects to emerge. That is, in the original sets of mixed materials that we created, the candidate is said to be excellent on two dimensions and to be quite poor on the other two dimensions – for example, in one version the candidate is said to be an excellent teacher and to have given an excellent job talk/interview, but to be rather poor in her/his social skills, and to be a below-average researcher. This relatively stark disparity between these individual qualification dimensions – with the candidate being excellent in certain aspects but poor in others – could be said to constitute a *particularly* ambiguous set of qualifications. It is possible, then, that highly ambiguous qualifications are necessary for the elimination of W & C's pro-woman bias. To test this possibility, we created two new sets of materials in which we manipulated the degree of ambiguity present in the candidate's qualifications. These new materials allow us to determine whether high ambiguity is a necessary condition for these effects to emerge, or whether *any* degree of ambiguity in the candidates' qualifications will be sufficient to eliminate W & C's pro-woman bias.

Pilot 4

Design

As in the previous studies, participants (N = 803; target sample of n = 100 per cell; 47% women, median age = 31) were randomly assigned to review either a male or female candidate. They were also randomly assigned to one of four different materials conditions. The first condition consisted of the original W & C materials used in the previous four pilots. The second condition was the "good-teacher/good job talk" set of mixed materials used in Studies 1 & 2. The latter two conditions were variations of this mixed materials condition, which were altered in order to either amplify or decrease the degree of ambiguity in the candidate's qualifications (i.e., the severity of the disparity between the different qualification dimensions, as discussed above). In all cases, the candidate was still said to be exceptional on the dimensions of teaching and the job talk/interview. In the high-ambiguity condition, however, the candidate was described as being very poor on the other two dimensions (e.g., as being a "below average" researcher and having "poor social skills"). Conversely, in the low-ambiguity condition, the candidate was described in more tepid – but not negative – terms on these latter two dimensions (e.g., being an "above average" researcher and having "mediocre social skills"). After evaluating the candidate, participants rated her/him on the 10-point scale used in the previous studies. They then completed a few exploratory dependent variables (described below) and provided demographic information.

Results

The results of this study provided clear support for our predictions. We found a significant interaction between candidate gender (male vs. female) and qualification strength (excellent vs. ambiguous; $F(1,626) = 6.78$, $p = .009$). Using W & C's original materials, we found that the exceptional woman candidate was rated directionally more positively than the

exceptional man (man M = 7.77; woman M = 8.27, $t(626) = 1.61$, $p = .11$). Conversely, in the ambiguous materials conditions, the woman was rated significantly less positively than the man (man M = 5.12; woman M = 4.68, $t(626) = 2.45$, $p = .015$).

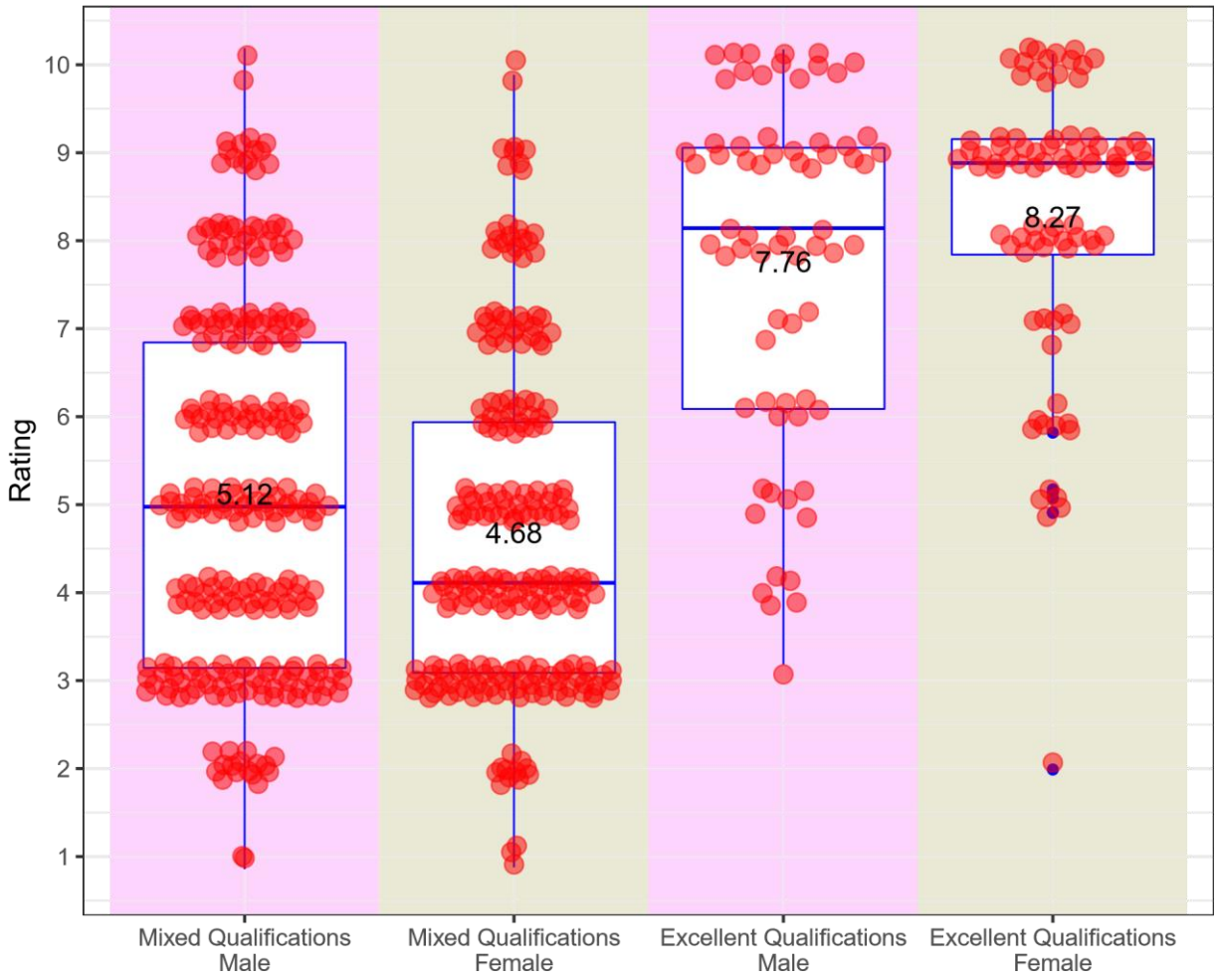

Fig 4 | Gender (man vs. woman) by qualification-level (mixed vs. excellent) interaction on 10-point ratings, Pilot 4.

Further, although we found an effect of ambiguity in general, we did not find significant differences between our three ambiguous materials conditions ($p = .85$), indicating that *degree* of ambiguity did not moderate this effect. In fact, the disadvantage for the female candidate was actually slightly larger in the low ambiguity condition (mean difference = .39) than in the high

ambiguity condition (mean difference = .15) or in the unaltered mixed materials condition (mean difference = .26, Fig. 5).

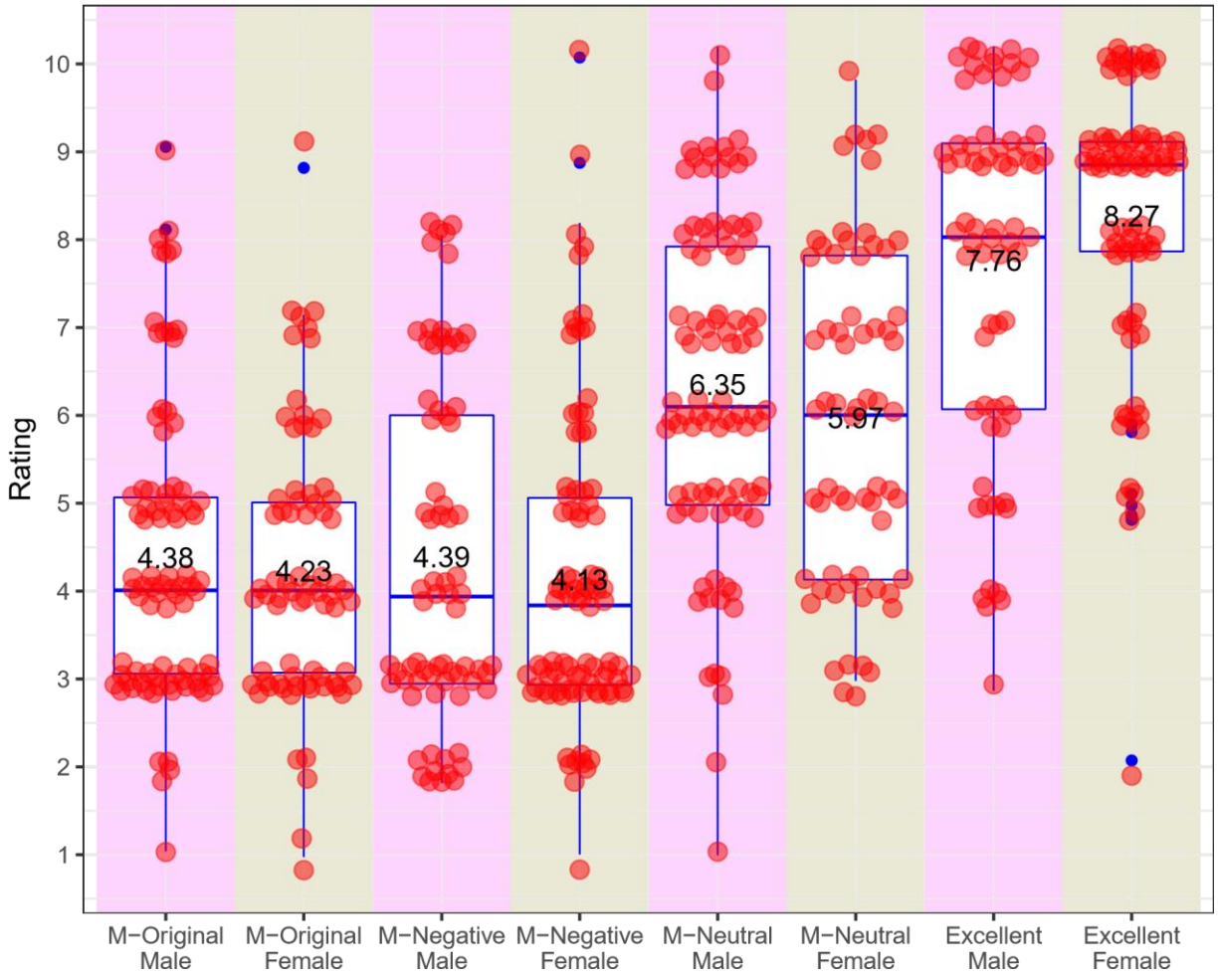

Fig 5 | Gender (man vs. woman) by qualification-level (mixed original vs. mixed negative vs. mixed neutral vs. excellent) interaction on 10-point ratings, Pilot 4.

Discussion

The results of this study provided tentative (though non-significant) support for the pro-woman bias for exceptional candidates that was observed by W & C. It also provided clear support for our prediction that adding any degree of ambiguity to the candidates' qualifications

would eliminate this (non-significant) female hiring advantage. In fact, in this study we found that not only were the less extraordinarily qualified female candidates not advantaged in these evaluations, they were actually significantly disadvantaged, with the woman candidate being rated more negatively than the man in all three of our mixed materials conditions. Further, we also found that a high degree of ambiguity is not necessary for this effect to emerge. Rather, it appears that adding any degree of ambiguity to the candidates' qualifications is enough to eliminate the pro-female bias and perhaps to even give rise to gender bias in evaluations of women.

Summary of Pilot Studies

In Pilot 1, we provided an initial test of whether adding ambiguity to the candidate's qualifications would attenuate or eliminate the pro-woman bias observed by W & C. To test this question, we compared W & C's original excellent-candidate materials with three novel sets of ambiguous materials that we created by slightly altering the wording of the original vignette. These new materials consisted of a "warm materials" condition in which the candidate was described in positive, but less hyperbolic, terms, as well as two "mixed materials" conditions in which the candidate was described as exceptional on two dimensions but more negative on the other two dimensions. We found a directional interaction between gender condition and materials condition in the predicted direction: with the original W & C materials, we observed a slight (non-significant) advantage for the woman candidate, but this effect was reversed in the ambiguous materials conditions, such that the woman was actually somewhat (non-significantly) disadvantaged in ratings.

In Pilot 2, we sought to replicate this effect with a larger sample, in order to determine whether these effects would emerge more clearly with additional statistical power. In this study

we compared W & C's original excellent materials with one of the mixed materials conditions from Pilot 1. We found a significant interaction in the predicted direction. With the original W & C materials, the woman candidate was significantly favored in evaluations. However, in the mixed qualifications condition this pro-woman bias disappeared completely, with no difference whatsoever between ratings of the woman and man candidate.

In Pilot 3, we sought to provide a high-powered replication of Pilots 1 and 2, using the original W & C materials and the second set of mixed materials from the first pilot. Unexpectedly, in this study we failed to replicate the pro-woman bias for the excellent woman candidate. In this study, we observed no difference in ratings of the woman versus man candidate in either condition.

In Pilot 4, we sought to provide an additional test of our hypothesis to help resolve the mixed findings from the previous pilots. Additionally, in this study we systematically manipulated the degree of ambiguity present in the candidate's qualifications. We found a significant interaction between candidate gender (man vs. woman) and qualification strength (excellent vs. mixed) in the predicted direction, such that the woman candidate was somewhat (but non-significantly) advantaged when the candidate's qualifications were excellent, but that she was significantly disadvantaged in the mixed qualifications condition. Further, we found that the degree of ambiguity did not moderate these effects, suggesting that *any* degree of ambiguity is sufficient to eliminate the pro-woman bias observed by W & C, at least among this participant sample.

Subjective Rating Scales: Shifting Standards and Social Desirability

The results of the four pilots discussed above provide no evidence of a general hiring advantage for women (i.e., no main effect of gender). Surprisingly, they also provide little support for a pro-woman bias even among exceptionally qualified candidates, as was documented by W & C. However, these pilot studies provide considerably stronger evidence that, at least in this sample, less exceptional woman candidates are *not* advantaged in evaluations, and may even be be the targets of gender bias. Below, we outline additional proposed research that will allow us to more decisively answer the question of whether, when, and why women might be (dis)advantaged in academic hiring. First, though, we will briefly discuss the second issue raised in the introduction to this paper: that the rating scale used by W & C may be particularly susceptible to shifting standards and social desirability. If this is the case, there is an additional reason to question whether W & C's findings truly indicate the existence of a pro-woman bias, or whether they may simply have been an artifact of the specific dependent variable that they designed.

Shifting standards

Past research has shown that subjective rating scales (i.e., those that rely on abstract positive/negative language rather than concrete objective standards) can lead to misleading conclusions when used for cross-group comparisons (e.g., when comparing male and female candidates; Biernat & Kobrynowicz, 1997; Biernat & Vescio, 2002). This is because people tend to evaluate individual members of a social group *relative to the other members of their group*, rather than to a constant objective benchmark (Biernat, Collins, Katzarska-Miller, & Thompson, 2009). That is, when evaluating the qualities or characteristics of a member of a group (particularly a minority group), people do not compare that individual's characteristics to the full

range of possible human qualities, but rather to their stereotypes regarding the prototypical member of that individual's social group. For example, as illustrated by Biernat and Vescio (2002), a woman who is 5'10 might be described as "tall," while a man who is 5'10 might be described as "average." In effect, when people are asked to judge a woman in a situation such as this, they do not ask the question of "Is this person tall?", but rather "Is this person tall *for a woman*?" (*ibid*.)

In this way, even when people's objective characteristics are identical, subjective ratings of members of different social groups can still diverge substantially. Further, stereotypes regarding the competence of different groups can create a similar divergence in subjective evaluations. For example, a woman who is promoted to middle management in a company might be described as "very successful," whereas a man who reaches the same level might be described as "somewhat successful." This can lead to the counterintuitive effect that members of groups that are stereotyped as less competent can themselves actually be rated as *more competent* on subjective measures (Biernat & Vescio, 2002). For example, because women are stereotypically portrayed as being less competent in math, a female student who earns an A- in her calculus class might be described as "good at math," whereas a male student with the same grade might be described as "average." Similarly – and more relevant to the current research – because women are stereotyped to be less competent in science, a woman candidate with five peer-reviewed publications might be described as an "excellent candidate," while a man with the same number of publications might only be described as "good."

This effect – referred to as "shifting standards" (Biernat & Manis, 1994) – tends to emerge only when the outcome measure is subjective and/or abstract, such as with "non-zero-sum" measures like verbal praise (Biernat & Kobrynowicz, 1997; Biernat & Vescio, 2002). On

these measures, women are often evaluated more positively, especially when compared to a member of a group that is stereotypically high in competence (e.g., men). However, gender bias against women is still likely to manifest on zero-sum measures, such as allocations of limited resources (e.g., salary or jobs). In this way, women can, ironically, suffer real disadvantage in the objective resources they are awarded (e.g., not being hired, lower salary, etc.), while still receiving greater subjective praise (Biernat & Kobrynowicz, 1997; Biernat & Vescio, 2002).

Based on the above research, we predicted that although the female candidate in our pilot studies (under some conditions) was rated somewhat more positively on the 10-point subjective measure, the same pro-female bias would not be evident in allocations of limited resources, and women might even be disadvantaged in the objective resources they received. To test this question, we included a measure of objective resource allocation in two of our previously described pilot studies, which asked participants to decide on the salary, start-up funding, teaching releases, and lab space that they would award to the candidate. Before reviewing these results, we briefly discuss the issue of social desirability bias, and how these same objective measures may help circumvent it.

Social desirability bias

A second issue with W & C's candidate-rating paradigm is the possibility that social desirability motivation may have skewed responses, artificially inflating participants' ratings of the woman candidate. Social desirability bias refers to the well-documented effect that research participants will often alter their survey responses in order to present themselves as morally upstanding individuals (Edwards, 1970; Furnham, 1986). This response bias is particularly problematic in research examining attitudes towards historically disadvantaged groups like women and racial minorities, where people are especially averse to appearing prejudiced

(Monteith, Mark, & Ashburn-Nardo, 2010). Research in this domain has revealed that social desirability bias can not only artificially attenuate evidence of prejudice against disadvantaged groups, it can even create the appearance that these groups are viewed *more favorably* than high-status groups (Norton, Sommers, Vandello, & Darley, 2006) – such as was the case with the pro-woman bias documented by W & C.

Although social desirability effects are always an issue when conducting research on bias and prejudice, W & C's studies featured a number of elements that might have made social desirability bias more likely to have affected their results. They provided no cover story to participants regarding the purpose of the candidate-rating task, and the situation described in the vignette was clearly fictional (unlike in other common paradigms for studying prejudice, such as audit studies, which measure real-world bias by assessing call-back rates for applications (ostensibly) from members of minority- vs. majority-groups; e.g., Bertrand & Mullainathan, 2004). Because this was a fictional task with no real consequences, participants may have provided a socially desirable response (Nederhof, 1985). Additionally, academic samples in the U.S. are predominately politically liberal (Abrams, 2016), and liberal individuals tend to be especially averse to appearing prejudiced against women and other minority groups (Winegard & Winegard, 2017), further increasing the possibility of socially desirable responding among this sample. Moreover, as academics themselves, W & C's participants may have been familiar with research on gender bias and thus possibly suspicious about the true purpose of the study – especially given that W & C, well-known gender bias researchers, recruited participants by writing personalized emails and solicited participants' responses directly via these emails, rather than allowing faculty to provide their responses anonymously. Abundant research and theory

from the social sciences suggests that these kinds of personally identifiable response formats are likely to elicit social desirability bias (Tourangeau & Yan, 2007).

For the above reasons, social desirability motivation could have shaped participants' responses in W & C's studies, perhaps explaining the apparent pro-female bias that they observed. Additionally, past research suggests that scales like the one used by W & C – which examine subjective, emotional reactions to members of outgroups – are especially likely to be influenced by concerns about appearing prejudiced. Specifically, research suggests that people are particularly averse to appearing biased when they are asked to make a personal, subjective, and valenced judgment of a member of a minority group (Dovidio & Gaertner, 2004) – for example, when they are asked to rate the positivity/negativity of a minority group member's abilities, or to rate their personal liking of that individual. In these situations, participants will often provide (artificially) positive evaluations of minority group members (Harber, 1998; Harber, Stafford, & Kennedy, 2010; Vanman et al., 1997). The 10-point rating scale designed by W & C requires people to make exactly this type of valenced judgment, asking them to rate the abilities and general quality of the candidate, as well as their personal positivity toward that individual.

For these reasons, in our pilot studies we sought to find a way of circumventing social desirability bias. Past research suggests that one way of doing so is to avoid abstract valenced statements and to instead use more objective measures, such as allocations of limited resources. Because resource allocation measures do not ask participants to rate their liking of a minority group individual or to evaluate how good/bad that person is, these measures should be less likely to make participants feel concerned about the possibility of appearing biased, and they should therefore be less likely to elicit social desirability bias in responding.

In sum, then, research suggests that an effective strategy for overcoming both shifting standards and social desirability bias is to use more objective dependent measures that require participants to allocate limited resources. Such measures better circumvent these response biases in order to determine whether there is truly a preference for/against a target individual, using outcomes that are likely to have real implications, rather than abstract verbal praise.

Design

Based on the above research, we included an exploratory objective resource allocation measure in our pilot studies. Specifically, in two of our pilot studies (Studies 2 and 4), participants were asked to assign the candidate a salary, start-up funding, teaching releases, and lab space. Because Mechanical Turk participants likely have little experience with the typical standards for these resources, the former two questions were asked using sliding scales with a fixed range of possible values (salary: $50,000-$150,000; start-up funding: $10,000-$200,000), and the latter two questions (teaching releases and lab space) were rated relative to "the average" amount (9-point scale ranging from "Much less than the average" to "Much more than the average.").[4]

Results

In both studies, we found a significant 3-way interaction between gender (male vs. female), materials condition (excellent vs. mixed) and response scale (subjective 10-point scale vs. objective resources measure; Pilot 2: $F(1,357) = 6.04$, $p = .014$; Pilot 4: $F(1,626) = 7.34$, $p = .007$). These results show that, consistent with previous research on questionnaire construction (Schwarz, 1999; Schwarz & Oyserman, 2001), the way that the question is asked strongly influences the answer: participants' relative ratings of the woman (vs. man) candidate differed as

[4] In Pilot 4, we also randomized the order of the objective and subjective scales to ensure that there were no order effects. There were none (all *p*s > .07)

a function of the specific response scale that they used. Examining the pattern of this interaction, it is clear that, as we predicted, there is no evidence of a general pro-female bias in allocations of objective resources – and, if anything, the woman candidate is generally disadvantaged in these allocation decisions.

In Pilot 2, we found a clear divergence between the subjective and objective measures: although the exceptionally qualified woman had been rated significantly higher on the 10-point subjective scale ($t(357) = 3.52$, $p < .001$), there was no such trend on the objective measures. In fact, the woman received slightly (though non-significantly) *fewer* objective resources in this condition ($t(357) = .70$, $p = .49$, Fig. 7). Overall, there was a significant main effect of gender on objective resource allocations, such that the woman candidate (in both the excellent and mixed materials conditions) received a significantly lower salary and start-up, fewer teaching releases, and less lab space ($t(359) = 2.24$, $p = .026$).

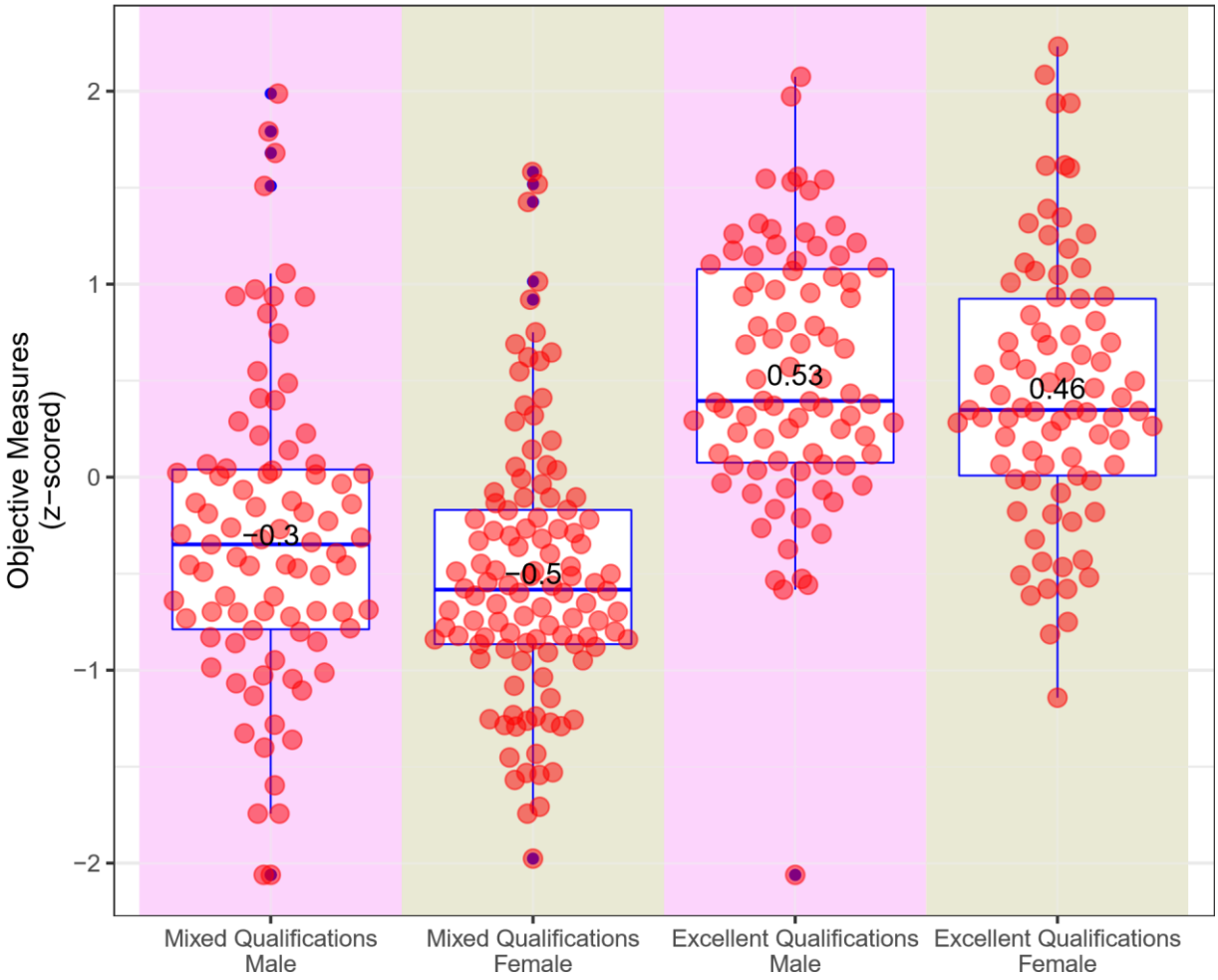

Fig 6 | Gender (man vs. woman) by qualification-level (mixed vs. excellent) interaction on objective ratings, Pilot 2.

In Pilot 4, we found a similar pattern of results. The trend towards higher ratings for the exceptional woman candidate on the subjective 10-point measure ($t(626) = 1.61$, $p = .11$) was completely eliminated on the objective measure, with no advantage for the woman whatsoever in objective resource allocation ($t(626) = .12$, $p = .91$, Fig. 6). Conversely, the relative disadvantage for the less exceptional woman candidate that we observed using the subjective rating scale ($t(626) = 2.45$, $p = .015$) was directionally consistent on the objective measure, with fewer resources being allocated to the woman candidate, although the difference on this measure was less pronounced and did not reach statistical significance ($t(626) = 1.5$, $p = .13$).

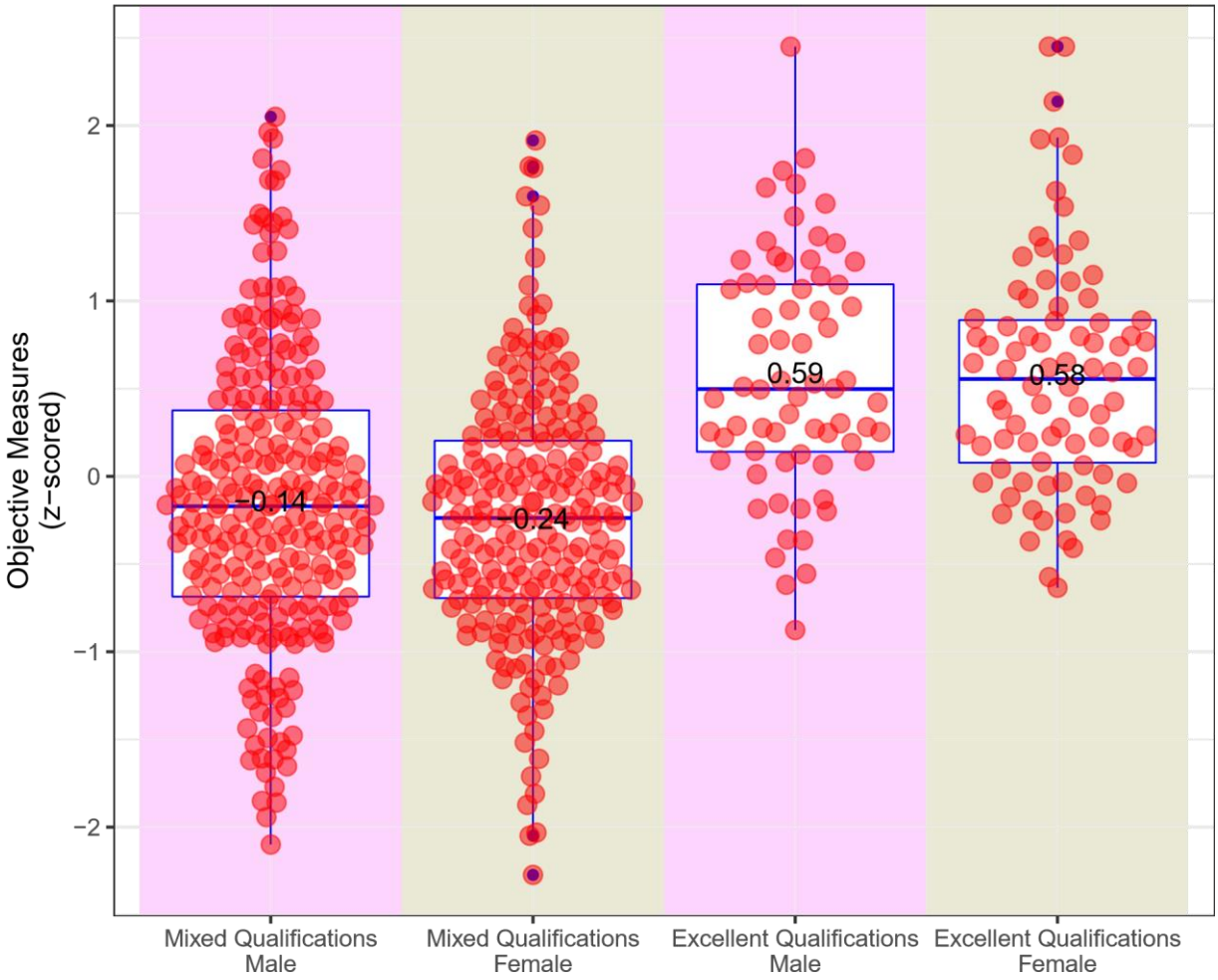

Fig 7 | Gender (man vs. woman) by qualification-level (mixed vs. excellent) interaction on objective ratings, Pilot 4.

Discussion

These results provide initial evidence that shifting standards and/or social desirability bias may have played a role in artificially creating the pro-female bias observed by W & C. Specifically, our findings suggest that while a pro-female advantage may emerge on more subjective measures of verbal praise – at least for exceptionally qualified woman candidates – when objective resource allocation measures are used, the woman is not advantaged, and may even suffer gender discrimination. In other words, participants may *report* liking women candidates more, but they nonetheless choose to pay them less.

This divergence between subjective and objective measures is notable given that in the real world, objective resources are likely to be more consequential for career success than is verbal praise – particularly when more positive verbal evaluations are paired with *fewer* objective resources. Although it may be beneficial for a woman candidate to be described as "outstanding" (the scale point corresponding to the average rating of the excellently qualified woman), if evaluators nonetheless choose not to award her the job, this verbal praise is unlikely to have any real import. Further, allocations of objective resources like start-up funding and lab space are likely to have important implications for later career success, such as research productivity, further grant funding, and tenure decisions (Martell, Lane, & Emrich, 1996; Merton, 1968). Therefore, even in the instances when the woman candidate *is* offered the job, if she is nonetheless disadvantaged in the objective resources that she receives, then she likely will not have the same level of career success as the (identically qualified but better compensated) male academic. Such disparities could potentially contribute to women's lower levels of career success in many academic domains (Shen, 2013; Weisshaar, 2017; West, 2013), and their greater propensity to leave certain domains of academia at later stages of their careers (Ceci & Williams 2010).

In our proposed research, outlined below, we again ask participants to complete these objective resource allocation measures in order to better assess whether women are truly advantaged in academia, or whether this apparent pro-female advantage may be explained by response biases elicited by the abstract and subjective dependent measure used in these studies. In line with the results of these pilot studies, we predict that even in instances that women are rated more positively on subjective measures of verbal praise, they may still be systematically disadvantaged in the resources that they are awarded.

Discussion and Registered Report Proposal

In sum, we have identified two core issues that appear likely to have contributed to the disparity between W & C's findings and the findings of past research on gender bias, potentially explaining why W & C observed an apparent pro-woman bias in their studies. First, the exceptional nature of the candidates' qualifications may have provided the one (possibly relatively rare) set of conditions under which gender discrimination is attenuated or eliminated. Second, the abstract and subjective rating scale that W & C designed may have allowed shifting standards and social desirability bias to shape participants' responses, artificially creating the appearance of a pro-woman bias.

Interestingly – and unexpectedly – in our pilot studies we also generally failed to replicate W & C's effects, observing little pro-female bias in evaluations, even for exceptionally qualified candidates. This failure to replicate raises the question of whether these divergent findings are due to differences in the participant samples (Mechanical Turk vs. tenure-track faculty) or other design factors (e.g., our studies using anonymous responding rather than personalized emails), or whether W & C's initial effect size may have been inflated, in keeping with the widespread effect-size inflation that has been well documented elsewhere in the field of experimental psychology (e.g., Klein et al., 2018).

Another interesting finding of these pilot studies is that the evidence of gender bias *against* women was also not as pronounced as that which has been observed in past research. That is, although we did observe gender bias against women in several of our mixed qualification conditions, the effect size of this anti-woman bias was also of a smaller size than that which has

been documented in past research (for a recent meta-analysis, see Koch et al., 2015). This raises the question of whether this discrepancy, too, may be due to publication bias in the literature, or whether bias against women in certain domains may be decreasing, in line with the changes in gender stereotypes that have been observed over the past several decades (Koenig et al., 2011; Lewis & Michalak, 2018; Sczesny et al., 2004).

In the research proposed below, we seek to provide more definitive answers to these questions. This work will constitute a replication and extension of Williams and Ceci (2015A), building on the results of the pilot studies discussed above. In this research, we will survey a large sample of tenure-track academics and ask them to evaluate a man or woman candidate for an assistant professor position. We will test whether qualification strength moderates ratings of the woman (vs. man) candidate by randomly assigning participants to view either an exceptionally qualified candidate (original W & C materials) or a more ambiguously qualified candidate. To test for possible differences between subjective and objective dependent measures, we will ask participants both to complete W & C's 10-point rating scale, as well as our pilot-tested resource allocation measure.

Registered Report Proposal

Methods

Power Analysis/Sample Size

To determine the necessary sample size for this study, we first meta-analyzed the results of our pilot studies in order to determine the average effect size of the gender (man vs. woman)

X qualifications (mixed vs. excellent) interaction. The estimated effect size of this interaction effect size was Cohen's d = .36 / f = .18 (se = .10, z = 3.73, p = .0002, 95% CI [.17,.54]). However, given the additional uncertainty of attempting to generalize from a Mechanical Turk sample to a sample of academics, we will power to an effect half this size: Cohen's d = .18 / f = .09. We conducted a power analysis based on 80% power to detect this estimated effect size, which resulted in a recommended sample size of 971 participants. This will be our target sample size for this study.

Sample Selection

We will adhere to the sample selection procedure employed by W & C, with the exception that we will select a larger number of universities in order to reflect our higher target sample size.

First, we will randomly select a sample of colleges and universities based on the Carnegie Classification system of institutions of higher education (http://carnegieclassifications.iu.edu/index.php). We will randomly select 400 institutions in total. 200 of these will be Ph.D-granting universities randomly selected from the "Doctoral Universities" classification category (collapsing across the three subdivisions of research activity: "moderate," "higher," and "highest"). The remaining 200 institutions will be randomly selected from 1) the "Master's Colleges and Universities" classification (collapsing across the "small," "medium," and "larger" program size subdivisions) and 2) the "Baccalaureate Colleges" classification (collapsing across the "diverse fields" and "arts and sciences focus" subdivisions). (W & C chose to oversample Ph.D.-granting institutions because of their higher prestige. We adhere to their sampling strategy for consistency).

As in W & C's studies, one additional criterion will be used to determine which institutions are selected: to be included in the final sample, the institution must have academic programs in at least 3 of the following 4 disciplines: engineering, economics, biology, and psychology. If a selected institution does not meet this requirement, it will be removed from the list and another institution from the same classification level will be randomly chosen to replace it. (W & C chose to limit their sample to the above four disciplines, which include two math intensive fields in which women are substantially underrepresented, engineering and economics, and two non-math-intensive fields in which women are well represented, biology and psychology; Ceci et al., 2014; Cheryan et al., 2017. We adhere to their sampling strategy for consistency.)

Next, the names of all tenure-track faculty from each of these institutions and disciplines will be collected from each institution's website. "Tenure-track faculty" will be defined as assistant professors, associate professors, professors, and department chair-holders, whose titles are not qualified by terms indicating that they have a reduced or temporary position (e.g., "visiting professors," "emeritus professors," "professors by courtesy"). Following the logic of W & C, adjunct faculty, lecturers, and other non-tenure-track positions are not included in the study because they typically play a less important role in hiring decisions.

Faculty names will then be added to randomly ordered lists, separated by university, discipline, and gender. For the initial round of recruitment, the first faculty member from each list (i.e., one male and one female faculty member from each department at each university) will be emailed a link to the survey, along with a short note requesting their participation in the study. Survey links will contain a unique code indicating the university, department, and sex of the target faculty member, which will allow us to track (using anonymized codes) which faculty

have responded to the survey. After 10 days, faculty who have not participated in the survey will be replaced by the next faculty member on the list, and a new batch of survey links will be emailed. Following this schedule, a new batch of survey links will be emailed every ten days. We will continue this process until we have reached our target sample size of N = 971. Ten days after sending the last batch of study links, we will end data collection and begin analysis. Any responses received after this date will be excluded from analyses.

We will follow the above sampling procedures in order to ensure that our sample is as consistent as possible with the sample collected by W & C. However, we also note that these procedures include some potential sources of variability that will shape the final number of participants that are included in our sample. The most important of these regards our stopping rule: although we have specified the exact point at which we will stop distributing survey links, as well as the exact point at which we will end data collection (at 11:59pm EST of the tenth day following the dispersal of the last batch of links), additional faculty members are likely to respond after the target number of participants is reached, which would increase our final sample size. However, despite this uncertainty regarding the total number of participants that will constitute our final sample, the sampling procedures themselves allow for no variability, and will be followed precisely. Therefore, although our final sample size may differ somewhat from our a priori target, we (i.e., the experimenters) will have no control over the size or composition of this sample.

Procedure

Faculty members who are selected for participation will be sent an email containing a link to the study, which will be conducted via the online survey platform Qualtrics

(www.qualtrics.com). Those who choose to follow the link will first be asked to provide informed consent. Individuals who consent to participate will then begin the study. As in W & C's original study, participants will be asked to evaluate a male or female candidate for a tenure-track assistant professor position. The candidate's gender will be indicated by gender pronoun alone (e.g., he vs. she; his vs. her). As in our pilot studies, participants will also be randomly assigned to one of two materials conditions. Those assigned to the "Excellent Qualifications Condition" will view the original W & C materials.[5] Those in the "Mixed Qualifications Condition" will evaluate a candidate who is described as excelling on two of the four qualifications dimensions, but as being less exceptional on the other two dimensions. We will use two different sets of mixed materials in order to counterbalance the specific dimensions on which the candidate is said to excel (however, we do not expect the specific set of mixed materials to moderate effects, and therefore do not include this condition assignment in our analyses). In the first set of mixed materials, the candidate will be described as an exceptional researcher and as having exceptional social skills, but as being a below average teacher and as having been rated poorly on the job talk/interview (mixed materials from Pilot 2). In the other set of mixed materials, the candidate will be described as an exceptional teacher and as having received an exceptional job talk/interview score, but as being an "above average" researcher and having mediocre social skills (the low ambiguity materials from Pilot 4).

Original W & C Materials (female candidate version)

> *"Dr. X impressed the entire search committee as a great potential hire. Based on her vita, letters of recommendation, and their own reading of her work, the search committee rated X's research record as "extremely strong." Letter-writers especially noted that X is highly creative and original in her approach to scholarship, with comments like "X is poised to break new ground with her unique and imaginative applications of her advisor's theory, and is sure to change how people think about her research area." They*

[5] The Excellent Materials Condition of this study constitutes a direct replication of W & C's original Experiment 5: all materials and procedure are identical up through the completion of the 10-point subjective scale.

also described X's impressive teaching abilities, mentioning that she was "widely considered an effective and supportive mentor by the junior graduate students and undergraduates she worked with." She also won a teaching award in graduate school. X's faculty job talk/interview score was 9.5/10. At dinner with the committee, she reached out to everyone, showing herself to be very likeable, kind, and socially skilled. During our private meeting, X was enthusiastic about our department, and there did not appear to be any obstacles if we decided to offer her the job. She mentioned that she is single with no partner/family issues. X said our department has all the resources needed for her research."

Mixed Materials 1 (excellent research and social skills, female candidate version)

"Dr. X seemed to the entire search committee to be an acceptable potential hire. Based on his vita, letters of recommendation, and their own reading of his work, the search committee rated X's research record as "extremely strong." Letter-writers especially noted that X is highly creative and original in his approach to scholarship, with comments like "X is poised to break new ground with his unique and imaginative applications of his advisor's theory, and is sure to change how people think about his research area." They noted, however, that X's teaching abilities were less impressive, mentioning that he was "widely considered a difficult and distant mentor by the junior graduate students and undergraduates he worked with." X's faculty job talk/interview was also not particularly well rated, scoring only 6/10. However, at dinner with the committee, he reached out to everyone, showing himself to be very likeable, kind, and socially skilled. During our private meeting, X was enthusiastic about our department, and there did not appear to be any obstacles if we decided to offer him the job. He mentioned that he is single with no partner/family issues. X said our department has all the resources needed for his research."

Mixed Materials 2 (excellent teaching and job talk/interview, female candidate version)

"Dr. X appeared to the entire search committee to be an acceptable potential hire. Based on her vita, letters of recommendation, and their own reading of her work, the search committee rated X's research record as "above average." Letter-writers noted that X is sufficiently creative and original in her approach to scholarship, with comments like "X is likely to expand the literature with her consistent and incremental applications of her advisor's theory, and is likely to add nuance to how people think about her research area." They noted X's impressive teaching abilities, mentioning that she was "widely considered an effective and supportive mentor by the junior graduate students and undergraduates she worked with." She also won a teaching award in graduate school. X's faculty job talk/interview score was 9.5/10. However, at dinner with the committee, she failed to impress, not coming off as particularly likable or kind, and seeming to have mediocre social skills. During our private meeting, X was enthusiastic about our department, and there did not appear to be any obstacles if we decided to offer her the job. She mentioned that she is single with no partner/family issues. X said our department has all the resources needed for her research."

After reviewing the materials, participants will be asked to rate the candidate on the 10-point rating scale designed by W & C. They will then answer our four objective resource allocation questions from Pilots 2 and 4:

1) *The typical starting salary for a new professor is between $50,000 and $150,000 What salary would you recommend that Dr. X receives?*
2) *The typical "start-up fund" (that is, funding for research, travel, etc.) for a new professor is between $10,000 and $200,000. How large of a start-up fund would you recommend that Dr. X receives?*
3) *A "teaching release" gives a professor one semester in which they do not need to teach, which allows them to be more productive with their research. How many teaching releases do you think Dr. X should receive?*
4) *Laboratory space is important for a researcher's productivity. How much laboratory space do you think Dr. X should receive?*

Following these dependent measures, participants will be asked to guess the purpose of the study, on an open-ended question asking: "In a few words, what do you think the true purpose of this study was?" Participants will then be asked to provide demographic information (age, gender, and political orientation). They will then answer the attention-check question from our pilot studies, in which they will be asked to report the gender of the candidate in the vignette that they read. Finally, participants will answer a question assessing whether they previously participated in W & C's study: "There is a small chance that you may have previously completed a similar study in the past. This study would have been nearly identical to the one that you just completed, and you would have received it via email. Did you participate in this study?"

Statistical Analyses

Exclusion Criteria

Participants will be excluded if 1) they do not accurately report the gender of the candidate in the vignette (i.e., indicating the incorrect gender or selecting "I don't know") and/or 2) they indicate that they participated in W & C's original study.

Hypothesis Testing

Hypothesis 1A: There will be no pro-woman bias in subjective evaluations.

To test this hypothesis, we will use an inferiority test, following the recommendations of Lakens, Scheel and Isager (2018). An inferiority test is a form of equivalence test that tests the probability that – given a certain observed effect size – there exists a true effect in the population that is at least as large as the *smallest effect size of interest* (SESOI; Lakens et al., 2018). (In other words, an inferiority test tests whether the null hypothesis that *there is an effect at least as large as SESOI* can be rejected.) For this test, we will define the smallest effect size of interest as a Cohen's *d* of .1, following past research that has argued that effects smaller than this are unlikely to have meaningful real-world impacts, either in general (Maxwell, Lau, & Howard, 2015) or in the domain of gender bias specifically (Hyde, Lindberg, Linn, Ellis, & Williams, 2008). In line with these arguments, we expect that this is the smallest effect that could realistically be expected to lead to a meaningful pro-woman bias in hiring or other outcomes.

We will first conduct a two-way ANOVA in which candidate gender (man vs. woman) and materials condition (excellent vs. mixed) will be entered as fixed predictor variables, and subjective 10-point rating will be entered as the dependent variable. Based on the results of this model, we will calculate the 90% confidence interval (one-tailed test; Lakens et al., 2018) around

the observed Cohen's *d* for gender condition. We will consider this hypothesis to be supported if this confidence interval does not include .1. In this case, we will conclude that the effect of gender on subjective evaluations either does not exist, or is too small to be of any practical importance (Lakens et al., 2018).

Hypothesis 1B: There will be no pro-woman bias in allocation of objective resources.

To test this hypothesis, we will first z-score and average across the four individual objective resource questions. We will then use an inferiority test to determine whether the effect of gender on objective resource allocation is at least as large as a Cohen's *d* of .1 (the smallest effect size of interest).

We will conduct a two-way ANOVA in which candidate gender (man vs. woman) and materials condition (excellent vs. mixed) will be entered as fixed predictor variables, and objective resources will be entered as the dependent variable. Based on the results of this model, we will calculate the 90% confidence interval around the observed Cohen's *d* for gender condition. We will consider this hypothesis to be supported if this confidence interval does not include .1. In this case, we will conclude that the effect of gender on objective evaluations either does not exist, or is too small to be of any practical importance.

Hypothesis 2: Subjective evaluations of women (versus men) will be more negative when the candidates' qualifications are more ambiguous.

To test this hypothesis, we will conduct a two-way ANOVA in which candidate gender (man vs. woman) and materials condition (excellent vs. mixed) will be entered as fixed predictor variables, and subjective 10-point rating will be entered as the dependent variable. We will consider this hypothesis to have been supported if the interaction term of gender and materials condition is significant at p <= .05, and if it is in the predicted direction, such that the woman (vs. man) is rated more negatively in the mixed (vs. excellent) materials condition.

Hypothesis 3: Women (versus men) will be rated more negatively on the objective (vs. subjective) measures.

To test this hypothesis, we will conduct a two-way mixed ANOVA in which candidate gender (man vs. woman) will be entered as a between-subjects fixed predictor variable, and measure type (subjective vs. objective) will be entered as a within-subjects predictor. We will consider our hypothesis to be supported if the interaction term of gender and measure type is significant at p <= .05, and if the pattern is in the predicted direction, such that ratings of the woman (versus man) are more negative on the objective (vs. subjective) measure.

Exploratory Supplemental Analyses

We will also examine the simple effect of gender within each materials condition, for both the subjective and objective measures. Specifically, we will examine (a) whether we replicate W & C's pro-woman bias in the excellent materials condition (i.e., whether there is a significant main effect of gender within this condition, such that women are generally favored).

We will also examine (b) whether there is a significant anti-woman bias in the mixed materials condition (i.e., a significant main effect of gender in this condition, such that women are generally disfavored).

We will also meta-analyze the results of all of our studies (the four pilot studies and the faculty sample) in order to determine whether sample type (Mechanical Turk vs. Academic) moderates these effects. We will first calculate the standardized mean difference (Cohen's *d*) between ratings of the woman (vs. man) candidate for both the mixed qualifications condition and the excellent-qualifications condition for each study. We will then fit a random effects meta-analysis model separately within each level of condition (mixed and excellent), adding sample type (Mechanical Turk vs. Academic) to the model as a predictor. We will then examine whether sample type significantly moderates the size of the effect (i.e., the difference in ratings between the woman and man candidate) in either qualifications condition.

Finally, we will also examine the number of participants who expressed suspicion that the purpose of the study was related to gender, sex, sexism, or gender bias.

Project Timeline

We anticipate completing this study 6 months after stage 1 acceptance: one month for amendment of the IRB with any changes that are recommended by reviewers, one month to collect the data, another month to complete all analyses, one month to ensure all data and materials are reproducible prior to uploading to the Open Science Framework, and two months to finalize manuscript revisions for Stage 2 submission.

References

Abrams, S. J. (2016). Professors moved left since 1990s, rest of country did not. *Heterodox Academy*. Retrieved from: *https://heterodoxacademy.org/professors-moved-left-but-country-did-not/*

Altmetric (2019). Overview of attention for "National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track" Retrieved on 1/14/2019 from https://pnas.altmetric.com/details/3903930/misc

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), 991-1013.

Biernat, M., Collins, E. C., Katzarska-Miller, I., & Thompson, E. R. (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin*, *35*(1), 16-28.

Biernat, M., & Fuegen, K. (2001). Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues*, *57*(4), 707-724.

Biernat, M., & Kobrynowicz, D. (1997). Gender-and race-based standards of competence: lower minimum standards but higher ability standards for devalued groups. *Journal of personality and social psychology*, *72*(3), 544.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of personality and social psychology*, *66*(1), 5.

Biernat, M., & Vescio, T. K. (2002). She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. *Personality and Social Psychology Bulletin*, *28*(1), 66-77.

Blau, F. D. & Kahn, L. M. (2016) : The Gender Wage Gap: Extent, Trends, and Explanations, IZA Discussion Papers, No. 9656, Institute for the Study of Labor (IZA), Bonn

Boynton, J. R., Georgiou, K. Reid, M., Govus, A. (2018). Gender bias in publishing. *The Lancet*. 392: 1514-1515.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on psychological science*, *6*(1), 3-5.

Burgess, D., & Borgida, E. (1999). Who women are, who women should be: Descriptive and prescriptive gender stereotyping in sex discrimination. *Psychology, public policy, and law*, *5*(3), 665.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection. *SAGE Open*, *7*(2), 2158244017712774.

1067 Ceci S. J., Ginther D. K., Kahn S., Williams W. M. (2014) Women in academic science: A
1068 changing landscape. Psychol Sci Publ Interest 15(3):75–141.

1069 Ceci, S. J., & Williams, W. M. (2010). Sex differences in math-intensive fields. *Current*
1070 *Directions in Psychological Science*, *19*(5), 275-279.

1071 Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's
1072 underrepresentation in science. *Proceedings of the National Academy of Sciences*,
1073 201014871.

1074 Cejka, M. A., & Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to
1075 the sex segregation of employment. *Personality and social psychology bulletin*, *25*(4), 413-
1076 423.

1077 Charlesworth, T. E. S., & Banaji, M. R. (in press). Patterns of Implicit and Explicit Attitudes: I.
1078 Long-Term Change and Stability From 2007-2016. *Psychological Science*

1079 Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields
1080 more gender balanced than others?. *Psychological Bulletin*, *143*(1), 1.

1081 Dasgupta, N., Scircle, M. M., & Hunsinger, M. (2015). Female peers in small work groups
1082 enhance women's motivation, verbal participation, and career aspirations in
1083 engineering. *Proceedings of the National Academy of Sciences*, 201422822.

1084 Dennehy, T. C., & Dasgupta, N. (2017). Female peer mentors early in college increase women's
1085 positive academic experiences and retention in engineering. *Proceedings of the National*
1086 *Academy of Sciences*, *114*(23), 5964-5969.

1087 Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in experimental social*
1088 *psychology*, *36*, 4-56.

1089 Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female
1090 leaders. *Psychological review*, *109*(3), 573.

1091 Edwards AL. 1970. The Measurement of Personality Traits by Scales and Inventories. New
1092 York: Holt, Rinehart & Winston

1093 Ferriman, K., Lubinski, D., & Benbow, C. P. (2009). Work preferences, life values, and personal
1094 views of top math/science graduate students and the profoundly gifted: Developmental
1095 changes and gender differences during emerging adulthood and parenthood. *Journal of*
1096 *personality and social psychology*, *97*(3), 517.

1097 Fleming, M. A., Petty, R. E., & White, P. H. (2005). Stigmatized targets and evaluation:
1098 Prejudice as a determinant of attribute scrutiny and polarization. *Personality and Social*
1099 *Psychology Bulletin*, *31*(4), 496-507.

1100 Francis, M. R. (2015). "A Surprisingly Welcome Atmosphere": A vaunted new study says
1101 women have it easy in STEM fields. Don't believe it. *Slate*. Retrieved from
1102 https://slate.com/human-interest/2015/04/no-sexist-hiring-in-stem-fields-a-vaunted-new-
1103 study-makes-that-claim-unconvincingly.html

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and individual differences*, *7*(3), 385-400.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, *101*(1), 109.

Hakim, C. (2006). Women, careers, and work-life preferences. *British Journal of Guidance & Counselling*, *34*(3), 279-294.

Harber, K. D. (1998). Feedback to minorities: Evidence of a positive bias. *Journal of personality and social psychology*, *74*(3), 622.

Harber, K. D., Stafford, R., & Kennedy, K. A. (2010). The positive feedback bias as a response to self-image threat. *British Journal of Social Psychology*, *49*(1), 207-218.

Haynes, C., & Sweedler, J. (2015). Are We There Yet? Biases in Hiring Women Faculty Candidates.

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. *Research in organizational behavior*.

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

Hoobler, J. M., Wayne, S. J., & Lemmon, G. (2009). Bosses' perceptions of family-work conflict and women's promotability: Glass ceiling effects. *Academy of management journal*, *52*(5), 939-957.

Hottenrott, H., & Lawson, C. (2017). Fishing for complementarities: Research grants and research productivity. *International Journal of Industrial Organization*, *51*, 1-38.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494–495. doi:10.1126/science.1160364

Institute for Women's Policy Research (2017). The Gender Wage Gap 2017: Earnings differences by gender, race, and ethnicity. Retrieved from: https://iwpr.org/wp-content/uploads/2018/09/C473.pdf

Johnson, S. K., Murphy, S. E., Zewdie, S., & Reichard, R. J. (2008). The strong, sensitive type: Effects of gender stereotypes and leadership prototypes on the evaluation of male and female leaders. *Organizational Behavior and Human Decision Processes*, *106*(1), 39-60.

Jussim, L. (2017). Gender Bias in Science? Double standards and cherry-picking in claims about gender bias. Psychology Today. Retrieved from: https://www.psychologytoday.com/us/blog/rabble-rouser/201707/gender-bias-in-science

Kite, M. E., Deaux, K., & Haines, E. L. (2008). Gender stereotypes. *Psychology of women: A handbook of issues and theories*, *2*, 205-236.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128.

Koenig, A. M., Eagly, A. H., Mitchell, A. A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological bulletin*, *137*(4), 616.

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*(5), 603-625.

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*(3), 278-292.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1 (2), 259–269.

Leahey, E. (2006). Gender differences in productivity: Research specialization as a missing link. *Gender & Society*, *20*(6), 754-780.

LeBel, E. P., McCarthy, R., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *OpenLeBel, Etienne P et al."A Unified Framework to Quantify the Credibility of Scientific Findings". PsyArXiv*, *13*.

Lewis, N. A., Jr., & Michalak, N. M. (2018). *Has Stereotype Threat Dissipated Over Time? A Cross-Temporal Meta-Analysis.* Manuscript Under Review.

Lyness, K. S., & Heilman, M. E. (2006). When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology*, *91*(4), 777.

Martell, Richard F., David M. Lane, and Cynthia Emrich. 1996. "Male-female Differences: A Computer Simulation." American Psychologist 51:157-8.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. doi:10.1037/a0039400

Merton, Robert K. 1968. "The Matthew Effect in Science." Science 159: 56-63

Monteith, M. J., Mark, A. Y., & Ashburn-Nardo, L. (2010). The self-regulation of prejudice: Toward understanding its lived character. *Group Processes & Intergroup Relations*, *13*(2), 183-200.

Mulligan, T. (2017). Uncertainty in Hiring Does Not Justify Affirmative Action. *Philosophia*, *45*(3), 1299-1311.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European journal of social psychology*, *15*(3), 263-280.

Norton, M. I., Sommers, S. R., Vandello, J. A., & Darley, J. M. (2006). Mixed motives and racial bias: The impact of legitimate and illegitimate criteria on decision making. *Psychology, Public Policy, and Law*, *12*(1), 36.

Okimoto, T. G. & Brescoll, V. L. (2010) "The price of power: Power seeking and backlash against female politicians." *Personality and Social Psychology Bulletin* 36, no. 7: 923-936.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of personality and social psychology*, *75*(3), 811.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex roles*, *57*(7-8), 509-514.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American psychologist*, *54*(2), 93.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, *22*(2), 127-160.

Sczesny, S., Bosak, J., Neff, D., & Schyns, B. (2004). Gender stereotypes and the attribution of leadership traits: A cross-cultural comparison. *Sex roles*, *51*(11-12), 631-645.

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

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128.

Stewart-Williams, S. & Halsey, L. G. (2018, August 26). Men, Women, and Science: Why the Differences and What Should Be Done?. https://doi.org/10.31234/osf.io/ms524

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*(5), 859.

United Nations Women. (2018). Facts and Figures: Economic Empowerment. Retrieved from http://www.unwomen.org/en/what-we-do/economic-empowerment/facts-and-figures

Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. *Journal of Personality and Social Psychology*, *73*(5), 941.

Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in Academia. *Social Forces*, *96*(2), 529-560.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PloS one*, *8*(7), e66212.

Williams, W. M., & Ceci, S. J. (2015A). National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, 201418878.

Williams, W. M., & Ceci, S. J. (2015B). Additional resources for: "National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track." *Proceedings of the National Academy of Sciences*, 201418878.

Williams, W. M., & Ceci, S. J. (2015C). Supplementary Information for "National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track." *Proceedings of the National Academy of Sciences*, 201418878.

Williams, J. C., & Smith, J. L. (2015). The myth that academic science isn't biased against women. *The Chronicle of Higher Education*.

Winegard, B., & Winegard, B. (2017). Paranoid egalitarian meliorism: An account of political bias in the social sciences. In J. T. Crawford & L. Jussim. *Frontiers of Social Psychology Series: The Politics of Social Psychology* (pp. 193-209). New York: Routledge.

Wood, W., & Eagly, A. H. (2010). Gender. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (pp. 629-667). Hoboken, NJ, US: John Wiley & Sons Inc.