Research in Higher Education
https://doi.org/10.1007/s11162-024-09779-6

Making the Band: Constructing Competitiveness in Faculty Hiring Decisions

Damani K. White-Lewis[1] · KerryAnn O'Meara[2] · Jennifer Wessel[3] · Julia Anderson[4] · Dawn Culpepper[5] · Lindsey Templeton[6]

Received: 8 August 2022 / Accepted: 6 February 2024

Keywords Faculty hiring · Equity · Experimental survey · Mixed methods

Introduction

Over the last thirty years, social scientists have consistently found that bias pervades and shapes faculty hiring decisions. Although multiple methods have been used, some of the most consistent and compelling evidence has come from experimental studies (e.g., Beattie et al., 2013; Eaton et al., 2020; Moss-Racusin et al., 2012; Sheltzer & Smith, 2014; Steinpreis et al., 1999). For example, Eaton and colleagues (2020) found that when faculty members reviewed curricula vitae (CVs), they viewed men, white, and Asian candidates as more competent and hireable compared to women, Black, and Latinx candidates with the same qualifications. Yet a few recent studies show evidence of no bias and/or slight bias towards women candidates (Bernstein et al., 2022; Williams & Ceci, 2015). These conflicting findings are interesting, and we return to them later in the manuscript; overall, a significant number of colleges and universities have responded by developing inclusive hiring guidelines and implementing inclusive hiring practices, such as rubrics, equity charges, and structured interviews (Culpepper et al., 2023; Liera, 2020; O'Meara et al., 2020).

✉ Damani K. White-Lewis
dkwlewis@upenn.edu

[1] Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA

[2] Teachers College, Columbia University, New York, NY, USA

[3] College of Behavioral and Social Sciences, Department of Psychology, University of Maryland, College Park, College Park, MD, USA

[4] College of Education, Department of Educational Administration, Michigan State University, East Lansing, MI, USA

[5] College of Education, Department of Counseling, Higher Education, and Special Education, University of Maryland, College Park, College Park, MD, USA

[6] Higher Education Resource Services, Denver, CO, USA

Published online: 04 March 2024

Interestingly, workshops to increase awareness of implicit bias cite the experimental studies we cite above, but there are key limitations to this research. The first is that in many of them (e.g., Eaton et al., 2020; Steinpreis et al., 1999), evaluators were sent a single CV to review. Though the number of applicants often differs by discipline, rank, institutional type, and geographic region, virtually all search committees review multiple CVs. Why might this difference between experimental studies and the naturalistic hiring setting be important? Some research shows that evaluation bias can be reduced when reviewers change how they approach candidate review, such as looking at files side-by-side, compared to scenarios wherein applicants are reviewed one at a time (Bohnet et al., 2016). Additionally, the time it takes to review multiple files may increase the prevalence of bias if raters feel rushed to complete review (Kahneman, 2011). The diversity of the candidate pool may also matter. For example, one study showed that when there are multiple candidates from historically marginalized groups in a candidate pool, the chances of hiring a minoritized candidate are enhanced beyond mere probability (Johnson et al., 2016). As such, empirically examining how bias manifests when evaluators have a pool of applicants, who vary in terms of race/gender and qualification, is needed.

Additionally, how evaluators indicate their preference for candidates may shape hiring decisions and this has not been studied in most experimental design studies. For example, in real searches, committees are typically asked by hiring officials to do one of two things at the end of a search: rate candidates and submit a few names of candidates who are qualified and "hireable" or rank a group of candidates and submit a ranked list (Fine & Handelsman, 2012). In real faculty searches, the process of ranking or identifying several candidates who are hireable (i.e., meet the hiring threshold) could interact with social biases and impact hiring outcomes. We assert each of these naturally occurring hiring contexts likely shape hiring outcomes. Further research is needed to examine the presence of social biases in hiring settings wherein evaluators examine a larger set of CVs, with different perceived identities and qualifications.

We address this gap using data from a self-designed and validated experimental survey of 315 mechanical engineering faculty members. To further understand the presence of either positive or negative biases by gender, race and the intersection of gender/race, we created CVs in two "bands" of candidates - those with average publication productivity and those with excellent publication productivity. We signaled candidate identity (gender/race), asked participants why they evaluated candidates as they did, and explored the factors most influential in their decision-making when reviewing ten CVs for a faculty position. We also collected the demographic information and expertise of participants to examine whether participant identities contribute toward decision outcomes. The research question that directed this work was: Were engineering faculty members' hiring-related decisions shaped by any of the following factors?

a. Candidate characteristics (gender identity, racial identity, and publication record)
b. Selection conditions (threshold list or ranked list)
c. Evaluator demographic factors (race, gender, and disciplinary expertise)

Our research makes several important contributions to the literature. First, it expands the research on implicit biases in faculty hiring using a within-subjects design, which replicates more closely the natural setting of hiring decisions (i.e., side by side comparisons of a

candidate pool). Second, we explore whether ranking candidates or identifying a group of candidates who meet a threshold matters to decision outcomes. The answer to this question has important implications for higher education institutions advising search committees on equitable practices. Third, our study was conducted during a time when there have been significant efforts toward racial justice and increased focus on diversifying the faculty. This allowed us to explore whether, all candidates being equal, and/or in cases with publication productivity being different, there was a preference for candidates of perceived identities. Finally, we considered whether aspects of evaluator identity and expertise impacts decision outcomes.

Conceptual Framework

This study was guided by theory and literature on cognitive and social bias from behavioral economics and decision-making science. Because much of the literature on faculty hiring is rooted in the conceptual understanding of cognitive and social biases, we first discuss our conceptual framework and then review the literature on how bias impacts decision-making.

Bias is defined as systematic and patterned ways of automatic thinking or behaving (Greenwald & Lai, 2020; Kahneman, 2011). Studies show that as humans, we often use short-cuts, or heuristics to make our everyday decision-making easier (Kahneman, 2011). While these short-cuts may be functional in some cases (e.g., helping us navigate social interactions; Norris & Epstein, 2011), they often cause us to make irrational and sub-optimal decisions (Greenwald & Lai, 2020; Milkman, 2021). Researchers have documented hundreds of different biases and heuristics that can influence our decisions in areas like health, financial wealth, environmental and civic engagement, and education (Kahneman, 2011; Milkman, 2021; Thaler & Sunstein, 2008). Some are cognitive such as confirmation bias that lead us to look and overvalue information that confirms our preexisting schemas, or social biases such as favoring a man for a lab position due to implicit associations between men and STEM.

Certain conditions exacerbate bias, and many of these conditions are present in faculty hiring (Moody, 2012; O'Meara et al., 2020). For instance, when decision-makers are rushed, tired, stressed, or lack complete information, bias is more likely to take over (Kahneman, 2011; Milkman, 2021; Thaler & Sunstein, 2008). In faculty hiring, search committees often evaluate hundreds of applications and make decisions based on a rather limited set of information (e.g., a CV and a cover letter), meaning that they may rely on their biases to make the decision-making process easier. Humans are also conservative in their decision-making: we are risk- and loss-averse, prefer the status quo, and place a high value on things we create ourselves (Kahneman, 2011; Milkman, 2021). Committees and hiring officials tend to view faculty hiring as a "high stakes" decision and therefore may go with candidates thought to be safer or known entities, which may also reproduce inequities (O'Meara et al., 2023). All said, there is substantial reason to believe that faculty hiring would be a context wherein multiple kinds of biases could manifest.

Recognizing the ubiquity of bias, researchers, policymakers, and other decision-makers have attempted to identify likely biases that emerge and constrain high quality decision-making, and then offer ways to change the context around those choices to make it less likely that bias negatively impacts decisions (Castleman & Page, 2014; Field, 2015; Thaler & Sunstein, 2008). The context surrounding decisions is known as "choice architecture"

(Thaler & Sunstein, 2008). Experimental and quasi-experimental studies show that small changes in context, such as altering how information is presented or framed, providing reminders about upcoming deadlines, or automatically opting someone into a policy, can reduce bias (Damgaard & Nielsen, 2018; Milkman, 2021; O'Meara et al., 2022; Thaler & Sunstein, 2008).

We wanted to test two important kinds of choice architecture. The first one is embedded in the very design of our study. Most of the previous experimental faculty hiring studies had participants rate only one CV. Although gender and race were signaled, each participant only reviewed a CV with one signaled identity. Yet social scientists working to improve equity in hiring have found that comparative evaluation, where candidates are reviewed in batches, can significantly reduce bias in faculty hiring (Bohnet, 2016). For example, Bohnet, van Geen & Bazerman (2016) found in one experiment that, "when evaluators looked at candidate profiles individually, men were more likely to be hired for the math task and women for the verbal task, including those who had performed below par. Our intervention, where evaluators were exposed to more than one candidate, was able to overcome stereotypical assessments. Comparative evaluation focused evaluators' attention on individual performance, instead of group stereotypes. When candidates were evaluated comparatively, not only did the gender gap vanish completely, but basically all evaluators now chose the top performer." (p. 127).

In this study, we wanted to understand how different contexts – perceived identity and qualifications of candidates; evaluator demographics and disciplinary expertise; and ranking versus threshold conditions – shaped hiring outcomes. We were also interested in whether bias against women and racially minoritized candidates would emerge as strong as it had before in prior studies if multiple CVs were analyzed. Our goal was to better understand how these hiring contexts could be altered such that bias is reduced and hiring decisions are more effective and inclusive.

Literature Review

As we sought to apply the vast literature on decision-making bias to faculty hiring, we were informed by several bodies of research. Prior research shows how implicit biases regarding race, gender, and their various intersections have shaped evaluations. There is also research explaining why ranking or rating candidates as above or below a hiring threshold might shape outcomes. Finally, we drew on research on the role of disciplinary expertise in decision-making.

Gender and Racial Biases in Hiring

Whether and how evaluators know a candidate's identity when making decisions is inconsistent, complicated, and varies across stages of the hiring process. On the one hand, equal employment laws and regulations prohibit evaluators from making hiring decisions[1] based on a candidate's race and gender. For this reason, search committees review candidate files without formal information on candidate identities. On the other hand, some candidates are known by evaluators through networks and professional relationships in their field. Some

[1] https://www.eeoc.gov/prohibited-employment-policiespractices.

candidates signal their identities in application materials. Evaluators may also assume identities based on candidate names widely associated with a particular gender and/or racial group. Indeed, studies show that when not known, faculty evaluators try to ascertain a candidate's race and gender and that their perceptions of a candidate's identity shape decision-making (Eaton et al., 2020; Liera, 2020; Rivera, 2017; White-Lewis, 2020). Race and gender play a role when candidates are considered equally qualified (White-Lewis, 2019) and when candidates excel in different areas (Liera, 2020). The effects of cognitive and social biases also play out in an intersectional way, advantaging, for example, white women and disadvantaging women of color (Eaton et al., 2020).

Moreover, evaluators' perceptions of what makes a candidate qualified may also be biased. For example, if an evaluator unconsciously associates publication productivity with a white and or male sounding name and then finds a white and male candidate with excellent publications, they may be exhibiting a confirmation bias wherein they were looking for something and found it (Kahneman, 2011). Alternatively, an evaluator may assume that a candidate of color and/or a white woman would not have these qualifications and therefore may not have recognized when candidates from these groups possess excellent publications. As such, we would expect that bias may look different for candidates that have excellent versus average qualifications, and that this would vary at the intersection of qualifications and identity factors.

We also know that the concepts of racism and sexism change over time in society and may operate differently than has been shown in past studies (Bonilla-Silva, 2006). For instance, although higher education institutions have moved away from anti-nepotism policies that tended to disadvantage women faculty members who were married to academic men (Shoben, 1997), women in dual-career academic couples still encounter significant gender bias (Culpepper, 2021). In the current climate, with the increased emphasis toward hiring a diverse faculty at many institutions, it is possible that bias in favor of marginalized groups could be at play in hiring, even if structural biases are still present. Yet, many if not all the faculty hiring studies on bias (e.g., Eaton et al., 2020; Steinpreis et al., 1999) were conducted prior to higher education's most recent reckoning with systemic racism (Perez, 2022). As a result, we were interested in examining intersectional identities across multiple CVs in this study.

The kinds of racial and gender biases we have been discussing to this point might be considered, "differential treatment" which is one of at least two ways in which candidates might be harmed. Cheryan and Marcus (2020) observe that differential treatment takes place when two candidates of similar qualifications are treated differently only because of their gender, race or other characteristic. However, there are of course other ways in which inequitable hiring processes occur. Cheryan and Marcus observe that sometimes the issue is not differential treatment but that the characteristics most important for the job are more likely to be associated with a particular group (a default characteristic). So, for example, the formal job criteria might preference excellent scholarship and doctoral training. Committee members associate Ivy League institutions with these criteria. Racially minoritized candidates are less represented in those institutions. In this way, the committee may employ seemingly neutral criteria that in fact reinforces a racist system (Ray, 2019).

Ranking versus Rating

In addition to our interest in asking evaluators to review a group of candidates of different qualifications and perceived identities, we were also motivated to understand whether ranking them or recommending a threshold list of candidates for hiring magnified or mitigated bias. There are reasons to believe what happens as search committees create a short-list and narrow their final candidates down at the end is important for hiring outcomes. At some universities, equity procedures and/or the preferences of the hiring official require that the search committee rank (e.g., first preference, second preference, etc.) their finalists as they submit them for hire (Fine & Handelsman, 2012). In others, committees provide a list of candidates they consider "hireable" (Candidate A and Candidate C are equally hireable) (Fine & Handelsman, 2012). In the latter situation, the search committee understands that if they leave a candidate off the list, that candidate will not be hired, whereas in the former, if the committee indicates a candidate as a third or fourth choice, there is still a chance that candidate could be hired.

In either situation, there are different stakes and a different choice architecture surrounding the decision (Kahneman, 2011; Thaler & Sunstein, 2008). In the threshold context, faculty members are being nudged to say a group of candidates are relatively equally qualified. When ranking, faculty evaluators are asked to put individuals in order, ostensibly in order of their qualifications. Although the evaluator would guess that those placed at the bottom of a ranked list would not be hired, it is unclear where the line would be drawn. That is, a hiring official may move down the ranked list to the third or fourth choice before finalizing a candidate. As such, the two decisions evoke different levels of risk-taking for the committee. In the first situation, the committee infers risk by leaving their specific preferences hidden: the risk is that the hiring official might select a candidate who they did not want as much as another. In the second situation, the risk is that the third or fourth ranked candidate is viewed as less qualified and might therefore encounter greater resistance in the department once hired. Thus, we sought to understand if employing one of these strategies was more associated with bias than the other.

Evaluator Demographics and Disciplinary Expertise

Finally, we were interested in whether the demographic characteristics of the evaluators would impact their evaluations of both competitiveness and ultimate selection. Studies of implicit bias typically show that cultural and social norms lead all humans – across race, gender, level of education, and other aspects of identity to have somewhat similar biases (Banaji & Greenwald, 2013). For instance, studies of bias in faculty hiring and letters of recommendation showed that both men and women evaluators demonstrate gender bias (Madera et al., 2019; Steinpreis et al., 1999).

In academe, studies show that disciplinary and field background is another salient identity that may shape bias (Posselt et al., 2020). Disciplinary and field background is important for two reasons. First, fields are differentiated by varying levels of progress toward racial equity in faculty selection. Some fields like education have greater levels of racial and gender diversity across multiple subfields, whereas some fields like psychology and biology have growing levels of diversity but is higher in some subfields than others (e.g., diversity being higher in social psychology versus clinical psychology). We were interested in the

field of mechanical engineering because it presented an interesting case: diversity has been discussed at length in professional settings and conferences (Matthews, 2020), but progress has been slow, especially in robotics where both racial and gender diversity has been stagnant (Patrida, 2022). The compositional diversity of mechanical engineering and robotics makes it like other STEM areas with stated commitments to equity and inclusion, but low compositional diversity. Secondly, the selection of a subfield (i.e., robotics) stems from the reality that most faculty searches will have evaluators whose primary expertise matches that of candidates, and those who are "one step or more" to the side, meaning in a related but secondary field. For instance, a clinical psychologist may serve as an "outside member" on a search committee for a social psychologist. Additionally in the departmental vote, all departmental faculty that represent a range of academic subfields contribute to the decision of whether or not to hire a candidate in the final round.

We see two potential ways that a faculty member's disciplinary background might shape their evaluation of a candidate. On the one hand, ambiguity can invite noise and guessing, as well as implicit biases to emerge (Kahneman, 2011). Outside members' lack of subfield knowledge may increase the ambiguity of their assessment and make them more prone to short-cuts and social biases in evaluating candidates. Alternatively, subject matter experts can become overconfident in their evaluation of material with which they are most familiar and be more prone to biases (Moore & Schatz, 2017) which causes them to overlook important contexts and information in their assessments. Knowing the field well may also make a faculty member more likely to evaluate candidates by relying on different factors (e.g., publications or postdoctoral experience) than those with less subject matter expertise, which could shape hiring outcomes. In other words, being a subject matter expert or from a secondary field may make faculty members pay more or less attention to content (e.g., publication number or quality) in their evaluation, and therefore more or less prone to rely on biases. Given outside members of search committees are also often added to increase the diversity of the committee, we were interested in identifying any interactions between primary and secondary field reviewers, hiring outcomes, and biases.

Methodology

We employed a within-subjects, convergent mixed methods experimental survey design (Creswell & Plano Clark, 2018) to understand how, if at all, candidate characteristics, selection conditions, and evaluator characteristics impact hiring outcomes. In our experimental survey, participants responded to Likert scales and qualitative text entry boxes to evaluate the competitiveness of ten fictitious candidate CVs that differed in their gender identity, racial identity, and publication record. After rating each candidate's competitiveness, participants were assigned to one of two initial conditions: a ranking condition to rank the candidates from most competitive to least competitive, or a threshold condition to send three CVs to the hiring official for top consideration. Both conditions had the option to input comments. Participants were then exposed to the alternate condition to ensure internal consistency between conditions. After completing the survey, participants provided optional demographic information such as their race, gender, and subfield specialization.

We designed our survey as a multiple method instrument to collect both quantitative and qualitative data. We conducted a concurrent mixed method design, and more specifi-

cally, the questionnaire variant (Creswell & Plano Clark, 2018). This style of concurrent design is when researchers use "both open- and closed-ended questions on a questionnaire and the results from the open-ended questions are used to confirm or validate the results from the closed-ended questions." (Creswell & Plano Clark, 2018, p. 73). Though this does not provide as rich of data as other types of qualitative methods (e.g., interviews), this method was still the most appropriate means of collecting participant's immediate evaluations. Moreover, our approach satisfies necessary hallmarks and conditions of convergent mixed methods designs, such as independent streams of data collection and analysis, the merging process, and combined interpretation. Using both forms of data in this way also more closely replicates actual hiring procedures: evaluators typically rate candidates based on a reading of their CVs, and then justify those ratings in search committee deliberations. In what follows, we describe the steps that went into designing each CV, the survey instrument, and the validation and administration of the survey instrument. We then describe our analyzing process, which involved first analyzing the quantitative data, then analyzing the qualitative data, then merging datasets, and interpreting the convergence and divergence of the data.

Curricula Vitae (CV) Creation

Prior to CV creation, we designated mechanical engineering as our primary discipline, and robotics as our primary subfield. Given that there are differing expectations for research, teaching, and service across fields and subfields (Posselt et al., 2020), we wanted a specific subfield to stabilize CV expectations. Next, in line with previous scholarship (e.g., Eaton et al., 2020; Moss-Racusin et al., 2012), we manipulated the perceived identities of each CV by using two different data sources. For first names, we used data from mortgage applications as guided by Tzioumis (2018). We used this database to identify popular recurring first names by race and gender since the U.S. Census only systematically collects data on last names. Thus, U.S. Census data was appropriate for identifying popular recurring surnames by race and gender. Using both datasets, we created fictitious names that represented four different profiles: Black women, Black men, white women, and white men. The intent herein was to identify if candidates with names that signaled varying identities but had similar qualifications would be evaluated differently. At the same time, we did not want to signal to participants that diversity was a key focus of our study, and we wanted the ten CVs to mirror the demographics of robotics faculty. Therefore, we included two additional white men to signal a gender and race balance that was closer to the current demographics of robotics to reduce assumptions regarding the intent of our study.

To create the components of each CV, we collected 20 CVs from real postdoctoral scholars and early-career faculty in robotics. We replicated content from these CVs to create authentic accolades that would be recognized by mechanical engineers. Since we chiefly wanted to manipulate publication records to understand bias or lack thereof (e.g., a male candidate being rated as more competitive than a woman candidate despite fewer publications), we used the real CVs to determine an aggregate "competitive" count that would be typical at the point of hiring an assistant professor. We manipulated publication record as our primary qualification for two reasons. First, number of publications is a very common concern for hiring in research universities; it would be rare for a tenure track search to not consider this, as well as authorship and journal quality. Second, we wanted to test publica-

tion productivity because there are equity concerns in using number of publications in hiring decisions. Prior research shows that access to prestigious institutions, networks, and awards constrains levels of productivity for minoritized scholars (e.g., Bendels et al., 2018; Lubienski, et al., 2018; Mendoza-Denton et al., 2017). Each profile (e.g., Black men, white women) had a CV with a "high productivity" count of ten publications, or a "low productivity" count of six publications, for a total of ten CVs. We also controlled for the number of first authored publications, and the journal quality by aggregating journal impact factor and ensuring each CV had stable aggregate scores in this area.

Other qualifications (e.g., educational background, appointments, awards, and teaching) were held constant across CVs. However, because this was a within-subjects experimental design wherein all participants read all CVs - as opposed to prior studies in which each participant only read one CV - this meant that the CVs could not have identical characteristics. To imitate hiring decisions in a more naturalistic setting, each candidate needed to have different publications, graduate from different institutions, receive different awards, and vary somewhat in courses taught. For items such as educational background, we used university rankings to create bands of similarly prestigious institutions and departments that would be considered on par with each other. For areas such as awards or publication titles where no such numerical ranking exists, we used a validation strategy from prior studies to ensure internal consistency of CV criteria (e.g., Eaton et al., 2020; Moss-Racusin et al., 2012). We used subject matter experts (SMEs) to provide feedback on whether backgrounds and experiences were comparable. We used 12 subject matter experts in an iterative three-wave process. That is, four SMEs reviewed the 10 CVs and identified credentials that seemed better or worse than the average intended in each band of candidates. We revised the CVs to equalize and sent them CVs to a second group of four different SMEs. These SME's reviewed and made minor recommendations for revisions (e.g. tweaking a paper title to be more realistic or description of an award). We then sent revised CVs to a final group of four SMEs in a third wave of validation. These 4 new subject matter experts confirmed that indeed we had achieved comparable qualifications across the CVs in the areas that were held constant and within the two productivity bands.

Survey Instrument Creation

All ten CVs were imputed into Qualtrics, an online platform to create and disseminate surveys. At the onset of the survey, participants were provided the following introductory prompt: "Please review 10 CVs of postdoctoral associates applying for a tenure-track assistant professor position at a research university…Please rate the CVs as if you were making actual hiring decisions." Participants evaluated all ten CVs in randomized order by rating the competitiveness for each along a 5-point Likert-scale, ranging from "not at all competitive," to "most competitive." They were also provided an option to comment on their scoring in a text entry box. After CV review, participants rated the relative importance of categories such as educational background and teaching experience as either "not important," "somewhat important," or "very important," and were provided an optional qualitative comment box to contextualize their responses. Participants were then randomly assigned to either a control condition of ranking all ten candidates, or the treatment condition of creating a threshold list by providing "1–3 names to the Dean for final hiring consideration," and then were exposed to the alternate condition after completing the primary condition. At the end of the

survey, participants were given the option to specify their race, gender, and subfield (e.g., thermodynamics, robotics).

Administration and Sample

To identify a list of mechanical engineering faculty for dissemination, we began by selecting a range of universities within the *U.S. News and World Report* top mechanical engineering graduate programs. We selected schools that were considered "high research activity" or "very high research activity," according to the Carnegie classification system. Once identified, we used web scraping techniques to comb through each institution's mechanical engineering program faculty, and selected tenure-track faculty for solicitation. We emailed each faculty member with an offer to participate in the survey without mentioning that the CVs were of fictitious candidates. Our strategy, like many other studies before, was not to extensively divert participants, but rather avoid blatantly sharing the multiple purposes of the study. Participants took the survey online on their own devices and received a $50 Amazon gift card after completion.

Between December 2020 and August 2021, we contacted 1,654 tenure-track faculty at 59 U.S. research universities. 320 faculty fully completed the survey - a 19.3% response rate. In terms of racial and ethnic diversity, 17.5% of participants identified as either "American Indian or Alaskan Native," "Black or African American," "Hispanic, Latinx, or of Spanish origin," "Middle Eastern or North African," "Native Hawaiian or Other Pacific Islander," or "Multiracial," with the remaining 49.5% identifying as "White," and "31.4% identifying as "Asian or Asian American." For gender, 17.5% of participants identified as women, 69.8% identified as men, 1% as non-binary, and 4.6% chose not to disclose. Finally, 21.5% of participants identified themselves as conducting research in robotics or related subfields that would designate them as subfield specialists for the purposes of our analyses. Five participants were excluded from analyses due to written comments that indicated they did not take the survey earnestly, bringing the final analytic sample to 315. We conducted a series of power analyses for each of our planned analyses, specifying a medium effect size and .80 power. We found that required total sample sizes ranged from 34 to 263, depending on the type of analysis (e.g., independent or paired, Z test or t test). Importantly, we are adequately powered with our sample size of 315.

Quantitative Measures and Variables

In this study, participants were asked to evaluate the candidate CVs via three quantitative outcome measures. First, they rated the perceived competitiveness of each candidate, ranging from 1 (not at all competitive), to 5 (most competitive). Half of the participants were first exposed to the ranking condition in which they first ranked all ten candidates on the basis of who they would most (ranking = 1) to least (ranking = 10) recommend for selection. The other half of the participants were exposed to the threshold condition first, in which they chose three candidates who they would move forward to the final round, which we coded as a binary variable (0 = Not Top Three; 1 = Top Three).

We also collected data on two participant characteristics and used those as independent variables: gender (0 = woman, 1 = man), and specialization, where 0 equated to not belonging to the field or robotics or a related subfield, and 1 signaled that they identified their

scholarship as within the field of robotics or a similar subfield (i.e., control and/or biomechanics) that would designate them as subfield experts. We also gathered information on whether the participant was randomly assigned to the ranking condition first (Order = 0), or threshold condition (Order = 1).

Quantitative Analyses

Once we identified our predictor and outcome variables, we conducted several different tests to answer our research questions. To understand how participants rated competitiveness, we conducted a paired-samples t-test to compare mean competitiveness ratings between all candidates. In order to identify how evaluator characteristics such as gender and subfield expertise shaped assessments, we used a series of Mann-Whitney U tests, Z tests, and independent samples t-tests to compare independent assessments. In regard to our last research question, we used two different statistical tests. For the ranking condition, we conducted a series of Wilcoxon Signed-Rank tests, a type of nonparametric test that is more appropriate than traditional parametric comparison tests (e.g., paired samples t-tests) when comparing ranked and ordinal data (Aron & Aron, 1999). For the threshold condition, we compared the proportion of positive responses (i.e., chosen as a top three finalist versus not) for each candidate CV using McNemar's test for comparing paired proportions (Adedokun & Burgess, 2012). To understand the differences between both conditions, we looked across both tests and located similarities and differences in the relative standing of candidates across both conditions.

Qualitative Analysis

After conducting quantitative analyses, we moved toward the qualitative portion of our study to reach greater depth and explanation of our survey data. We used directed content analysis to analyze participants' text-based responses (Hsieh & Shannon, 2005). Directed content analysis derives from content analysis, which is a method to analyze text-based data through a systematic classification process of coding and identifying patterns. Whereas conventional content analysis uses inductive reasoning to generate categories sans prior knowledge, directed content analysis is heavily influenced by existing literature, theory, and prior research to form codes and identify patterns in textual data (Hsieh & Shannon, 2005). As was the case in our study, we were aware of the prior studies in this area and leveraged our quantitative findings to generate a priori codes such as "importance of publications," and "weight assigned to perceived identity as a criterion." But we also developed new codes as well when text segments did not fit neatly into pre-prescribed categories. Throughout the coding process we used structural coding to organize codes by research question, and magnitude coding to determine the magnitude or prevalence of codes in these domains (Saldaña, 2016). These deductive procedures brought the qualitative data closer to the quantitative data to answer our research questions.

Mixed Methods Integration

Mixed methods designs require explicit integration of quantitative and qualitative data to achieve a more nuanced understanding of the phenomena that neither can convey sepa-

rately. That is, there is unique explanatory power from each method, and *shared* explanatory power once brought together. To reach such an understanding, data must be brought together through an integration process (Creswell & Plano Clark, 2018). The first point of integration was through making joint display tables to identify how certain qualitative trends (e.g., considering one's own identity in justifying an evaluation) varied by evaluator demographics – one of our research questions. Examining both qualitative and quantitative data simultaneously in this way provides more explanatory power, and makes the additional contribution of showcasing convergent, rather than divergent, rationales. Another way in which we combined the data was by organizing text entry comments by competitiveness score. For example, this allowed us to examine the qualitative rationales of survey participants who rated Essie as the most competitive candidate, versus those who rated her as less competitive in descending order. Organizing the data in this way helped us address the first research question, and a similar approach was used to reach a more nuanced understanding of selection conditions (i.e., ranked versus threshold) as well. Overall, the concurrent mixed methods approach was the ideal methodological strategy to formulate valuable insights into how faculty participants responded to our experimental questionnaire.

Limitations

There is an important limitation in creating the CVs that merits consideration when interpreting results. One potential limitation are perceived differences between the different CVs based on conditions we did not purposefully manipulate, such as small differences in publication titles, institutions, or specific names of awards. Though we did conduct a rigorous, three-wave validation process with SMEs, each CV was not identical to preserve the semblance of authentic candidate review. As a result, some participants may hold idiosyncratic preferences for certain institutions over others, despite our SMEs confirming that they are similarly ranked and evoke relatively similar perceptions of prestige within the discipline. Although this opens the possibility that other factors explain the differences we observed in candidate evaluation, we do not believe that this was a significant concern for two reasons. First, the validation process with SMEs, as we previously discussed. But secondly, the qualitative response trends do not indicate that things like publication titles or institutional names between the CVs were deciding factors. Second, even in real evaluation settings, details like institutional type are never constant for all applicants. Finally, as we discuss in the results, publication count (which were consistent across CVs within the same competitiveness band), and race/gender identity (which we purposefully manipulated) were cited as the most important factors. Future research could bolster the interpretation of our findings by randomizing CVs instead of tying them to specific candidates.

Finally, there are valid concerns of social desirability when asking participants to rate applicants in an experimental design. Specifically, being racist or sexist are widely seen as undesirable character flaws, which one could argue would influence raters to rate our women and Black applicants more favorably in an experiment, but perhaps not in real life (Luke & Grosche, 2018). However, our results do not support this explanation, as we did not find that Black applicants were universally rated above white applicants, nor that women applicants were universally rated above male applicants. Specifically, our Black female applicant with an average CV was consistently rated lower than white and male applicants with average CVs. If social desirability to be seen as non-racist and non-sexist were strongly influencing

our results, we would expect both Black female applicants to be rated most highly, or at least most highly within their respective bands. While we cannot fully rule out socially desirable responding from this study, our findings suggest it was not the primary determinant.

Findings

In this section we share findings regarding the influence of perceived candidate identity, rating or ranking, and participant identities on hiring decisions. Each section begins with the quantitative findings, and then we supplement those with qualitative data where appropriate to illustrate convergence and divergence in answering the research questions.

Candidate Competitiveness: Who was Hired? What Mattered most to Evaluators?

Participants were given ten CVs, five of which were manipulated to have an above-average number of publications and five of which were designed to have an average number of publications. In each "band" or group of five CVs, names were manipulated to signal particular identities. All other factors such as academic appointments, awards, and teaching were controlled to be equal through multiple rounds of validation by mechanical engineering subject matter experts. In this first section we show which candidates were considered most competitive and ultimately selected, and the factors that were most prominent in their decision-making.

Table 1 displays the mean competitiveness rating of each of the ten candidates. A higher mean indicates the candidate was seen as more competitive. Results show that the Black and white female applicants with high publication records received the highest ratings, and that the Black female and white male applicants with an average publication record received the lowest competitiveness ratings. These results largely match the relative standing of applicants for ranking (Table 2) and rating (Table 3) selection methods.

When asked to rate different criteria, in terms of how important they were to participants in making these types of decisions, it is clear that publication record (both quality and quantity was seen as the most important (See Table 4).

Given that these findings depart from many previous studies showing bias against marginalized candidates, we were interested in what qualitative comments might reveal about

Table 1 Competitiveness ratings

Applicant Characteristics				Competitiveness Ratings	
Race	Gender	Record	Name	M	SD
Black	Female	Outstanding	Ayanna	3.66_{a}	0.839
White	Female	Outstanding	Kathleen	3.58_{b}	0.859
Black	Male	Outstanding	Cedric	3.57_{b}	0.829
White	Male	Outstanding	Doug	3.42_{c}	0.820
White	Male	Outstanding	H. Neil	3.32_{d}	0.901
Black	Male	Average	Jermaine	2.93_{e}	0.912
White	Female	Average	Emily	2.91_{e}	0.903
White	Male	Average	John	2.85_{f}	0.889
Black	Female	Average	Essie	2.83_{f}	0.889
White	Male	Average	Rick	2.81_{f}	0.890

Note. Different subscripts indicates significantly different at $p < .05$ level

Table 2 Applicant average rankings (out of 10)

Applicant Characteristics				Rankings	
Race	Gender	Record	Name	M	SD
Black	Female	Outstanding	Ayanna	2.94_{a}	2.104
White	Female	Outstanding	Kathleen	3.45_{b}	2.472
Black	Male	Outstanding	Cedric	4.25_{c}	2.595
White	Male	Outstanding	Doug	4.37_{c}	2.167
White	Male	Outstanding	H. Neil	$5.92_{d,e}$	3.067
White	Female	Average	Emily	5.96_{d}	2.396
White	Male	Average	John	6.19_{d}	2.475
Black	Male	Average	Jermaine	6.36_{e}	2.281
Black	Female	Average	Essie	7.73_{f}	2.185
White	Male	Average	Rick	7.83_{f}	1.970

Note. Different subscripts indicates significantly different at $p < .05$ level

Table 3 Applicant placement in top three group

Applicant Characteristics				Finalist Frequencies and Percentage		
Race	Gender	Record	Name	Top Three	Not Top Three	% in Top Three
Black	Female	Outstanding	Ayanna	210	99	0.68_{a}
White	Female	Outstanding	Kathleen	190	119	0.61_{a}
Black	Male	Outstanding	Cedric	141	168	0.46_{b}
White	Male	Outstanding	Doug	98	211	0.32_{c}
White	Male	Outstanding	H. Neil	80	229	0.26_{c}
White	Female	Average	Emily	49	260	0.16_{d}
Black	Male	Average	Jermaine	40	269	$0.13_{d,e}$
White	Male	Average	John	27	282	$0.09_{e,f}$
Black	Female	Average	Essie	21	288	0.07_{f}
White	Male	Average	Rick	14	295	0.05_{f}

Note. Different subscripts indicates significantly different at $p < .05$ level

Table 4 Importance of different criteria for evaluating candidates

Criteria	Very Important		Somewhat Important		Not Important	
	#	%	#	%	#	%
Publication Quality	270	85.71%	38	12.06%	7	2.22%
Publication Quantity	233	73.97%	74	23.49%	8	2.54%
Educational Background	170	53.97%	122	38.73%	23	7.30%
Teaching Experience	80	25.40%	162	51.43%	73	23.17%
Honors	75	23.81%	177	56.19%	63	20.00%
Post-Doctoral Institution	73	23.17%	190	60.32%	52	16.51%
Dissertation Topic	66	20.95%	144	45.71%	105	33.33%

decision-making. Participants' comments in the competitiveness rating and selection decision exercises show that publication count was the foremost factor used to differentiate candidates. One participant explained that the "number and quality of publications was my first criteria…honestly it was hard to make a decision because it seems that all these candidates are very competitive." Another participant's comment encapsulates how numerous participants went about making their competitiveness rating and selection decisions:

> I could only meaningfully designate between the applicants based on their publication records…there were a "top 5" with better publication records than the "bottom 5". Beyond this, it was very hard to meaningfully distinguish between the quality of the applicants…If the survey constraints did not exist, and these were the only applicants, I would schedule the top 5 for screening interviews.

Either intentionally or unintentionally, most participants ordered their assessments using this logic, establishing a "top 5" pool of candidates with ten publications and a "bottom 5" pool of candidates with fewer publications. It was also evident that participants spent much more time and attention on the top five candidates. One participant said that they only carefully considered the top five "because below that wouldn't make any difference anyway in real life," whereas another explained that it was "very difficult to rank the candidates in the lower half of the list."

After identifying differences between bands, we probed for differences within bands to understand how and why Ayanna and Kathleen received the highest competitiveness ratings. Given that the top five candidates all had the same number of publications, participants used other factors to inform their decisions. One prominent factor was perceived identity. Comments pertaining to the salience of perceived identity accounted for 12.6% of all comments in the ranking condition, and 17.3% of all comments in the threshold condition. Using structural and magnitude coding, we identified three different types of comments: (1) non-descriptive comments that highlighted the fact that certain candidates would increase the racial and/or gender diversity of their departments, (2) more detailed comments that suggested women and men of color and white women should receive a "second (or closer) look," and (3) comments that suggested that white women and racially minoritized women would or should receive a "boost" in competitiveness due to their identity-based characteristics. The first category was most common (e.g., "diversity potential," "very competitive as a woman"), whereas the second category had more nuance but stopped short of an action or "boost" (e.g., "Ayanna would be looked at since underrepresented candidates are given a closer look"). The third type of comment were more direct on how perceived identity made certain candidates more competitive than others (e.g., "I bumped her by a rank because of her being a minority in MechE").

Comments regarding perceived identity show that it was a salient criterion in competitiveness ratings and selection decisions, but this does not mean that it was the most salient selection criterion. A greater percentage of overall comments (28%) pertained to other factors. From most recurring to least recurring, faculty used other information on awards, institutional type, teaching, area of expertise, and information on mentoring and service to decide between the top five candidates. Despite controlling for the quality level of these to be equal, participants may have had idiosyncratic attachments to certain accolades over others as would happen in any real search (e.g., I went to Ohio State and so did this candidate). A significant number of comments (21%) described a desire for more information outside of the bounds of our survey to make their decisions. In order of most recurring to least, faculty wanted interviews, research statements, letters of recommendation, teaching statements, information on the department's research needs, and input from other committee members. Considerations of perceived identity typically came *after* an assessment of publications and other factors that constituted academic excellence or quality. One participant stated that when the "quality of the candidates was comparable, I chose women's names over men's

names because we are usually told to prioritize diversity in our hiring choices. When quality was comparable, I also chose names that indicated underrepresented groups over others." Another participant wrote,

> I am looking not just for academic achievement, but for some diversity at least in gender in my final selection since this group of CVs seemed relatively equivalent. I would be looking for other types of diversity were I provided that information because I recognize that often, if the CVs look the same, the candidates from minority groups (whether female-identifying, BIPOC, disabled or otherwise) are likely to have had challenges that those from the majority didn't have.

Overall, the preference for the Black woman and White woman candidates in the highest publication band may be explained by how participants understood the weight and ordering of secondary criterion after applying the primary selection criterion of publication productivity as a narrowing agent. Data show that considerations and contributions of diversity were among those factors but did not surface as most important through qualitative magnitude coding. These data illustrate the contours of how participants constructed ideal logic models of competitiveness.

Ranking and Thresholding: Does it Matter?

Participants were randomly assigned to either rank all candidates (ranking) first and then identify a list of three to move forward (threshold) or complete threshold ratings first and then rank all candidates. The effect of this order of ranking assignments was tested via a Mann-Whitney U Test comparing the rankings of each candidate dependent on whether the participants ranked candidates first or after they completed the threshold ranking. Threshold ratings were compared by conducting a series of Z tests comparing the independent proportions of participants who put each candidate in the top three, dependent on the order in which they completed the threshold ratings. Order had no influence over either rankings or threshold ratings.

We then turned to understand whether the choice architecture of ranking or rating shapes hiring decisions. Table 4 displays the mean *ranking* of each of the ten candidates evaluated by each participant. A lower ranking indicates the person was more highly recommended. Of note, the Black female applicant with an outstanding record had the highest average ranking, followed by the White female applicant with the outstanding record. Both applicants' rankings were significantly higher than all other candidates. Regarding the candidates with average records, one of the two average White male applicants and the average Black female applicant received rankings significantly lower than everyone else.

We then turned to see how candidates did if they were *rated*, the threshold condition. Table 2 displays the percentage of participants who recommended that a particular candidate be moved on to the final round, as well as the frequencies indicating how many times each candidate was or was not placed in the "Top Three" finalist group. We compared the proportion of positive responses (i.e., chosen as a top three finalist versus not) for each candidate using the McNemar's test for comparing paired proportions (Adedokun & Burgess, 2012). Of note, the Black and White female candidates with outstanding records were significantly more likely to have been placed in the finalist group, compared to all other

candidates. One of the White male candidates with an average record and the Black female candidate with an average record were the least likely to be placed in the finalist group. However, their proportional "success rate" was not significantly lower than the other White male with an average record.

Overall, results suggest very few differences between selection methods (ranking, threshold), in terms of the relative standing of applicants. For outstanding applicants, the Black female applicant was always evaluated most highly, although not significantly differently from the White female applicant under the threshold rating method. For average applicants, one of the White male applicants and the Black female applicant were always evaluated most poorly, although under the threshold rating method the other White male average applicant was evaluated similarly poorly. One of the White male applicants with an outstanding record benefited more from the rating than ranking, dropping to a similar standing as some of the candidates with an average record when ranked. But overall, there were very few advantages to either evaluation method for any of the applicants.

Who is Looking? Evaluator Demographic Characteristics

We next analyzed results based on participant gender, race and specialization. Rankings for each applicant were compared by gender or race or specialization of participant using a Mann-Whitney U test to compare independent rankings. Threshold ratings were compared based on participant race using a series of Z-tests of independent proportions. Competitiveness ratings based on participant race were compared by conducting a series of independent samples t-tests. Due to low diversity within the sample, participant race was coded as a binary variable (0 = White, 1 = POC). Specialization of the participant was coded as a binary variable as well (1 = those most closely aligned with Robotics, 0 = those who are one or more steps to the side in terms of expertise).

In comparing participant gender, one of the average White male applicants was ranked more highly by men ($M = 6.07$) than by women ($M = 7.15$, $p = .004$). Also, the average Black female applicant was ranked lower by men ($M = 7.80$) than by women ($M = 6.98$, $p = .015$). There were no significant differences by participant gender for either threshold ratings or competitiveness ratings for any of the applicants.

Four candidates received significantly different rankings dependent on participant race. The two White male applicants with an average record both received lower rankings ($M = 6.62$, $M = 8.16$) from White participants than they did from POC participants ($M = 5.88$, $p = .013$; $M = 7.50$, $p = .008$; respectively). The White female applicant with an outstanding record received higher rankings from White participants ($M = 3.13$) than she did from POC participants ($M = 3.81$, $p = .006$). The Black male applicant with an outstanding record also received higher rankings from White applicants ($M = 3.66$) than he did from POC applicants ($M = 4.80$, $p < .001$). Threshold ratings were largely unrelated to participant race, with one exception. The Black male applicant with an outstanding record was significantly more likely to be put in the top three by White participants (53%) than POC participants (40%, $p = .021$). One of the White male applicants with an outstanding record, the White female applicant with an outstanding record, the Black male applicant with an outstanding record, the White female applicant with an average record, and one of the White male applicants with an average record all were rated as significantly more competitive by White participants ($M = 3.43$, $M = 3.74$, $M = 3.70$, $M = 3.03$, $M = 2.94$, respectively) than they were by

POC participants ($M=3.20$, $p=.031$; $M=3.41$, $p=001$; $M=3.43$, $p=.007$; $M=2.80$, $p=.032$; $M=2.73$, $p=.046$; respectively).

There were very few significant differences by participant specialization, but both the outstanding White female ($M=3.31$) and one of the outstanding White male ($M=5.72$) applicants were ranked more highly by those who did not specialize in Robotics than by those who did ($M=3.91$, $p=.005$; $M=6.63$, $p=.028$; respectively). Also, the average White female applicant ($M=6.11$) was ranked lower by those who did not specialize in Robotics than by those who did ($M=5.18$, $p=.005$). There were also very few differences by participant specialization in threshold ratings. The Black female applicant with an average record was significantly less likely to be put in the top three group by those who did not specialize in Robotics (5%) than by those who did specialize in Robotics (13%, $p=.015$). In contrast, one of the White male applicants with an average record was significantly more likely to be put in the top three groups by those who did not specialize in Robotics (11%) than by those who did specialize in Robotics (3%, $p=.026$). There were no significant differences in competitiveness ratings by participant specialization.

Overall, quantitative findings indicate there were very few differences in evaluations by participant gender. Participant race difference trends indicate that White participants evaluated the outstanding Black male, outstanding White male, and outstanding White female applicants generally more positively than did POC participants. Specialization analyses indicate non-specialized participants gave outstanding White female, outstanding White male, and average White male applicants a boost in some evaluations (compared to specialized participants), while putting average White female and average Black female applicants at a relative disadvantage on some evaluations.

We examined the qualitative data to determine if there were any trends across participant race, gender, and specialization, and their relation to competitiveness ratings or positioning within either the ranking or threshold conditions. Regarding specialization, no participant explicitly commented on how their expertise influenced their decision-making. In terms of race and gender, a small fraction of participants (4%) discussed how their own identity shaped their decision-making. This was most often women of color, men of color, and white women who described how their own identity and lived experiences helped them see that candidates like themselves had to navigate racist and sexist academic environments, making those candidates more competitive in their eyes. For example, one participant said, “I am a female faculty in engineering. I do not know a single female or underrepresented minority professor who did not have to face discrimination in their studies and their careers. So, those two female candidates succeeded notwithstanding many obstacles thrown their ways, and I consider this a badge of honor that needs to be rewarded.” Occasionally white men discussed DEI as well, but this was less often from a personal experience perspective and more often from a compliance perspective, such as “meeting diversity requirements for the college.”

Discussion

The prevalence and magnitude of bias in faculty hiring is a pressing topic in higher education, and there are generally two prevailing positions on the matter. Many argue that implicit and explicit bias pervade faculty hiring in ways that negatively impact historically minori-

tized groups, and various experimental studies, longitudinal career studies, and qualitative studies support this claim (e.g., Beattie et al., 2013; Eaton et al., 2020; Liera & Hernandez, 2021; Moss-Racusin et al., 2012; Rivera, 2017; Sheltzer & Smith, 2014; Steinpreis et al., 1999; White-Lewis, 2020). As a result, colleges and universities over the last two decades have implemented implicit bias trainings and guides to make their faculty more aware of their biases, and to provide them with the tools to mitigate them (Fine & Handelsman, 2012; Fine et al., 2014).

Yet it is for this very same reason that others argue that bias in faculty hiring is minimal, if not entirely non-existent. Either by virtue of institutional trainings showing intended positive results, or the assumption that historically marginalized groups are advantaged in today's job market, some argue that the kind of bias shown in prior studies is an old issue, and there is supportive empirical literature in this direction as well (Bernstein et al., 2022; Williams & Ceci, 2015). Given design limitations in extant experimental studies not using comparative evaluation, and potential shifts in search committee thinking and priorities, we wanted to understand the complicated and nuanced issues of faculty selection set in a more recent landscape and realistic setting. In what follows, we outline our three primary areas of contribution and relate them to prior literature on faculty hiring and decision-making, and then discuss what this means for future practice and research in prioritizing equity and excellence in faculty hiring.

Implicit Bias and Choice Architecture

Our study considered issues of choice architecture (e.g., the context surrounding decision-making) in two ways. We explicitly tested whether rating or ranking candidates shaped hiring outcomes and found that it did not. We were surprised that there were very marginal differences between participants that rated candidates and those that ranked them. Tables 2 and 3 show stable distributions across selection techniques, with the Black woman and white woman with outstanding publication records being considered most competitive and selected for hire most often. We also implicitly tested whether comparative evaluation shaped hiring outcomes. It is true that we did not conduct two studies, one with comparative evaluation and another with single CV review. However, by requiring participants to rate and rank, and by virtue of their qualitative comments showing their work, we replicated many of the previous studies but altered the terrain slightly by making participants compare CVs, which yielded very different results from similar prior work that found gender and racial bias when participants reviewed one CV.

In this study we did not find explicit negative bias against historically marginalized candidates. As shown in Table 1, the Black woman candidate with an outstanding publication record and the White woman candidate with an outstanding publication record were deemed the most competitive candidates and were selected for hire most often as shown in Tables 2 and 3. There may be a few explanations of this; we believe that this is due, in part, to the choice architectural design of comparative evaluations. Comparative evaluation techniques, when used as an equity tool, have been shown to reduce bias in decision-making (Bohnet, 2016). This is because evaluators can consider files side by side and use consistent metrics as benchmarks to inform their decisions, rather than rely on intuition, faulty retrospective thinking, and/or group stereotypes (Bohnet et al., 2016). When evaluators in this study considered applicants side by side, it may have provided counter-stereotypical information that

shifted their attitudes about Black and white women in mechanical engineering from what has been found in previous studies. Comparative evaluation may have operated as a form of choice architecture, reconstituting the decision in a way that reduced bias against these candidates. But before equity-minded advocates raise the alarm on this result, or those that argue implicit bias is a thing of the past celebtrate, we must ask: what mattered most to our participants in their hiring decisions that led to these results? We turn next to how they constructed competitiveness.

Construction of Competitiveness

By examining the construction of competitiveness, we mean to describe the process by which evaluators use criteria and cues to discern the perceived competitiveness of a candidate. Table 4 and the qualitative comments show that this process markedly begins with the evaluation of scholarly productivity, or the number and quality of peer-reviewed publications. This makes sense given that our participants were mechanical engineering faculty at research-intensive institutions, and that peer-reviewed publications are widely perceived to be most important criteria for hiring, promotion, and tenure in these settings. We did not tell participants this, but we had ourselves created two bands of candidates: those with an "outstanding" publication record and those with an "average" publication record.

There are two important points to be made here about how faculty decided who "made the band" (Pearlman et al., 2000–2009). First and foremost, it should be noted that participants were given all ten CVs to evaluate and were not interested in moving candidates from the lower strata to the upper strata based on any factor, including perceived identity. Thus, despite beliefs that affirmative action policies motivate evaluators to hoist lesser qualified members of minoritized groups into more qualified strata (Harrison et al., 2006; Jaschik, 2017), we did not find evidence of this. We instead found that the productivity bands were extremely durable, such that the five candidates with the highest number of publications were consistently rated as more competitive, and were selected more often, than those with fewer publications. Our second point relates to how and when perceived identity mattered. It mattered after research productivity was well-established, and after many other factors were considered as well. Evaluation in this manner resembles the "equalizer" perspective in faculty hiring (White-Lewis, 2019). White-Lewis (2019) found that faculty had different approaches for weighing the perceived identity and DEI contributions of candidates in hiring; whereas many faculty did not consider it all or were directly opposed, some believed that if two candidates were "equal," but one satisfied a departmental aim such as diversification, then they would prefer that person over the other. Given that perceived identity is what separated the three most competitive candidates from the bottom two White male candidates in that band, we believe that this played an important role in selection decisions. The results from our study show that the equalizer perspective may be becoming increasingly common in faculty hiring settings.

This perspective leads to two essential points about the construction of competitiveness. While we found what Cheryan and Marcus (2020) refer to as "differential treatment," or bias that favored that Black and white woman candidates with publications being held constant, we also found the presence of strong default characteristics, given that scholarly productivity was the most important deciding factor. The dilemma with the equalizer perspective is that it maintains the status-quo in the assessment of academic excellence without

interrogating structural constraints to achieving excellence for marginalized groups. Given previous research showing differential access to mentoring networks (Kachchaf et al., 2015; Weeden et al., 2017), labs and research funding (Chen et al., 2022; Hoppe et al., 2019), disproportionate DEI-related service and teaching loads that impact productivity (Jimenez et al., 2019; O'Meara et al., 2017) and differential rates of publication (Mitchneck, 2020) – especially during pivotal societal challenges such as the pandemic these structural default characteristics would likely have constrained the likelihood that our top Black and White female candidates would be in the top band for top consideration. It is clear that participants viewed these factors in isolation, rather than in combination, and applied them as blocks akin to building a pyramid: productivity as the base, other qualifications throughout, and identity as the tipping point when all things are considered equal. Perceived identity only mattered at the end, after the "band" had been brought together.

Role of Evaluator Identity

Finally, there was mixed evidence on the impact of participant identities on decision-outcomes. By and large, participant gender had little to no effect on competitiveness scores or selection decisions, but race had a greater effect. White male participants rated White male applicants with average records more harshly compared to faculty of color, and they ranked the Black male applicant with an outstanding record higher than faculty of color. Moreover, both White men and women ranked the White female applicant with an outstanding record (Kathleen) higher than participants of color and were statistically significantly more likely to put the Black male applicant with an outstanding record (Cedric) into the top three compared to participants of color. These results indicate that White faculty may be making progress toward more fairly evaluating CVs of candidates from marginalized groups, or were more cautious when evaluating said candidates, either by virtue of the comparative nature of our study (Bohnet, 2016) or for some other reason. This also emphasizes that all faculty benefit from interventions and training to reduce bias, which is not isolated to any one or two groups (Banaji & Greenwald, 2013).

Concerning participant specialization, we found evidence that participants outside of the field exhibited more bias against minoritized candidates than those inside the subfield. Those who did not specialize in robotics rated the outstanding White male applicants more highly and were more likely to put one of the White male applicants with an average record into the top three group compared to those who did specialize in robotics. Though we are unsure as to why non-experts exhibited more bias against the minoritized candidates, it may be that their lack of specific subfield knowledge (e.g., journal quality, subfield trends, etc.) means their decisions were made with less clarity about criteria, and thus they relied on short cuts and social biases, whereas subject matter experts were more familiar with the criteria and benchmarks and had more calibrated assessments in this study. Given that most search committees have both subject matter experts from the department, and faculty, administrators, and/or equity advocates who are non-experts, it is important to continue to study the role they play in making selection decisions.

Implications for Faculty Selection: Research and Practice

Our study makes three important contributions to research on faculty selection and efforts to make faculty hiring more effective and inclusive. First, we studied faculty selection using an experimental method with at least two components that more closely mirror actual hiring. We forced comparative evaluation of candidates with different qualifications and found less overt bias against historically marginalized candidates using this method. This should be examined and replicated in future studies. Second, we tested whether rating or ranking candidates, two common outcomes requested of search committees, resulted in different outcomes for candidates, and it did not. This is useful to know because ranked lists can unintentionally reinforce negative stigmas of those ranked 2nd or 3rd during onboarding and socialization processes. Thus, if the difference in outcomes is negligible, threshold lists prevent that and are advisable. Third, we found the very rigid construction of competitiveness bands and how perceived identity emerged as relevant only after all other factors had been considered equal, and in isolation of those factors. This is important because equity-minded scholarship reminds us that these factors operate in tandem to create unique opportunity structures that include some and exclude others. We explore implications for future research and practice building from these contributions.

The results from this study lead us to believe that we need greater clarity on evaluating candidates. There must be an explicit discussion of the equalizer perspective (White-Lewis, 2019) in search committee trainings. This perspective does not fit neatly into the "bias or no bias" discourse, yet it is becoming increasingly common in faculty hiring settings. Some may argue that it constitutes progress, given that it is not entirely orthogonal to increasing diversity. However, the equalizer perspective does relatively little to unsettle the default characteristics that embed discrimination into evaluation criteria (Posselt et al., 2020). There will also be renewed and heightened public interest in how candidates are evaluated, and which factors matter most given the Supreme Court's recent ruling on affirmative action. Thus, administrators must be proactive in ensuring that search paradigms comply with the law, but do not overcorrect in ways that reduce equity. Relatedly, we also urge those responsible for search trainings to examine default characteristics more critically, since they have been documented in the assessment of grantsmanship (Chen et al., 2022; Ginther et al., 2011; Hoppe et al., 2019) and scholarly productivity (Settles et al., 2020). We argue that this integrated approach must be intertwined in the faculty hiring landscape to increase awareness of the opportunity structures inherent in academic careers (White-Lewis et al., 2022).

For faculty search committees in particular, faculty need to work through if, and how, they are going to construct their bands, and when perceived identity and DEI contributions will matter before reviewing candidates. This is akin to calibration exercises (Culpepper et al., 2023; White-Lewis, 2020), where search committees identify the criteria and determine how they will measure excellence in those areas using a batch of sample files. One positive example of this is UCLA Life Sciences division's Mentor Professor program, which aims to increase faculty diversity by recalibrating when and how contributions to DEI are evaluated in faculty searches (UCLA Life Sciences, 2024). In addition to conducting division-wide searches, contributions to mentoring minoritized students is evaluated first, even prior to scholarly productivity, which then shapes how candidates are recruited and evaluated. Evaluation of the program is ongoing, but this illustrates how taken-for-granted aspects of

the hiring process, such as evaluating scholarly productivity first, can be reimagined to challenge default hiring criteria and change faculty hiring.

Next, we found that the ranking and threshold selection mechanisms led to very similar selection trends. Since we did not find any differences in candidate selection, we consider other aspects important to faculty professionalization, socialization, and onboarding. There is a concern that the threshold method for selecting candidates may challenge the search committee's ability to indicate who they want hired. That is, some faculty may perceive that their ability to make autonomous decisions will deteriorate in the presence of a hiring official, and that the official's decision is an encroachment on faculty power. However, there is an opposing concern that the ranking condition positions candidates as second-class citizens within their departments, such that candidates who were originally ranked second or third will be met with less enthusiasm in the department. Given we did not see differences in either the creation of bands or from an equity perspective, we recommend threshold lists be recommended to searches and hiring officials to avoid assigning stigma to candidates.

This study also speaks to the widespread practice of assigning outside faculty members onto search committees in order to increase compositional diversity. We found that subfield experts exhibited less bias compared to non-subfield experts. We see our finding as incongruent with the current practice, but also advise caution given that outside members do typically increase the compositional diversity of search committees, which is significantly related to more diverse recruitment (Kazmi et al., 2021). At minimum, these findings suggest that faculty outside of the department should not only attend search trainings but be calibrated to the norms and trends of the discipline in which they are contributing toward.

There are several implications for future research, both for future studies of this design and those of any design that examine faculty hiring. In conducting these types of experiments, there is a tension between creating a survey experience that is manageable but also mirrors real-life faculty hiring. We are well aware of the fact that faculty review more than a CV to make hiring decisions, but an experiment that asks participants to review letters of recommendation, multiple statements, and interview performances not only introduces more noise, but makes it more difficult for participants to actually take the survey instrument as well. In addition, at least in the first round of evaluation, in creating the lower band most search committees rely primarily on the CV in evaluation, so our method is defensible in that it parallels common practice in the first stage of review. Future studies should consider the balance of what makes the study feasible versus most naturalistic.

In sum, we see excellent opportunities for furthering this work in other directions as well. Future research could use actual hiring data to see if minoritized candidates have to have higher qualifications to get into the final band. In Bertrand and Mullainathan's (2004) seminal study, they not only found that Lakisha Washington and Jamal Jones received fewer callbacks than identical white candidates but found that the Black candidates needed eight more years of work experience to receive similar attention (Bertrand & Mullainathan, 2004). Using actual hiring data would also create opportunities to understand how search committees construct bands and study the characteristics of candidates that are passed over for the most competitive group within a defined field such as psychology. These data would also combat limitations of experimental studies (e.g., social desirability, risk, etc.), and would provide a more nuanced window into how faculty construct competitiveness and make complex decisions regarding faculty applicants.

Declarations

This work was supported by the National Science Foundation Division of Human Resource Development (grant number 1820975).
The authors have no relevant financial or non-financial interests to disclose.

References

Adedokun, O. A., & Burgess, W. D. (2012). Analysis of paired dichotomous data: A gentle introduction to the McNemar test in SPSS. *Journal of MultiDisciplinary Evaluation*, *8*(17), 125–131.
Aron, A., & Aron, E. N. (1999). *Statistics for psychology*. Prentice-Hall, Inc.
Banaji, M. R., & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Bantam.
Beattie, G., Cohen, D., & McGuire, L. (2013). An exploration of possible unconscious ethnic biases in higher education: The role of implicit attitudes on selection for university posts. *Semiotica*, *2013*(197), 171–201.
Bendels, M. H. K., Muller, R., Brueggmann, D., & Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PLOS ONE, 13*(1): e0189136.
Bernstein, R. H., Macy, M. W., Williams, W. M., Cameron, C. J., Williams-Ceci, S. C., & Ceci, S. J. (2022). Assessing gender Bias in Particle Physics and Social Science Recommendations for Academic Jobs. *Social Sciences*, *11*(2), 74.
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.
Bohnet, I. (2016). *What works: Gender equality by design*. Harvard University Press.
Bohnet, I., Van Geen, A., & Bazerman, M. (2016). When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, *62*(5), 1225–1234.
Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield.
Castleman, B. L., & Page, L. C. (2014). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *Journal of Human Resources*, *51*(2), 389–415.
Chen, C. Y., Kahanamoku, S. S., Tripati, A., Alegado, R. A., Morris, V. R., Andrade, K., & Hosbey, J. (2022). Decades of systemic racial disparities in funding rates at the National Science Foundation. https://doi.org/10.31219/osf.io/xb57u.
Cheryan, S., & Markus, H. R. (2020). Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review*, *127*(6), 1022–1052. https://doi.org/10.1037/rev0000209.
Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd Edition). Sage.
Culpepper, D. (2021). We have a partner hire situation: The personal and professional lives of dual-career academic couples. [Unpublished doctoral dissertation]. University of Maryland.
Culpepper, D., White-Lewis, D., O'Meara, K., Templeton, L., & Anderson, J. (2023). Do rubrics live up to their promise? Examining how faculty search committees use rubrics in candidate evaluation and selection. *The Journal of Higher Education*, *94*(7), 823–850.
Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313–342.
Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How gender and race stereotypes impact the advancement of scholars in STEM: Professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, *82*(3), 127–141.
Field, J. (2015). Improving student performance using nudge analytics. Proceedings of the 8th *International Conference on Educational Data Mining* https://files.eric.ed.gov/fulltext/ED560905.pdf.
Fine, E., & Handelsman, J. (2012). Searching for excellence and diversity: A guide for search committees (2nd Ed.). WISELI. https://wiseli.wisc.edu/wp-content/uploads/sites/662/2018/11/SearchBook_Wisc.pdf.
Fine, E., Sheridan, J., Carnes, M., Handelsman, J., Pribbenow, C., Savoy, J., & Wendt, A. (2014). Minimizing the influence of gender bias on the faculty search process. In V. Demos, C. W. Berheid, & M. T. Segal (Eds.), Gender research: Gender transformation in the academy (Vol. 19, pp. 267–289). Emerald Insight. https://doi.org/10.1108/S1529-212620140000019012.
Ginther, D. K., Shaffer, W. T., Schnell, J., Masimore, B., Liu, F., Hakk, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science*, *333*(6045), 1015–1019.
Greenwald, A. G., & Lai, C. K. (2020). Implicit social cognition. *Annual Review of Psychology*, *71*, 419–445.

Harrison, D. A., Kravitz, D. A., Mayer, D. M., Leslie, L. M., & Lev-Arey, D. (2006). Understanding attitudes toward affirmative action programs in employment: Summary and Meta-Analysis of 35 years of research. *Journal of Applied Psychology*, *91*(5), 1013–1036.

Hoppe, T. A., Litovitz, A., Willis, K. A., Meseroll, R. A., Perkins, M. J., Hutchins, A., Davis, A. F., Lauer, M. S., Valantine, H. A., Anderson, J. M., & Santangelo, G. M. (2019). Topic choice contributes to the lower rate of NIH awards to African-American/Black scientists. *Science Advances*, *5*(10), 1–12.

Hsieh, H., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Jaschik, S. (2017). White perceptions of affirmative action. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/admissions/article/2017/10/30/survey-draws-attention-white-perceptions-affirmative-action.

Jimenez, M. F., Laverty, T. M., Bombaci, S. P., Walkins, K., Bennett, D. E., & Pejchar, L. (2019). Underrepresented faculty play a disproportionate role in advancing diversity and inclusion. *Nature Ecology & Evolution*, *3*, 1030–1033.

Johnson, S. K., Hekman, D. R., & Chan, E. T. (2016). If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, 26(04).

Kachchaf, R., Ko, L., Hodari, A., & Ong, M. (2015). Career–life balance for women of color: Experiences in science and engineering academia. *Journal of Diversity in Higher Education*, *8*(3), 175–191.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kazmi, M., Spitzmueller, C., Yu, J., Madera, J. M., Tsao, A. S., Dawson, J. F., & Pavlidis, I. (2021). Search committee diversity and applicant pool representation of women and underrepresented minorities: A quasi-experimental field study. *Journal of Applied Psychology*, *107*(8), 1414–1427.

Liera, R. (2020). Equity advocates using equity-mindedness to interrupt faculty hiring's racial structure. *Teachers College Record*, *122*(9), 1–42.

Liera, R., & Hernandez, T. E. (2021). Color-evasive racism in the final stage of faculty searches: Examining search committee hiring practices that jeopardize racial equity policy. *The Review of Higher Education*, *45*(2), 181–209.

Lubienski, S. T., Miller, E. K., & Saclarides, E. S. (2018). Sex differences in doctoral student publication rates. *Educational Researcher, 47*(1), 76–81.

Luke, K., & Grosche, M. (2018). What do I think about inclusive education? It depends on who is asking. Experimental evidence for a social desirability bias in attitudes towards inclusion. *International Journal of Inclusive Educationm*, *22*(1), 38–53.

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2019). Raising doubt in letters of recommendation for academia: Gender differences and their impact. *Journal of Business and Psychology*, *34*(3), 287–303. https://doi.org/10.1007/s10869-018-9541-1.

Matthews, K. (2020, September 24). *Why is diversity in engineering a major opportunity?* The American Society of Mechanical Engineers [ASME]. https://www.asme.org/topics-resources/content/why-is-diversity-in-engineering-a-major-opportunity.

Mendoza-Denton, R., Patt, C., Fisher, A., Eppig, A., Young, I., Smith, A., & Richards, M. A. (2017). Differences in STEM doctoral publication by ethnicity, gender and academic field at a large public research university. *PLOS ONE, 12*(4): e0174296.

Milkman, K. (2021). *How to change: The science of getting from where you are to where you want to be*. Penguin.

Mitchneck, B. (2020). *Synthesizing research on gender biases and intersectionality in citation analyses and practices*. ADVANCE Resource and Coordination (ARC) Network.

Moody, J. (2012). *Faculty diversity: Removing the barriers*. Routledge.

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, 11(8), e12331.

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

Norris, P., & Epstein, S. (2011). An experiential thinking style: Its facets and relations with objective and subjective criterion measures. *Journal of Personality*, *79*(5), 1043–1080.

O'Meara, K., Kuvaeva, A., & Nyung, G. (2017). Constrained choices: A view of campus service inequality from annual faculty reports. *The Journal of Higher Education*, *88*(5), 672–700. https://doi.org/10.1080/0022.

O'Meara, K., Culpepper, D., & Templeton, L. L. (2020). Nudging toward diversity: Applying behavioral design to faculty hiring. *Review of Educational Research*, *90*(3), 311–348.

O'Meara, K., Culpepper, D., Lennartz, C., & Braxton, J. (2022). Leveraging nudges to improve the academic workplace: Challenges and possibilities. *Higher education: Handbook of theory and research: Volume 37* (pp. 277–346). Springer.

O'Meara, K., Templeton, L., White-Lewis, D., Culpepper, D., & Anderson, J. (2023). The safest bet: Identifying and assessing risk in faculty selection. *American Educational Research Journal*, *60*(2), 330–366.
Patrida, D. (2022, August 11). *Promoting diversity and inclusion in robotics education*. RobotLab. https://www.robotlab.com/blog/promoting-diversity-and-inclusion-in-robotics-education.
Pearlman, L., Combs, S., Murray, J., & Bunim, M. E. (2000–2009). Making the band [TV series]. Bunim/Murray Productions; Trans Continential Pictures; Bad Boy Films; The Ted & Perry Company; MTV Series Entertainment.
Perez, A. B. (2022, January 27). A racial reckoning for college access in America. Forbes. https://www.forbes.com/sites/angelperez/2022/01/27/a-racial-reckoning-for-college-access-in-america/?sh=4b9c50ba5f8c.
Posselt, J., Hernandez, T. E., Villarreal, C. D., Rodgers, A. J., & Irwin, L. N. (2020). Evaluation and decision making in higher education: Toward equitable repertoires of faculty practice. *Higher Education: Handbook of Theory and Research: Volume*, *35*, 1–63.
Ray, V. (2019). A theory of racialized organizations. *American Sociological Review*, *84*(1), 26–53. https://doi.org/10.1177/0003122418822335.
Rivera, L. A. (2017). When two bodies are (not) a problem: Gender and relationship status discrimination in academic hiring. *American Sociological Review*, *82*(6), 1111–1138.
Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd ed.).). Sage.
Settles, I. H., Jones, M. K., Buchanan, N. T., & Dotson, K. (2020). Epistemic exclusion: Scholar (ly) devaluation that marginalizes faculty of color. *Journal of Diversity in Higher Education*, *14*(4), 493–507.
Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences*, *111*(28), 10107–10112.
Shoben, E. (1997). From antinepotism rules to programs for partners. In M. Ferber, & J. W. Loeb (Eds.), *Academic couples: Problems and practices* (pp. 226–247). Board of Trustees of the University of Illinois.
Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*(7), 509–528.
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
Tzioumis, K. (2018). Demographic aspects of first names. *Scientific Data*, *5*(180025), 1–9. https://doi.org/10.1038/sdata.2018.25.
University of California, Los Angeles [UCLA]. (2024). *Life sciences: The mentor professor initiative*. https://equity.ucla.edu/initiatives/life-sciences-the-mentor-professor-initiative/
Weeden, K. A., Thébaud, S., & Gelbgiser, D. (2017). Degrees of difference: Gender segregation of US doctorates by field and program prestige. *Sociological Science*, *4*(6), 123–150.
White-Lewis, D. (2019). The facade of fit and preponderance of power in faculty search processes: Facilitators and inhibitors of diversity. [Unpublished doctoral dissertation]. University of California, Los Angeles.
White-Lewis, D. (2020). The facade of fit in faculty search processes. *The Journal of Higher Education*, *91*(6), 833–857.
White-Lewis, D., Bennett, J., & Redd, K. (2022). *Setting a national agenda for systemic reform in postsecondary faculty careers*. Association of Public and Land-Grant Universities (APLU).
Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. Proceedings of the National Academy of Sciences, 112(17), 5360–5365.