The Journal of Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uhej20

Do Rubrics Live up to Their Promise? Examining How Rubrics Mitigate Bias in Faculty Hiring

Dawn Culpepper, Damani White-Lewis, KerryAnn O'Meara, Lindsey Templeton & Julia Anderson

To cite this article: Dawn Culpepper, Damani White-Lewis, KerryAnn O'Meara, Lindsey Templeton & Julia Anderson (2023): Do Rubrics Live up to Their Promise? Examining How Rubrics Mitigate Bias in Faculty Hiring, The Journal of Higher Education, DOI: 10.1080/00221546.2023.2168411

To link to this article: https://doi.org/10.1080/00221546.2023.2168411

Published online: 27 Jan 2023.

THE JOURNAL OF HIGHER EDUCATION
https://doi.org/10.1080/00221546.2023.2168411

Routledge
Taylor & Francis Group

Do Rubrics Live up to Their Promise? Examining How Rubrics Mitigate Bias in Faculty Hiring

Dawn Culpepper[a], Damani White-Lewis[b,c], KerryAnn O'Meara[b], Lindsey Templeton[d], and Julia Anderson[e]

[a]ADVANCE Program for Inclusive Excellence, University of Maryland College Park, College Park MD, USA; [b]College of Education, University of Maryland College Park, College Park MD, USA; [c]Graduate School of Education, University of Pennsylvania, Philadelphia, Pennsylvania, USA; College of Education, University of Maryland College Park, College Park MD, USA; [d]Higher Education Resource Services, Denver, Colorado, USA; [e]Higher, Adult, and Lifelong Education, Michigan State University, East Lansing, Michigan, USA

ABSTRACT

Many colleges and universities now require faculty search committees to use rubrics when evaluating faculty job candidates, as proponents believe these "decision-support tools" can reduce the impact of bias in candidate evaluation. That is, rubrics are intended to ensure that candidates are evaluated more fairly, which is then thought to contribute to the enhanced hiring of candidates from minoritized groups. However, there is scant — and even contradictory — evidence to support this claim. This study used a multiple case study methodology to explore how five faculty search committees used rubrics in candidate evaluation, and the extent to which using a rubric seemed to perpetuate or mitigate bias in committee decision-making. Results showed that the use of rubrics can improve searches by clarifying criteria, encouraging criteria use in evaluation, calibrating the application of criteria to evidence, and in some cases, bringing diversity, equity, and inclusion work (DEI) into consideration. However, search committees also created and implemented rubrics in ways that seem to perpetuate bias, undermine effectiveness, and potentially contribute to the hiring of fewer minoritized candidates. We conclude by providing stakeholders with practical recommendations on using rubrics and actualizing DEI in faculty hiring.

ARTICLE HISTORY
Received 28 January 2022
Accepted 1 December 2022

KEYWORDS
Higher education; faculty hiring; rubric; evaluation; equity

Many U.S. colleges and universities require faculty search committees to use rubrics to evaluate candidates. A rubric is a "decision-support tool" that contains quantitative and/or qualitative items imbued with job- and/or organizational-level criteria that are uniformly applied to evaluate candidates (Isaac et al., 2009; Uhlmann & Cohen, 2005). The underlying premise is that rubrics force search committees to apply evaluation criteria more fairly and effectively to candidates, as compared to relying upon their instinctual and biased evaluations (Moody, 2012).

CONTACT Dawn Culpepper dkculpep@umd.edu
Dawn Culpepper and Damani White-Lewis are co-first authors on this article.

Although the benefits of using rubrics seem obvious, limited evidence substantiates this premise. Some studies show positive results from applying rubrics in faculty hiring (e.g., Blair-Loy et al., 2022). Yet, in these studies rubrics are used in concert with other inclusive hiring strategies. Rubrics can breed implicit biases under certain conditions, especially if they include purportedly neutral criteria evaluated through subjective lenses of merit, quality, or promise (White-Lewis, 2020; Uhlmann & Cohen, 2005). How rubrics are used (e.g., flexibly or rigidly) similarly shapes evaluation effectiveness (White-Lewis, 2020). The diffusion of rubrics over the last decade suggests that many institutions view them as attainable, low-cost solutions to making more effective and less biased hiring decisions. However, we lack strong empirical evidence to support their efficacy or a comprehensive understanding of their impact on hiring practices.

Lack of concrete evidence is important because progress on improving the representation and meaningful inclusion of minoritized scholars (i.e., Black, Latinx, and Indigenous scholars spanning different gender identities; white women in some STEM fields; and Asian and international faculty members who experience bias and discrimination) in tenure-track roles and academic leadership remains problematic (National Science Foundation [NSF] 2021). Identifying just how and why — if at all — rubrics lead to better decision-making processes and outcomes for hiring minoritized groups is therefore a compelling need. Thus, the goal of this study was to examine how faculty search committees used rubrics in candidate evaluation and examine the extent to which rubrics seemed to perpetuate or mitigate bias in committee decision-making.

Literature review: Faculty diversity and faculty hiring

Numerous researchers document the ways that systemic racism and sexism undermine efforts to recruit, hire, advance, and retain faculty from minoritized groups. From bias in graduate admissions and mentoring (Posselt, 2016); microaggressions and discrimination (Harris et al., 2021); invisible and uncredited labor (O'Meara et al., 2021; Misra et al., 2021), and devaluation of scholarship (Settles et al., 2020), the empirical evidence is conclusive: academe's structures, systems, and processes are ill-suited to promote a more diverse faculty (Griffin, 2020).

Many institutions focus attention on strategies for enhancing the recruitment of faculty from minoritized groups, including pipeline programs (Culpepper et al., 2021) and target of opportunity hiring initiatives (Muñoz et al., 2017). Other efforts focus on interventions related to faculty search committees, with some evidence that equity coaches who embed with search committees (Liera, 2020); implicit bias trainings for faculty search committees (Devine et al., 2017); removing candidate names from application materials

(Langin, 2021); candidate diversity statements (Schmaling et al., 2015) can help reduce bias and increase the hiring of minoritized candidates. Still, our understanding of how, why, and in what contexts these practices matter is limited.

Within this broader suite of interventions, rubrics are often suggested as a strategy to enhance fairness in evaluation (Liera, 2020; Liera & Ching, 2019; O'Meara et al., 2020). Advocates suggest that using rubrics facilitates more accurate employment decisions, since committees will identify and apply job-specific criteria to each candidate (White-Lewis, 2020). Rubrics also force individuals to rationalize why, based on the criteria, certain candidates are more qualified than others (O'Meara et al., 2020). Rubrics may assist committee members in evenly applying the same criteria to each candidate based on their application materials, thus yielding more consistent judgments with less variance between evaluators (White-Lewis, 2020). Search committees that use rubrics may be better able to assess each candidate's qualifications in different domains (Moody, 2012). For instance, committees that incorporate experience with diversity, equity, and inclusion (DEI) into the rubric separately from teaching, service, and research may be able to elevate the contributions of minoritized faculty, who disproportionately bring DEI knowledge and skills to their workplaces (Liera, 2020; Liera & Ching, 2019). Rubrics can therefore bring visibility or "framing" to different candidate qualifications that might be obscured if the committee instead based their evaluation on assessments of a candidate's general hireability (O'Meara et al., 2022).

Two recent studies provide more insight into how rubrics enhance fairness and reduce bias in faculty evaluation. Rivera and Tilcsik (2019) used quasi-experimental methods to understand how different rating scales in a rubric affected the presence of gender bias using teaching evaluations from one research institution. They found that when faculty instructors' teaching performance were rated on a six-point scale (as compared to a 10-point scale), faculty members were evaluated more consistently across genders. The researchers concluded that the six-point scale was less likely to activate gender stereotypes about excellence and brilliance that typically thwart women from being evaluated fairly. Similarly, Blair-Loy et al. (2022) compared the hiring outcomes of one engineering department before and after the department's search committees started using a rubric, finding that more women were hired after implementation. However, even with rubrics, gender bias persisted when evaluating candidates on criteria like research productivity and impact, though negative scores in those domains were offset by women's higher average scores on criteria related to contributions to diversity. As such, Blair-Loy et al. (2022) argued that "rubric usage be accompanied by strategic application in departmental meetings to counteract individual bias and check interactional bias" (p. 37). Such studies show that the structure (i.e., rating scale) and content (i.e., criteria) of rubrics shape candidate evaluation, but give little insight into how

search committees use rubrics in evaluation and the process by which evaluations become less biased. Moreover, both studies looked specifically at gender bias, which may operate differently than racial bias. Overall, there is an assumption that committees that use rubrics will be more accurate, consistent, and fair, and thus candidates from minoritized groups will be evaluated with less bias and hired more frequently. Yet, we lack empirical evidence that using a rubric accomplishes these goals, consistently, and/or across contexts.

Conceptual framework: Hiring biases and nudges

We draw from concepts from behavioral economics and social psychology to guide this study. Research from these fields demonstrate how our automatic cognitive processes and cultural socialization primes us to think and act in irrational and suboptimal ways. We define cognitive bias as the systematic yet often flawed ways individuals make decisions, based on heuristics, or mental short-cuts (2011). Additionally, we define social bias as the ways in which social role expectations and norms, and stereotypes shape our perceptions of individuals from different social groups (e.g., by race, gender, class) in ways that are unfair and/or prejudicial (Banaji & Greenwald, 2016).

Both kinds of bias have been observed in faculty hiring. Racial and gender stereotypes about competence and hireability disadvantage women and racially minoritized scholars, and especially women of color (Beattie et al., 2013; Eaton et al., 2020). Search committee members' perceptions about the moveability of women candidates with partners represent another form of gender bias (Rivera, 2017), while racial bias also shapes search committee perceptions of the quality or merit of scholarship of racially minoritized candidates (Settles et al., 2020). Cognitive biases also play a role in faculty hiring decisions. Faculty members may use the prestige of a candidate's doctoral institution as a proxy for quality (Posselt, 2016; Posselt et al., 2020), or view candidates who do research that is more familiar to them more favorably (Moody, 2012). All said, faculty members, like all humans, use heuristics to make assessments about candidates, and often bias influences these heuristics.

Bias is magnified in certain conditions. When evaluators are rushed in deciding or must sift through hundreds of applications, they are more likely to use their instinctual reactions to make decisions (O'Meara et al., 2020; Posselt et al., 2020). When the criteria for evaluation are ambiguous (e.g., use "fit" without clearly defining it), bias is also more likely to play a role (White-Lewis, 2020). Many of these conditions exist in faculty hiring (Moody, 2012; O'Meara et al., 2020), and thus, these contexts are primed for bias and less effective decision-making.

Bias can be mitigated with the introduction of certain interventions, including those which behavioral economists term "nudges" (Thaler & Sunstein,

2008). Nudges are small changes to the context in which decisions are made, intended to prompt the decision-maker to render a more accurate decision while still allowing the decision-maker to go their own way (Thaler & Sunstein, 2008). Decision-support tools like rubrics potentially nudge an evaluator's decision-making in a few ways. Rubrics may invoke more active decision-making, requiring evaluators to input criteria, make assessments about which criteria matter, rate all candidates against the criteria, and provide a rationale for their score (Damgaard & Nielsen, 2018; Isaac et al., 2009), thereby slowing down the decision-maker and allowing them to make a better decision (2011). Rubrics add structure to the decision-making process, forcing evaluators to focus on the salient aspects of the candidate's application and ignore aspects not related to the criteria (Damgaard & Nielsen, 2018; Isaac et al., 2009). In some cases, rubrics may alter the decision-making context altogether, by allowing for systematic comparisons across candidates along various dimensions where before there were none (Isaac et al., 2009).

Reasons to be cautious about the efficacy of a rubric to mitigate the effects of bias in hiring exist. Committees may create rubrics that do not contain criteria related to the qualifications they seek or may create criteria that are not reflective of future performance (Sheppard et al., 2011). Evaluators can interpret criterion differently or may not understand them, leading to inconsistent assessments (Goldhaber et al., 2017). A rubric may also contain criteria that are intrinsically biased. For instance, because minoritized scholars receive less federal grant funding (Chen et al., 2022) and are cited less frequently (Mitchneck, 2020), an evaluator who uses rubric containing a criterion like "experience with grants" or "h-index" could make an accurate assessment, but accuracy in this case would likely advantage white and men candidates.

Research shows conditions wherein nudges are more likely to fail. A decision-maker may reject a nudge if it subverts social or professional norms, personal preferences, or are viewed to infringe too much on autonomy (Sunstein, 2021). Given that departments are typically given autonomy in hiring decisions, rubrics may be viewed as overly prescriptive and therefore resisted by search committees (O'Meara et al., 2022). Nudges typically address simple binary problems (Sunstein, 2021). Faculty hiring is complex and multi-faceted, with many potential outcomes. A single nudge, such as rubric, may not have a large impact on the results. Effective nudges require an accurate diagnosis of the context(s) that hold an individual back from making the better choice (Sunstein, 2021). Such context(s) could include bias(es), but also aspects of the process surrounding the decision, such as inadequate information or cumbersome bureaucratic procedures (Sunstein, 2021). Given the idiosyncratic nature of hiring (White-Lewis, 2020), it is difficult to assess with any certainty the specific kinds of biases that may emerge within any specific hiring scenario.

Nudges have also been widely critiqued. Many scholars are concerned with the philosophical and ethical underpinnings of nudges, arguing that such interventions, particularly when deployed by governments, limit free choice and constrain civil liberties (Glod, 2015). Some assert that nudge creators make normative assumptions about what constitutes a "good" or "better" outcome without a recognition of their own biases or motivations, leading to ethical concerns about the direction(s) individuals are nudged toward (Brown, 2012), particularly when there is a lack of transparency and accountability about when nudges are being used (Baldwin et al., 2011). There are also more practical concerns about nudges. Some scholars suggest that nudge interventions have been ill-defined (Kosters & Van der Heijden, 2015), leading to uncertainty about their impact, while other researchers suggest that nudges have been limited in their effectiveness in producing desired outcomes in the short and long-term (e.g., Hummel & Maedche, 2019; Kosters & Van der Heijden, 2015).

While we recognize these critiques of nudges, this study's design and purpose engaged with nudges in a way that allowed us to navigate some of these issues. As discussed below, our study was not an experiment where we compared search committees that used rubrics to those that did not; each committee had adopted a rubric on their own, without prompting from our research team, with the goal of adding some kind of structure to their decision-making process. As such, we were interested in observing, using qualitative methods, if and how a rubric might produce some of the effects of a nudge (e.g., invoke active decision-making), in essence studying the context in which rubrics were used and link their use to reduced bias (and therefore more diverse hiring), rather than trying to make a quantitative assessment of whether search committees that used rubrics produced "better" outcomes than those that did not. We return to these issues in our discussion, but felt it was important to address these concerns from the outset.

Methodology

Epistemological approach and positionality

Our approach to this study was constructivist, assuming that decision-makers construct multiple realities, and act on those constructions throughout the faculty hiring process. Our team, comprised of a Black, cis-gender man, a Black, biracial cis-gender woman; a Multiracial (white and Asian) cis-gender woman, and two white cis-gender women, bring different personal and professional experiences with DEI in the academy. As a group with representation of professors, practitioners, and a doctoral student, we have worked to mitigate the role of bias and enhance DEI in academia through the use of various strategies, rubrics among them. These perspectives motivate our

study; that is, of wanting to interrogate hiring processes using empirical evidence and identify strategies to enhance hiring procedures in academe. Though a (post-)positivist lens would assume that we are therefore biased and require us to bracket these experiences, this is not necessary through a constructivist approach. In fact, our personal experiences strengthen the work: we have seen rubrics used in detrimental and self-serving ways that undermine DEI. Thus, we do not believe that they are a panacea. But given the multitude of ways to use rubrics, we sought to understand how they were used in ways that facilitate and/or inhibit DEI in faculty hiring, which requires a level of expert discernment our team brings to the work.

Multiple case study

We utilized multiple case methods in this study. The multiple case study method is a variation of the case study tradition that situates multiple "bounded" cases within their real-life contexts to understand their uniqueness and interplay (Stake, 2005; Yin, 2018). In this study, we examined five search committees (the cases), each of which was bounded by its own disciplinary and institutional context (See Table 1 for a description of each case). Multiple case study was an appropriate method for several reasons. This method is concerned with "how" and "why" questions, especially concerning real-life processes whose parts cannot be narrowly parsed-out and/or manipulated (Yin, 2018). Multiple case studies examine situations where several cases experience the same phenomenon, allowing researchers to interrogate how certain phenomena manifest within and across different contexts, which provides more analytic depth and stability of findings (Stake, 2005; Yin, 2018).

There are three approaches to case studies (Yazan, 2015). Whereas Yin (2018) is considered more positivist, Stake (2005) aligns more closely with constructivism, while Merriam and Tisdell (2016) operate between them yet still closer to constructivism. These differences are important given our own epistemologies. In terms of case study design, collection, and analysis, we

Table 1. Case descriptions.

Committee	# of Observations	Demographic of Candidate(s) Forwarded to Hiring Official
Microbiology, Southern State University (Teaching Intensive)	7 (3 committee meetings, 3 job talks, 1 department meeting)	1 Latino Man
Chemical Engineering, University of Redwood (Research Intensive)	14 (9 committee meetings, 4 job talks, 1 department meeting)	2 White Women
Environmental Engineering, University of Redwood (Research Intensive)	7 (5 committee meetings, 2 job talks)	Search Failed
Developmental Psychology, Hudson University (Research Intensive)	19 (11 committee meetings, 5 job talks, 3 department meetings)	White Woman, Asian Woman*
Plant Biology, Hammond University (Teaching Intensive)	17 (12 committee meetings, 4 job talks, 1 department meeting)	1 White Man

opted for a blended approach, given that scholars can "either choose to utilize the tools offered by one methodologist or construct an amalgam of tools from two or three of them" (Yazan, 2015, p. 135). For the initial case study design we were primarily informed by Yin's (2018) work, since it creates an instructive roadmap for how to design case studies and create comparison matrices. However, we departed from Yin as we approached data collection, and followed the work of Merriam and Tisdell (2016) more closely given their emphasis on constructivism and more defined roadmap as compared to Stake (2005). In what follows, we describe how the different approaches synergized and yielded a data collection and analysis strategy that connected our research questions and extant literature to findings and discussion.

Case selection and sample

We recruited from a group of universities involved in a multi-institutional effort to enhance faculty diversity. Each institution had a general espoused concern for increasing faculty diversity, as indicated in their participation in the initiative, but different policies and practices related to faculty diversity. Academic leaders involved with the initiative helped our research team identify departments within their respective universities that had been authorized to do a faculty search. Our research team then met with department and/or search chairs to explain the goals of the study, gain access to their committee meetings, and assuage any concerns of search interference and emphasize our role as passive observers. Next, all search committee members completed an IRB-approved informed consent form. To participate, the search committee needed to have an open, faculty search for a tenured or tenure-track faculty member during the two-year period in which we collected data. It is important to emphasize that the searches were *not* involved with the institutional-level initiative (which was concerned with pre-professoriate diversity). Thus, these searches were still considered "typical" cases (Yin, 2018), resembling searches that might occur elsewhere throughout the U.S., which helped increase the "transferability" of findings at similar institutional types (Merriam & Tisdell, 2016). Overall, our recruitment strategy was partially informed by convenience (i.e., composed of committees to which we had access through the diversity initiative, met broad criteria, and whose departmental leaders agreed to participate) but also partially purposeful (i.e., open searches for tenure and tenure-track positions in universities).

Multiple case studies are strongest when the study contains at least four cases (Stake, 2005); our team recruited five search committees in four different universities to participate in the study (See Table 1). Our specific design was that of a *multiple holistic design*, which emphasizes attention to the whole case as opposed to differentiating between embedded units within each case (Yin, 2018). The departments were two engineering departments (Chemical and

Environmental) that came from a large, four-year research-intensive institution, a psychology department located at another doctoral-granting university that had less research infrastructure and greater faculty teaching loads (Developmental Psychology search), a biology department at a predominantly teaching university (Microbiology search), and another biology department at a different predominantly teaching university (Plant Biology search). To ensure confidentiality, we do not provide the actual subfield but swap them for peripheral subfields, and all participant and institution names are pseudonyms. Each search committee used a rubric at some point during their search. However, the criteria, scoring strategy, application, and prior experience with rubrics varied substantially (Table 2 contains a brief description of each rubric and the process by which each committee applied their rubric to candidates).

Data collection

Our data sources included observations, document analysis, and interviews with department and search committee chairs. Though this aligns with Yin's (2018) emphasis on collecting multiple sources of data to triangulate findings, we did so not as a means to confirm a single truth, but to create robust case narratives that revealed different angles of the faculty hiring process (Merriam & Tisdell, 2016). With some notable exceptions (e.g., Liera, 2020; Rivera, 2017), researchers have not used observations of search committees to understand faculty hiring processes. Our study's use of observations makes a unique contribution in that we observed firsthand how different departmental personnel shape faculty hiring.

We engaged with search committees over the course of their search, which typically launched in the fall semester and concluded in the spring. One researcher collected observational data of search meetings, using an observational protocol each meeting. Over the course of the observation, the researcher sat away from the committee and remained quiet to not interfere with search processes or serve as a reminder of their presence. Observations and interviews took place in-person and over Zoom (Table 1). Alongside observational data, we collected documents that were relevant to the search, such as (1) position descriptions, (2) rubrics used for the general pool, (3) rubrics used for phone interviews, and (4) e-mail correspondences as appropriate. We conducted five interviews with search chairs to recall significant events, clarify discipline-specific norms, share preliminary themes, and confirm who was ultimately hired after departmental votes.

Table 2. Rubric descriptions.

Committee	Summary of Rubric*	Summary of Search Committee Process for Using Rubric
Microbiology, Southern State University (Teaching Intensive)	(1) Minimum qualifications (1–4) (2) Collaborative skills (1–4) (3) Demonstrated teaching effectiveness (1–4) (4) Evidence of scholarly potential (5) Strong foundation in microbiology (1–4) (6) Potential to develop a research program that involves students (1–4) (7) Teaching experience (1–4) (8) Postdoctoral experience (1–4) (9) Grant experience (1–4) (10) Experience with evidence-based teaching practices (1–4) (11) Experience working with diverse students (1–4) (12) Potential to contribute to diversity and recruitment (1–4) *This rubric also contained 3 criteria related to analysis. Each candidate was given an overall score (maximum score of 60).*	Search committee used the rubric to evaluate all candidates. Two search committee members reviewed and provided scores for an assigned number of candidates. The committee used the overall candidate scores to determine the candidates who advanced to the short-list.
Chemical Engineering, University of Redwood (Research Intensive)	Research (1–5) (1) Productivity: Large number of recent first author papers and large number of contributing author papers (2) Impact: Several papers in high impact journals and large number of citations. The candidate has also received research awards (3) Funding: Received a graduate or post-doctoral fellowship, or has an agency-funded grant award or other career transition grant (4) Vision: Clear, innovative research direction. Teaching (1–5) (1) Demonstrated use of evidence-based teaching. (2) Teaching/mentoring experience in classroom and lab Institutional stewardship (1–5) (1) Leadership and service experience (2) Commitment to diversity and inclusion Fit (1–5) (1) Fit and synergies with existing departmental research strengths (2) Ability to teach a large variety of courses *Each area was equally weighted with maximum possible points of 20.*	Search committee used the rubric to evaluate all candidates. Three committee members reviewed and provided scores for an assigned number of candidates. The committee deliberated all candidates who scored 14 points or higher or if they received a 5 in any of the four categories.

(*Continued*)

Table 2. (Continued).

Committee	Summary of Rubric*	Summary of Search Committee Process for Using Rubric
Environmental Engineering, University of Redwood (Research Intensive)	(1) Minimum Qualifications (0–1) (2) Applied before best consideration date (0–1) (3) Research Field Fit (0–2) (4) Research Productivity (0–3) (5) Research Impact (0–3) (6) Teaching (0–2) (7) Mentoring (0–2) (8) Service (0–1) (9) Promise (0–2) (10) Other (0–1) *Each candidate received an overall score from (maximum score 18).*	Search committee used the rubric to evaluate all candidates. All committee members reviewed all candidates. The committee calculated the average score of each candidate and the candidates that received the absolute highest scores and who the committee members were most excited about were brought in for immediate final interviews, while the committee invited candidates who received high (but not the absolute highest) were invited for screening interviews.
Developmental Psychology, Hudson University (Research Intensive)	Research (1–5) (1) Program of research (2) Potential for attracting funding Teaching (1–5) (1) Potential to teach students within the department (2) Commitment to social justice *Each candidate received an overall average (Maximum score of 5).*	Search committee used the rubric to evaluate all candidates. Two search committee members reviewed each candidate's application materials and provided scores based on the rubric. The committee reviewed in more depth candidates who received average scores of 3.5 or higher and the committee additionally reviewed all minoritized candidates regardless of score to generate their short-list of candidates.
Plant Biology, Hammond University (Teaching Intensive)	Teaching (1) Recent teaching experience (1–10) (2) Ability to teach courses (1–10) (3) Potential for contributing novel courses (1–10) Research (1) Recent research productivity (1–10) (2) Experience in research mentoring (1–10) (3) Recent grant activities (1–10) *Candidates were given a score of 1–10 for each criterion (Maximum score of 60).*	Search committee used a rubric to evaluate all candidates. All committee members reviewed all candidates. Rubric scores helped committees create a personal ranking of each committee member's top 5 candidates; candidates who had multiple committee members ranked in their top 5 were then invited to the interview round.

*Criteria have been summarized to enhance confidentiality.

Data analysis

Yin (2018) provides three foundations of data gathering: gathering multiple sources of evidence, creating a case study database, and developing a chain of evidence. As described above, we collected multiple sources of evidence, but not for triangulation purposes. We also created a case study database for each committee which was practically useful for organizing our data. But we did not create a chain of evidence, which was more positivist than our intended approach. After the data were organized, we engaged in several analytic procedures with multiple research team members to analyze layers of interlocking data. In lieu of a chain of evidence that begets linear propositions akin to hypothesis testing (Yin, 2018), we instead used the conceptual framework of bias and nudges as a guide to create broad categories to parse through the data

(Merriam & Tisdell, 2016; Saldaña, 2016). This conceptual coding strategy resulted in categories such as kinds of criteria included, or ambiguity about criteria and scoring inconsistencies. However, our coding procedure was not exclusively static — we intentionally left our coding procedure open enough to leave space for emergent themes (Saldaña, 2016). This open strategy resulted in codes such as weighting strategies and diverging from the rubric criteria. Multiple research team members read through transcripts, observation notes, and reviewed documents with this deductive and inductive approach, coding and re-coding the data until we reached a final codebook.

Next, we created single case studies for each committee. Each case was analyzed using the constant-comparative method, comparing within-group clusters of coded data to bring about greater complexity (Merriam & Tisdell, 2016). This yielded more nuanced themes within theoretical categories, such as how differences in rubric scoring between committee members were reconciled in real-time between different committee members, or how the rubric criteria seemed to influence the evaluative results. Once theoretical themes had more nuance from single-case analyses, we then conducted cross-case analysis (Merriam & Tisdell, 2016), to examine the similarities and differences between themes across cases. This analytic process yielded themes that spoke to the ways in which bias seemed to be at times perpetuated and other times mitigated by rubrics across all five faculty searches.

Data trustworthiness

Several aspects of our data collection procedure strengthened the study's trustworthiness. Conducting interviews with search committee chairs served as a member-checking tool, but also helped consider rival explanations, wherein we considered alternative interpretation(s) of our preliminary results (Yin, 2018). We also created robust case narratives using multiple sources of data to generate findings (Merriam & Tisdell, 2016). For example, we reviewed the rubrics committee used and compared those to how the criteria were discussed to assess gaps between the stated criteria and the implicit criteria.

Limitations

In this study we were concerned with the social construction of reality. That is, we observed how faculty search committee members assigned racial and gender categories to candidates based on their own assembly of cues and markers available in candidate files. For reasons of confidentiality, we did not receive the demographic information of applicants. Moreover, even search committee members did not even have systemic data on candidate identity, as that would violate the institutions' nondiscrimination policies. As such, we do not purport to fully understand the myriad of social identities held by

candidates, and we cannot systematically connect rubric scores to candidates with certain identities. Instead, we report on how faculty applied rubrics to candidates they believed held certain identities based on their own social constructions and understandings of race and gender. What aided our understanding of how faculty understood candidate identity were their discussions of certain pieces of evidence and cues (e.g., candidate self-disclosure, professional affiliation with an identity-based organization, first and/or last names, and personal knowledge) to infer candidate identity. Furthermore, we did not always have access to the actual rubric scores committee members gave to candidates. In some cases, we received these scores after the meeting and in some cases these scores were not shared, which somewhat limited our ability to link candidate scores to the specific candidates under discussion during our observations.

Findings

This study's findings focus on themes that emerged across cases. We provide a description of how search committees broadly developed and used their rubrics, organized into the themes of: defining the criteria; scoring candidates; assigning weights to criteria; counting diversity, equity, and inclusion (DEI) experience; and the declining utility of rubrics. Within each theme, we consider examples of how committees used rubrics to mitigate evaluation bias and counterexamples of rubric use that seemed to perpetuate bias.

Defining the criteria

The first step in candidate evaluation was creating and defining the criteria that would be used in the rubric to evaluate candidate files. All the committees in our study included criteria related to the domains of research, teaching, and service (Table 2). However, the number and kinds of criterion committees used varied by discipline/field as well as institutional type, with the rubrics for Microbiology and Plant Biology committees including more teaching-focused criteria and the Chemical Engineering, Environmental Engineering, and Developmental Psychology rubrics including more research-focused criteria.

Rubrics seemed to mitigate bias and enhance decision-making when committees took time to deliberate the criteria, ensuring that all committee members agreed about what the criteria meant and what evidence they would use to assess whether candidates met the criteria. For example, the Plant Biology committee clearly measured "recent grant-seeking activities," defining what constituted both "recency" and "grant activity." They assessed these for each candidate individually, rather than comparing metrics across files. When one committee member indicated that a particular candidate "hadn't gotten huge grants," others emphasized that the candidate's research

area did not necessitate significant grant funding. This collectively brought the search committee back to the parameters expressed in their own rubric, rather than simply identifying candidates with large grants. Likewise, Development Psychology spent a substantial amount of time in their preliminary meetings developing precise wording and definitions they would use in their rubrics. The committee deliberated whether they would consider individuals who had Ph.D's versus clinical degrees as meeting minimum qualifications and whether candidates who described their attempts to get grants, regardless of outcome, as counting toward the criteria for "potential to get funding." These initial conversations about the criteria helped the committee achieve greater consensus about how the criteria would be applied as they discussed candidates.

One of the ways that bias seemed to influence the process was when committee members had different opinions on what kinds of evidence they would use to determine if candidates met the criteria. Ensuring that committee members were on the same page about what each criterion meant and what constituted evidence was a major stumbling block for several committees. For example, the Chemical Engineering rubric contained criteria related to research fit. The former criterion identified that research fit would be exemplified by "clear synergies with existing departmental research strengths." Yet, the distinction between what constituted excellent or adequate research field fit were nebulous as written in the rubric. During deliberations, some committee members considered candidates to be poor fits because "nobody is doing [that kind of research here]," while others viewed candidates poor fits because "we already have an expert in that area." Similar issues of research fit were also present in the Environmental Engineering search, where the committee's rubric lacked any kind of definitions or examples of the research area they were hoping to fill. As such, committee members often clashed, as exemplified in the following exchange we observed:

> Committee Member 1: Why was Candidate A scored so low by others? [They] have a PhD in Environmental Engineering and over 30 pubs.
>
> Committee Member 2: Candidate A is super focused on just one or two topics.
>
> Committee Member 3: They were too one dimensional.

As this deliberation reveals, some committee members had a concrete idea about what kinds of research the department needed, and Candidate A did not fit the bill, whereas other committee members seemed to be looking at broader qualifications such as general research area and past productivity. In both searches, differing definitions meant that candidates received different scores depending on the reviewer's implicit definition of research fit, introducing an area where bias could impact evaluation.

Scoring candidates

Prior to evaluating candidates, search committees created and applied different kinds of scoring strategies within their rubric. By scoring, we refer to the process by which committee members applied criteria to the evidence presented in candidate application materials and used said evidence to generate numeric rating. Scoring strategies ranged, with some committees creating points systems with different maximum ranges and others creating scores based on averages in different domains. Our results showed that the process committees used to generate their scores substantially impacted how evaluations were formed.

We found that committees that took time to discuss how to score candidates seemed to render less bias in their decision-making. In Chemical Engineering, the chair instructed the committee to review several example applicants using the scoring rubric, and then convened the committee to review their evaluative tendencies. This helped increase consistency between raters and initiated a discussion on how each commitee member assigned points, which ultimately determined which candidates were discussed and advanced. For example, the chair explained their stance by stating, "one recent, first author paper gets at least a three in terms of productivity." The other committee members agreed with that standard, and then incorporated it into their future ratings.

Likewise, committees that used time during their committee meetings to discuss scoring inconsistencies also strengthened the use of rubrics. When one Chemical Engineering committee member gave a candidate a score of 46 and the other four search committee members gave them 30, 30, 30, and 26, respectively, the committee paused. The faculty member that scored the candidate higher explained their scoring rationale and the committee collectively identified points of discrepancy. This meant that subsequent evaluations became more normalized toward the group mean, which reduced inconsistencies and increased the odds that candidates would not fall through the cracks. Environmental Engineering likewise initiated a discussion when it became clear that the committee had inconsistencies in their candidate evaluation. For example, one committee member rated a candidate highly in research whereas the rest of the committee rated the same candidate lower. The committee paused to discuss what they were looking for in research, and during this discussion, the committee member who provided the higher rating admitted that they had propped the candidate "up intentionally in order to generate discussion because they do [research] like me." This committee member also acknowledged that the candidate "may be better for another division in the department." The committee member purposely gave the candidate a higher score in research even though they knew the candidate did not meet the criteria. In the end, this candidate was not advanced. In this scenario, the rubric served as a "check" on this committee member's bias:

without a rubric, the committee member's score may have gone uninterrogated. In all, discussion about scoring rubrics in Environmental Engineering and Chemical Engineering slowed down the evaluation process in a positive way, ensuring that committees reviewed candidates more consistently.

Committees that did not take time to deliberate their scoring strategy encountered greater challenges. In Microbiology, there were several areas that generated wide disagreements based on broad criteria such as "evidence of scholarly potential," and "potential to develop a research program that involves undergraduate and graduate students." Because these parameters were discussed without any type of threshold or examples, the committee struggled to arrive at any type of consensus: a search committee member would highly rate one candidate with 20 publications while another committee member gave a low score. Environmental Engineering likewise struggled to discern a coherent scoring strategy regarding candidate service, most likely because they did not provide any examples or definitions of what exemplary service experience entailed. One committee member stated, almost jokingly, that they gave "0′s on all people's service" experience, and this comment was left uninterrogated by the committee. Overall, committees struggled at certain points to calibrate their scores, often influenced by personal preferences or biases that went unchecked within the committee, even with a rubric in place.

Committees that generated strategies for considering candidates who scored well across areas *as well as in specific domains* seemed to advance a broader pool of candidates. Microbiology, Chemical Engineering, and Developmental Psychology used a "cut-off" score, wherein the committee deliberated at-length all candidates who received a minimum score based on the rubric. However, each of these committees also identified additional ways to consider candidates who may have been strong in a certain area but who had not scored well overall. For instance, Developmental Psychology decided that they would re-review and discuss all candidates from minoritized groups, regardless of score, as a sort of bias check that may have emerged during the initial scoring process. Similarly, Chemical Engineering deliberated candidates who met the minimum score as well as candidates who had scored a "5" in the categories of research, teaching, or fit. These strategies seemed to broaden the types of candidates considered, elevating candidates who may have not risen to the top in the aggregate but had unique strengths in certain areas.

Environmental Engineering and Plant Biology used a more ad hoc approach in using scores to determine who was advanced, which seemed to invite greater opportunities for bias. Both committees evaluated all candidates using the rubrics, but committee members had wide discretion to elevate candidates based on individual preferences. In Environmental Engineering, candidates who received the absolute highest scores were advanced, but the committee also decided that a few candidates who "they were really excited about" would also be interviewed (regardless of their rubric score). The criteria

by which "excitement" was evaluated was never defined. In Plant Biology, each committee member generated a list of their top five candidates, ostensibly informed at least in part by their rubric scores. Yet, each member's top five list was not cross-checked to ensure that the highest scoring candidates were advanced. In this way, the process by which the committee used the rubric scores — whether committee members were asked to justify their scores; whether scores were used rigidly or in context with other factors; and whether rubric scores had any bearing on who was advanced — shaped the hiring outcomes, and some processes left open greater room for bias to emerge.

Assigning weights to criteria

How and if committees assigned and made explicit the weights associated with certain criteria also impacted how committees used rubrics. Weights in this case refer to the extent to which committees determined that certain criteria (e.g., grant activity) would count more as compared to other criteria (e.g., DEI experience) in the rubric.

Committees that explicitly weighted criteria from the beginning seemed to navigate candidate evaluation more effectively. In Microbiology, it was clear in the rubric that teaching was the most heavily weighted criterion, as evidenced by the number of teaching and advising-related items in their rubric. The chair explained the importance of evaluating quality teaching in candidate review when they said, "I think it's safe to say that for most faculty here the teaching and student learning experience is the most important." This was a shared commitment across the committee, as demonstrated in the importance of examining teaching quality during each committee meeting. The committee seemed to be successful in applying these specific teaching criteria to the candidates, as all three finalists, according to the committee, delivered high-quality teaching demonstrations. According to the search chair, the Latino man candidate that was ultimately forwarded to the department for a vote had "several students come up and talk to him [after his teaching demonstration] … A lot of faculty liked that." In all, Microbiology's rubric criteria and weighting strategy allowed them to effectively identify candidates who had the qualifications they desired, leaving less room for ambiguity in terms of who would be advanced and why.

In contrast, bias seemed to play a role when committee members had implicit weighting strategies. For instance, the Chemical Engineering committee included in their criteria related to research the receipt of a prestigious research award. As the committee deliberated which candidates to advance to the short-list, multiple members led with statements like "Candidate A has [the prestigious award]" as a blanket justification for advancement to the next round. In this case, the criterion was quite clear — either a candidate had the award or not. However, this particular award seemed to automatically boost

some candidates to the next evaluation round, even if they had not scored highly in other areas — a sort of halo effect for some candidates that went relatively uniexamined by the committee. The Plant Biology committee also struggled with implicit weights. For example, the Plant Biology position description clearly articulated the committee's and institution's emphasis on high-quality teaching. The rubric likewise contained "recent teaching experience," "potential for contributing novel courses in the curriculum," and "ability to teach." However, subsequent deliberations and interviews with candidates revealed that the committee was not interested in candidate's teaching experience or abilities, but rather their ability to teach courses on specific plant biology topics, even though all three criteria were evenly weighted. The issue was not that these committees had inappropriate weights or should not weigh research or teaching more heavily, but rather, these weights were implicit, undiscussed, and thus more prone to bias.

Counting DEI experience and qualifications

Committees varied in the extent to which DEI-related criteria were present in their rubrics, with Microbiology, Chemical Engineering, and Development Psychology having specific DEI-related qualifications present. In contrast, Plant Biology included criteria related to teaching and mentoring of students intended to elevate candidates with diversity experience but did not specifically use these terms in their criteria, and Environmental Engineering's rubric contained mentions of teaching and service but no criterion related to DEI.

Committees that integrated DEI-related criteria into their rubrics and concretely discussed and defined DEI activities experienced greater success in elevating minoritized candidates. The Chemical Engineering rubric required candidates to have "demonstrated commitment to diversity and inclusion, which constituted "listed [DEI] activities in the research/teaching statements." This meant that candidates were encouraged to integrate diversity and inclusion into both their research and teaching areas. The rubric used in Developmental Psychology was also integrative, including definitions and categories for what constituted DEI in different evaluation areas (e.g., research, teaching, service). When one committee member described their fear that it would "unfairly penalize candidates that did not engage in critical scholarship," the chair re-emphasized his commitment to awarding points for such scholarship because "it [was] in the job description that we value this kind of work." This latter example highlights how including and meaningfully weighting DEI experience in teaching and in research on a rubric helped mitigate bias against minoritized candidates. Altogether, these factors likely contributed to each committee's success in identifying and advancing candidates from minoritized groups, as Chemical Engineering ultimately hired two White women

and Developmental Psychology hired one Asian woman and one White woman.

Even with this commitment to DEI, these same committees still experienced challenges in determining how, if at all, what kind, and when DEI qualifications would be "counted" in their evaluations. For example, Chemical Engineering had well-defined DEI criteria in their rubric. Even so, as the committee deliberated, it became clear that the research criteria were more important than the DEI-related ones. In the preliminary applicant review, the committee initially agreed that all candidates who received a high score in any of the four rubric domains (research, teaching, institutional stewardship, and fit) would merit deliberation from the full committee. However, many candidates received a five in at least one the domains and the committee realized that they would not have time to review each of these candidates in depth. At this point, the chair commented that "getting a 5 in the diversity statement [under institutional stewardship] shouldn't warrant as much discussion as getting a 5 in the research domains." In other words, at least in the initial evaluation stage, research was the criteria that mattered most to this committee, even though DEI-related criteria seemed to have equal weight based on the rubric. In addition, the devaluing of DEI-related experience in the initial round also represented a potential structural roadblock for minoritized candidates, in that applicants who excelled in this area were not necessarily getting the same second-look as candidates who excelled in other areas. Similarly, Developmental Psychology sometimes struggled to evaluate candidates' potential ability to contribute to departmental DEI activities based on their application materials. As the committee reviewed candidates in the initial round, some members observed that one candidate had not "taken their diversity statement to the next level" because they had not specifically addressed racial diversity, while other committee members felt the same candidate had a strong DEI statement. Different perspectives on what constituted DEI experience, and failure to define the specific DEI-qualifications the committee sought, therefore undermined the ability of the committee to agree on which candidates should be advanced.

Neither Environmental Engineering nor Plant Biology's rubrics included specific DEI-related criteria; as such, we observed that their committee deliberations related to DEI were much more rooted in considering a candidate's identity (i.e., their ability to contribute to the department's demographic diversity) rather than candidate's experiences with DEI. For example, throughout the Environmental Engineering search, the committee often debated whether certain candidates, including women and international candidates, "counted" as underrepresented. In one meeting, a search committee member asked, "if this [international] person doesn't count as a minority, do they count as a woman?" The committee continued to express confusion about this topic as the search went on. However, in the interview stage, the committee did not

ask candidates about their experiences with diversity, equity, and inclusion (e.g., mentoring of students, inclusive pedagogy), but instead continued to emphasize aspects of the candidate's identities as the factor being considered. We observed similar conversations in the Plant Biology search, including a heated discussion among committee members about a candidate that they perceived to be Black. In this exchange, some committee members wanted to advance this candidate because of their potential to contribute to the diversity of the department, whereas other committee members argued vehemently that this candidate did not meet the department's stated criteria in the rubric and was therefore being elevated because of race alone. Although the committee members agreed that this candidate had the potential to contribute as a teacher that would support a diverse student body, this was not part of their criteria; they ultimately did not advance the candidate. While we observed that the committees that did have DEI-related criteria in their rubric likewise had conversations about which candidates "counted as diverse," because they had criteria that operationalized to some extent how a candidate could contribute to DEI in the department, they seemed to rely more on qualifications than identity alone. All said, committees that lacked DEI criteria in their rubric seemed to root their discussions of diversity in the candidate's identity, rather than taking into account how candidates might engage with DEI in their teaching, research, and/or service.

Declining utility of rubrics

All search committees in our study developed and applied the rubric to candidates at the beginning stages of the search (e.g., as they sifted through curriculum vitae). However, as search committees moved further into the hiring process, they tended to rely on the rubrics less and less to guide their evaluations

At the later stages, evaluations were more likely to be based on interpersonal interactions and inferences about candidate interest and "hireability" as compared to candidate qualifications and performance based on the rubric criteria. In Developmental Psychology, one finalist gave a good teaching demonstration according to the committee, which based on the rubric seemed to be an important criterion. However, the chair explained that the committee ultimately did not favorably evaluate her, saying, "What really ranked her down was her interview. She was just lackluster … Several people had the same comment … that she didn't engage or she didn't have questions or didn't want to learn about the position." It is clear from this statement that departmental members experienced this candidate as disengaged or not that interested. Members of the Chemical Engineering search also expressed similar concerns about a woman candidate's "energy levels." Similarly, in the Environmental Engineering, Chemical Engineering, and Plant Biology searches, committee members viewed

candidates with Western accents more favorably compared to the international candidates with non-Western accents, citing "language barriers" and other communication concerns after interviews with international candidates, even though communication style was not a part of the rubric.

Similarly, committees seemed to discuss different aspects of each candidate randomly at the final stages. When Chemical Engineering deliberated their top four candidates, they discussed whether candidates "had champions" in the department, whether their interactions with staff had been positive, and whether candidates were "too green" and therefore underprepared or not, but these criteria were neither in the rubric nor discussed systematically for each candidate. In Developmental Psychology, research area, potential for attracting funding, and teaching (three of the criterion listed on the rubric) were discussed for some candidates, whereas for others, department members debated why a candidate wanted to leave their current position, whether another candidate had collaboration potential, and how another candidate's postdoc placement at a prestigious university "didn't seem to make a huge difference" in their productivity. Rubrics seemed to carry greater weight and bring more structure at the beginning of each search, whereas by the end of the search, deliberations were less grounded in the rubrics and criteria within them.

Discussion

This study sought to explore how five faculty search committees used rubrics in candidate evaluation and examine the extent to which rubrics seemed to perpetuate or mitigate bias in committee decision-making. We considered rubrics through the lens of nudge theory, which suggests that interventions, or nudges (i.e., a rubric), to decision-making contexts, can help reduce bias and produce better, more effective outcomes. We found wide variation in how search committees created, designed, and used their rubrics, as well as in the criteria within each rubric. Moreover, consistent with past studies (e.g., Blair-Loy et al., 2022; Rivera & Tilcsik, 2019), our results suggest that rubrics did not *remove* social bias from influencing committee evaluations, but rather, encouraged or nudged committee members to engage in decision-making processes that seemed to reduce errors and bias. In what follows, we examine the four aspects of rubric use that seemed most beneficial for search committees to advance equity, place them alongside extant literature, and examine potential pitfalls that coincide with each aspect. We conclude by making general observations about the state of rubrics in faculty hiring to advance DEI and offer areas for future practice and research.

First, our data showed that rubrics seemed to slow down search committees' decision-making process. Committees invested time in determining the criteria and associated weights that went into the rubric, reviewing candidates with those criteria in mind, and discussing the candidates to be advanced

through the hiring process based on the criteria. This has the potential to reduce bias and therefore aid in the hiring of minoritized candidates, as behavioral economists suggest that slower, more deliberative decision-making processes are critical for rendering more effective judgments (2011) and studies within higher education recommend discussion and calibration exercises are important steps for search committees to take if they want to enhance DEI in hiring (White-Lewis, 2020). However, we can also see how search committees might view a slower process to be a hindrance to their ability to hire top candidates (White-Lewis, 2022) and therefore cause them to reject and/or undermine (Sunstein, 2021) the use of rubrics and other decision-support tools in hiring.

Second, like nudges that focus attention on salient information (Damgaard & Nielsen, 2018; Thaler & Sunstein, 2008), our data showed that rubrics helped committees concentrate on the criteria and reduced the extent to which committee members considered extraneous or irrelevant information in their assessments, but only at the earlier stages of the evaluation process. Such focus can help mitigate bias, as prior work shows that evaluators will often deviate from the criteria and bring in irrelevant and typically negative information when evaluating candidates from minoritized groups (Rivera, 2017). On the other hand, focusing on the criteria was still problematic when the criteria itself reflected structural inequality within the academy (Posselt et al., 2016). The prime example here was the Chemical Engineering committee's use of an award to indicate research productivity, when much research shows bias in the award-making process (Chen et al., 2022). Thus, we witnessed scenarios in which committees believed they were "doing the right thing" by sufficiently calibrating their rubrics, but nevertheless calibrated on criteria that perpetuated structural inequities (White-Lewis et al., 2022).

Third, rubrics seemed to enhance consistency in applying the criteria across candidates, particularly when committees had calibrated their scoring strategies and deliberated what each criteria meant. Rendering more consistent evaluations across candidates has the potential to reduce bias, as prior research shows that minoritized candidates are often subject to increased scrutiny or shifting standards (Biernat et al., 2009; Moody, 2012), for instance, needing more publications to demonstrate their research competency compared to white and/or men candidates. While we observed this to be mainly a good thing in this study, similar to issues with focusing attention described above, consistent application of criteria could present problems for DEI if it does not allow evaluators to consider candidates in the context(s) of their opportunities (Bastedo, 2021). For example, a criteria related to lab experience could be applied consistently and disadvantage candidates who went to universities with less access to labs.

Finally, rubrics brought in a more inclusive, holistic view of candidates strengths and weaknesses, including in diversity, equity, and inclusion, if such

criteria were included in the rubric. Such benefits potentially contribute to the hiring of minoritized candidates in at least two ways. One way is by highlighting specific contributions in specific areas, as research shows that when faculty members are asked to give overall assessments of general hireability (Eaton et al., 2020) or fit (White-Lewis, 2020), biases are more likely to disadvantage candidates from minoritized groups. Similar to work on inclusive admissions (Bastedo, 2021; Posselt, 2016), rubrics also forced evaluators to pay attention to a candidate's "whole file" rather than rely on only one criterion (i.e., research as is most often the case at universities). However, as our results showed, incorporating DEI into the evaluation process was not without its problems, as committees often did not specify what kinds of DEI contributions they were looking for or how much they would be weighted relative to other criteria. On the other hand, we observed that when committees lacked concrete DEI criteria in their rubric, they were left with relatively few ways, beyond guessing a candidate's gender and/or race, to assess how applicants might contribute to DEI in the department. Both scenarios are problematic, but, consistent with past research (Liera, 2020; Liera & Ching, 2019) we view the incorporation of DEI related criteria into the rubric as an important step in the right direction.

We wanted to call particular attention to our finding regarding the declining utility of the rubric over the course of the search. In some ways, these results are to be expected, as research on rubrics and their use in hiring show that they are most often applied at the beginning of evaluation stages (White-Lewis, 2020). Indeed, an argument could be made that a rubric is being used successfully when all final candidates are (relatively) equally qualified. In such a case, a committee would naturally need to identify new criteria to make differential assessments among the candidates. The DEI problem in this scenario is that these new criteria are implicit, not discussed, thereby allowing evaluators to relying on their instincts and inferences (Posselt, 2016), as we observed when committee members tried to make meaning out of energy levels or communication abilities, qualities that are often infused with social stereotypes (Rivera, 2017). However, research on inclusive admissions (Bastedo, 2021; Posselt, 2016 problematizes the entire notion of "highly qualified": if a relatively narrow set of criteria are applied at the outset, it makes sense that the candidates in the final pool would have the same qualifications, and therefore using holistic review at the later stages will have diminishing returns because the pool will already be less diverse. As such, attention must be paid to implicit and explicit criteria used across all stages of a search.

This study serves as an important reminder that rubrics and other decision-support tools are exactly that: tools. As behavioral economists note, nudges such as rubrics have reduced impact when they are (a) deployed in contexts where the problems are complex and (b) the root cause of the error or bias is difficult to pinpoint or inaccurately diagnosed (Sunstein, 2021; Thaler &

Sunstein, 2008). Hiring is a complex human process with multiple actors bringing their own biases to the decision outcome. There is a constant interaction between the tools and the people that use them. We have witnessed a significant clamoring for new tools and best practices across disciplines, as if the solution to equity exists "out there" somewhere yet to be discovered. Indeed, there is a wealth of literature in organizational psychology on optimal rubric scaling and categorization (e.g., Goldhaber et al., 2017), and researchers and practitioners would do well to leverage this scholarship to further enhance rubric use. But what we learned from this study is that we need a both/and approach. Optimizing rubrics means both enhancing their development and application, *and* not losing sight of the equity-minded principles that must undergird them (Bensimon, 2007; Liera, 2020; Liera & Ching, 2019 O'Meara et al., 2021). One such principle is accountability (O'Meara et al., 2021). Even with a rubric, we witnessed discrimination against international candidates of color, entrenching inequities into evaluation through overreliance on grantsmanship, and a lack of rubrics in later stages that promoted the use of biased personality judgments. Given that there is no "accountability" category on any rubric, we must remain vigilant of how we use tools to advance (in)equity. Next, we provide recommendations for future practice and research on rubrics in faculty hiring.

Recommendations for practice and research

This study was animated by a critical concern of helping faculty, department chairs, and administrators make research-driven decisions around the use and substance of rubrics to make faculty hiring processes fairer and more equitable. Based on the results of this study, we outline five practical recommendations for these groups to use rubrics to sustainably improve future faculty search and selection processes:

- Develop rubric subcomponents that define what constitutes excellence in research, teaching, and service (e.g., breaking down research productivity by number and quality of publications).
- Conduct calibration exercises prior to reviewing candidates to enhance consistency among committee members (White-Lewis, 2020).
- Embed DEI criteria and ensure that said criteria are meaningfully weighted (Liera & Ching, 2019; Posselt et al., 2016).
- Develop a process for how the committee will use rubric scores to make decisions. Consider strategies for evaluating candidates from minoritized groups a second time with bias in mind.
- Create different rubrics for different phases of the search (one for the review of all candidates, another for those invited for interviews).

There are also implications for future research. Although our study provided in-depth information on how a small number of committees used rubrics, larger quasi-experimental studies wherein some committees use a rubric and others do not would enhance our ability to understand the impact of rubrics on hiring outcomes. We also recommend studies across different disciplines and institutional types to understand how disciplinary logics and institutional resources affect how faculty apply rubric criteria to candidates overall.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was funded by the National Science Foundation under AGEP Award 1820975.

ORCID

Dawn Culpepper http://orcid.org/0000-0002-3547-4615

References

Baldwin, R., Cave, M., & Lodge, M. (2011). *Understanding regulation: Theory, strategy, and practice*. Oxford University Press.

Banaji, M. R., & Greenwald, A. G. (2016). *Blindspot: Hidden biases of good people*. Bantam.

Bastedo, M. (2021). Holistic admissions as a global phenomenon. In H. Eggins, A. Smolentseva, & H. de Wit (Eds.), *Higher education in the next decade* (pp. 91–114). Brill.

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. https://doi.org/10.1515/sem-2013-0087

Bensimon E.M. (2007). The underestimated significance of practitioner knowledge in the scholarship on student success. *The Review of Higher Education*, *30*(4), 441–469. https://doi.org/10.1353/rhe.2007.0032

Biernat, M., Collins, E. C., Katzarska-Miller, I., & Thompson, E. R. (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin*, *35*(1), 16–28. https://doi.org/10.1177/0146167208325195

Blair-Loy, M., Mayorova, O. V., Cosman, P. C., & Fraley, S. I. (2022). Can rubrics combat gender bias in faculty hiring? *Science*, *377*(6601), 35–37. https://doi.org/10.1126/science.abm2329

Brown, P. (2012). A nudge in the right direction? *Social Policy & Society*, *11*(3), 305–317. https://doi.org/10.1017/S1474746412000061

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

Culpepper, D., Reed, A. M., Enekwe, B., Carter-Veale, W., LaCourse, W. R., McDermott, P., & Cresiski, R. H. (2021). A new effort to diversify faculty: Postdoc-to-tenure track conversion models. *Frontiers in Psychology*, Article #4992. https://doi.org/10.3389/fpsyg.2021.733995

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313–342. https://doi.org/10.1016/j.econedurev.2018.03.008

Devine, P. G., Forscher, P. S., & Cox, W. T. (2017). A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology*, *73*, 211–215. https://doi.org/10.1016/j.jesp.2017.07.002

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. https://doi.org/10.1007/s11199-019-01052-w

Glod, W. (2015). How nudges often fail to treat people according to their own preferences. *Social Theory and Practice*, *41*(4), 599–617. https://doi.org/10.5840/soctheorpract201541433

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2017). Screen twice, cut once: Assessing the predictive validity of applicant selection tools. *Education Finance and Policy*, *12*(2), 197–223. https://doi.org/10.1162/EDFP_a_00200

Griffin, K. A. (2020). Institutional barriers, strategies, and benefits to increasing the representation of women and men of color in the professoriate. In L. W. Perna (Ed.), *Higher education: Handbook of theory and research* (Vol. 35, pp. 1–73). Springer.

Harris, J. C., Snider, J. C., Anderson, J. L., & Griffin, K. A. (2021). Multiracial faculty members' experiences with multiracial microaggressions. *American Journal of Education*, *127*(4), 531–561. https://doi.org/10.1086/715004

Hummel, D., & Maedche, A. (2019). How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of Behavioral and Experimental Economics*, *80*, 47–58. https://doi.org/10.1016/j.socec.2019.03.005

Isaac, C., Lee, B., & Carnes, M. (2009). Interventions that affect gender bias in hiring: A systematic review. *Academic Medicine*, *84*(10), 1440–1446. https://doi.org/10.1097/ACM.0b013e3181b6ba00

Kahneman, D. (2011). *Thinking: Fast and slow*. Farrar, Straus and Giroux.

Kosters, M., & Van der Heijden, J. (2015). From mechanism to virtue: Evaluating nudge theory. *Evaluation*, *21*(3), 276–291. https://doi.org/10.1177/1356389015590218

Langin, K. (2021, August 3). Can anonymous faculty searches boost diversity? *Science*. https://www.science.org/content/article/can-anonymous-faculty-searches-boost-diversity

Liera, R. (2020). Equity advocates using equity-mindedness to interrupt faculty hiring's racial structure. *Teachers College Record*, *122*(9), 1–42. https://doi.org/10.1177/016146812012200910

Liera, R., & Ching, C. (2019). Reconceptualizing "merit" and "fit": An equity-minded approach to hiring. In A. Kezar & J. Posselt (Eds.), *Administration for social justice and equity in higher education: Critical perspectives for leadership and decision-making* (pp. 111–131). Routledge.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). Jossey-Bass.

Misra, J., Kuvaeva, A., O'Meara, K., Culpepper, D. K., & Jaeger, A. (2021). Gendered and racialized perceptions of faculty workloads. *Gender & Society*, *35*(3), 358–394. https://doi.org/10.1177/08912432211001387

Mitchneck, B. (2020). *Synthesizing research on gender biases and intersectionality in citation analyses and best practices*. ARC Network. https://uploads-ssl.webflow.com/60ceadbbd1b31b75588b6cd7/616b4be13010b4087717f037_Mitchneck-ARC-VVS-Final-Report-Updated.pdf

Moody, J. (2012). *Faculty diversity: Removing the barriers* (2nd ed.). Routledge.

Muñoz, S. M., Basile, V., Gonzalez, J., Birmingham, D., Aragon, A., Jennings, L., & Gloeckner, G. (2017). Critical perspectives from a university cluster hire focused on diversity, equity, and inclusion. *Journal of Critical Thought and Praxis*, *6*(2), 1–21. https://doi.org/10.31274/jctp-180810-71

National Science Foundation. (2021). *Women, minorities, and persons with disabilities in science and engineering*. National Center for Science and Engineering Statistics (NCSES) https://ncses.nsf.gov/pubs/nsf21321/report

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

O'Meara, K., Culpepper, D., Misra, J., & Jaeger, A. (2021). *Equity-minded faculty workloads: What we can and should do now*. American Council on Education. https://www.acenet.edu/Documents/Equity-Minded-Faculty-Workloads.pdf

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. https://doi.org/10.3102/0034654320914742

Posselt, J. (2016). *Inside graduate admissions: Merit, diversity, and faculty gatekeeping.*Harvard University Press.

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. In L. Perna (Ed.), *Higher education: Handbook of theory and research* (Vol. 35, pp. 1–63). Springer.

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. https://doi.org/10.1177/0003122417739294

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review*, *84*(2), 248–274. https://doi.org/10.1177/0003122419833601

Saldaña, J. (2016). *The coding manual for qualitative researchers*. Sage.

Schmaling, K. B., Trevino, A. Y., Lind, J. R., Blume, A. W., & Baker, D. L. (2015). Diversity statements: How faculty applicants address diversity. *Journal of Diversity in Higher Education*, *8*(4), 213–224. https://doi.org/10.1037/a0038549

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. https://doi.org/10.1037/dhe0000174

Sheppard, L. D., Goffin, R. D., Lewis, R. J., & Olson, J. (2011). The effect of target attractiveness and rating method on the accuracy of trait ratings. *Journal of Personnel Psychology*, *10*(1), 24–33. https://doi.org/10.1027/1866-5888/a000030

Stake, R. E. (2005). *Multiple case study analysis*. The Guilford Press.

Sunstein, C. R. (2021). *Sludge: What stops us from getting things done and what to do about it*. MIT Press.

Thaler, R. H., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science*, *16*(6), 474–480. https://doi.org/10.1111/j.0956-7976.2005.01559.x

White-Lewis, D. K. (2020). The facade of fit in faculty search processes. *The Journal of Higher Education*, *91*(6), 833–857. https://doi.org/10.1080/00221546.2020.1775058

White-Lewis, D. (2022). The role of administrative and academic leadership in advancing faculty diversity. *The Review of Higher Education*, *45*(3), 337–364. https://doi.org/10.1353/rhe.2022.0002

Yazan, B. (2015). Three approaches to case study methods in education: Yin, Merriam, and Stake. *The Qualitative Report*, *20*(2), 134–152. doi:https://doi.org/10.46743/2160-3715/2015.2102

Yin, R. K. (2018). *Case study research and applications: Design and methods*. Sage.