POLICY FORUM

DIVERSITY

Can rubrics combat gender bias in faculty hiring?

Some bias persisted, but rubric use should be encouraged

By **Mary Blair-Loy[1], Olga V. Mayorova[1], Pamela C. Cosman[2], Stephanie I. Fraley[3]**

Research has documented the presence of bias against women in hiring, including in academic science, technology, engineering, and mathematics (STEM). Hiring rubrics (also called criterion checklists, decision support tools, and evaluation tools) are widely recommended as a precise, cost-effective remedy to counteract hiring bias, despite a paucity of evidence that they actually work (see table S8). Our in-depth case study of rubric usage in faculty hiring in an academic engineering department in a very research-active university found that the rate of hiring women increased after the department deployed rubrics and used them to guide holistic discussions. Yet we also found evidence of substantial gender bias persisting in some rubric scoring categories and evaluators' written comments. We do not recommend abandoning rubrics. Instead, we recommend a strategic and sociologically astute use of rubrics as a department self-study tool within the context of a holistic evaluation of semifinalist candidates.

Although academic STEM aspires to be a meritocracy, its taken-for-granted cultural schemas of merit smuggle in biases (*1*), which contribute to a dearth of diversity that undermines scientific innovation and impact (see table S8). In academic engineering, one of the most male-dominated STEM fields, on average 17.6% of engineering faculty positions are held by women (*2*). Although the percent of women engineering doctorates increased from 15.8% in 2000 to 24% in 2019 (*3*), these increases will not be matched by gains in the professoriate if women face unfair barriers at hiring.

Academic policy-makers and EDI (equity, diversity, and inclusion) specialists strongly encourage rubric usage, in which faculty evaluators systematically rate each candidate on a set of previously agreed-on criteria. This process is believed to counteract the bias of individual evaluators by promoting slower, more deliberative, and analytical thinking and by focusing them on skill sets that directly affect job performance rather than on impressions and intuitions (*4*, *5*). However, we are aware of only one study, conducted in a laboratory setting, in which participants rated candidate summaries, which shows that agreeing on rubric criteria in advance reduces evaluation bias (*6*). We are aware of no studies that analyze the effect of rubric use on bias in real-world hiring, in which actual evaluators assess voluminous candidate files to make actual high-stakes decisions. Real-world case studies are important because the effectiveness of interventions depends on the social context and the identities of all involved (*7*).

Despite this paucity of evidence, many fields have developed rubrics to standardize candidate assessment and have promoted rubrics as a best practice for EDI in hiring. In policy guides for academic hiring, several applied treatises and websites provide sample rubrics (see table S8) (*5*). A recent influential review of faculty hiring lists "mandatory use" of rubrics as one of the interventions that the authors "view as having the most promise for [university] institutions seeking to improve inclusivity in hiring across disciplines" (*4*).

CASE STUDY OF RUBRIC USAGE

To help address this knowledge and policy gap, we developed an in-depth case study of an engineering department in a research-intensive (Carnegie classification R1), highly ranked university. Like other R1s, this department strongly values research productivity when evaluating faculty candidates (*8*). Like most academic engineering departments, our case department was male dominated; women composed 18% of the faculty, which is close to the 17.6% national average (*2*).

[1]Department of Sociology, University of California, San Diego, La Jolla, CA, USA. [2]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA. [3]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. Email: mblairloy@ucsd.edu; sifraley@ucsd.edu

Rubric usage and hiring patterns

We started with a faculty candidate evaluation template from the University of Michigan STRIDE program (Strategies and Tactics for Recruiting to Improve Diversity and Excellence), which is funded by the National Science Foundation (NSF). This template includes widely accepted criteria for faculty at research-intensive universities (*5*).

We worked with the department under study to adapt the template to fit its searches. Its rubric evaluated faculty candidates across six dimensions: research productivity, research impact, teaching ability, contributions to diversity, potential for collaboration, and overall impression (see fig. S1). Rubric scores ranged from excellent to poor for each evaluation category (we translated these ratings into a numerical variable: excellent = 4, good = 3, neutral = 2, fair = 1, and poor = 0). Written commentary was also encouraged.

Department faculty agreed to fill out the rubric as a tool in their evaluation of the semifinalist list compiled by the recruitment committee. In four faculty search cycles over four recent academic years, faculty used the rubric to evaluate written materials supporting applications of 62 semifinalists (32 women and 30 men; gender was self-reported by candidates in their application file). At the beginning of each faculty meeting that was focused on selecting finalists, a faculty member summarized and presented rubric scoring results and commentary, with evaluators anonymized, filtering out any inaccurate or off-topic content.

Our analysis next compared the proportion of women hired during the 8-year period immediately before rubric use (which we refer to as "Phase One") with the proportion hired during the four academic years of rubric use ("Phase Two"). Near the outset of Phase One, the campus implemented three EDI interventions: Faculty serving as equity advisers took on administrative oversight of shortlisted candidates, applicants' "Contributions to Diversity Statements" (C2D) were added to application files, and equity advisers began giving diversity training to search committee members. Such training focused on evaluating C2D statements and an overview of research on implicit biases. Nonetheless, during Phase One, the department conducted eight searches and hired eight men and one woman.

At the outset of Phase Two, the department introduced an additional intervention: rubrics used as described above. During Phase Two, the department conducted four faculty searches and hired three women and six men. The number of women hired increased from only one per nine hires in Phase One to three per nine hires in Phase

Two. The phase with the increased hiring of women coincides with the period when rubrics were used. (We cannot fully attribute this increase to rubric utilization because unmeasured factors may also have changed.)

Analysis of rubric scoring

All 62 semifinalist candidates received rubric scores from 6 to 21 faculty, with a mean of 13.5 and a median of 12. There was no statistically significant gender difference in the number of scores received.

Analysis of the rubric scoring patterns for men and women candidates revealed statistically significant differences in three of the six evaluation categories (see table S1.A): Women were scored lower than men in research productivity and research impact but higher than men in contributions to diversity. In the other categories, including the overall impression category, scores for men and women were not significantly different.

To determine whether gender bias was incorporated into rubric scores, we analyzed the research productivity category because it can be most directly compared with external metrics. We chose two metrics calculated for the candidate's application year. First, we tallied from candidate curricula vitae the number of articles published (and confirmed this tally in the Web of Science database). Second, we pulled from the Web of Science the H-index, a dominant measure of researcher output that incorporates productivity and impact in a single number that can be compared across faculty of all seniority levels (*9*). We call research productivity, a rubric category that can be measured independently, a "calibration category" (see table S2, footnote).

To test for gender bias, we constructed ordinary least squares (OLS) regression models to predict rubric scores of research productivity, controlling for the independently measured categories of seniority and number of articles or H-index (see table S2). We found that women candidates, on average, received statistically significantly lower productivity rubric scores than those of men, even after controlling for seniority (measured as number of years since PhD) and number of articles published [unstandardized β coefficient (B) = −0.36, $P \leq 0.01$] (see table S2, model 1). Similarly, women receive significantly lower scores on average than men while controlling for seniority and H-index (B= −0.29, P< 09.05) (table S2, model 3). Thus, rubric scoring alone did not appear to fully mitigate gender bias. These findings mirror the gender bias detected in other academic peer-review processes (*10*, *11*). Because the H-index itself has been found to incorporate bias against women (see table S8), our findings should be interpreted as additional bias.

Social psychology literature on double standards finds that among more junior candidates, women are often held to higher standards of competence than men in ways that sometimes change for candidates with more experience (*12*). We thus tested whether the effect of gender on rubric scores is contingent on the value of the external metric (see table S2, models 2 and 4).

Women's productivity rubric scores are consistently below those of men who have the same number of articles and seniority (see the first figure, left). Women face an average 0.36-point penalty, which remains the same across the range of number of articles published. (The rubric scores are mostly clustered at the middle to high portion of the scale, between 2 and 4; within this range, there is an approximately 18% penalty for being a woman.)

Yet when controlling for H-index and seniority, the gender penalty is harshest among candidates with the lowest H-indices (see the first figure, right), who are disproportionately junior (see table S4, footnote). At the lowest tail of the H-index distribution, men receive research productivity rubric scores that are on average 0.7 points higher than the scores of women with the same seniority. At this end of the H-index distribution, the rubric scores are mostly clustered between 1 and 3; for these candidates, there is an approximately 35% penalty for being a woman.

The gender difference in rubric scoring gradually decreases by about 0.01 point for each 1 point of H-index gained. Yet women do not catch up to men in how productive they are rated in the rubrics until reaching an H-index of 17.5, a productivity index well above the 12.8 average and achieved by only a handful of candidates. Because rubric scores for men and women in the overall impression category were not statistically significantly different (see table S1), evaluators may have combined category scores so that women's higher average scores on contributions to diversity offset their lower average scores on productivity and impact.

Content analysis of qualitative comments

Next, we conducted what to our knowledge is the first content analysis of qualitative rubric comments in a real faculty search context. Candidates received written commentary alongside their rubric scores. An average of three and a maximum of nine evaluators wrote comments on each candidate. The number of comments received did not differ by candidate gender.

In our content analysis, we prepared a dataset of comments, in which candidate gender was concealed, by removing gender indicators such as pronouns. We then combined an inductive exploration for emergent themes with deductive searches for specific patterns found in previous literature on letters of recommendation. We conducted hand coding, which some research suggests is superior to computer-assisted coding for studies such as ours with new analyses (*13*).

Many comments contained evaluative notes on the quality, number, authorship order, or impact of the publications. Inductively, we coded these as either negative (for example, "some gaps in the pubs." and "only one paper from a postdoc of three-plus years") or positive (for example, "strong publication record, letters attest to research

Indication of gender bias in rubric scores on research productivity

Graphs show predicted values by gender from ordinary least squares regression models regressing rubric scores of research productivity on independent productivity metrics, controlling for seniority. n = 62 semifinalist candidates (see table S2, models 1 and 4).

GRAPHIC: K. FRANKLIN/*SCIENCE*

output and collaboration"). We also searched deductively for the presence of two themes mentioned in previous literature on other types of evaluation. One code is "standout" language (*13*, *14*). Examples in the rubric comments include "outstanding productivity and quality" and "probably the best [search specialty] candidate out there this cycle." The last code is "doubt raisers" (*15*), when a seemingly positive or neutral comment is accompanied by language that minimizes the accomplishment or raises concerns (for example, "several publications, but... some impact factors are very modest"). We turned these four codes into four dichotomous measures of the presence or absence of comments in each category. (See table S3 for details on coding methods, intercoder reliability, dichotomous measure rationale, and robustness check.)

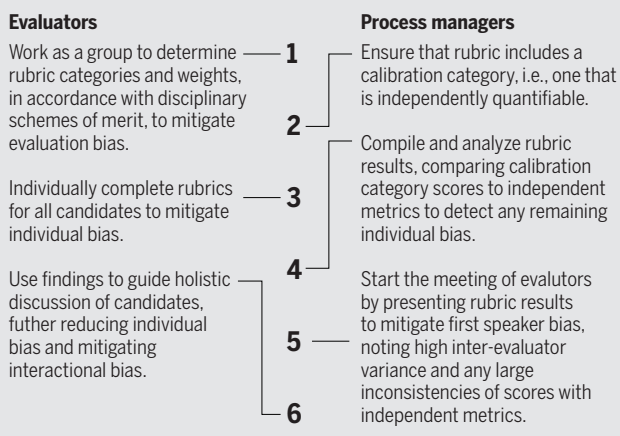
Candidate gender was subsequently unmasked, and the percentages of women and men who had received at least one positive comment, negative comment, standout term, and doubt raiser were calculated (see table S3). We found that 86% of men but only 63% of women candidates received at least one positive comment. Men were half as likely to receive a negative comment (25%) compared with women (50%). Men were 3.5 times more likely to receive standout language (32%) compared with women (9%). A χ^2 test indicates statistically significant gender differences for these three variables. Thirteen percent of women and 25% of men received a doubt-raising comment. This pattern was not expected, yet this gender difference is not statistically significant (see table S3). Overall, the gendered patterns in rubric quantitative scores align with the qualitative comments and with previous research on other evaluative language (*13*–*15*).

Survey of faculty

At the conclusion of Phase Two, we conducted an anonymous survey of department faculty, to which 56% of the professors responded. Our check for selection bias in responses revealed no statistically significant difference in the two indicators we have for respondents and nonrespondents: gender and tenure status (see table S6). Most reported that rubric usage helped them evaluate candidates in a more organized fashion (78%) and may have helped them be more objective (78%), potentially reducing individual bias (see fig. S2 and table S5).

Using rubrics for EDI hiring

Recommended process for rubric use to improve EDI in faculty hiring. Steps 1, 3, and 6 are conducted by faculty evaluators while steps 2, 4, and 5 are conducted by faculty on the search committee with administrative support.



GRAPHIC: C. SMITH/*SCIENCE*

At the beginning of the meeting to discuss the semifinalist list and choose finalists to invite for interviews, a faculty member presented rubric numeric scores and comments with the identity of the evaluators anonymized. It was explained to faculty that this step aimed to enable all viewpoints to be heard while avoiding first-speaker bias, in which initial speakers set the tone for discussion. However, faculty were not discouraged from making additional comments or claiming or echoing their support for particular comments after the presentation. Most survey respondents said that this practice prompted meeting attendees to focus on more objective criteria (78%), improved the climate of the meeting (80%), and reduced the first speaker effect (67%) (see fig. S2).

Additionally, some commented in open-ended survey responses that meeting time was used efficiently to quickly identify candidates with strong consensus and then spend more time on those with high variance in ratings [whom we determined were disproportionately women (see table S1.B)]. Taken together, these results suggest that beginning the faculty meeting with rubric results reduced interactional bias emergent in the faculty meeting. By opening with a neutral reading of the full set of both positive and negative rubric comments, the impact was blunted of any first speakers, often senior men, attempting to vociferously promote or shoot down a candidate. The faculty meeting format may have mitigated gender bias in the research productivity scores and the selection of finalists; 47% of women semifinalists and 37% of men semifinalists advanced to the finalist stage.

POLICY TEMPLATE

In light of our findings that gender bias remains endemic even in this seemingly objective evaluation process, it is vital that rubric usage be accompanied by strategic application in departmental meetings to counteract individual bias and check interactional bias during the discussion of candidates. Our results suggest that using rubrics according to this process framework can improve diversity in hiring. Thus, we recommend a strategic and sociologically astute use of rubrics as a department self-study tool within the context of a holistic evaluation of the short-listed candidates (see the second figure).

We have studied this process with regard to gender. Given the otherwise limited diversity among candidates in our study, we were unable to address whether rubrics could also be a tool to promote and check on the fairness of evaluations with regard to race/ethnicity or other minoritized identities. This suggests priorities for future research. ■

REFERENCES AND NOTES

1. M. Blair-Loy, E. Cech, *Misconceiving Merit: Paradoxes of Excellence and Devotion in Academic Science and Engineering* (Univ. Chicago Press, 2022).
2. J. Roy, C. Wilson, A. Erdiaw-Kwasie, C. Stuppard, "Engineering and engineering technology by the numbers 2019" (American Society for Engineering Education, 2020).
3. J. Falkenheim, "Doctoral Recipients from U.S. Universities: 2019" (National Science Foundation, 2019).
4. K. O'Meara, D. Culpepper, L. L. Templeton, *Rev. Educ. Res.* **90**, 311 (2020).
5. University of Michigan, in *Advance Program,* University of Michigan, Ed. (University of Michigan Office of the Provost, 2018).
6. E. Uhlmann, G. L. Cohen, *Psychol. Sci.* **16**, 474 (2005).
7. R. H. Thaler, C. R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale Univ. Press, 2008).
8. National Research Council, *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty* (National Academies Press, 2010).
9. V. Koltun, D. Hafner, *PLOS ONE* **16**, e0253397 (2021).
10. C. Wenneras, A. Wold, *Nature* **387**, 341 (1997).
11. E. R. Andersson, C. E. Hagberg, S. Hägg, *Front. Res. Metr. Anal.* **6**, 594424 (2021).
12. M. Foschi, *Annu. Rev. Sociol.* **26**, 21 (2000).
13. S. J. Correll, K. R. Weisshaar, A. T. Wynn, J. D. Wehner, *Am. Sociol. Rev.* **85**, 1022 (2020).
14. T. Schmader, J. Whitehead, V. H. Wysocki, *Sex Roles* **57**, 509 (2007).
15. J. M. Madera, M. R. Hebl, H. Dial, R. Martin, V. Valian, *J. Bus. Psychol.* **34**, 287 (2019).

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, grant 1661306 (M.B.-L.). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. S.I.F. is a cofounder and scientific adviser for MelioLabs and holds equity in the company.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm2329

10.1126/science.abm2329

Science
AAAS

Supplementary Materials for

Can rubrics combat gender bias in faculty hiring?

Mary Blair-Loy *et al.*

Corresponding authors: Mary Blair-Loy, mblairloy@ucsd.edu; Stephanie I. Fraley, sifraley@ucsd.edu

Science **377**, 35 (2022)
DOI: 10.1126/science.abm2329

The PDF file includes:

Supplementary Materials for

Title: Gender Bias and Hiring Rubrics Evaluating Engineering Faculty Candidates

Authors: Mary Blair-Loy[1]*, Olga V. Mayorova[2], Pamela C. Cosman[3], Stephanie I. Fraley[4]*

Correspondence to: mblairloy@ucsd.edu, sifraley@ucsd.edu

This PDF file includes:

Materials and Methods

The Materials and Methods are explained within the main text, figures and supplemental. Given the small sample size, the limited pool of potential subjects, and the risk of re-identification of subjects, our agreement with the IRB at our university to protect human subject confidentiality precludes making the full raw datasets available. If scholars would like to access the full data sets for research purposes, the two-step procedure for them would be as follows. (1) They should contact their own IRB to see if they would need to seek approval from that body. (2) After that approval has been granted (or has been determined not to be required), they should then contact one of the corresponding authors (Mary Blair-Loy or Stephanie Fraley) and request access. Fraley or Blair-Loy would then refer the inquiring researcher to the university's contracting office, which would then work with the researcher to establish a data transfer agreement in which they would agree not to attempt to re-identify the subjects, thus preserving the protections for the data subjects.

Fig. S1.

I have read the candidate's: (check all that apply)

CV Research statement Letters of recommendation Teaching statement
Contributions to diversity statement

Rank the Candidate on evidence of research productivity.

Excellent Good Neutral Fair Poor
Comment:

Rank the candidate on evidence of high impact scholarship.

Excellent Good Neutral Fair Poor
Comment:

Rank the candidate's ability to teach undergraduate and graduate courses in [redacted] engineering and its fundamentals. (List courses that the applicant would be well qualified to teach based on their training.)

Excellent Good Neutral Fair Poor
Comment:

Rank the candidate on their demonstrated service contributions that promote diversity and equal opportunity (e.g. developing strategies or programs to engage/recruit/mentor underrepresented groups, advance equitable access, etc.).

Excellent Good Neutral Fair Poor
Comment:

Rank the candidate on their potential for collaboration. (List likely collaborators, areas of collaboration, or departments that the candidate's research would likely promote interaction with.)

Excellent Good Neutral Fair Poor
Comment:

Overall impression of the candidate.

Excellent Good Neutral Fair Poor
Comment:

Figure S1: Rubric for faculty evaluation of semi-finalist job candidates

Fig. S2.

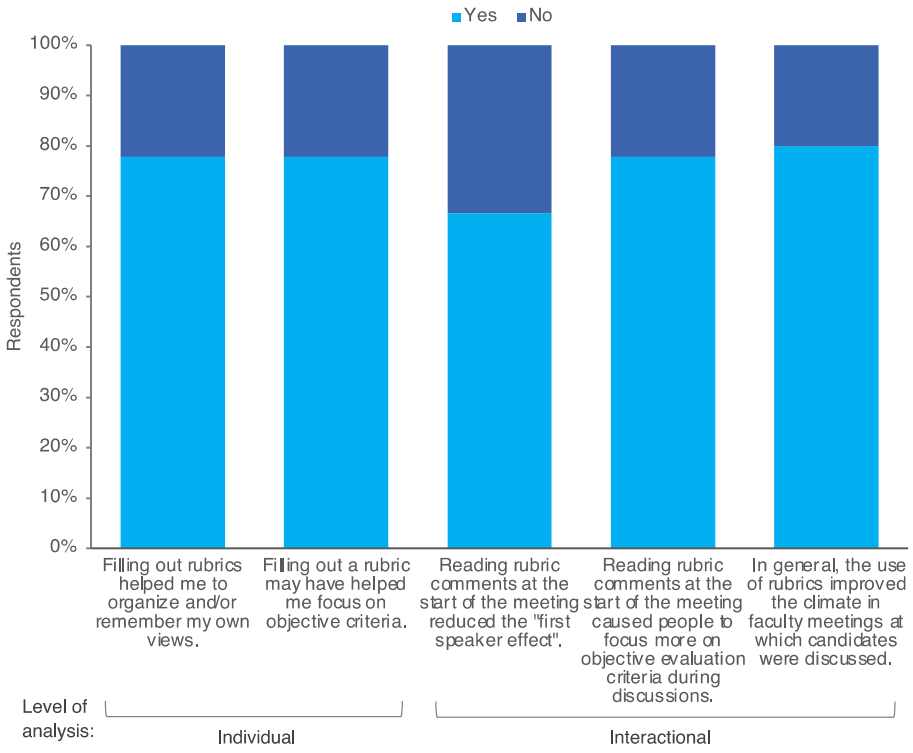


Figure S2: Results of faculty survey

Graph shows frequencies of responses from faculty survey (N=19).

Table S1.A Comparison of rubric category scores for men and women candidates[1]

	Mean (Std. Deviation) Men	Women	t-value
Research Productivity Score[2]	**3.5 (0.34)**	**3.1 (0.60)**	**3.066****
Research Impact Score	3.3 (0.37)	2.9 (0.41)	4.258***
Teaching Ability Score	2.7 (0.72)	2.7 (0.68)	-0.099
Contributions to Diversity Score	2.5 (0.42)	2.8 (0.47)	-2.521*
Potential for Collaboration Score	2.7 (0.49)	2.8 (0.55)	-0.588
Overall Impression Score	3.1 (0.71)	3.0 (0.62)	0.631
N	*30*	*32*	

*** p<0.001, ** p<0.01 (Two-tailed t-tests). Note: Equal variances are assumed when calculating t-value, except for Research Productivity score (based on Levene's test of equality of variances).

5

Table S1.B Comparison of variances of raters' research productivity scores for men and women candidates[3]

	Mean (Std. Deviation) Men	Women	t-value
Research Productivity Score Raters' Variance for each candidate	0.4 (0.29)	0.7 (0.54)	2.608*
N	*30*	*32*	

* p<0.05 (Two-tailed t-tests). Note: Equal variances are not assumed when calculating t-value (based on Levene's test of equality of variances).

[1] IBM SPSS 28.0.0.0 syntax: T-TEST GROUPS=Gender(1 2) /MISSING=ANALYSIS /VARIABLES=R1resprod_score R1resim_score R1teach_score R1divers_score R1collab_score R1overall_score /ES DISPLAY(TRUE) /CRITERIA=CI(.95).

[2] For the research productivity category, the number of raters providing scores per candidate ranges from 6 to 21, with a mean of = 13.5, and a median of 12. There is no statistically significant gender difference in the number of raters per candidate. We analyze the research productivity category in detail; see Table S2 and Figure 1.

[3] IBM SPSS 28.0.0.0 syntax: T-TEST GROUPS=Gender(1 2) /MISSING=ANALYSIS /VARIABLES=R1resprod_variance /ES DISPLAY(TRUE) /CRITERIA=CI(.95). Each candidate received their own variance score based on their own raters. We then compared means of these variances by candidate gender.

Table S2: OLS regressions predicting rubric scores of research productivity of candidates[4] [5]

	Model 1	Model 2	Model 3	**Model 4**
	B	B	B	**B**
	(SE)	(SE)	(SE)	**(SE)**
Female	**-0.36** **	**-0.53** *	**-0.29** *	**-0.7** **
	(0.12)	(0.22)	(0.13)	**(0.25)**
Years post-Ph.D.	**-0.04**	-0.04	-0.02	**-0.05**
	(0.03)	(0.03)	(0.03)	**(0.03)**
CV Measures of Productivity:				
No. articles	**0.02** **	0.01	-	**-**
	(0.01)	(0.01)		
H-index	-	-	0.03 **	**0.01**
			(0.01)	**(0.01)**
Interactions:				
Female*No. articles	**-**	0.01	-	**-**
		(0.01)		
Female*H-index	**-**	-	-	**0.04** *
				(0.02)
Constant	**3.35** ***	3.49 ***	3.21 ***	**3.51** ***
	(0.15)	(0.21)	(0.17)	**(0.23)**
Model Fit Statistic (F-test)	**6.7** ***	5.2 ***	6.2 ***	**5.9** ***
R-square	**0.26**	0.27	0.23	**0.28**
R-square change (F-test)		0.01		0.05 *
N	***62***	*62*	*62*	***62***

*** p<0.001, ** p<0.01, * p<0.05 (One-tailed t-tests for regression coefficients and F-tests for model fit and R-square change). Models 1 and 4 are best-fitting and are graphed in Fig 1. In Model 4, the significant interaction between Female and H-index indicates that the effect of gender on rubric scores is contingent upon the values of H-index.

[4] IBM SPSS 28.0.0.0 syntax: Models 1 and 2: REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS CI(95) BCOV R ANOVA COLLIN TOL CHANGE /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT R1resprod_score /METHOD=ENTER Female post_phd_yrs N_artic /METHOD=ENTER Female_Nartic. Models 3 and 4: REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /MISSING LISTWISE /STATISTICS COEFF OUTS CI(95) BCOV R ANOVA COLLIN TOL CHANGE /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT R1resprod_score /METHOD=ENTER Female post_phd_yrs WSCI_hindex /METHOD=ENTER Female_Hindex.

[5] We chose to analyze the "research productivity" rubric evaluation category because it is a significant dimension of quality for R1 universities and, in our data, was significantly correlated with the calibration metrics with broad salience to our case department: number of articles (which was also discussed in the written comments) and H-index, a widely used metric of productivity and impact. (Despite critique and proposed alternative indices, the H-index remains standard because of its simplicity of calculation and interpretation, its provision of a single number that can be compared across faculty of all seniority levels, and its dominant position in citation databases. See reference (9)). We encourage future research to examine the usefulness of other metrics. We do not analyze the rubric category "research impact" because it was not significantly correlated with either calibration metric presented here. We suspect that evaluators' judgements of "research impact" for the predominately junior candidates in this study are subjectively based upon a guess of impact potential.

Table S3: Chi-square tests of rubric comment categories by candidate gender[6] [7] [8]

	Men %		Women%		
	At least One	None	At least One	None	Chi-Square
Positive Comment(s)	86%	14%	63%	38%	4.115*
Negative Comment(s)	25%	75%	50%	50%	3.948*
Standout Language	32%	68%	9%	91%	4.838*
Doubt-Raiser(s)	25%	75%	13%	87%	1.558
N	*28* [9]		*32*		

* $p<0.05$ (Two-tailed tests).

[6] IBM SPSS 28.0.0.0 syntax: CROSSTABS /TABLES=PosCommentY NegCommentY StandoutLang DoubtRaiser BY Gender /FORMAT=AVALUE TABLES /STATISTICS=CHISQ /CELLS=COUNT COLUMN /COUNT ROUND CELL.

[7] We chose hand coding, as research suggests that hand coding is superior to computer assisted coding for studies like ours with novel analyses (see reference 13). Our coding process proceeded as follows. First, gender indicators, such as pronouns, were redacted from the comment data set. Next, the PI -- sociologist MBL -- and engineers SIF and PCC constructed coding categories. Then, MBL coded the comments. She consulted with Co-PIs SIF and PCC when comments or categorizations were unclear to her, and the three reached consensus. As a check, OVM separately coded the comments. An inter-coder reliability analysis revealed statistically significant ($p \leq 0.001$), strong to moderate agreement. Kappa coefficients for each comment type follow: Positive comment(s) 0.822; Negative comment(s) 0.859; Standout Language 0.722; Doubt-raiser(s) 0.624. We determined that the original coding was accurate.

[8] We analyze qualitative comment types as dichotomous variables (At least one vs. None). These variable values are nominal summaries of qualitative information. The number of comments ranges from 0 to 9, but most candidates received three or fewer comments. By gender, candidates receive similar medians (3 for men and 2 for women) and similar means of numbers of comments (3.3 for men and 2.8 for women; this difference of means is not statistically significant). As a robustness check we conducted a difference of means by gender analysis on the count variables. This yielded the same results as our original dichotomous variable Chi-square analysis: Mean differences by gender for normalized comment count in each category (proportion of comments in a category of all comments received the candidate) are statistically significant for all comment categories except doubt-raisers. However, we believe the conservative estimates of the dichotomous data are more appropriate due to the skewness of the normalized comment counts.

[9] Two men candidates are missing because the department that provided anonymized data had lost the records of those 2 qualitative comment fields.

Table S4: Descriptive statistics for semi-finalist candidates by gender[10]

	Male		Female		Difference of Means	All	
	Mean	SD	Mean	SD		Mean	SD
Years post-Ph.D. (Seniority)	5.0	2.0	5.5	3.0	-0.5	5.3	2.5
No. of peer-reviewed articles	16.8	9.0	16.7	13.8	0.1	16.7	11.6
Average No. of peer-reviewed articles per yr	3.1	1.8	2.8	2.1	0.3	2.9	2.0
H-Index[11]	14.0	7.4	10.9	6.7	3.1 *	12.4	7.1
Rubrics score for research productivity	3.5	0.3	3.1	0.6	0.4 **	3.3	0.5
Total N	30		32			62	

** $p<0.01$, * $p<0.05$ (Two-tailed t-tests).

5

[10] IBM SPSS 28.0.0.0 syntax: T-TEST GROUPS=Gender(1 2) /MISSING=ANALYSIS /VARIABLES=post_phd_yrs N_artic Articles_per_year WSCI_hindex R1resprod_score /ES DISPLAY(TRUE) /CRITERIA=CI(.95). FREQUENCIES VARIABLES=post_phd_yrs N_artic Articles_per_year WSCI_hindex R1resprod_score /FORMAT=NOTABLE /STATISTICS=STDDEV MEAN /ORDER=ANALYSIS.

[11] Candidates at the lower end of the H-Index distribution (0 to 5 points) tend to have less seniority; for these eight candidates, the median is 2.0 and the mean is 2.5 years since graduation). In contrast, candidates at the higher end of the H-Index distribution H-index (15 to 20 points) have higher seniority; for these ten candidates, the median is 5 and the mean is 5.3 years since graduation. More senior candidates have had twice as much time to publish paper *and* to collect citations from their earlier papers.

Table S5: Faculty Survey Frequencies

	N Replies Tenured	N Replies Non-Tenured	N Replies - All Faculty	% Tenured	% Non-Tenured	% All Faculty
1. For each item below, mark whether this statement is true for you:						
Filling out rubrics helped me to organize and/or remember my own views.	10	4	14	77%	80%	78%
Filling out a rubric may have helped me focus on objective criteria.	10	4	14	77%	80%	78%
I wrote comments because that allowed me to convey my opinions to the search committee.	11	4	15	85%	80%	83%
I did not find filling out the rubrics to be valuable.	0	0	0	0	0	0
Total Faculty Voted	*13*	*5*	*18*	*13*	*5*	*18*
2. In faculty meetings, as in many other discussion and debate settings, people who speak first or early in the meeting tend to have an outsized impact on the outcome of the discussion. This effect can be called the "first speaker effect". For those faculty who participated in faculty meetings about candidates BOTH BEFORE AND AFTER RUBRICS WERE INSTATED, mark the following statement you most agree with for this question and the next two questions. For faculty who did not participate in faculty meetings both before and after rubrics were introduced, mark "not applicable" for this and the following two questions. Reading rubric comments at the start of the meeting...						
Reduced the "first speaker effect"	8	2	10	62%	40%	56%
Increased the "first speaker effect"	0	0	0	0%	0%	0%
Had no effect on the "first speaker effect"	5	0	5	38%	0%	28%
Not applicable	0	3	3	0%	60%	17%
Total Faculty Voted	*13*	*5*	*18*	*13*	*5*	*18*
3. **Reading rubric comments at the start of the meeting...**						
Caused people to focus more on objective evaluation criteria	10	4	14	77%	80%	78%
Caused people to focus less on objective evaluation criteria	0	0	0	0%	0%	0%
Had no effect on whether people focused on objective criteria	3	1	4	23%	20%	22%
Not applicable	0	0	0	0%	0%	0%
Total Faculty Voted	*13*	*5*	*18*	*13*	*5*	*18*
4. 'Climate' can be defined as "Behaviors in a workplace, ranging from subtle to cumulative to dramatic, that can influence whether an individual feels listened to, valued, and treated fairly and with respect. In general, the use of the rubrics...						
Improved the climate in faculty meetings at which candidates were discussed.	10	2	12	83%	40%	71%
Worsened the climate in faculty meetings at which candidates were discussed.	0	0	0	0%	0%	0%
Had no effect on the climate in faculty meetings at which candidates were discussed.	1	2	3	8%	40%	18%
Not applicable	1	1	2	8%	20%	12%
Total Faculty Voted	*12*	*5*	*17*	*12*	*5*	*17*

Table S6: Faculty Survey Response Analysis

	Survey Participants	Non-Participants	Chi-Square	P-Value [a]	Fisher's Exact Significance [a]
Women	26.30%	7.10%	1.992	0.158	0.209
Tenured	78.90%	78.60%	0.001	0.979	1.000
N	*19*	*14*			

[a] Two-tailed tests.

Table S7: Regression analysis guided by the grant proposal that has funded this research

Our research design and hypotheses for the regression analysis are guided by Aim 1 of the grant proposal that has funded this research (National Science Foundation #1661306).

Aim 1 includes the following research question: Are women more likely to receive lower ratings than comparable men in the areas of research productivity and research impact?

Aim 1 hypotheses include the following:

- We hypothesize that despite the added structure of focused candidate evaluation rubrics, gender bias will be detected in candidate [ratings].
- We hypothesize that gender bias occurs within rubric evaluations in a systematic way, mirroring stereotypical assumptions about gendered competence.

Table S8: Supplemental References

Topic	Reference
Gender Inequality in hiring	C. Isaac, B. Lee, M. Carnes, Interventions That Affect Gender Bias in Hiring: A Systematic Review. *Acad Med* **84**, 1440-1446 (2009).
	C. T. Begeny, M. K. Ryan, C. A. Moss-Racusin, G. Ravetz, In Some Professions, Women Have Become Well Represented, Yet Gender Bias Persists—Perpetuated by Those Who Think It Is Not Happening. *Science Advances* **6**, (2020).
	National Research Council, *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty*. (The National Academies Press, Washington, DC, 2010), pp. 384.
	K. Hamrick, "Women, Minorities, and Persons with Disabilities in Science and Engineering: 2021," (National Science Foundation, Alexandria, VA, 2021).
Innovation and other positive associations with diversity of STEM units	C. Puritty *et al.*, Without Inclusion, Diversity Initiatives May Not Be Enough. *Science* **357**, 1101-1102 (2017).
	C. C. Perez, *Invisible Women: Data Bias in a World Designed for Men*. (Harry N. Abrams, 2019), pp. 272.
	J. Price, The Effect of Instructor Race and Gender on Student Persistence in STEM Fields. *Economics of Education Review* **29**, 901-910 (2010).
Case study methodology	M. Lamont and P. White. "Workshop on Interdisciplinary Standards for Systematic Qualitative Research." National Science Foundation. (2005). https://www.nsf.gov/sbe/ses/soc/ISSQR_workshop_rpt.pdf (accessed March 27, 2022)
Studies, treatises and websites promoting use of rubrics in candidate assessment	S. D. Gumbert *et al.*, Reliability of a Faculty Evaluated Scoring System for Anesthesiology Resident Applicants (Original Investigation). *Journal of Clinical Anesthesia* **31**, 131-136 (2016).
	D. Doyle, G. Locke, "Lacking Leaders: The Challenges of Principal Recruitment, Selection, and Placement," (Thomas B. Fordham Institute, Washington, DC, 2014).
	M. Braileanu *et al.*, Structured Curriculum Vitae Scoring as a Standardized Tool for Selecting Interview Candidates for Academic Neuroradiology Faculty Positions. *Current Problems in Diagnostic Radiology* **49**, 377-381 (2020).

	E. Fine, J. *et al.*, in *Gender research: Gender transformation in the academy,* V. V. Demos, C. W. Berheid, M. T. Segal, Ed. (Emerald Insight, 2014), vol. 19, pp. 267–289.
	D. LaVaque-Manty, A. J. Stewart, in *Gendered innovations in science and engineering* L. Schiebinger, Ed. (Stanford University Press, Palo Alto, CA, 2008), pp. 165–181.
	E. Fine, J. Handelsman, *Searching for Excellence and Diversity: A Guide for Search Committees*. (Women in Science & Engineering Leadership Institute, 2012).
	University of Maryland ADVANCE. Search Committee Resources: Constructing Clear Candidate Evaluation Criteria & Using A Rubric. https://advance.umd.edu/node/210. Accessed 3/27/2022.
	University of Washington, "Handbook for Best Practices for Faculty Searches: Online Toolkit. https://www.washington.edu/diversity/faculty-advancement/handbook/toolkit/. Accessed 3/27/2022.
Studies finding gender bias in metrics such as H-index	Matthew R.E. Symonds, et al. Gender Differences in Publication Output: Towards an Unbiased Metric of Research Performance. *PLoS ONE* **1**, e127 (2006).
	A. A. Pashkova *et al.*, Gender in Academic Anaesthesiology. *Acta Anaesthesiol Scand* **57**, 1058-1064 (2013).
	E. B. Holliday *et al.*, Gender Differences in Publication Productivity, Academic Position, Career Duration, and Funding among U.S. Academic Radiation Oncology Faculty. *Acad Med* **89**, 767-773 (2014).
Study finding gender bias in interruptions of faculty candidate job talks	M. Blair-Loy, L.E. Rogers, D. Glaser, Y. L. A. Wong, D. Abraham and P.C. Cosman. Gender in Engineering Departments: Are There Gender Differences in Interruptions of Academic Job Talks? *Social Sciences 6*(1): 1-19 (2017). https://www.mdpi.com/2076-0760/6/1/29