JOURNAL OF WOMEN'S HEALTH
Volume 26, Number 4, 2017
© Mary Ann Liebert, Inc.
DOI: 10.1089/jwh.2016.6021

Original Article

Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers' Critiques

Wairimu Magua, PhD, MS,[1] Xiaojin Zhu, PhD, MS,[2] Anupama Bhattacharya,[3] Amarette Filut, BS,[3]
Aaron Potvien, MS,[4,5] Renee Leatherberry,[3] You-Geon Lee, PhD,[6] Madeline Jens, BA,[3]
Dastagiri Malikireddy, BS,[3] Molly Carnes, MD, MS,[3,7,8] and Anna Kaatz, PhD, MPH[3]

Abstract

Background: Women are less successful than men in renewing R01 grants from the National Institutes of Health. Continuing to probe text mining as a tool to identify gender bias in peer review, we used algorithmic text mining and qualitative analysis to examine a sample of critiques from men's and women's R01 renewal applications previously analyzed by counting and comparing word categories.

Methods: We analyzed 241 critiques from 79 Summary Statements for 51 R01 renewals awarded to 45 investigators (64% male, 89% white, 80% PhD) at the University of Wisconsin-Madison between 2010 and 2014. We used latent Dirichlet allocation to discover evaluative ''topics'' (*i.e.*, words that co-occur with high probability). We then qualitatively examined the context in which evaluative words occurred for male and female investigators. We also examined sex differences in assigned scores controlling for investigator productivity.

Results: Text analysis results showed that male investigators were described as ''leaders'' and ''pioneers'' in their ''fields,'' with ''highly innovative'' and ''highly significant research.'' By comparison, female investigators were characterized as having ''expertise'' and working in ''excellent'' environments. Applications from men received significantly better priority, approach, and significance scores, which could not be accounted for by differences in productivity.

Conclusions: Results confirm our previous analyses suggesting that gender stereotypes operate in R01 grant peer review. Reviewers may more easily view male than female investigators as scientific leaders with significant and innovative research, and score their applications more competitively. Such implicit bias may contribute to sex differences in award rates for R01 renewals.

Keywords: women's career advancement, NIH funding, gender differences

Introduction

Advancing women to institutional leadership in academic medicine is important because diversity drives innovation and productivity,[1–4] and because female leaders are more likely to promote research on women's health.[5–10] Despite the clear benefits of female leaders, women remain disproportionately underrepresented in leadership positions across medical specialties, even among those such as obstetrics and gynecology that attract higher numbers of women than men in early career stages.[11–18] Reflecting this, a 2014 report on the state of women in academic medicine by the

Departments of [1]Population Health Sciences and [2]Computer Science, University of Wisconsin-Madison, Madison, Wisconsin.
[3]Center for Women's Health Research, University of Wisconsin-Madison, Madison, Wisconsin.
[4]Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin.
[5]Health Innovation Program, University of Wisconsin-Madison, Madison, Wisconsin.
[6]Wisconsin Center for Education Research, University of Wisconsin-Madison, Madison, Wisconsin.
[7]Departments of Medicine, Psychiatry, and Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, Wisconsin.
[8]William S. Middleton Veterans Hospital, Madison, Wisconsin.

Association of American Medical Colleges (AAMC) shows that women make up just 21% of associate professors, 34% of full professors, 15% of department chairs, and 16% of academic deans in academic medicine, even though women have been near parity with men as medical students for the past 20 years (45.9%–49.7%).[18,19] Multiple reports show that higher attrition and lower promotion rates for women than men perpetuate the gender/leadership gap.[5,11–13]

Career persistence and advancement in academic medicine, particularly to leadership, is predicated on obtaining funding for research, and the National Institutes of Health's (NIHs) R01 grant is the ''gold standard'' for research awards.[20–22] R01s provide 3–5 years of funding and can be renewed. Since 1998, the NIH has reported lower funding success rates for renewal applications of R01 (or equivalent) awards for female than male principal investigators (PIs).[23] A study by Pohlhaus et al. on NIH award outcomes confirmed significantly lower funding success rates for women than men who submitted R01 renewals.[22] The NIH leadership acknowledges that unconscious bias may operate in the peer review process.[24,25] For example, Collins (NIH Director) and Tabak (Deputy Director) suggest that the ''Matthew effect''—a combination of accumulated advantage and status that disproportionately propels elite scientists toward success—could explain gaps in NIH award rates.[24] Other studies by Carnes et al. on the nomination, evaluation, and selection processes for high status NIH awards, such as the Director's Pioneer Award and the Clinical and Translational Science Awards (CTSAs), make a strong theoretical case that female applicants may be disadvantaged in NIH peer review from bias due to gender stereotypes that may implicitly impact the judgment of reviewers.[26,27] Science has historically been a male dominated field, and gender stereotypes characterize men as intrinsically endowed with the traits associated with ability in science.[28–31] Several studies have found that stereotype-based bias can influence the written evaluations of men and women in such male-typed fields by engendering doubt about women's ability and competence.[32–35]

Our prior work suggests that text analysis of NIH critiques can provide a window into peer reviewers' decision-making processes and show evidence of cognitive bias, if it occurs.[10,36,37] In studies of NIH outcomes spanning 2007–2009 (NIHs former peer review criteria) and 2010–2014 (NIHs enhanced peer review criteria), we found that critiques of female PIs' R01 applications, particularly renewals, were more favorable despite the receipt of similar or worse proposal scores than comparable men.[38,39] An important limitation of this work is that it relied on simple word count software that assessed the percentage of words in documents from predefined categories,[40] which did not permit examination of reviewers' statements verbatim, or the context in which words were used.

Using state-of-the-art text mining algorithms to discover verbatim linguistic patterns in reviewers' critiques of R01 renewal applications provides an innovative solution to address these limitations.[41] In this study, we analyzed a sample of critiques from unfunded and funded R01 renewal applications spanning from 2010 to 2014 using a data-driven algorithm called latent Dirichlet allocation (LDA).[42] Also referred to as ''topic modeling,'' LDA is an algorithmic Bayesian statistical model that can extract themes or topics in documents.[42] Since human cognition remains ''the gold standard'' for interpreting the contextual meaning of words,[43,44] we developed a technique to extract the text surrounding evaluative words identified with LDA and engaged experts in stereotype-based bias to qualitatively analyze their meaning.[45] The purpose of this study was to explore the extent to which contextually meaningful evaluative differences in critiques of male and female investigators' applications exist, and whether linguistic patterns would differ in ways that suggest gender bias may operate in NIH peer review process for R01 renewals. We also analyzed scores assigned to male and female investigators' renewal applications for significant differences in models that controlled for productivity (*i.e.*, impact factors for publications, and numbers of prior NIH awards).

Data analysis

Data

All aspects of this research were approved by the University of Wisconsin-Madison Institutional Review Board (IRB). The data used in this study are a subset of NIH Summary Statements (*i.e.*, files that contain reviewers' critiques) previously collected by Kaatz et al.[37] Using NIH publicly accessible database, Research Portfolio Online Reporting Tools (*i.e.*, project ''RePORTER''), Kaatz et al. identified PIs at the University of Wisconsin-Madison with R01 awards funded on the initial submission or after revision between 2010 and 2014. Of the 278 PIs identified, 132 (47%) donated their Summary Statements. NIH only provides information about funded applications, so Kaatz et al. could not request Summary Statements from terminally unfunded applications or assess success rates. From this data set, we extracted Summary Statements from R01 renewal (Type 2) applications (49 PIs). We excluded four PIs' data from analysis because their Summary Statements were dated before the NIH altered its peer review process (*i.e.*, in 2009–2010) or had incomplete information.[38,39] Of the remaining 45 PIs, 64% were men ($N = 29/45$), 89% were white ($N = 40/45$), 80% held PhDs ($N = 36/45$), and 91% were ''experienced'' investigators[46] ($N = 41/45$), meaning they had previously obtained an R01 or equivalent award at the time of application (Table 1). The participating PIs represented 23 different departments; 61% were in the School of Medicine and Public Health (SMPH; $N = 28/46$), and 22% were in the College of Agriculture and Life Sciences (CALS; $N = 10/46$). We found no significant differences in the distribution of participants (P) who donated their Summary Statements and nonparticipants (NP) who did not by sex (P, male: 63%, $N = 29/46$; NP, male: 76%, $N = 45/59$) or by race (P, white: 87%, $N = 40/46$ and NP, white: 86%, $N = 51/59$).

Our final analytic sample included 79 Summary Statements from 51 R01 awards from these 45 PIs. Approximately one-quarter of grants ($N = 14/51$; 27%) were for clinical research. Twenty-four of the 51 grants (47%) were funded as resubmissions. For these R01s, we had 28 Summary Statements from the initially unfunded submission, and 24 Summary Statements from the subsequent funded submission. The remaining 27 Summary Statements in our sample derived from R01s funded on the first submission ($N = 27/51$, 53%). Our sample had a total of 241 critiques (85 from unfunded, 156 from funded application Summary Statements). Generally, each Summary Statement contained three to five sets of reviewers' critiques regarding the ''strengths'' and ''weaknesses'' of the proposed work's overall impact, significance, investigator(s), innovation, approach, and environment; Summary Statements also included an impact/priority score and scores for each of the other criteria (Table 2).[47,48] Summary Statements were de-identified using R software.[49]

Table 1. Descriptive Statistics for Background Characteristics of Participating Male and Female PIs

	Male PIs, N=*29*	*Female PIs,* N=*16*
Background characteristics, *n* (%)		
White	26 (90)	14 (88)
PhD	20 (69)	16 (100)
Experienced PI[a]	29 (100)	15 (94)

[a] A principal investigator (PI) submitting an NIH R01 application is considered experienced if they have received a prior R01 or equivalent major award.

We extracted scores from Summary Statements and entered them into a database of applicant and application information. Kaatz et al. previously identified each PI sex, experience level, and training background.[36,37] For this study, we added productivity information, regarding PIs' previous NIH awards, and their publication impact. We obtained data on previous NIH awards by searching NIH RePORTER database for each PIs total numbers of NIH grants, by mechanism and year, received before the year of their R01 in our database. We grouped these data into eight categories: Initial R01s (Type 1); Renewal R01s (Type 2), all other R(esearch)-Level Awards; P(rogram) awards; T(raining) awards; Mid-Career Development (K) awards and Mentored Career Development (K) awards; and Pre/Post-Doctoral F(ellowship) awards. We collected information about the impact of each PIs' publications from two sources: from Thomson Reuters *Web of Science*[50] we retrieved the h-index (a widely used metric based on the number of highly cited articles and the number of citations for each highly cited article, where a higher score indicates higher productivity/impact), and from NIH public-access iCite website (https://icite.od.nih.gov/), we retrieved NIH Relative Citation Ratio (RCR; a measure of the number of articles, and citations for articles, adjusted for field norms, for publications indexed in PubMed, where a higher score means higher productivity/impact).[51]

Text analysis

Quantitative method—LDA to identify evaluative topics in critiques. Using the R open source statistical software program,[49] we separated each Summary Statement file into two documents, one containing the combined strengths sections and the other containing the combined weaknesses sections from all critiques because the same words may contextually vary based on the assumed positivity and negativity of these sections. For topic modeling with LDA, we used the R package ''mallet'' that implements the Machine Learning for Language Toolkit (MALLET).[52,53] Before submitting documents in our corpus to LDA, we removed uninformative words called ''stop words'' using a predefined list, provided in MALLET, of common English adverbs, conjunctions, pronouns, and prepositions. Multiple LDA models were fit, using a range of 2 to 40 topics. We used 1 million sampling iterations for model convergence. We chose the optimal model with 26 topics, using the maximization of the harmonic mean of the model log likelihoods as our model selection criterion.[54]

The number of topics generated by LDA can be specified by the user. Topics reflect words with the highest probability of co-occurrence in a collection of documents. Each topic reveals a different theme. It is standard practice for researchers to select a subset of topics relevant to their research question.[55] Using a modified Delphi procedure,[56–59] we asked a panel of seven peer review experts to evaluate the interpretability of the 26 topics and came to a consensus that one topic was representative of reviewers' evaluative commentary in critiques (*i.e.*, it contained adjectives characterizing the qualifications of PIs and quality of the proposed research); we labeled this topic ''sentiment.'' The remaining 25 topics were about specific areas of research (*e.g.*, dysphagia, molecular biology) and were deemed irrelevant to our research question.

To interpret the meaning of words in our sentiment topic, particularly when they co-occur, we first combined the top 30 words from the sentiment topic into all possible pairs ($N=435$; *e.g.*, ''highly'' + ''innovative''). Then, we used the chi-square test to rank co-occurring words based on how disproportionately they were used in critiques of men's and women's applications; we retained the top 100 co-occurring words used in higher proportions in critiques of men's, and of women's applications, respectively (Table 3). Next, we wrote a program to extract the sentence(s) surrounding co-occurring words, which we called ''context windows.'' In this technique, we allowed co-occurring words to appear sequentially or up to five words apart. We imposed a five-word distance limit because as the distance between co-occurring words increases, it becomes less likely that they are meaningfully related. We used the R packages ''stringr''[60]

Table 2. Average Scores (and Standard Deviations) for Male and Female PIs' Unfunded and Funded R01 Type 2 Applications

(N=*applications and critiques)*	*Male PIs*		*Female PIs*	
	Unfunded (N=*13/39)*	*Funded* (N=*32/95)*	*Unfunded* (N=*7/24)*	*Funded* (N=*18/54)*
Priority score	33.85 (8.37)	18.84 (6.18)	34.71 (6.95)	23.00 (6.89)
Approach score	3.05 (1.26)	2.18 (1.04)	3.54 (1.38)	2.63 (0.94)
Significance score	2.26 (1.04)	1.75 (0.94)	3.00 (1.35)	2.00 (0.78)
Innovation score	2.54 (1.33)	1.82 (0.86)	3.08 (1.32)	2.31 (0.97)
Investigators score	1.44 (0.68)	1.22 (0.55)	1.42 (0.50)	1.37 (0.62)
Environment score	1.26 (0.50)	1.29 (0.48)	1.48 (0.59)	1.35 (0.52)

Table 3. The Top 30 Words from the Sentiment Topic, and Top 20 Co-Occurring Words (Out of 100) from Chi-Square Rankings That Occurred in Higher Proportions in Critiques of Male and Female PIs' R01 Renewal Applications, Respectively

	Co-occurring words from sentiment topic[a]	
Top words from sentiment topic	*Critiques of male PIs' applications*	*Critiques of female PIs' applications*
Proposed	Highly, proposal	Excellent, work
PI	Innovative, proposal	Dr, research
Important	Field, PI	Excellent, research
Innovative	Innovative, work	Development, mechanisms
Highly	Approaches, potential	Approach, excellent
Research	Important, study	Dr, proposal
Mechanisms	PI, studies	Mechanisms, proposed
Studies	Application, PI	Environment, innovative
Strong	Approach, highly	Established, research
Excellent	Highly, innovative	Investigator, mechanisms
Environment	Potential, proposal	Model, strong
Model	Application, hypothesis	Model, understanding
Outstanding	Application, established	Development, potential
Potential	Established, investigator	Expertise, proposal
Dr	Important, outstanding	Proposal, proposed
Approach	Approach, strong	PI, project
Application	Outstanding, study	Important, investigator
Significant	Field, work	Approach, model
Work	Innovative, PI	Outstanding, proposed
Understanding	Highly, PI	Project, proposed
Proposal		
Field		
Development		
Study		
Hypothesis		
Established		
Project		
Expertise		
Approaches		
Investigator		

[a]Co-occurring words appeared within five words of each other in the text of critiques.

and "tm"[61,62] for working with character-classed variables (*i.e.*, language) and for text extraction, and the package "openxlsx"[63] for formatting and exporting context windows for analysis.

We implemented a word intrusion test developed by Chang et al. to evaluate the semantic cohesiveness of our sentiment topic.[55] The premise of this method is to see if an outside observer can detect an outlying word in a list of otherwise cohesive words. Within our topic, the top 25 words with the highest probability of co-occurrence were randomly ordered to form five sets of five words each. We then randomly selected an intruder word candidate from each of the remaining 25 topics, for a total of 25 intruder words. These words were randomly inserted into the sets of nonintruder words so that each set contained five nonintruder words and one intruder word. Each of the 15 judges reviewed five such sets and attempted to identify the intruder word in each set.

Qualitative method—thematic analysis. We imported all context windows for our sentiment topic into the NVivo (version 11) Software program for qualitative analysis.[64] To inductively develop cohesive themes,[10,45,65,66] four authors (W.M., A.K., A.F., M.C.) first read through a sample of context windows and generated initial codes through discussion and consensus. Using these codes, two of the authors (A.K., A.F.), blinded to applicant gender and application funding, then coded the remaining context windows. On a randomly drawn subset of 20 context windows, the coders had 95%–99% agreement in code assignment (kappa = 0.96).[10,45,65,66] In an iterative "constant comparison" process, the initial codes assigned to all context windows were collapsed into conceptually congruent themes.[10,45,65,66] The coders then explored thematic patterns for male and female PIs' unfunded and funded applications.

Analysis of scores and productivity data

To test for differences in scores assigned to male and female PIs' applications, we transformed all scores to a logarithmic scale, because they were skewed, and submitted them as dependent variables for ordinary least squares linear regression with PI sex (M vs. F) as a predictor variable. Models used standard errors clustered at the applicant level, and adjusted for experience level (new vs. experienced investigator); application funding outcome (unfunded vs. funded); h-index; RCR; and eight types of NIH awards (detailed above under the "Data" section). Because PI age could bias productivity information, for each PI, we subtracted the year of

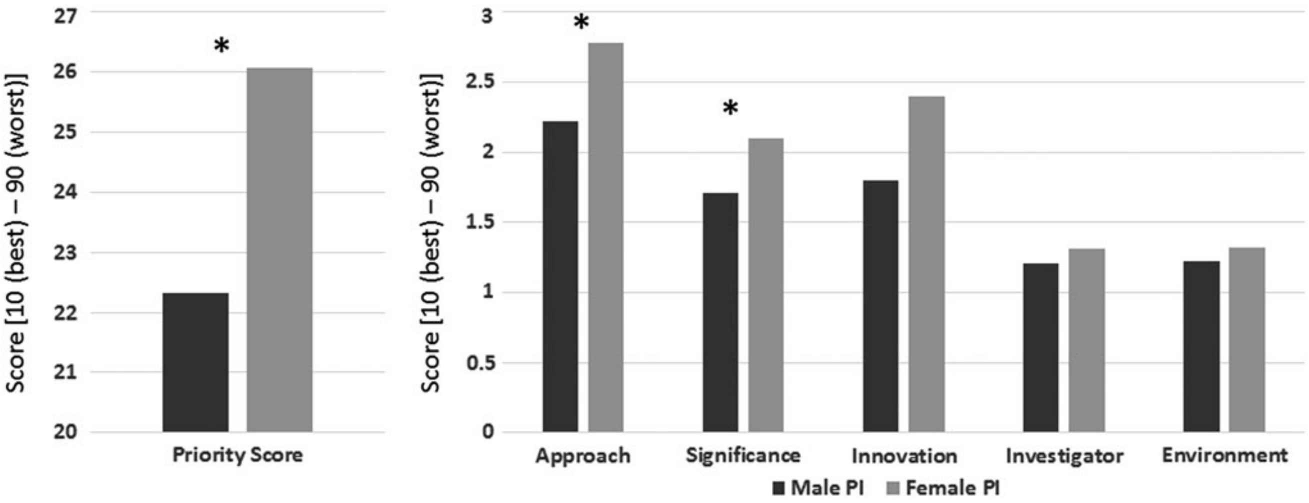

FIG. 1. Scores assigned to male and female PIs' Type 2 R01 applications. Estimated priority, approach, and significance scores assigned to male and female investigators' Type 2 applications from a study of 241 NIH R01 grant critiques and scores, University of Wisconsin–Madison, fiscal years 2010–2014. Priority scores (scale: 10 [best] to 90 [worst]) are modeled at the application Summary Statement level, $N = 68$; and criteria scores (scale: 1 [best] to 9 [worst]) are modeled at the critique level, $N = 206$. Regression models adjusted for application funding outcome, PI experience level, and productivity measures [h-index, NIH Relative Citation Ratio (RCR), prior NIH Grants (Type 1 and Type 2 R01, other R(esearch) awards, P(rogram) awards, T(raining) awards, Career Development (K) awards, F(ellowship) awards)]. NIH, National Institutes of Health; PI, principal investigator. *Difference between scores assigned to male and female PIs' Type 2 applications is significant at the $p < 0.05$ level.

his or her first NIH award from the year (*i.e.*, 2010–2014) of his or her award in our database. Based on this measure, we found no significant difference in the number of years male and female PIs had engaged in NIH funded research (M = 14.2 vs. F = 12.78).

Results

Applicant sex differences in scores

Regression models showed that female PIs' R01 renewal applications were assigned significantly worse (*i.e.*, higher) priority ($b = 0.28$, SE = 0.11), approach ($b = 0.23$, SE = 10), and significance scores ($b = 0.22$, SE = 0.10; $ps < 0.05$) even after controlling for experience level, application funding outcome, and productivity measures (Fig. 1).

Latent Dirichlet allocation

The sentiment topic was rated as semantically coherent as indicated by results from the topic intrusion test showing an acceptable accuracy of 85.3%. For female PIs, co-occurring terms such as ''excellent'' + ''work,'' ''environment'' + ''excellent,'' and ''expertise'' + ''research'' had high chi-square ranks, while terms such as ''highly'' + ''proposal,'' ''field'' + ''PI,'' and ''highly'' + ''innovative'' were ranked highly for male PIs (Table 3).

Qualitative thematic analysis

Thematic analysis of context windows generated 308 initial codes, which were ultimately collapsed into three themes—evaluative remarks about the (1) qualifications of the PI, (2) quality of the proposed work, and (3) quality of the research environment. Exploration of themes showed unique linguistic patterns for critiques of male and female PIs' unfunded and funded applications (Table 4).

Theme one—evaluative remarks about the qualifications of the PI. Exploration of the co-occurring words, ''field'' and ''PI,'' with high chi-square rankings for male investigators revealed that reviewers generally characterized male ''PIs'' as ''well-regarded,'' ''respected'' ''leaders'' and ''pioneers'' in their ''fields.'' These descriptors appeared more often in critiques of funded than unfunded applications. The following illustrative quotes are from critiques of different male PIs' applications:

Male PIs:

> ''The PI is a leader in the field with a strong record of research […].''
>
> ''The applicant is an internationally recognized leader in this field.''
>
> ''The PI is a recognized leader in the field.''
>
> ''[The PI] is internationally known as the pioneer of […] who is still extending the envelope of research.''
>
> ''[The PI] is a respected, productive investigator who is qualified to oversee the proposed research program.''

By comparison, exploration of words with high chi-square rankings, such as ''expert(ise),'' and standout adjectives, such as ''excellent'' and ''outstanding'' for female PIs, revealed that reviewers generally characterized female PIs as having ''expertise'' in critiques of both their unfunded and funded applications. Reviewers also sometimes characterized them as ''outstanding,'' ''excellent,'' and ''exceptional,'' especially in critiques of funded applications. Illustrating these patterns are quotes from critiques of different female PIs' applications:

Table 4. Themes in Critiques of R01 Renewal Applications and Subtle Variations for Male and Female PIs' Applications and Unfunded and Funded Applications

		Variation in themes	
Theme	*Subthemes*	*Applicant sex*	*Application funding outcome*
1. Qualifications of the PI.	None.	Reviewers described male PIs as ''leaders'' and ''pioneers'' in their fields.	Appeared to occur more in critiques of funded applications.
		Reviewers described female PIs as having ''expertise'' in their fields.	Appeared to occur similarly in critiques of unfunded and funded applications.
		Reviewers described female PIs as ''outstanding,'' ''excellent,'' and ''exceptional.''	Appeared to occur more in critiques of funded applications.
2. Quality of the Proposed Work.	Innovation, significance, and impact of the proposed work.	Reviewers described male PIs' research as ''highly significant,'' ''highly innovative,'' and ''high impact.''	Appeared to occur more in critiques of funded applications.
	Weaknesses and reviewers' concerns about the approach.	Reviewers described problems with male PIs' applications, but balanced criticism with praise.	Appeared to occur similarly in critiques of unfunded and funded applications.
		Reviewers described male PIs' research as ''overambitious'' or ''risk,'' but balanced these concerns with assurance of PIs' competence.	Appeared to occur similarly in critiques of unfunded and funded applications.
		Reviewers described female PIs' research as having only minor or negligible weaknesses.	Appeared to occur similarly in critiques of unfunded and funded applications.
3. Quality of the Research Environment.	None.	Reviewers described female PIs' laboratories, programs, and environments as ''strong,'' ''outstanding,'' and ''excellent.''	Appeared to occur more in critiques of funded applications.

Female PIs:

> ''The […] expertise of the PI is excellent.''
>
> ''The PI, […] has the extensive expertise in […].''
>
> ''Expertise of the PI is excellent.''
>
> ''[…] PI, is an outstanding investigator who has contributed numerous seminal papers on […].''
>
> ''The PI is exceptionally strong and the environment outstanding.''
>
> ''[The] PI, is an outstanding investigator […].''

Theme two—evaluative remarks about the quality of the proposed research. This theme contained two subcategories: (1) innovation, significance, and impact of the proposed research and (2) weaknesses and reviewers' concerns about the proposed approach (*e.g.*, concerns about methods, sample, analysis, or design). In these subcategories, many of the top words in our sentiment topic (*e.g.*, ''innovative,'' ''approach,'' ''significance,'' ''proposal,'' and ''studies'') were used to describe both highly positive and negative aspects of proposals. Different patterns within these subcategories surfaced for male and female PIs.

Reflecting the high chi-square rankings of words such as ''highly,'' ''innovative,'' and ''proposal'' for male PIs, these descriptors appeared to abundantly co-occur in critiques of male PIs' applications, particularly when they were funded. These words appeared to be less likely to co-occur for female PIs. The following quotes illustrate this semantic pattern observed for male PIs:

Male PIs:

> ''This is a highly significant proposal that addresses an exciting […] that has not been as well evaluated.''
>
> ''The use of […] is highly innovative.''
>
> ''[…] offer a highly significant arena for developing methods for […].''
>
> ''The application is clearly written, compelling and draws on several highly innovative advancements to accomplish the work.''

For the second subtheme—weaknesses and concerns about the approach—reviewers used many sentiment topic words, such as ''potential,'' ''studies,'' ''innovative,'' ''significant,'' and ''impact,'' in a negative context. This observation illustrated the importance of examining the context in which words are used because such descriptors on their own would likely be interpreted as positive. Within this subcategory, we identified different patterns for male and female PIs. We observed that reviewers offered male PIs criticism about their proposals almost reluctantly, as a *concession*, usually after first describing positive facets about the proposed work's ''innovation'' or ''significance.'' In this concessional sentence structure, top ranked words for male PIs, such as ''studies,'' ''innovative,'' ''significant,'' and ''highly,'' were used by reviewers to temper the impact of critical remarks. This semantic trend occurred in critiques of both unfunded and funded applications for male PIs (as illustrated by examples from different male PIs' application critiques below), but appeared to rarely surface for female PIs.

Male PIs:

> ''This is a potentially powerful component of the proposed studies, but the narrative was not developed beyond a basic description of the procedure.''
>
> ''Highly innovative, but there are no preliminary data for […].''

Only in critiques for male PIs (unfunded and funded applications) did reviewers comment about the proposed work being ''overambitious'' or ''risky,'' and these remarks were always tempered by assurance of the PIs' competence. Two quotes from critiques of different male PIs' applications illustrate this semantic pattern:

Male PIs:

> ''Overly ambitious and elaborate proposal, however investigator has excellent track record in executing and completing his proposed work.''
>
> ''[…] is more risky, but PI has demonstrated work with […] already.''

By comparison, concerns and criticism in both unfunded and funded application critiques of female PIs' applications were generally more mild in tone (*e.g.*, ''this is a minor concern''), and included top ranked words for female PIs, such as ''work'' and ''strong.'' These two quotes from critiques of different female PIs' applications illustrate this finding:

Female PIs:

> ''A negligible weakness is that the […] is not explained; it would help to […]. A related negligible weakness is that the […] group has already done significant work in characterizing the […].''
>
> ''Overall these minor weaknesses did not lower my enthusiasm for this very strong application.''

Theme three—evaluative remarks about the quality of the research environment. This theme dominated female PIs' critiques of funded applications, but appeared less likely to occur for male PIs. In critiques of female PIs' applications, top chi-square ranked words, such as ''environment,'' ''strong,'' ''outstanding,'' ''excellent,'' ''proposed,'' ''studies,'' and ''research,'' were used to describe aspects of the research environment. Illustrating this trend are several illustrative remarks from different female PIs' applications:

> ''The environment for conducting the proposed work is excellent […].''
>
> ''This is a well written application from a strong investigator in an excellent environment […].''
>
> ''[…] is an outstanding environment for conducting the proposed experiments.''

Discussion

In this mixed methods study, we demonstrated that latent Dirichlet allocation (LDA, ''topic modeling'') may be useful for analyzing scientific peer reviewers' remarks especially when combined with qualitative thematic analysis. We found different thematic patterns of evaluative remarks in reviewers' critiques of male and female PIs' unfunded and funded applications. We also found that female PIs' applications were assigned significantly worse (*i.e.*, higher) significance, approach, and overall priority scores (Fig. 1). Importantly, productivity/impact measures (*i.e.*, h-index, RCR, and NIH awards), which we used as a proxy for the quality of PIs' research, did not account for disparities in male and female PIs' application scores. This suggests that some other factors may have influenced reviewers' decision-making.

A large body of experimental and observational research spanning the past 40 years shows that gender stereotypes can influence reviewers to differentially evaluate men's and women's work and their qualifications in male-typed fields, such as science.[5,12,13,26,27,32,37,67–72] Our results strongly align with this work. Specifically, stereotypes that men are agentic (*e.g.*, leaders, logical, independent) and women are communal (*e.g.*, supportive, sensitive, dependent) influence reviewers to assume men have high competence, will be more likely to succeed, and are a better ''fit'' for male-typed jobs and roles (*e.g.*, such as scientists) that are assumed to require agentic skills and traits.[5,12,13,26,27,32,37,67–72] As R01 renewals are high status awards in a male-typed field and have criteria that align with performance expectations for men and leaders [*i.e.*, applicants are expected to have ''an ongoing record of accomplishments that have advanced their field(s)''],[73] we posit that gender bias may have impacted NIH reviewers' judgment and could explain our results in several ways:

First, gender bias can lead reviewers to more easily perceive men as ''leaders[74–77] and assign greater value to men's than women's credentials (even when credentials are similar).[78] This may explain why, in our study, reviewers appeared to more easily perceive male PIs as ''pioneers'' and ''leaders in the field.'' Although reviewers characterized female PIs as having ''expertise,'' such descriptors have less positive valence as they are more frequently used to describe the value of consultants, collaborators, or contributors—roles that more closely align with stereotypes about women (*e.g.*, women are ''communal,'' ''helpful,'' ''cooperative'').[67,69,79] Moreover, someone with expertise may know a lot about a subject, but a respected leader is regarded as a strong contributor and driver of change—a very agentic role that is congruent with stereotypes about men.[67,69,79] Although subtle, if gender bias leads reviewers to differently interpret male and female PIs' qualifications and accomplishments, as our study suggests, it could translate to penalized scoring and funding outcomes for female applicants because PIs of R01 renewals are expected to be scientific leaders.[73] This would explain worse (*i.e.*, higher) priority, approach, and significance scores assigned to female than male PIs' applications with similar productivity.

Second, such bias can lead reviewers to perceive the same work as more valuable when thought to be performed by a man.[1,80,81] This may explain why reviewers in our study appeared to more easily view research proposed by male PIs as ''highly significant,'' ''highly innovative,'' and ''high impact.'' Because significance, innovation, and impact are important criteria for determining scores and funding outcomes, female applicants could face disadvantages if gender bias leads reviewers to more easily view research proposed by male PIs as higher in value for NIH.

Third, gender bias can lead reviewers to be more willing to provide constructive critical feedback to men than women, which has been shown to bolster career development and success.[82] This may explain why reviewers in our study appeared to provide male PIs criticism accompanied by positive remarks—as such a style of feedback is more often used for members of positively stereotyped groups to help improve future performance.[82] Moreover, Correl and Simard found that women are more likely to receive vague instead of constructive feedback.[83] If this occurs in NIH peer review, as our results suggest, female applicants would be less likely to benefit from peer reviewers' expertise, which would disadvantage them, particularly if their proposals were unfunded and required revision.

Several other studies on gender bias in evaluation processes may help explain our observation that some women and their research were described as ''outstanding'' and ''exceptional.'' For example, experiments show that assumptions that women are less competent than men can lead reviewers to hold female applicants to higher ability standards to confirm their competence.[68,72,84] Thus, words such as outstanding and exceptional used to describe women and their work may be evidence of reviewers requiring higher quality work and stronger track records of past performance for female than male PIs. Other studies show that when pro-male gender bias operates in a review process, it can lead to men's advantage on the criteria that matter most (*e.g.*, numerical rankings) for obtaining a tangible reward (*e.g.*, a job, promotion, funding), which is compensated for by advantaging women on criteria that matter least (*e.g.*, written commentary).[32,68,77,85–87] Importantly, results from our previous studies found that more critiques of female PI applications for R01 renewals contained standout adjectives (*e.g.*, outstanding, excellent), despite assignment of worse or similar priority, approach, and significance scores,[36,37] which are tightly linked to funding outcomes at NIH. Taken together, these and our current findings suggest that the latter interpretation (pro-male bias in consequential numerical rankings compensated for by pro-female bias in less consequential commentary) may be most likely. Future experiments are needed to establish causal reasons for the greater occurrence of standout adjectives in critiques of female PI R01 renewal applications.

Our final finding that female PI research environments were characterized as ''outstanding'' and ''exceptional,'' predominately when their applications were funded, adds to literature regarding the relationship between institutional, departmental, and laboratory climate; and women's career persistence and advancement in academic medicine. This body of work links poor climate to attrition from research careers.[1] Findings from our study suggest that women who successfully persist in research may be in better environments or that reviewers are likely to assume there is a strong environment if women are successful. Alternatively, environments may be positively impacted when women persist in research careers. Future studies are needed to more concretely examine the direction of the relationship between the climate of the environment in which research is conducted and women's persistence and success in research careers.

Implications and future directions

This study shows scoring disparities and complex linguistic differences in critiques of male and female PIs' R01 renewal applications. Such differences may be consequential to funding outcomes. If women systematically earn worse scores on their R01 renewals, as we found in our study, this might explain why they have lower R01 renewal success rates, nationally. Our study builds on previous studies by Carnes et al. on other high status NIH awards, such as the Director's Pioneer Award and the Clinical Translational Science Award, which suggest that NIH review criteria and male-typed descriptors may inadvertently favor men.[26,27] If future work identifies gender bias as a factor for female PIs' lower funding success rates for R01 renewals and other high status NIH awards, interventions that approach bias as a changeable habit may be helpful for NIH reviewers, as they have proven to be effective at reducing stereotype-based bias for faculty in science fields.[88] If text analysis proves useful as a flag for bias in peer review, it has the potential to be of use to funding agencies and policy makers interested in leveraging the substantial untapped information available in written critiques both to test for bias in peer review and to test the impact of bias reduction interventions. It may also prove useful to funding agencies to better inform themselves and applicants about reasons for scoring and funding decisions.

Limitations

This study has limitations. We examined data from a single site, and a single NIH award type, so results may not generalize to the population of investigators who obtained R01 renewals between 2010 and 2014 and those who had other types of awards. Due to the small sample size (and consequential identifiability) of underrepresented racial/ethnic minorities in our sample, our analyses only targeted applicant sex. LDA does not consider the context in which words are used, but we overcame this limitation by using thematic analysis. We did not have access to terminally unfunded applications (*i.e.*, applications that were unfunded and never revised) since all initially unfunded applications in our sample were funded in subsequent resubmissions. Because of this we could not estimate funding success rates or test for differences in funding success rates. We did not have access to raw numbers of male and female PIs' publications; instead, we used h-index and RCR scores, and raw numbers of NIH awards as surrogate estimates for productivity/impact. Consequently, we could not assess the extent to which raw numbers of publications differed for male and female PIs in our sample, and whether this might explain scoring differences. There is research showing that men publish more than women, which may be a consequence of differences in resources or differential impact of work/life balance and family responsibilities.[24,89–99] However, this publishing gap disappears when controlling for career stage/rank,[100] and almost all PIs in our sample (98%, N = 44/45) were experienced investigators (Table 2) and had held NIH funding for similar lengths of time (M = 14.2 vs. F = 12.78).

Conclusion

Combining data-driven algorithmic text mining methods with qualitative thematic analysis allowed us to detect different patterns of evaluative feedback in critiques of male and female PI R01 renewal applications. Results suggest that subtle implicit gender bias may be operating in peer review.

Acknowledgments

We are grateful to Eve Fine from Women in Science and Engineering Institute at the University of Wisconsin-Madison for her generous consultations. This research was funded by the University of Wisconsin-Madison Department of Medicine and the National Institutes of Health Grant # R01 GM111002.

Ethical Approval

The University of Wisconsin Institutional Review Board approved all aspects of this study. Protocol # SBS2012-1177.

Author Disclosure Statement

No competing financial interests exist.

References

1. Committee on Maximizing the Potential of Women in Academic Science EU, Committee on Science, & Public Policy (US). Beyond bias and barriers: fulfilling the potential of women in academic science and engineering. Washington, DC: National Academy Press, 2007.
2. Herring C. Does diversity pay?: Race, gender, and the business case for diversity. Am Sociol Rev 2009;74:208–224.
3. Hong L, Page SE. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc Natl Acad Sci U S A 2004;101:16385–16389.
4. Desvaux G, Devillard-Hoellinger S, Baumgarten P. Women matter: Gender diversity, a corporate performance driver. Paris: McKinsey, 2007.
5. Carnes M, Morrissey C, Geller SE. Women's health and women's leadership in academic medicine: Hitting the same glass ceiling? J Womens Health 2008;17:1453–1462.
6. Ruzek SB, Becker J. The women's health movement in the United States: From grass-roots activism to professional agendas. J Amer Med Women's Assoc 1999;54:4–8.
7. Nicolette J, Jacobs MB. Integration of women's health into an internal medicine core curriculum for medical students. Acad Med 2000;75:1061–1065.
8. Weiss LB, Levison SP. Tools for integrating women's health into medical education: Clinical cases and concept mapping. Acad Med 2000;75:1081–1086.
9. Rosser SV. Women's health—Missing from US medicine. Washington, DC: Georgetown University Press, 1994.
10. Kaatz A, Dattalo M, Regner C, Filut A, Carnes M. Patterns of feedback on the bridge to independence: A qualitative thematic analysis of NIH mentored career development award application critiques. J Womens Health 2016;25:78–90.
11. Bickel J, Wara D, Atkinson BF, et al. Increasing women's leadership in academic medicine: Report of the AAMC Project Implementation Committee. Acad Med 2002;77:1043–1061.
12. Wright AL, Schwindt LA, Bassford TL, et al. Gender differences in academic advancement: Patterns, causes, and potential solutions in one US College of Medicine. Acad Med 2003;78:500–508.
13. Carr PL, Gunn CM, Kaplan SA, Raj A, Freund KM. Inadequate progress for women in academic medicine: Findings from the national faculty study. J Womens Health 2015;24:190–199.
14. Hofler L, Hacker MR, Dodge LE, Ricciotti HA. Subspecialty and gender of obstetrics and gynecology faculty in department-based leadership roles. Obstet Gynecol 2015;125:471–476.
15. Hofler LG, Hacker MR, Dodge LE, Schutzberg R, Ricciotti HA. Comparison of women in department leadership in obstetrics and gynecology with those in other specialties. Obstet Gynecol 2016;127:442–447.
16. Jena AB, Khullar D, Ho O, Olenski AR, Blumenthal DM. Sex differences in academic rank in US medical schools in 2014. JAMA 2015;314:1149–1158.
17. Burden M, Frank MG, Keniston A, et al. Gender disparities in leadership and scholarly productivity of academic hospitalists. J Hosp Med 2015;10:481–485.
18. Lautenberger DM, Dandar VM, Raezer CL, Sloane RA. The State of Women in Academic Medicine: The Pipeline and Pathways to Leadership. Washington, DC: Association of American Medical Colleges, 2014.
19. Association of American Medical Colleges (AAMC). Medical Students, Selected Years, 1965–2013. Available at: www.aamc.org/download/411782/data/2014_table 1.pdf Accessed May 26, 2016.
20. Svider PF, Mauro KM, Sanghvi S, Setzen M, Baredes S, Eloy JA. Is NIH funding predictive of greater research productivity and impact among academic otolaryngologists? Laryngoscope 2013;123:118–122.
21. King A, Sharma-Crawford I, Shaaban AF, et al. The pediatric surgeon's road to research independence: Utility of mentor-based National Institutes of Health grants. J Surg Res 2013;184:66–70.
22. Pohlhaus JR, Jiang H, Wagner RM, Schaffer WT, Pinn VW. Sex differences in application, success, and funding rates for NIH extramural programs. Acad Med 2011;86:759–767.
23. NIH Data Book—U.S. Department of Health and Human Services. R01-Equivalent grants: Success rates, by gender and type of application. Available at: https://report.nih.gov/nihdatabook/ Accessed March 9, 2016.
24. Tabak LA, Collins FS. Weaving a richer tapestry in biomedical science: NIH leadership discusses the need for renewed efforts to increase diversity in the US biomedical research workforce. Science (New York, NY) 2011;333:940–941.
25. Valantine HA, Collins FS. National Institutes of Health addresses the science of diversity. Proc Natl Acad Sci 2015;112:12240–12242.
26. Carnes M, Geller S, Fine E, Sheridan J, Handelsman J. NIH directors pioneer awards: Could the selection process be biased against women? J Womens Health 2005;14:684–691.
27. Carnes M, Bland C. Viewpoint: A challenge to academic health centers and the National Institutes of Health to prevent unintended gender bias in the selection of clinical and translational science award leaders. Acad Med 2007;82:202–206.
28. Carli LL, Alawa L, Lee Y, Zhao B, Kim E. Stereotypes about gender and science: Women $\neq$ scientists. Psychol Women Q 2016;40:1–17.
29. Nosek BA, Smyth FL, Sriram N, et al. National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceed Natl Acad Sci 2009;106:10593–10597.
30. Leslie S-J, Cimpian A, Meyer M, Freeland E. Expectations of brilliance underlie gender distributions across academic disciplines. Science 2015;347:262–265.

31. Meyer M, Cimpian A, Leslie S-J. Women are underrepresented in fields where success is believed to require brilliance. Front Psychol 2015;6:235.
32. Biernat M, Tocci M, Williams JC. The language of performance evaluations gender-based shifts in content and consistency of judgment. Soc Psychol Person Sci 2012;3: 186–192.
33. Isaac C, Chertoff J, Lee B, Carnes M. Do students' and authors' genders affect evaluations? A linguistic analysis of Medical Student Performance Evaluations. Acad Med 2011;86:59–66.
34. Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in medical student performance evaluations. Eval Health Prof 2010;33:365–385.
35. Trix F, Psenka C. Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse Soc 2003;14:191–220.
36. Kaatz A, Magua W, Zimmerman DR, Carnes M. A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. Acad Med 2015;90:69–75.
37. Kaatz A, Lee YG, Potvien A, et al. Application critiques, impact and criteria scores: Does the sex of the principal investigator make a difference? Acad Med 2016;91:1080–1088.
38. National Institutes of Health: Side-by-side comparison of enhanced and former review criteria (research grants and cooperative agreements). Available at: http://grants.nih.gov/grants/peer/guidelines_general/comparison_of_review_criteria.pdf Accessed April 10, 2016.
39. National Institutes of Health: Enhancing peer review at NIH. Available at: http://enhancing-peer-review.nih.gov/index.html Accessed April 10, 2016.
40. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. Austin, TX: University of Texas at Austin, 2015.
41. Aggarwal CC, Zhai C. Mining text data. New York, NY: Springer Science & Business Media, 2012.
42. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Machine Learn Res 2003;3:993–1022.
43. Dreyfus HL. What computers still can't do: a critique of artificial reason. Cambridge, MA: MIT Press, 1992.
44. Bellotti V, Edwards K. Intelligibility and accountability: Human considerations in context-aware systems. Hum Comput Interact 2001;16:193–212.
45. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3:77–101.
46. National Institutes of Health: Frequently Asked Questions. NIH new and early stage investigator (ESI) policies. Available at: http://grants.nih.gov/grants/new_investigators/investigator_policies_faqs.htm Accessed December 11, 2016.
47. Office of Extramural Research. National Institutes of Health: NIH Research Project Grant Program (R01). Available at: https://grants.nih.gov/grants/funding/r01.htm Accessed April 9, 2016.
48. Office of Extramural Research. National Institutes of Health: Peer Review Process. Available at: https://grants.nih.gov/grants/peer_review_process.htm Accessed April 9, 2016.
49. R Core Team: R: A language and environment for statistical computing. Vienna, Austria: R Core Team. R Foundation for Statistical Computing, 2014.
50. Hirsch J. An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A 2005;102: 16569–16572.
51. Hutchins BI, Yuan X, Anderson JM, Santangelo GM. Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. PLoS Biol 2016;14:e1002541.
52. McCallum AK. MALLET: A machine learning for language toolkit, 2002. http://mallet.cs.umass.edu Accessed February 14, 2017.
53. Mimno D. Vignette-mallet: A wrapper around the Java machine learning tool MALLET. 2013. https://CRAN.R-project.org/package=mallet Accessed February 14, 2017.
54. Ponweiser M. Latent Dirichlet allocation in R, 2–21. Austria: Institute for Statistics and Mathematics, WU (Wirtschaftsuniversitat Wien), 2012.
55. Chang J, Gerrish S, Wang C, Boyd-Graber JL, Blei DM. Reading tea leaves: How humans interpret topic models. Adv Neural Inf Process Syst 2009;31:288–296.
56. De Villiers MR, De Villiers PJ, Kent AP. The Delphi technique in health sciences education research. Med Teach 2005;27:639–643.
57. Gupta UG, Clarke RE. Theory and applications of the Delphi technique: A bibliography (1975–1994). Technol Forecast Soc Change 1996;53:185–211.
58. Hsu C, Sandford B. The Delphi technique: Making sense of consensus. Pract Assess Res Eval 2007;12:1–8.
59. Landeta J, Barrutia J, Lertxundi A. Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. Technol Forecast Social Change 2011;78:1629–1641.
60. Wickham H. Stringr: Make it easier to work with strings. R package version 0.6. 2012;2:96–7.
61. Feinerer I, Hornik K. tm: Text mining package. R Package, 2014.
62. Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. J Stat Software 2008;25:1–54.
63. Walker A. openxlsx: Read, write and edit XLSX Files. R Package, 2015.
64. NVivo qualitative data analysis Software. Victoria, Australia: QSR International Pty Ltd, 2002.
65. Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. Int J Qual Methods 2006;5:80–92.
66. Boyatzis RE. Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, CA: Sage, 1998.
67. Eagly AH, Karau SJ. Role congruity theory of prejudice toward female leaders. Psychol Rev 2002;109:573–598.
68. Biernat M, Kobrynowicz D. Gender-and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. J Pers Soc Psychol 1997;72:544–557.
69. Ridgeway CL. Gender, status, and leadership. J Soc Issues 2001;57:637–655.
70. Rudman LA, Kilianski SE. Implicit and explicit attitudes toward female authority. Pers Soc Psychol Bull 2000;26: 1315–1328.
71. Marchant A, Bhattacharya A, Carnes M. Can the language of tenure criteria influence women's academic advancement? J Womens Health 2007;16:998–1003.
72. Heilman ME, Haynes MC. Subjectivity in the appraisal process: A facilitator of gender bias in work settings. In: Borgida E, Fiske ST, eds. Beyond common sense: Psychological science in the courtroom. Malden, MA: Blackwell Publishing, 2008:127–155.

73. National Institutes of Health (NIH). Office of Extramural Research. Grants and funding: Definitions of criteria and considerations for research project grant (RPG/X01/R01/R03/R21/R33/R34) critiques. Available at: http://grants.nih.gov/grants/peer/critiques/rpg_D.htm-rpg_overall Accessed April 13, 2016.
74. Jackson D, Engstrom E, Emmers-Sommer T. Think leader, think male and female: Sex vs. seating arrangement as leadership cues. Sex roles 2007;57:713–723.
75. Porter N, Geis FL, Jennings J. Are women invisible as leaders? Sex Roles 1983;9:1035–1049.
76. Watts JH. 'Now you see me, now you don't': The visibility paradox for women in a male-dominated profession. Revealing and concealing gender. New York, NY: Springer, 2010:175–193.
77. Biernat M. Stereotypes and shifting standards: Forming, communicating, and translating person impressions. In: Devine PGP, E.A., ed. Advances in Experimental Social Psychology. San Diego, CA: Academic Press, 2012;45:1.
78. Uhlmann EL, Cohen GL. Constructed criteria redefining merit to justify discrimination. Psychol Sci 2005;16:474–480.
79. Eagly AH, Wood W, Diekman AB. Social role theory of sex differences and similarities: A current appraisal. In: Eckes T, Trautner HM eds. The Developmental Social Psychology of Gender. Mahwah, NJ: Lawrence Erlbaum Associates, 2000:123–174.
80. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci 2012;109:16474–16479.
81. Steinpreis RE, Anders KA, Ritzke D. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. Sex Roles 1999;41:509–528.
82. King EB, Botsford W, Hebl MR, Kazama S, Dawson JF, Perkins A. Benevolent sexism at work gender differences in the distribution of challenging developmental experiences. J Manage 2012;38:1835–1866.
83. Correll S, Simard C. Vague Feedback Is Holding Women Back. Harvard Business Review. 2016.
84. Foschi M. Double standards in the evaluation of men and women. Soc Psychol Q 1996;59:237–254.
85. Castilla EJ. Gender, race, and meritocracy in organizational careers. Am J Sociol 2008;113:1479–1526.
86. Vescio TK, Gervais SJ, Snyder M, Hoover A. Power and the creation of patronizing environments: The stereotype-based behaviors of the powerful and their effects on female performance in masculine domains. J Person Soc Psychol 2005;88:658–672.
87. Wilson KY. An analysis of bias in supervisor narrative comments in performance appraisal. Hum Relat 2010;63:1903–1933.
88. Carnes M, Devine PG, Manwell LB, et al. The effect of an intervention to break the gender bias habit for faculty at one institution: A cluster randomized, controlled trial. Acad Med 2015;90:221–230.
89. Jagsi R, Guancial EA, Worobey CC, et al. The ''gender gap'' in authorship of academic medical literature—A 35-year perspective. N Engl J Med 2006;355:281–287.
90. Cole JR, Zuckerman H. The productivity puzzle: Persistence and change in patterns of publication of men and women scientists. Adv Motiv Achiev 1984;2:217–258.
91. Toutkoushian RK, Bellas ML. Faculty time allocations and research productivity: Gender, race and family effects. Rev High Educ 1999;22:367–390.
92. Fox MF. Research, teaching, and publication productivity: Mutuality versus competition in academia. Sociol Educ 1992:293–305.
93. Stack S. Gender, children and research productivity. Res High Educ 2004;45:891–920.
94. Settles IH, Cortina LM, Malley J, Stewart AJ. The climate for women in academic science: The good, the bad, and the changeable. Psychol Women Q 2006;30:47–58.
95. Leahey E. Gender differences in productivity research specialization as a missing link. Gender Soc 2006;20:754–780.
96. Allison PD, Stewart JA. Productivity differences among scientists: Evidence for accumulative advantage. Am Sociol Rev 1974;39:596–606.
97. Fox MF. Gender, family characteristics, and publication productivity among scientists. Soc Stud Sci 2005;35:131–150.
98. Carr PL, Ash AS, Friedman RH, et al. Relation of family responsibilities and gender to the productivity and career satisfaction of medical faculty. Ann Intern Med 1998;129:532–538.
99. Hunter L, Leahey E. Parenting and research productivity: New evidence and methods. Soc Stud Sci 2010;40:433–451.
100. Diamond SJ, Thomas CRJ, Desai S, et al. Gender differences in publication productivity, academic rank, and career duration among U.S. academic gastroenterology faculty. Acad Med 2016;91:1158–1163.

Address correspondence to:
Anna Kaatz, PhD, MPH
Center for Women's Health Research
University of Wisconsin-Madison
700 Regent Street, Suite #301
Madison, WI 53715

E-mail: akaatz@wisc.edu