A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques From Investigators at One Institution

Anna Kaatz, PhD, MPH, Wairimu Magua, MS, David R. Zimmerman, PhD,
and Molly Carnes, MD, MS

Abstract

Purpose

Career advancement in academic medicine often hinges on the ability to garner research funds. The National Institutes of Health's (NIH's) R01 award is the "gold standard" of an independent research program. Studies show inconsistencies in R01 reviewers' scoring and in award outcomes for certain applicant groups. Consistent with the NIH recommendation to examine potential bias in R01 peer review, the authors performed a text analysis of R01 reviewers' critiques.

Method

The authors collected 454 critiques (262 from 91 unfunded and 192 from 67 funded applications) from 67 of 76 (88%) R01 investigators at the University of Wisconsin–Madison with initially unfunded applications subsequently funded between December 2007 and May 2009. To analyze critiques, the authors developed positive and negative grant application evaluation word categories and selected five existing categories relevant to grant review. They analyzed results with linear mixed-effects models for differences due to applicant and application characteristics.

Results

Critiques of funded applications contained more positive descriptors and superlatives and fewer negative evaluation words than critiques of unfunded applications. Experienced investigators' critiques contained more references to competence. Critiques showed differences due to applicant sex despite similar application scores or funding outcomes: more praise for applications from female investigators, greater reference to competence/ability for funded applications from female experienced investigators, and more negative evaluation words for applications from male investigators (all $P < .05$).

Conclusions

Results suggest that text analysis is a promising tool for assessing consistency in R01 reviewers' judgments, and gender stereotypes may operate in R01 review.

An important determinant of career advancement in academic medicine is the ability to compete for research support.[1] The National Institutes of Health (NIH) is the largest funder of research at U.S. academic medical centers, and its R01 award is considered the "gold standard" of an independent research program.[1] The NIH's two-stage system of peer review determines R01 funding.[2] In the first phase, approximately three peer reviewers assign preliminary scores and write critiques based on the proposed work's overall impact/priority, significance, innovation, approach, investigators, and environment.[2,3] Applications with approximately the top 50% of preliminary impact/priority scores are discussed at review group meetings where all members contribute a final score.[3] Applicants receive a Summary Statement with individual reviewers' critiques, a summary paragraph, and the average final impact/priority score for discussed applications. In the second stage, NIH staff and the advisory council of each NIH institute and center (IC) weigh peer review outcomes and IC priorities to make final funding recommendations to IC directors.[2,3] The NIH continually evaluates its review process. Although NIH peer review is considered one of the best systems in the world, studies have identified inconsistencies among R01 reviewers' scores[4–6] and unexplained differences in award outcomes for some groups of R01 applicants.[3,7–13] If bias unrelated to the quality of the proposed science negatively impacts the outcome of a grant review, it runs counter to the NIH's goal to fund the best science, threatens scientific workforce diversity, and undermines the competitiveness of the U.S. scientific enterprise.[7,14–16]

An advisory committee to the NIH director,[17] followed by a plan for action by the NIH's deputy director,[18] made recommendations to examine the NIH's grant review process for bias. Recommendations included the need for "text-based analysis of the commentary on individual grant reviews."[17] Prior studies of R01 peer review outcomes have analyzed application success rates or applicant funding rates,[8–12,19] award probabilities,[7,13] or reviewer-assigned scores.[4,5,8,12] These methods can effectively identify award or scoring disparities between certain groups of applicants or proposal types but provide little insight into reviewers' reasoning for scoring or award recommendations. Text analysis of reviewers' critiques would be novel to the study of scientific review because it can provide a window into reviewers' decision-making processes.[20–26] When used in combination with traditional comparisons, text analysis would permit testing whether reviewers' judgments are congruent with scores and funding outcomes.[21] Our research aligns with the NIH's call for action and, to our knowledge, is the first text analysis of the written critiques of R01 applications. In this study, we empirically link the contents of critiques to funding outcomes and scores to test for consistency in reviewers'

Please see the end of this article for information about the authors.

Correspondence should be addressed to Dr. Carnes, Center for Women's Health Research, University of Wisconsin–Madison, 700 Regent St., Suite 301, Madison, WI 53715; telephone: (608) 263-9770; fax: (608) 265-6423; e-mail: mlcarnes@wisc.edu.

Acad Med. 2015;90:69–75.
First published online August 19, 2014
doi: 10.1097/ACM.0000000000000442

judgments across different categories of R01 investigators.

Method

The institutional review board at the University of Wisconsin–Madison (UW-Madison) approved all aspects of this study. In June 2009, we searched the NIH's Computer Retrieval of Information on Scientific Projects (CRISP) database for all principal investigators (PIs) at UW-Madison with R01 awards funded with amended status (i.e., initially unfunded and subsequently revised applications) in the NIH's 2008–2009 fiscal years. Award dates spanned December 2007 through May 2009. We invited PIs, via three rounds of e-mail letters, to send us electronic PDF copies of all Summary Statements from unfunded and funded award cycles. R01 recipients provided consent by e-mailing PDF copies of their Summary Statements and could request withdrawal of their materials from the study.

We assigned identifiers (IDs) to applicants, Summary Statements, and critiques (as surrogate for the anonymous reviewer). We recorded the following (if present): application funding outcome, application type (Type 1/new R01 or Type 2/renewal), impact/priority score, the NIH IC, and use of human subjects. To ascertain applicant sex, we searched the Internet using a strategy similar to that employed by Jagsi and colleagues.[19] We also examined home institution Web sites to ascertain each applicant's training background. We searched CRISP for all grants held by investigators prior to the study award period, and classified investigators as "experienced" according to NIH criteria.[27] We electronically formatted and deidentified each critique prior to text analysis.

We analyzed critiques with the Linguistic Inquiry Word Count (LIWC) text analysis software program (LIWC 2007, Austin, Texas), which calculates the percentage of words from predefined linguistic categories in written documents.[28] We examined words in the LIWC's 80 default word categories and 7 others developed for use with LIWC,[23–25] and identified 5 categories relevant to scientific grant review (Table 1). These word categories are "ability" (e.g., skilled, expert, talented), "achievement" (e.g., honors, awards, prize), "agentic" (e.g., accomplish, leader, competent), "research" (e.g., scholarship, publications, grants), and "standout adjectives" (e.g., exceptional, outstanding, excellent). We developed 2 categories that reflect "positive evaluation" (e.g., groundbreaking, solid, comprehensive) and "negative evaluation" (e.g., illogical, unsubstantiated, diffuse) of a grant application.

Employing a modified Delphi technique,[29,30] we solicited lists of positive and negative evaluation words relevant to NIH grant applications from four local experienced NIH grant reviewers. We collated and resent lists for feedback,

Table 1

Words in Linguistic Categories Used With the LIWC Software Program, From a Text Analysis Study of 454 Critiques of R01 Applications From Male and Female Investigators, University of Wisconsin–Madison, Fiscal Year 2008–2009

Category	Word or root word[a]
Ability words[24,25]	abilit*, able, adept*, adroit*, analy*, aptitude, brain*, bright*, brillian*, capab*, capacit*, clever*, compet*, creati*, expert*, flair, genius, gift*, inherent*, innate, insight*, instinct*, intell*, knack, natural*, proficien*, propensity, skill*, smart*, talent*
Achievement words[28,b]	abilit*, able*, accomplish*, ace, achiev*, acquir*, acquisition*, adequa*, advanc*, advantag*, ahead, ambiti*, approv*, attain*, attempt*, authorit*, award*, beat*, best, better, bonus*, burnout*, capab*, celebrat*, challeng*, champ*, climb*, closure, compet*, conclud*, conclus*, confidence, confident*, conquer*, conscientious*, control*, creat*, crown*, defeat*, determin*, diligen*, domina*, demote*, driven, dropout*, earn*, effect*, efficien*, effort*, elit*, enabl*, endeav*, excel*, fail*, finaliz*, first, firsts, found*, fulfill*, gain*, goal*, hero*, hon*, ideal*, importan*, improv*, inadequa*, incapab*, incentive*, incompeten*, ineffect*, initiat*, irresponsible*, king*, lazie*, lazy, lead*, lesson*, limit*, los*, master*, medal*, mediocr*, motiv*, obtain*, opportun*, organiz*, originat*, outcome*, overcome, overconfiden*, overtak*, perfect*, perform*, persever*, persist*, plan*, potential*, power*, practice, prais*, presiden*, pride, prize*, produc*, proficien*, progress, promot*, proud*, purpose*, queen, queenly, quit*, rank*, recover*, requir*, resolv*, resourceful*, responsib*, reward*, skill*, solution*, solv*, strateg*, strength*, striv*, strong*, succeed*, success*, super, superb*, surviv*, team*, top, tried, tries, triumph*, try, trying, unable, unbeat*, unproduc*, unsuccessful*, victor*, win*, won, work*
Agentic words[23]	accomplish*, achiev*, active*, agentic, agress*, ambiti*, analy*, assert*, assiduous*, assurance, blunt*, bold*, candid*, compete*, competi*, confident*, conscientious*, daring, decisive*, defend*, direct*, domina*, driv*, dynamic*, forc*, forthright*, frank*, hardwork*, hostil*, independen*, individualistic, influence*, initiat*, intellectual, lead*, manage*, masculine, master*, mechanic*, mechanistic*, noetic, organiz*, originat*, outspoken, perform*, perserver*, power, produc*, rational*, reliabl*, risk, solid, start*, strength, strong*, success*, suggest*, superior, sure, worldly
Negative evaluative words	bias*, concern*, deficient, dependent, detract*, diffuse*, diminish*, fail*, ill*, inaccura*, inadequate*, inappropriate*, incomplete*, insignificant, insufficient, lack*, limit*, missing, narrow*, need*, not, omission, omit*, overambitious, overly, overstat*, poor*, question*, quo, shaky, simplistic, tentative*, unacceptabl*, unclear, underdevelop*, unproductive*, unproven, unsubstantiated, unsupported, weak*
Positive evaluative words	accept*, accomplish*, advanc*, ambitio*, appropriat*, art, believabl*, best, breadth, clear*, commit*, competitive*, complet*, comprehensive, convinc*, creat*, detail*, didactic, efficac*, energ*, enthus*, exceptional*, expan*, expla*, fascinat*, feasib*, focus*, groundbreaking, high*, impact*, impress*, includ*, indisputabl*, innovat*, interest*, logic*, mechanistic, meticulous*, new*, nice*, novel, obvious*, original, outstanding, pioneer*, productiv*, provocative*, quality, reasonabl*, reliab*, rigor*, significan*, solid*, sophistic*, sound*, specif*, stellar, strength, strong*, success*, superior*, support*, tailor*, target*, thought*, transform*, unimpeachable, unique*, valid*, valu*, well*
Research words[24,25]	contribution*, data, discover*, experiment*, finding*, fund*, grant*, journal*, manuscript*, method*, project*, publication*, publish*, research*, result*, scholarship*, scien*, studies, study*, test, tested, testing, tests, theor*, vita, vitas
Standout adjectives[24,25]	amazing, *excellen*, exceptional*, extraordinar*, fabulous*, magnificent, most, outstanding, remarkable, suberb*, suprem*, terrific*, unique, unmatched, unparalleled, wonderf*

Abbreviation: LIWC indicates Linguistic Inquiry Word Count text analysis software program.

[a] An asterisk (*) indicates root words counted with any ending; some words appear in more than one category.

[b] Category condensed to accommodate space; full category available from authors on request.

and gathered experts for a final vote.[29,31] To further validate these categories, we recruited two students to rate on Likert scales the levels of negative or positive evaluation words in critiques. Students' ratings and LIWC output were correlated for both positive ($r = 0.22$) and negative ($r = 0.24$) evaluation words (all $P < .01$).

We imported LIWC results and priority scores into IBM SPSS statistical software, version 20.0 (Armonk, New York; IBM Corp., 2011), and matched these data to applicant IDs and Summary Statement and applicant information. We analyzed all data with linear mixed-effects models and deemed P values ≤ .05 as statistically significant.

Results

Out of 76 eligible PIs identified from the NIH's CRISP database, 67 (88%) participated. Of these 67 participants, 44 (66%) were male and 23 (34%) were female; 59 (88%) held PhDs; 17 (25%) proposed clinical research, 12 (27% of 44) male, 5 (22% of 23) female; and 54 (80%) were experienced investigators. Our final sample included 454 critiques (262 from 91 unfunded and 192 from 67 funded applications). Investigators were from 45 different departments, and 15 of the NIH's 27 ICs funded their applications. There were between 2 and 5 critiques from each unfunded and between 2 and 4 critiques from each funded application; 28 investigators (42%) had two unfunded applications.

We computed the intraclass correlation coefficient (ICC) for each linguistic variable[32,33] and identified significant between-subject variation in each word category (word count = 11.3%; achievement = 14%; ability = 18.3%; agentic words = 22.3%; negative evaluation = 22.8%; positive evaluation words = 11%; research = 37%; and standout = 41%). We modeled each linguistic word category as a dependent variable with application funding outcome (unfunded versus funded), applicant experience level (new versus experienced investigator), and applicant sex (M versus F) as fixed effects. Models included applicant IDs as a random effect and used restricted maximum likelihood (REML) estimation. Initial models assessed main effects, and subsequent models included interaction terms.

Models showed a main effect for funding outcome for five word categories. Critiques of funded applications contained significantly more ability, agentic, standout, and positive evaluation words, and significantly fewer negative evaluation words, than critiques of unfunded applications (Table 2) (all $P < .05$). There were also main effects for experience level and applicant sex for four word categories. Critiques of experienced investigators' applications contained significantly more ability, agentic, standout, and positive evaluation words than critiques of new investigators' applications (all $P < .05$). Critiques of female investigators' applications contained significantly more words from the ability, agentic, and standout categories and significantly fewer negative evaluation words than those of male investigators (all $P < .05$).

Main effects were qualified by significant three-way interactions between funding outcome, investigator experience level, and applicant sex for ability ($\beta = 0.40$, $t[397] = 2.76$, $P = .006$), agentic ($\beta = 0.70$, $t[402] = 2.62$, $P = .009$), positive evaluation ($\beta = -0.97$, $t[397] = -2.67$, $P = .008$), and standout words ($\beta = 0.11$, $t[391] = 2.64$, $P = .009$); and two-way interactions between funding outcome and applicant sex ($\beta = -0.55$, $t[412] = -2.73$, $P = .007$) and experience level and applicant sex ($\beta = -0.32$, $t[166] = -2.17$, $P = .032$) for negative evaluation words. To probe these results we performed pairwise comparisons, based on estimated marginal means (Table 3), on three-way interaction terms. We used the Bonferroni correction to adjust P values.

There were no significant linguistic category differences between male and female new investigators' unfunded application critiques (Table 3). However, critiques of funded applications from female new investigators contained significantly more positive evaluation ($F[1, 175] = 13.1$, $P < .001$) and standout words ($F[1, 126] = 7.74$, $P = .006$) and significantly fewer negative evaluation words than those from male new investigators ($F[1, 272] = 19$, $P < .001$) (Figure 1).

Pairwise comparisons showed significantly more standout and significantly fewer

Table 2

The Average Percentage of Words in Unfunded Versus Funded, New Versus Experienced Investigators', and Male Versus Female Investigators' R01 Application Critiques, From a Text Analysis Study of 454 Critiques, University of Wisconsin–Madison, Fiscal Year 2008–2009[a]

	Funding outcome		Experience level		Applicant sex	
Linguistic category	Unfunded (n = 262)	Funded (n = 192)	New (n = 86)	Experienced (n = 368)	Male (n = 292)	Female (n = 162)
Ability	0.49 (0.03)	0.63 (0.03)[b]	0.49 (0.05)	0.63 (0.03)[c]	0.48 (0.03)	0.64 (0.04)[d]
Achievement	2.83 (0.08)	2.95 (0.08)	2.81 (0.13)	2.98 (0.07)	2.82 (0.09)	2.70 (0.11)
Agentic	1.04 (0.05)	1.35 (0.06)[b]	1.04 (0.09)	1.34 (0.04)[c]	0.92 (0.06)	1.46 (0.07)[d]
Negative	1.78 (0.04)	1.61 (0.04)[b]	1.72 (0.05)	1.67 (0.03)	1.97 (0.04)	1.42 (0.04)[d]
Positive	2.24 (0.07)	2.60 (0.07)[b]	2.24 (0.11)	2.61 (0.06)[c]	2.41 (0.07)	2.44 (0.09)
Research	2.67 (0.12)	2.66 (0.12)	2.58 (0.20)	2.75 (0.10)	2.62 (0.13)	2.71 (0.16)
Standout	0.14 (0.01)	0.20 (0.01)[b]	0.15 (0.02)	0.19 (0.01)[c]	0.09 (0.01)	0.24 (0.01)[d]

[a] Numbers in table reflect estimated marginal means (and standard errors) of words by linguistic category in critiques of applications from 67 investigators. N = 454 total critiques; n in table = number of critiques per category.
[b] Difference in means of unfunded versus funded grant critiques is significant at $P < .05$.
[c] Difference in means of new versus experienced investigators grant critiques is significant at $P < .05$.
[d] Difference in means of male versus female investigators' grant critiques is significant at $P < .05$.

negative evaluation words in female than in male experienced investigators' critiques from both unfunded and funded applications. Female experienced investigators' critiques from funded applications also contained significantly more ability and agentic words (all $P < .01$). Experienced investigators can submit either Type 1 (new R01) or Type 2 (renewal) applications, so we segregated their text analysis results and computed another set of linear mixed-effects models using REML estimators for each linguistic category with funding outcome (unfunded versus funded), applicant sex (M versus F), and application type (Type 1 versus Type 2) as fixed effects; we included interaction terms. Models used applicant IDs as a random effect. Models showed a significant three-way interaction effect between funding outcome, applicant sex, and application type for ability ($\beta = 0.39$, $t[320] = 2.91$, $P = .004$), agentic ($\beta = 1.16$, $t[329] = 4.76$, $P < .001$), and standout words ($\beta = -0.10$, $t[319] = -2.67$, $P = .008$); and a significant two-way interaction effect between funding outcome and applicant sex for negative evaluation words ($\beta = -0.44$, $t[360] = -2.77$, $P = .006$).

Pairwise comparisons showed that compared with critiques of applications from equivalent male investigators, only critiques of funded Type 2 applications from female experienced investigators contained significantly more ability words (F[1, 114] = 50.61, $P < .001$), and only critiques of unfunded Type 2 applications from female experienced investigators contained significantly more standout words (F[1, 80] = 41, $P < .001$) (Figure 1). Critiques of both Type 1 and Type 2 funded applications from female experienced investigators contained significantly more standout and agentic words (all $P < .01$). Negative evaluation words occurred significantly more often in male than female experienced investigators' Type 1 and Type 2 critiques from both unfunded and funded applications (all $P < .01$). Models showed no significant differences in word counts or in research or achievement words.

We found no significant correlation between study variables and an LIWC category called "negate" (e.g., not, never). This suggests that words from each linguistic category do not co-occur in critiques with words that would reverse their meaning (e.g., "not" enthusiastic).

Table 3

The Average Percentage of Words From Linguistic Categories in Critiques of Unfunded and Funded R01 Applications From Male and Female Investigators by Experience Level and Application Type, From a Text Analysis Study of 454 Critiques, University of Wisconsin–Madison, Fiscal Year 2008–2009[a]

					Experienced investigators							
	New investigators				Type 1				Type 2			
	Male		Female		Male		Female		Male		Female	
Linguistic category	Unfd n = 27	Fd n = 18	Unfd n = 22	Fd n = 19	Unfd n = 70	Fd n = 56	Unfd n = 28	Fd n = 17	Unfd n = 70	Fd n = 51	Unfd n = 45	Fd n = 31
Word count	1,031 (134)	1,096 (154)	767 (145)	828 (153)	1,107 (88)	915 (93)	1,024 (143)	837 (168)	1,211 (88)	1,026 (97)	993 (111)	777 (124)
Ability	0.41 (0.08)	0.49 (0.09)	0.46 (0.08)	0.57 (0.09)	0.60 (0.48)	0.56 (0.05)	0.57 (0.08)	0.70 (0.09)	0.45 (0.05)	0.53 (0.05)	0.50 (0.06)	1.14[c] (0.07)
Achievement	2.60 (0.22)	2.87 (0.24)	2.97 (0.23)	2.83 (0.24)	2.68 (0.97)	3.04 (0.97)	2.94 (0.98)	3.55 (0.99)	3.03 (0.97)	2.85 (0.97)	3.00 (0.97)	2.96 (0.98)
Agentic	0.96 (0.13)	0.90 (0.15)	0.96 (0.14)	1.26 (0.15)	1.02 (0.07)	1.04 (0.08)	1.19 (0.11)	1.52[c] (0.15)	1.00 (0.07)	1.01 (0.08)	1.24 (0.09)	2.72[c] (0.10)
Negative	1.96 (0.09)	1.87 (0.11)	1.85 (0.10)	1.21[c] (0.11)	1.98 (0.05)	1.95 (0.06)	1.60[b] (0.09)	1.13[c] (0.11)	1.97 (0.05)	1.97 (0.06)	1.50[b] (0.07)	1.16[c] (0.08)
Positive	1.76 (0.17)	2.25 (0.20)	1.82 (0.20)	3.26[c] (0.20)	2.46 (0.11)	2.73 (0.11)	2.26 (0.70)	2.41 (0.21)	2.58 (0.11)	2.74 (0.12)	2.53 (0.13)	2.74 (0.15)
Research	2.54 (0.30)	2.22 (0.31)	2.74 (0.32)	2.81 (0.32)	2.61 (0.20)	2.78 (0.20)	2.67 (0.25)	2.23 (0.30)	2.80 (0.20)	2.86 (0.21)	2.96 (0.21)	2.92 (0.24)
Standout	0.11 (0.03)	0.11 (0.03)	0.15 (0.03)	0.21[c] (0.03)	0.12 (0.01)	0.11 (0.01)	0.12 (0.02)	0.36[c] (0.03)	0.11 (0.01)	0.11 (0.01)	0.24[b] (0.02)	0.39[c] (0.02)

Abbreviations: Unfd indicates unfunded; Fd, funded.

[a] Numbers in table reflect estimated marginal means (and standard errors) of words by linguistic category in critiques of unfunded (Unfd) and funded (Fd) applications from new (n = 13) and experienced Type 1 (n = 24) and Type 2 (n = 30) investigators. N = 454 total critiques; n in table = number of critiques.

[b] Male versus female investigators' unfunded grant critiques, $P < .01$.

[c] Male versus female investigators' funded grant critiques, $P < .01$.

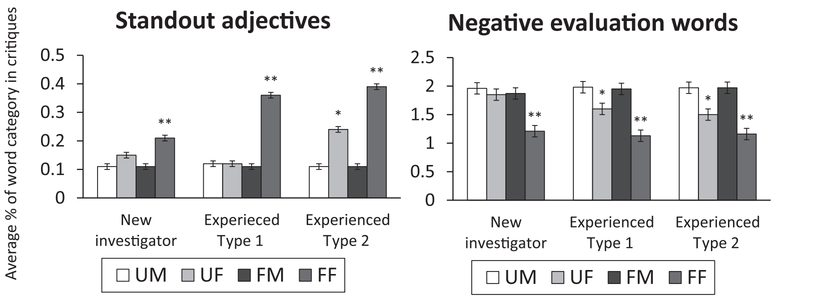

Figure 1 Average percentage of standout adjectives and negative evaluation words in National Institutes of Health R01 grant application critiques, from a text analysis study of 454 critiques, University of Wisconsin–Madison, 2008–2009. Figure reflects estimated marginal means (and standard error bars) of standout and negative words in critiques of unfunded and funded applications submitted by male and female applicants as new or experienced investigators of Type 1 or Type 2 R01s. UM indicates unfunded male; UF, unfunded female; FM, funded male; FF, funded female.
*UM versus UF, $P < .01$.
**FM versus FF, $P < .01$.

Priority scores assigned to 118 applications from 53 (98%) experienced investigators were available for analysis (54 scores from Type 1 [30 unfunded, 24 funded] and 64 from Type 2 [35 unfunded, 29 funded] applications). We computed a linear mixed-effects model with priority score as the dependent variable—and funding outcome, applicant sex, and application type as fixed effects—and included interaction terms. Models used applicant IDs as a random effect to account for repeated measures per applicant. Results showed a significant main effect only for funding outcome (i.e., funded applications had better scores) ($\beta = -51$, $t[68] = -7.65$, $P < .01$).

Priority scores and linguistic word categories showed significant correlations indicating that lower (i.e., more competitive) scores were associated with critiques containing more words in the ability, agentic, positive evaluation, and standout categories, but fewer negative evaluation words (all $P < .05$). Separate correlations of female and male experienced investigators' data showed that correlations were significant only for female investigators (Table 4).

To test whether high or low levels of words from each category in critiques predicted priority scores and whether this differed by applicant sex, we split each LIWC word category at its median to create dichotomous indicators of "high" versus "low" levels of words. We then analyzed experienced investigators' priority scores using a set of linear mixed-effects models with each word category indicator variable (high versus low) and applicant sex (M versus F) as fixed effects; we included interaction terms. Models used applicant IDs as a random effect, and REML estimators. Models showed significant two-way interactions between applicant sex and the high/low indicators of ability ($\beta = 20$, $t[327] = 2.60$, $P = .010$), agentic ($\beta = 20$, $t[326] = 2.45$, $P = .015$), standout ($\beta = 23$, $t[317] = 2.57$, $P = .011$), and negative evaluation words ($\beta = -19$, $t[318] = -1.89$, $P = .05$). Pairwise comparisons performed on the interaction terms showed that when critiques of female experienced investigators' applications contained high levels of words from the standout ($F[1, 317] = 7.58$, $P = .006$), ability ($F[1, 323] = 8.36$, $P = .004$), and agentic categories ($F[1, 322] = 8.17$, $P = .05$) and low levels of negative evaluation words ($F[1, 319] = 4.69$, $P = .031$), their proposals were assigned significantly lower (i.e., more competitive) scores. By comparison, male experienced investigators' priority scores did not differ significantly by the levels of linguistic category words in their critiques.

Discussion

Our findings suggest that text analysis of application critiques is a promising tool for evaluating potential bias in peer review of NIH R01 grant applications. Text analysis appropriately sorted R01 applications that were unfunded from those that were funded as well as those from new investigators versus experienced investigators. Overall, critiques of funded R01s contained more positive evaluation and standout words, more references to ability and competence (e.g., agentic words), and fewer negative evaluation words than critiques of unfunded applications. Critiques of experienced investigators' applications contained more words from the ability, agentic, standout adjective, and positive evaluation categories than critiques of new investigators' applications. However, these patterns were not consistent across critiques of applications from male and female investigators, suggesting that text analysis may be able to uncover discrepancies in reviewers' judgments that are masked when only scores or funding outcomes are compared. We identified three patterns of differences in R01 critiques by applicant sex that occurred despite similar scores or funding outcomes: more positive descriptors, praise, and acclamation for funded applications from all types of female investigators; greater reference to competence and ability for funded applications from female experienced investigators, particularly for renewals; and more negative evaluation words for applications from all types of male investigators. Subanalyses of experienced investigators' data again confirmed the potential of text analysis to uncover discrepancies in reviewers' judgments masked by comparing scores and funding outcomes alone. High levels of standout, ability, and agentic words and low levels of negative evaluation words in critiques predicted more competitive priority scores—as one would expect—but only for female experienced investigators.

When taken together, our findings suggest that subtle gender bias may operate in R01 peer review. Such gender bias may be unconscious and derives from pervasive cultural stereotypes that women have lower competence than men in fields like academic medicine and science where men have historically predominated.[14,34,35] A large body of experimental research concludes that in such male-typed domains, gender stereotypes lead evaluators to give a woman greater praise than a man for the same performance.[21,36–38] By comparison, the assumption of men's competence in male-typed domains leads evaluators to more often notice and document negative performance from men because it is not expected.[37,39] This could be one interpretation of the comparable scores and funding outcomes for male and female investigators despite the greater occurrence of negative evaluation words in critiques of men's proposals and the

Table 4

Correlations Between Priority Scores and Words in LIWC Categories in Critiques of R01 Applications From Experienced Male and Female Investigators, From a Text Analysis Study of 454 Critiques, University of Wisconsin–Madison, Fiscal Year 2008–2009[a]

	Variable	1	2	3	4	5	6	7	8
1	Priority scores	—	−.417[b]	.008	−.406[b]	.290 [b]	−.199[c]	−.071	−.425[b]
2	Ability	−.061	—	.073	.575[b]	−.336[b]	.072	−.041	.395[b]
3	Achievement	−.070	.075	—	.161	−.005	.300[b]	−.066	.078
4	Agentic	−.089	.430[b]	.293[b]	—	−.130	.275[b]	−.063	.386[b]
5	Negative	.020	.052	.083	.074	—	−.087	.171	−.346[b]
6	Positive	−.112	−.007	.204[b]	−.014	−.003	—	.024	.176
7	Research	.101	.076	.173[b]	.167[b]	−.013	.101	—	.022
8	Standout	−.110	−.017	.030	.051	.330[b]	−.119	−.038	—

Abbreviation: LIWC indicates Linguistic Inquiry Word Count text analysis software program.

[a] Correlations between experienced female investigators' scores and word categories in critiques are above the diagonal; experienced male investigators' are shown below.

[b] Correlation is significant, $P < .01$.

[c] Correlation is significant, $P < .05$.

apparent stronger critiques of women's proposals. Male and female evaluators are equally prone to such gender-biased judgments.[40,41]

Paradoxically, gender stereotypes also lead reviewers to require more proof of ability from a woman than a man prior to confirming her competence,[37,42] and greater proof to confirm men's incompetence in male-typed domains.[37,39] This may also explain why men's versus women's proposals were funded despite more negative critiques (i.e., higher standards for incompetence), and why there were more references to ability and competence in critiques of applications from female versus male experienced investigators. Being an experienced investigator, particularly one renewing an R01, conflates two male-typed domains: science and leadership.[43,44] Therefore, women scientific leaders would be held to the highest ability standards to confirm competence.[21,37]

Despite the more laudatory critiques of women's applications, particularly for renewals, we cannot conclude from our study that women needed to outperform men to receive an R01. If this were the case, we would expect to find that women had to earn more competitive scores than men to have their applications funded, which we did not. We think it is more likely that the same level of performance was interpreted in gender-stereotypic ways, leading to more positive commentary about women's applications. This does not fully rule out the possibility that gender stereotypes may also inadvertently influence reviewers to hold female investigators to a higher standard of competence. Such gender bias could help explain results from prior studies showing lower success rates for female versus male R01 applicants, particularly for renewals.[9,10,45]

A potential limitation of our text analysis is that although we selected word categories relevant to grant review, we did not include all possible categories. Another limitation is that the LIWC software program does not account for the context in which words are used, but word categories showed no correlation with "negate" words (i.e., that would reverse their meaning), and positive and negative ratings of critiques from our sample correlated with LIWC output for our positive and negative evaluation categories. Our study is limited to a single site, but UW-Madison is similar to other large public research institutions, and proposals represented 45 departments and were reviewed across 15 NIH ICs. Participant selection bias is possible, although we had an 88% response rate from all eligible PIs. We studied no critiques from initially funded or from unfunded, unresubmitted applications. We cannot rule out the possibility that observed differences in critiques occurred because of differences in background qualifications or because men and women are engaged in different research areas, but about 90% of our sample held PhDs, and similar proportions of male and female PIs performed clinical research.

We studied R01 critiques before the NIH implemented its streamlined review process. Priority scores and commentary based on the five criteria areas continue to be used to evaluate R01 applications; however, changes include a broader range for impact/priority scores, scoring for the five criteria, use of a bullet-point format instead of a narrative format for critiques, and limit of a single resubmission for unfunded applications. These changes might reduce the amount of text available for analysis, but we have no reason to believe that they would change the impact of cultural stereotypes on judgment. Findings from our study may provide a useful comparison point for future studies of the impact of streamlining on R01 peer review. Our results should encourage future experimental studies. If stereotype-based bias is confirmed, promising interventions that foster behavioral change could be studied in the context of R01 peer review.[46,47]

The NIH R01 is critical for launching the scientific careers of new investigators and in maintaining the research programs of experienced investigators. Promoting an equitable peer review process will ensure that the best and most innovative research will advance. Our text analysis of R01 critiques suggests that gender stereotypes may infuse subtle bias in the R01 peer review process.

Funding/Support: This work was supported by National Institutes of Health (NIH) grants R01 GM111002, R01 GM088477, DP4 GM096822, and R25 GM083252.

Other disclosures: None reported.

Ethical approval: The institutional review board at the University of Wisconsin–Madison approved all aspects of this study.

Previous presentations: Data were presented in part at the NIH workshop on Brainstorming Ideas for Conducting Studies With the Peer Review System: Strategies for Enhancing the Diversity of the Biomedical Research Workforce, March 28, 2012, Bethesda, Maryland.

Dr. Kaatz is assistant scientist, Center for Women's Health Research, University of Wisconsin–Madison, Madison, Wisconsin.

Ms. Magua is a doctoral candidate, Department of Industrial and Systems Engineering, University of Wisconsin–Madison, Madison, Wisconsin.

Dr. Zimmerman is professor, Department of Industrial and Systems Engineering, University of Wisconsin–Madison, Madison, Wisconsin.

Dr. Carnes is director, Center for Women's Health Research, and professor, Departments of Medicine, Psychiatry, and Industrial and Systems Engineering, University of Wisconsin–Madison, Madison, Wisconsin; and part-time physician, William S. Middleton Memorial Veterans Hospital Geriatric Research Education and Clinical Center, Madison, Wisconsin.

References

1 National Research Council of the National Academies. Bridges to Independence: Fostering the Independence of New Investigators in Biomedical Research. Washington, DC: National Academies Press; 2005.

2 Office of Extramural Research, National Institutes of Health. Grants and Funding: Peer Review Process. http://grants.nih.gov/grants/peer_review_process.htm. Accessed June 9, 2014.

3 Office of Extramural Research, National Institutes of Health. Grants and Funding: NIH Research Project Grant Program (R01). http://grants.nih.gov/grants/funding/r01.htm. Accessed June 9, 2014.

4 Johnson VE. Statistical analysis of the National Institutes of Health peer review system. Proc Natl Acad Sci U S A. 2008;105:11076–11080.

5 Kaplan D, Lacetera N, Kaplan C. Sample size and precision in NIH peer review. PLoS One. 2008;3:e2761.

6 Martin MR, Kopstein A, Janice JM. An analysis of preliminary and post-discussion priority scores for grant applications peer reviewed by the Center for Scientific Review at the NIH. PLoS One. 2010;5:e13526.

7 Ginther DK, Schaffer WT, Schnell J, et al. Race, ethnicity, and NIH research awards. Science. 2011;333:1015–1019.

8 Kotchen TA, Lindquist T, Miller Sostek A, Hoffmann R, Malik K, Stanfield B. Outcomes of National Institutes of Health peer review of clinical grant applications. J Investig Med. 2006;54:13–19.

9 Ley TJ, Hamilton BH. The gender gap in NIH grant applications. Science. 2008;322:1472–1474.

10 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.

11 Dickler HB, Fang D, Heinig SJ, Johnson E, Korn D. New physician–investigators receiving National Institutes of Health research project grants. JAMA. 2007;297:2496–2501.

12 Kotchen TA, Lindquist T, Malik K, Ehrenfeld E. NIH peer review of grant applications for clinical research. JAMA. 2004;291:836–843.

13 Ginther DK, Haak LL, Schaffer WT, Kington R. Are race, ethnicity, and medical school affiliation associated with NIH R01 type 1 award probability for physician investigators? Acad Med. 2012;87:1516–1524.

14 National Academy of Sciences, National Academy of Engineering, Institute of Medicine of the National Academies. Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering. Washington, DC: National Academies Press; 2007.

15 Chubin DE, Hackett EJ. Peerless Science: Peer Review and US Science Policy. Albany, NY: SUNY Press; 1990.

16 National Academies of Sciences. Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future: Washington, DC: National Academies Press; 2007.

17 Working Group on Diversity in the Biomedical Research Workforce (WGDBRW), Advisory Committee to the Director (ACD), National Institutes of Health. Draft Report of the Advisory Committee to the Director Working Group on Diversity in the Biomedical Research Workforce. http://acd.od.nih.gov/Diversity%20in%20the%20Biomedical%20Research%20Workforce%20Report.pdf. Accessed June 9, 2014.

18 Tabak LA. An Initiative to Increase the Diversity of the NIH-Funded Workforce. http://acd.od.nih.gov/Diversity-in-the-Biomedical-Workforce-Implementation-Plan.pdf. Accessed June 9, 2014.

19 Jagsi R, Motomura AR, Griffith KA, Rangarajan S, Ubel PA. Sex differences in attainment of independent funding by career development awardees. Ann Intern Med. 2009;151:804–811.

20 Biernat M, Eidelman S. Translating subjective language in letters of recommendation: The case of the sexist professor. Eur J Soc Psychol. 2007;37:1149–1175.

21 Biernat M, Tocci MJ, Williams JC. The language of performance evaluations: Gender-based shifts in content and consistency of judgment. Soc Psychol Pers Sci. 2012;3:186–192.

22 Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in medical student performance evaluations. Eval Health Prof. 2010;33:365–385.

23 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.

24 Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: Agentic and communal differences. J Appl Psychol. 2009;94:1591–1599.

25 Schmader T, Whitehead J, Wysocki VH. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex Roles. 2007;57:509–514.

26 Trix F, Psenka C. Exploring the color of glass: Letters of recommendation for female and male medical faculty. Discourse Soc. 2003;14:191–220.

27 Office of Extramural Research, National Institutes of Health. Grants and Funding: New and Early Stage Investigator Policies. http://grants.nih.gov/grants/new_investigators/. Accessed June 9, 2014.

28 Pennebaker J, Chung C, Ireland M, Gonzales A, Booth R. The Development and Psychometric Properties of LIWC2007. Austin, Tex: LIWC; 2007.

29 Custer RL, Scarcella JA, Stewart JA. The modified Delphi technique—a rotational modification. J Vocat Technol Educ. 1999;15(2):50–58.

30 de Villiers MR, de Villiers PJ, Kent AP. The Delphi technique in health sciences education research. Med Teach. 2005;27:639–643.

31 Landeta J, Barrutia J, Lertxundi A. Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. Technol Forecast Soc Change. 2011;78:1629–1641.

32 Hox JJ. Multilevel modeling: When and why. In: Balderjahn I, Mathar R, Schader M, eds. Classification, Data Analysis and Data Highways. New York, NY: Springer; 1998:147–154.

33 West BT, Welch KB, Galecki AT. Linear Mixed Models: A Practical Guide Using Statistical Software. Boca Raton, Fla: Chapman & Hall/CRC; 2007.

34 Handelsman J, Cantor N, Carnes M, et al. Careers in science. More women in science. Science. 2005;309:1190–1191.

35 Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci U S A. 2012;109:16474–16479.

36 Biernat M. Stereotypes and shifting standards: Forming, communicating and translating person impressions. In: Devine PGP, Plant A, eds. Advances in Experimental Social Psychology. Vol 45. San Diego, Calif: Academic Press; 2012:1–50.

37 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.

38 Biernat M, Vescio TK. She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. Pers Soc Psychol Bull. 2002;28:66–77.

39 Biernat M, Fuegen K, Kobrynowicz D. Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. Pers Soc Psychol Bull. 2010;36:855–868.

40 Devine PG. Stereotypes and prejudice: Their automatic and controlled components. J Pers Soc Psychol. 1989;56:5–18.

41 Nosek BA, Smyth FL, Hansen JJ, et al. Pervasiveness and correlates of implicit attitudes and stereotypes. Eur Rev Soc Psychol. 2007;18:36–88.

42 Foschi M. Double standards in the evaluation of men and women. Soc Psychol Q. 1996:59:237–254.

43 Marchant A, Bhattacharya A, Carnes M. Can the language of tenure criteria influence women's academic advancement? J Womens Health (Larchmt). 2007;16:998–1003.

44 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.

45 National Institutes of Health Research Portfolio Online Reporting Tools. R01-Equivalent Grants: Success Rates, by Gender and Type of Application. National Institutes of Health. NIH IMPAC, Success Rate File. http://report.nih.gov/NIHDatabook/Charts/Default.aspx?showm=Y&chartId=178&catId=15. Accessed June 9, 2014.

46 Carnes M, Devine PG, Isaac C, et al. Promoting institutional change through bias literacy. J Divers High Educ. 2012;5:63–77.

47 Devine PG, Forscher PS, Austin AJ, Cox WTL. Long-term reduction in implicit race prejudice: A prejudice habit-breaking intervention. J Exp Soc Psychol. 2012;48:1267–1278.