

1-28-2019

Little race or gender bias in an experiment of initial review of NIH R01 grant proposals

Patrick S. Forscher

University of Arkansas, Fayetteville, forscher@uark.edu

William T.L. Cox

University of Wisconsin-Madison


Markus Brauer

University of Wisconsin - Madison

Patricia G. Devine

University of Wisconsin-Madison

Follow this and additional works at: <https://scholarworks.uark.edu/psycpub>

 Part of the [African American Studies Commons](#), [Asian American Studies Commons](#), [Chicana/o Studies Commons](#), [Latina/o Studies Commons](#), and the [Other Feminist, Gender, and Sexuality Studies Commons](#)

Recommended Citation

Forscher, Patrick S.; Cox, William T.L.; Brauer, Markus; and Devine, Patricia G., "Little race or gender bias in an experiment of initial review of NIH R01 grant proposals" (2019). *Psychological Science Faculty Publications and Presentations*. 2.
<https://scholarworks.uark.edu/psycpub/2>

This Article is brought to you for free and open access by the Psychological Science at ScholarWorks@UARK. It has been accepted for inclusion in Psychological Science Faculty Publications and Presentations by an authorized administrator of ScholarWorks@UARK. For more information, please contact ccmiddle@uark.edu.

Title: Little race or gender bias in an experiment of initial review of NIH R01 grant proposals

Authors: Patrick S. Forscher^{1,2}, William T. L. Cox¹, Markus Brauer¹, Patricia G. Devine¹

Affiliations: ¹University of Wisconsin – Madison; ²University of Arkansas; Correspondence to: Patrick Forscher at schnarrd@gmail.com and/or Patricia Devine at pgdevine@wisc.edu.

Abstract: Many granting agencies allow reviewers to know the identity of a proposal's Principal Investigator (PI), which opens the possibility that reviewers discriminate on the basis of PI race and gender. We investigated this experimentally with 48 NIH R01 grant proposals, representing a broad spectrum of NIH-funded science. We modified PI names to create separate White male, White female, Black male, and Black female versions of each proposal, and 412 scientists each submitted initial reviews for three proposals. We find little to no race or gender bias in initial R01 evaluations, and additionally find that any bias that might have been present must be negligible in size. This conclusion was robust to a wide array of statistical model specifications. Pragmatically important bias may be present in other aspects of the granting process, but our evidence suggests that it is not present in the initial round of R01 reviews.

Preprint revised 2018-11-01

Main Text:

Grants are the engine of scientific innovation. As such, the fair evaluation of grant proposals has implications for both the speed of scientific discovery and the career trajectories of individual scientists. In the United States, the National Institutes of Health's (NIH's) R01 is the primary mechanism through which grants are awarded.

In the R01 review process, reviewers know the identity of each application's Principal Investigator (PI) and are explicitly required to evaluate the PI as one of the review criteria. It is therefore possible that personal characteristics of the PI that are irrelevant to the proposal's scientific merit, such as their race or gender, affect reviewers' evaluations. Indeed, Black PIs are funded at lower rates than White PIs¹, and although initial submissions of male and female PIs are funded at similar rates², pro-male gaps emerge in resubmissions³.

However, gaps in funding rates could be caused by many other processes besides reviewer discrimination. Compared to White male PIs, PIs who belong to other social categories could, for example, experience less effective mentoring or have access to fewer resources during grant preparation. They could also use less bold language in their applications or apply to more competitive research areas. Establishing whether the perceived race or gender of a PI exerts a causal influence on a reviewer's proposal evaluations requires a randomized, controlled experiment in which the PI's perceived social category is manipulated and all other factors are held constant.

We conducted just such an experiment to examine one particular stage of the NIH review process: the initial round of reviews. In the initial round, reviewers independently read and evaluate around 10 proposals, a third of which are their primary responsibility. We obtained both 24 funded and 24 unfunded R01 grant proposals for scientists to evaluate as primary reviewers. At NIH, proposals are reviewed by study sections and funded by institutes. Our proposals came from twelve study sections that, together, broadly represented the science funded by the four largest institutes at NIH (see Table S1). We collected four proposals per study section; two were high quality, funded proposals with strong Priority Scores, and two were moderate quality, unfunded proposals with relatively weak Priority Scores.^a Our 48 stimulus proposals captured a broad range of quality. The mean Priority Score of the high quality proposals ($M = 1.9$, $SD = .65$, $Min = 1.4$, $Max = 2.7$)^b was a full two points better on the 1 (exceptional) to 9 (poor) Priority Score scale than the moderate quality proposals ($M = 3.9$, $SD = .36$, $Min = 2.7$, $Max = 5.7$; four not discussed and therefore unscored). We removed the real PI's identifying information and created multiple versions of each proposal by assigning it one of several fictitious names. These versions implied the PI was a White male, White female, Black male, or Black female.

We used the NIH RePORTER database to recruit scientists whose expertise matched the content of the grant proposals and supplemented these with suggestions from the prospective

^a The "moderate quality" proposals, although initially unfunded, were eventually funded after one or more rounds of revision and resubmission. The participant-reviewers in our study only evaluated the initial, unfunded versions of these proposals. See the Methods for more details.

^b Priority Scores are formed by multiplying averaged Overall Impact scores by 10, yielding a 10-90 scale. However, we use a 1-9 scale throughout the text for comparability with the scale used at the stage of reviews.

reviewers. We attempted to screen out scientists who were familiar with the original PIs and proposals through an “eligibility survey”. Scientists who had previously served as NIH reviewers were asked whether they had served on the NIH study section(s) that had previously reviewed our proposals when the proposals were under review, allowing us to exclude people who may have encountered the proposals during their NIH service. We also asked reviewers to look at a list of researchers’ names containing both the original PIs and our fictitious PIs, and, on the pretext of avoiding conflicts of interest, asked the prospective reviewers to select the PIs with whom they were familiar. This strategy allowed us to avoid assigning the reviewers proposals written by PIs with whom they were familiar.

We did not request demographic information from our reviewers. As inferred from their institutional websites, our final sample ($N = 412$ reviewers) was predominantly White (59%) and male (76%); almost half (45%) were both White and male. NIH does not publicly release demographic information about its reviewers, so we do not know how the demographics of our reviewers compare to NIH’s pool. However, the majority (58%) of our reviewers reported past review experience on an NIH study section, suggesting that the two pools are demographically similar.

Reviewers were informed that we were studying the NIH review process and that they would evaluate modified versions of actual R01 proposals, though we did not tell reviewers the nature of these modifications. In exchange for \$300, each reviewer evaluated a set of three proposals as primary reviewer, a number similar to the number of first-stage primary reviews requested by NIH. Two of the set of three proposals were ostensibly written by White male PIs (one high quality and one moderate quality). The third proposal was either high or moderate quality and, depending on experimental condition, was ostensibly written by a White female, Black male, or Black female PI. To avoid arousing suspicion as to the purpose of the study, no reviewer was asked to evaluate more than one proposal written by a non-White-male PI. This design allows us to isolate the causal role of perceived PI demographic characteristics on scores and written critiques, independent of the characteristics of reviewers or proposals.

Reviewers used the official NIH rubric, and hence provided critiques and scores on a 1 (exceptional) to 9 (poor) scale on each proposal’s Overall Impact, Significance, Investigator, Innovation, Approach, and Environment. To mitigate the possibility of reviewers searching for the fake PI name on the internet, reviewers were instructed not to use outside sources when reviewing the grant proposal. Despite our instructions, 139 of our reviewers told us that they used PubMed and/or looked up a paper mentioned in one of the proposals. We eliminated from analysis 34 of these reviewers who either mentioned that they learned that one of the named personnel was fictitious or who mentioned that they looked up a paper from a PI biosketch. We retained the remaining 105 reviewers for analysis, but examine how sensitive our results are to their exclusion as part of a sensitivity analysis described below.

We preregistered our analysis plan at the end of data collection (but prior to viewing any data) at <https://osf.io/vhwnd/>. In the standard NIH review process, initial Overall Impact scores are used to determine whether a proposal is discussed by the full study section, shape subsequent discussion, and provide an anchor for post-discussion Overall Impact scores, which are averaged together to form the Priority Scores that determine funding decisions. For these reasons, our primary outcome was each proposal’s Overall Impact scores.

As shown in Fig. 1, we found no evidence that White male PIs received different Overall Impact scores than PIs who were not White male. In other words, when the same proposal had a White male PI, it was evaluated no differently than when that proposal had a White female, Black male, or Black female PI. As also shown in the Methods, this pattern does not vary by grant proposal quality, scientific topic area, or whether the reviewer was a White male. Although Fig. 1 shows some indications that the variance in the average reviewer and proposal scores differs by PI race and gender, we show in the Methods that the differences in these variances were no greater than one would expect due to sampling error.

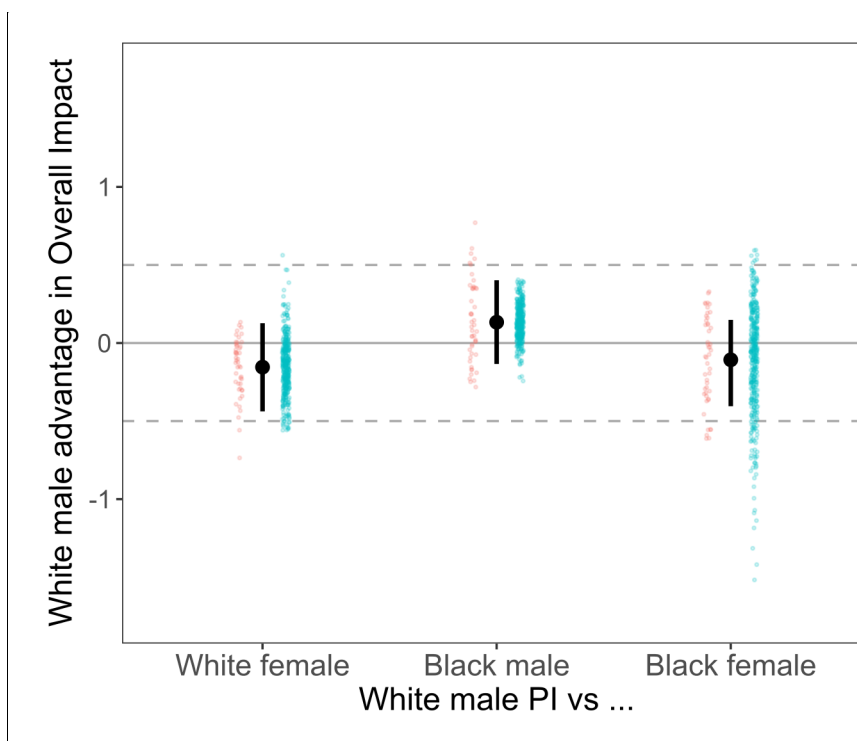


Fig. 1. Estimated differences between the Overall Impact scores given to proposals with White male PIs and each of White female, Black female, and Black female PIs. Black dots represent the mean differences, lines their bootstrapped 95% CIs. Dotted lines encompass the region bounded by a half-point difference in Overall Impact scores, which we defined as the smallest social-category-based gap that is pragmatically important. Points to the left and right of each bar represent proposal-level and participant-level random effects, respectively.

Despite this result, it is possible that pragmatically important bias is present but is too small for our experiment to detect. We assessed this possibility using an equivalence test⁴. An equivalence test involves defining the threshold above which effects are considered “pragmatically important”; the analysis then tests whether the observed effects are smaller than this threshold. We defined the threshold of “pragmatic importance” as .5 on the 1-9 Overall Impact scale because it is (1) a relatively small fraction (one quarter) of the total Priority Score gap between our groups of moderate and high quality proposals and (2) halfway between the two adjacent verbal descriptors on NIH’s 1-9 rating scale. As is shown in more detail in the Methods, our effects were significantly smaller than .5, suggesting that any bias that is present in our data is below this threshold.

	Point of flexibility	Justification
<i>Outcome</i>	(1) Overall Impact	Most pragmatically important: It has the greatest impact on funding lines
	(2) Significance	Reviewers may feel issues of non-White PIs are more "niche"
	(3) Investigator	We manipulated PI race/gender, which are investigator characteristics. Bias may therefore show up strongest here
	(4) Innovation	Judgments of innovation are highly subjective
	(5) Approach	This is an evaluation of the actual science described in the proposal
	(6) Environment	Reviewers may assume non-White PIs have fewer institutional resources with which to complete their proposals
<i>Condition</i>	(1) One dummy-coded variable	Treats each race/gender combination as unique
	(2) Separate race and gender variables	Focuses on overall race and gender categories
<i>Quality</i>	(1) Dichotomous	Design assumes proposals in dichotomous quality categories
	(2) Quantitative, grand mean centered	Dichotomous quality misses variation in the Priority Scores within the "moderate" and "high" categories
	(3) Quantitative, cluster centered	Proposal mean centering conflates between- and within-participant variation in proposal quality
<i>Proposal-level random effects</i>	(1) Intercept only	Proposals likely vary in average scores they receive
	(2) Intercept, condition slopes	Proposals likely vary in size of condition effect
	(3) Intercept, condition slopes, intercept-slope correlations	There may be a relationship between a proposal's average scores and the size of its condition effect
<i>Reviewer-level random effects</i>	(1) Intercept only	Reviewers likely vary in the average scores they assign
	(2) Intercept, condition slopes	Reviewers likely vary in their susceptibility to condition
	(3) Intercept, condition & quality slopes	Reviewers likely vary in their susceptibility to proposal quality
	(4) Intercept, condition, quality, interaction slopes	Reviewers likely vary in their susceptibility to the condition by quality interaction
	(5) Intercept, condition slope, intercept-slope correlations	There may be a relationship between a reviewer's average scores and their susceptibility to condition
	(6) Intercept, condition & quality slopes, intercept-slope correlations	There may be a relationship between a reviewer's average scores and their susceptibility to condition, quality
	(7) Intercept, condition & quality slopes, intercept-slope correlations	There may be a relationship between a reviewer's average scores and their susceptibility to condition, quality, and the condition-quality interaction
<i>Observations</i>	(1) All	Use all reviewers who completed their reviews
	(2) Remove people who read a biosketch paper	These reviewers very likely realized some elements of study were fictitious
	(3) Remove people who read any proposal paper	These reviewers may have realized some elements of study were fictitious
	(4) Remove ratings of different-quality proposals	Different-quality proposals may have stood out and attracted different reviews
	(5) Remove both people who read a biosketch paper and different-quality proposals	Combine (2) and (4)
	(6) Remove both people who read any proposal paper and different-quality proposals	Combine (3) and (4)

Table 1. Reasonable alternatives for how to analyze our data to test for bias in review scores. In combination, the alternatives yield 4,536 analytic models.

Other researchers could reasonably disagree with the decisions we made when specifying our statistical model. For example, our preregistered analysis treated each race/gender combination as unique (i.e., White male vs. Black male vs. White female vs. Black female), but one could argue we should look at race (Black vs. White) and gender (Male vs. Female) as separate variables in our model. In addition, others could argue that any reviewer who used outside resources should be excluded from analysis due to the possibility these reviewers discovered that the proposal PIs were fictitious. Table 1 lays out these and other reasonable alternatives for how to analyze our data; in combination, these alternatives yield 4,536 analytic models.

To assess the degree to which our results change under different models, we re-analyzed our data using all 4,536 of them. The observed pattern was highly similar across models: across the coefficients that tested for pro-White, pro-male, or pro-White-male bias, 99.7% showed no statistically significant bias favoring the non-stigmatized group and 97.1% stayed significantly below the threshold of half a point, representing our definition of a pragmatically important effect (see Fig. 2 and the Methods).

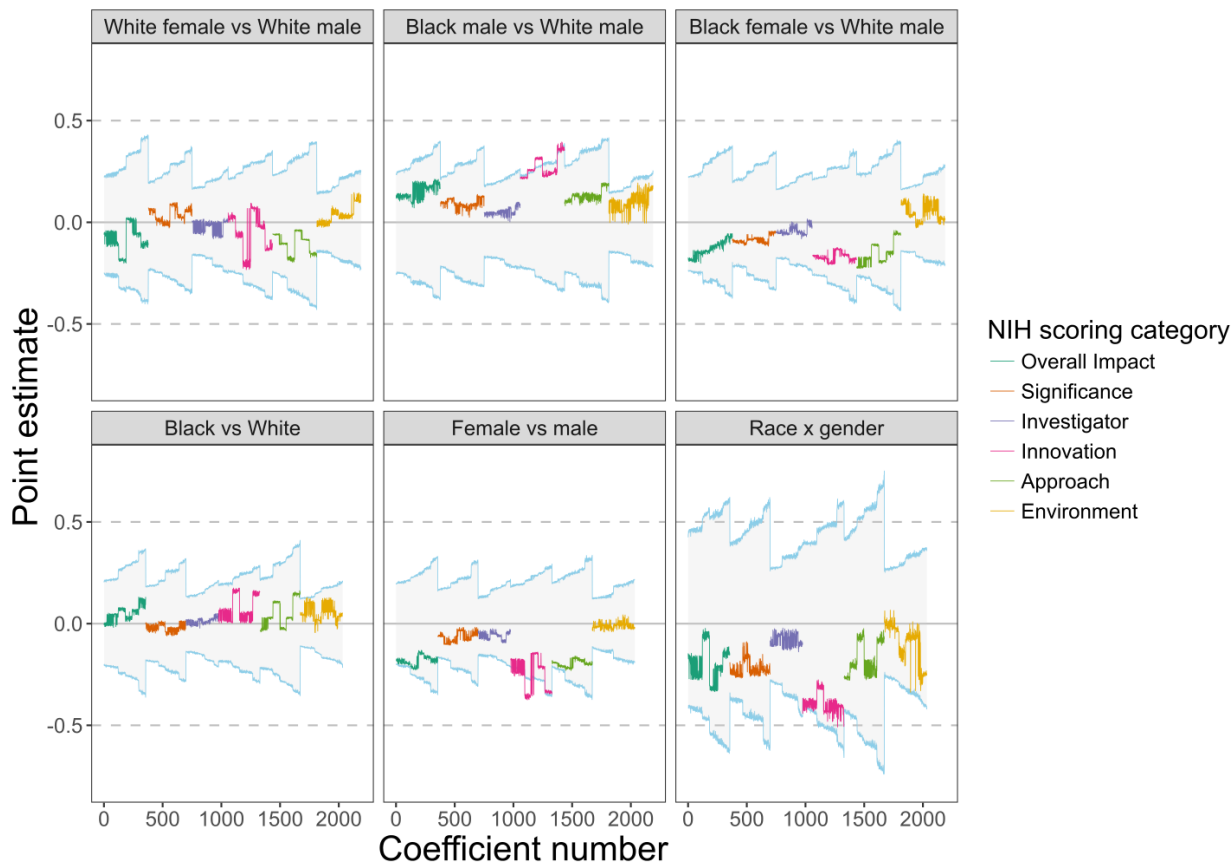


Fig. 2. Sensitivity of our results to alternative analytic models. Each multi-colored line represents a set of coefficient point estimates. Each point estimate comes from one of 4,536 analytic models. The six panels are grouped by the type of bias assessed by its point estimates (though note that the race x gender interaction cannot directly be interpreted as a test of bias). Blue lines represent the 2.5% and 97.5% quantiles, obtained via a permutation test using 500 randomly shuffled datasets, for how set of point estimates behaves under the null hypothesis. If a given point from the multi-colored lines is within the grey region bounded by the blue lines, that suggests that the point estimate value is not substantially different from what one would obtain under the null hypothesis.

Finally, we conducted exploratory analyses to examine whether reviewers used different language in their written critiques to describe PIs based on their demographics. Some past research has found that the written critiques evaluating female PIs contain language that is more positive despite similar scores, which may indicate that women need to meet higher standards to achieve the same scores⁵. To this end, we calculated, for each critique, the proportion of words falling into each of nine word categories (see Table S4) that past research has argued are relevant to grant proposal evaluation⁵. We then tested whether each of these proportions differs based on the social category membership of the PI. As shown in Fig. 3, we found no differences in any of the nine categories, and we show in the Methods section that the lack of bias was consistent across different levels of grant proposal quality.

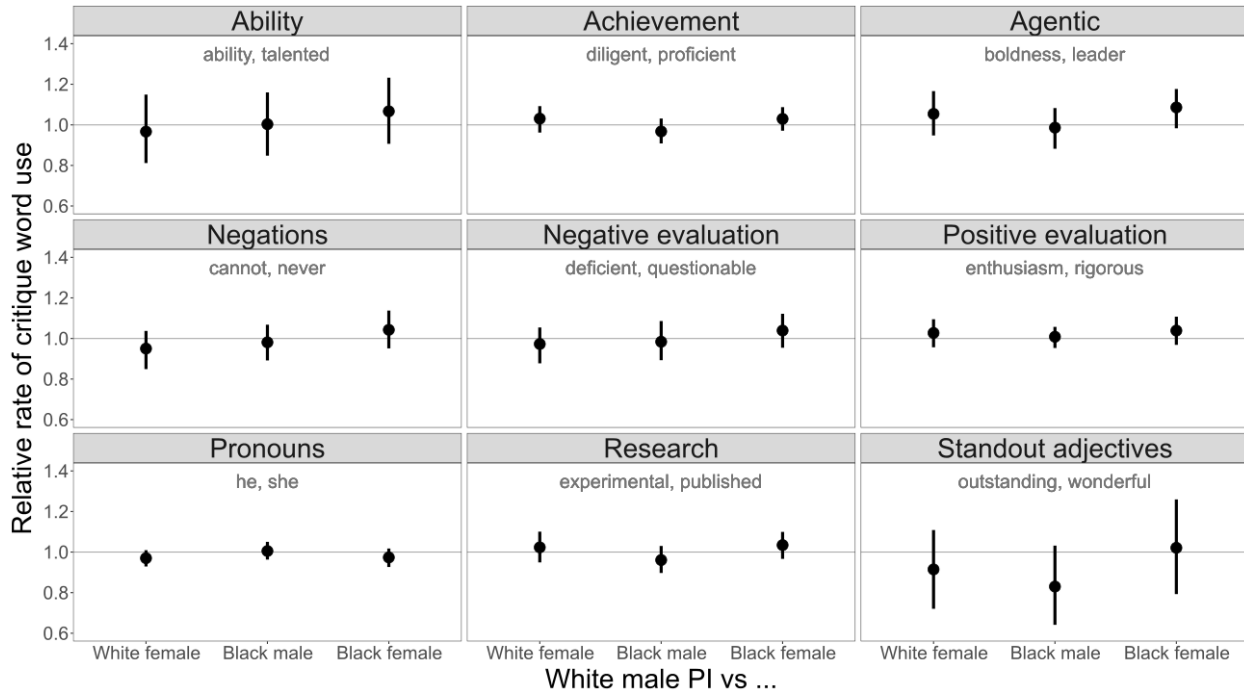


Fig. 3. The relative rate of word use in the written critiques given to White male PIs and each of White female, Black female, and Black female PIs. Each panel represents a different category of word that could plausibly be relevant to proposal evaluation. Values less than one indicate that the critiques of the non-White-male PIs used less of the word category, values above one indicate the critiques used more of the word category. Bars represent bootstrapped 95% CIs.

A skeptic of our findings might put forward two criticisms: first, our findings of little to no bias might be caused by low statistical power to detect bias rather than no bias in reviews, and second, our study bears little similarity to the true NIH review process and therefore cannot generalize to it. Our study is not vulnerable to the first criticism. As shown in the Methods, we conducted an a priori power analysis that showed that our power was very high to detect differences as small as half a point. Moreover, low statistical power decreases one's ability to reject the hypothesis of a substantively large effect in an equivalence test, and yet we were able to reject the hypothesis of a White male advantage of half a point or more.

As for the second criticism, there are indeed some real differences between the experiences of our reviewers and reviewers in the true NIH process. Our reviewers completed a similar number of primary reviews as true NIH reviewers, but did not complete secondary or tertiary reviews; it is possible that the lower workload in our study allowed reviewers to be relatively thoughtful in their reviews, decreasing bias. Our reviewers also knew they were in a study rather than an NIH study section, the consequences of which are unclear: although it could decrease bias due to the desire of reviewers to be on their best behavior⁶, this interpretation does not explain why other studies of bias have been able to demonstrate demographic-based discrimination despite telling their faculty participants that they are in a study^{7,8}. Moreover, despite these two differences, our study does bear many important similarities to the true NIH process: we used real R01 proposals that had either been funded or unfunded, the same training materials and criteria used by real reviewers, and recruited actual NIH grant-holders, the

majority of whom had reviewed for NIH in the past. We contend that our study is similar enough in its most critical features to speak to the initial stages of the true NIH review process.

Others could also disagree with the threshold that we used to define the scoring gap that is “pragmatically important”. There is some inevitable subjectivity in this assessment; there is no objectively correct threshold defining “pragmatic importance” in this or any other context. What our analysis does provide is some boundaries around how much race and gender bias could exist in the initial stage of R01 proposal review. Our evidence suggests that the amount of bias in initial reviews is smaller than half a point, which we believe is relatively small.

We can only speculate about why we found little to no race or gender bias in initial reviews. Reviewers must deliberately and systematically process a massive amount of information to adequately evaluate the grant proposals. Each reviewer must also justify their scores to the full study section during the latter stages of review, which forces them to be accountable for their scores⁹, a feature that should also work to mitigate the influence of bias. Limiting our attention to gender discrimination, ours is not the first study to have found little evidence of a pro-male preference¹⁰⁻¹⁶, suggesting that reviewers may have little gender bias that influences their reviews. Discovering why initial reviews do not appear to be subject to race or gender bias may help researchers and policy-makers build bias-mitigating features into other review processes.

Our conclusion of little to no bias in initial reviews does not imply that bias is absent from all other stages of the granting process. Before NIH even receives a grant proposal, the preparation of proposals requires great deal of mentorship and institutional support. After initial reviews are submitted, the full study section must discuss the initial reviews to come to a decision about each proposal’s Priority Score, after which NIH determines funding lines. Even if initial proposal reviews are unbiased, bias in these other stages of the granting process could produce disparities in funding rates.

Moreover, a lack of race and gender bias in initial reviews also does not mean that reviewers do not show bias on the basis of other PI characteristics. For example, a prior audit of the conference submission peer review process¹⁰ suggests that famous authors and authors from prestigious institutions are reviewed more favorably than authors without these advantages. A similar dynamic could afflict PIs from the Global South (i.e., South and Central America, Africa, South and Southeast Asia, Oceania)¹⁷.

Nevertheless, our evidence does suggest some good news: any name-based race or gender discrimination that is present in the initial review of R01 grant proposals is likely small, below half a point. If we want to understand differential funding rates based on race and gender, the present evidence suggests that we look beyond the initial review of grant proposals.

Methods:

Prior to the collection of any materials or human participant data, the University of Wisconsin-Madison Institutional Review Board reviewed our full research protocol. We conducted all procedures in accordance with their approved protocol. All recruited participants ($N = 446$) provided informed consent prior to participation and received \$300 in compensation.

We performed our preregistered analyses on 412 (59% White, 79% male, 58% experienced reviewers) of the recruited participants for reasons outlined in the “Deviations from preregistration” section. Data were collected blind to condition; the code for our main analysis was written using simulated data, so this analysis was also blind. However, follow-up analyses were not blind.

We created five versions of each of 48 grant proposals: two control versions (White male PI) and three experimental versions (White female, Black male, and Black female PI). To test whether any bias that we observe occurs for proposals that are judged to be high or moderate quality, half of our proposals are high quality and half are moderate quality.

Selecting names that connote identities. We manipulated PI identity by assigning proposals names from which race and sex can be inferred^{7,18}. We chose the names by consulting tables compiled by Bertrand and Mullainathan¹⁸. Bertrand and Mullainathan compiled the male and female first names that were most commonly associated with Black and White babies born in Massachusetts between 1974 and 1979. A person born in the 1970s would now be in their 40s, which we reasoned was a plausible age for a current Principal Investigator. Bertrand and Mullainathan also asked 30 people to categorize the names as “White”, “African American”, “Other”, or “Cannot tell”. We selected first names from their project that were both associated with and perceived as the race in question (i.e., > 60 odds of being associated with the race in question; categorized as the race in question more than 90% of the time).

We selected six White male first names (Matthew, Greg, Jay, Brett, Todd, Brad) and three first names for each of the White female (Anne, Laurie, Kristin), Black male (Darnell, Jamal, Tyrone), and Black female (Latoya, Tanisha, Latonya) categories. We also chose nine White last names (Walsh, Baker, Murray, Murphy, O’Brian, McCarthy, Kelly, Ryan, Sullivan) and three Black last names (Jackson, Robinson, Washington) from Bertrand and Mullainathan’s lists. Our grant proposals spanned 12 specific areas of science; each of the 12 scientific topic areas shared a common set of White male, White female, Black male, and Black female names. First names and last names were paired together pseudo-randomly, with the constraints that (1) any given combination of first and last names never occurred more than twice across the 12 scientific topic areas used for the study, and (2) the combination did not duplicate the name of a famous person (i.e., “Latoya Jackson” never appeared as a PI name).

Obtaining grant proposals for review. Our goal was to compare high-quality funded proposals and moderate-quality unfunded proposals. However, NIH only provides information about proposals that have been funded, so to obtain stimuli, we needed to start with proposals that had been funded. To get the desired range of quality, we solicited both proposals that were funded on their first submission and also proposals that were funded after one or more revisions and resubmissions. For the resubmitted proposals, we asked PIs to supply the original, unfunded proposal for use in the study. We intentionally selected proposals that maximized the gap in Priority Scores between our sets of high- and moderate-quality proposals. The stimuli seen by participants, therefore, were always an initial submission that was either funded with relatively high Priority Scores (scores between 1.4 and 2.7, $M = 1.9$) or not funded with middling Priority Scores (scores between 2.7 and 5.7, $M = 3.9$, four not discussed and therefore unscored).

We also wanted our proposals to broadly represent the science funded by the NIH. We selected the four institutes that contribute the most money to scientific funding: the National Cancer Institute (NCI), the National Institute of General Medical Sciences (NIGMS), the National Heart, Lungs, and Blood Institute (NHLBI), and the National Institute of Allergy and Infectious Diseases (NIAID). Reviewing, however, occurs at the level of study sections rather than at the level of institutes. To choose study sections that represent the funding priorities of these institutes, we selected the three study sections that reviewed the greatest number of funded grants per each of the four institutes from the 2013 Fiscal Year (see Table S1), resulting in 12 specific areas of science. We then collected email addresses of PIs whose funded proposals were reviewed by these study sections and sent requests for the original submissions and summary scores of these proposals.

Study section	NHLBI	NCI	NIGMS	NIAID
Vascular Cell and Molecular Biology Study Section	152	2	3	0
Myocardial Ischemia and Metabolism Study Section	130	0	1	0
Atherosclerosis and Inflammation of the Cardiovascular System Study Section	129	0	3	3
Basic Mechanisms of Cancer Therapeutics Study Section	1	187	4	0
Cancer Molecular Pathobiology Study Section	5	174	5	2
Tumor Progression and Metastasis Study Section	0	163	0	0
Macromolecular Structure and Function A, B, C, D, and E	13	27	489	31
Molecular Genetics A, B, and C	2	23	364	2
Synthetic and Biological Chemistry A and B	1	42	214	17
Cellular and Molecular Immunology A and B	3	13	18	206
Virology A and B	2	48	9	169
Bacterial Pathogenesis Study Section	1	0	4	133

Table S1. The number of proposals for each of 12 study sections that were eventually funded by the National Heart, Lungs, and Blood Institute (NHLBI), the National Cancer Institute (NCI), the National Institute of General Medical Sciences (NIGMS), and the National Institute of Allergies and Infectious Disease (NIAID) for Fiscal Year 2013. Only the three study sections that reviewed the greatest number of funded proposals per institute are shown in the rows. The content areas reviewed by these study sections can be seen as broadly representative of the funding priorities of the four institutes.

We did not reach our goal of 48 proposals after sending our first round of requests. To obtain the remaining proposals, we identified study sections that were highly similar to our target study sections. We quantified similarity using the topic terms applied to each proposal listed in the NIH RePORTER grant database. For the 2014 and 2015 fiscal years, we calculated the number of times each study section reviewed a proposal that was tagged by each of the 2,823 topic terms that was applied to at least 100 grants. We used this matrix of topic term counts for each study section to calculate the cosine similarity between study sections. After identifying study sections that were highly similar to our study sections with missing proposals (similarity $\geq .80$), we gathered emails from these similar study sections and requested proposals until we obtained a full 48 proposals. In some cases, our target study sections were already quite similar, which enabled us to use proposals reviewed by one study section as part of a set for another.

	NHLBI study sections			NCI study sections			NIGMS study sections			NIAID study sections		
	MIM	AICS	VCM	CAMP	TPM	BMCT	SBC	MG	MSF	BACP	VIR	CMI
<i>Proposals needed after round one</i>	1	2	2	1	1	2	2	0	1	0	0	0
Cardiac Contractility, Hypertrophy, and Failure	0.96	0.78	0.82	0.75	0.72	0.72	0.62	0.69	0.63	0.69	0.64	0.71
Vascular Cell and Molecular Biology	0.83	0.94	--	0.79	0.78	0.76	0.64	0.70	0.64	0.73	0.66	0.78
Hemostasis and Thrombosis	0.76	0.86	0.86	0.77	0.77	0.75	0.72	0.75	0.73	0.78	0.74	0.80
Tumor Microenvironment	0.71	0.79	0.81	0.89	0.96	0.88	0.67	0.69	0.61	0.70	0.66	0.76
Molecular Oncogenesis	0.72	0.77	0.79	0.97	0.92	0.90	0.66	0.77	0.63	0.72	0.71	0.77
Developmental Therapeutics	0.68	0.71	0.73	0.85	0.88	0.94	0.69	0.65	0.60	0.65	0.63	0.69
Macromolecular Structure and Function	0.61	0.64	0.64	0.66	0.61	0.64	0.84	0.78	--	0.75	0.74	0.69
Biochemistry and Biophysics of Membranes	0.66	0.69	0.68	0.69	0.65	0.67	0.80	0.79	0.85	0.78	0.81	0.76

Table S2. Cosine similarities of study sections with our target study sections. Bolded similarities represent values greater than or equal to .80. MIM = Myocardial Ischemia and Metabolism; AICS = Atherosclerosis and Inflammation of the Cardiovascular System; VCM = Vascular Cell and Molecular Biology; CAMP = Cancer Molecular Pathobiology; TPM = Tumor Progression and Metastasis; BMCT = Basic Mechanisms of Cancer Therapeutics; SBC = Synthetic and Biological Chemistry; MG = Molecular Genetics; MSF = Macromolecular Structure and Function; BACP = Bacterial Pathogenesis; VIR = Virology; CMI = Cancer and Molecular Immunology.

Because our fictional PIs were all US-born, we preferentially selected proposals from US-born PIs to simplify the proposal de-identification process. We selected proposals from foreign PIs for which we judged that it would be straightforward to replace foreign-identifying details (e.g., undergraduate experience at a foreign institution) with US-equivalents. We also preferentially selected proposals authored by a single investigator. Eight of our proposals were written by foreign PIs and seven of our proposals had a co-investigator.

Our selection process resulted in 48 proposals, 4 per specific area of science and 12 per institute. Half the proposals were high quality and half moderate. Characteristics of our final proposals are shown at <https://osf.io/c5csm/>.

Modifying the proposals. We conducted all modifications using Adobe Acrobat. We replaced all instances of each proposal's PI name with each of five constructed names (two White male, one White female, one Black male, one Black female). PI names appear in many places throughout a proposal, including in bibliographies, the biosketch, and in the form of nicknames in letters of support. We maintained the middle initials, if any, from the original PIs. We also changed any pronouns referring to the PI (e.g., in the letters of support) to the appropriate gender. If the PI was foreign-born and mentioned foreign institutions that they attended as part of their training (e.g., graduate school) in their biosketch, we changed these to US-equivalents.

We followed a similar process to deidentify the proposal's remaining named personnel. For each of the remaining names of personnel listed on the proposal, we created new names that roughly matched the old ones in length and country of origin. We then replaced all instances of the old names with the new, fabricated names, including in the proposal's bibliographies. We replaced signatures using fonts that look hand-drawn. We changed specific addresses, phone

numbers, and email addresses while preserving general institutional affiliations; one of the main criteria of review is whether PIs are located at an institution with the necessary resources to accomplish a project.

After we had all five of our proposal versions (two White male, one each of White female, Black male, and Black female), a second person who did not complete the original modifications checked each proposal for mentions of the original personnel. If the second person found any listings of the original personnel, these were removed and the proposal was checked again until there were no remaining modification issues.

Constructing proposal lists. We did not want any given reviewer to review multiple proposals written by non-White-male PIs (i.e., White female, Black male, or Black female PIs) because we judged that exposure to multiple non-White-male PIs would render the aims of our study too obvious. We also judged that asking our reviewers to review more than three proposals would result in an undue burden. We therefore limited the number of grant proposal reviews per reviewer to three: two control proposals (written by White male PIs) and one experimental proposal (written by a PI who was either female or Black or both).

		Compares White males with. . .			
		. . . White females	. . . Black males	. . . Black females	
Proposal 1 (<i>high quality</i>) is non-White-male	List 1	<i>Proposal 1, White female</i>	Proposal 1, Black male	Proposal 1, Black female	Four proposals define a set. Two of the proposals (1-2) are high quality and two of the proposals (3-4) are moderate quality. Each proposal is manipulated into White male, White female, Black male, and Black female versions. The proposal versions are used to create 12 lists, which together test experimental comparisons between White males and each of the three other social categories.
		Proposal 2, White male	Proposal 2, White male	Proposal 2, White male	
		Proposal 3, White male	Proposal 3, White male	Proposal 3, White male	
Proposal 2 (<i>high quality</i>) is non-White-male	List 4	Proposal 1, White male	Proposal 1, White male	Proposal 1, White male	
		Proposal 2, White female	Proposal 2, Black male	Proposal 2, Black female	
		Proposal 4, White male	Proposal 4, White male	Proposal 4, White male	
Proposal 3 (<i>moderate quality</i>) is non-White-male	List 7	Proposal 1, White male	Proposal 1, White male	Proposal 1, White male	
		Proposal 3, White female	Proposal 3, Black male	Proposal 3, Black female	
		Proposal 4, White male	Proposal 4, White male	Proposal 4, White male	
Proposal 4 (<i>moderate quality</i>) is non-White-male	List 10	Proposal 2, White male	Proposal 2, White male	Proposal 2, White male	
		Proposal 3, White male	Proposal 3, White male	Proposal 3, White male	
		Proposal 4, White female	Proposal 4, Black male	Proposal 4, Black female	

Each list of proposals is sent to 3 reviewers, yielding 36 reviewers per set of proposals. There are 12 sets of proposals, yielding 432 reviewers.

Fig. S1. A set of proposals and proposal versions, which are used to obtain the reviews from a cohort of 36 reviewers. Moderate quality proposals are shown in red, high quality proposals in blue, and, within each list, the proposal that has a non-White-male PI is italicized.

Within each specific topic area that we studied, we collected four proposals; we defined each grouping of four proposals as a set. As mentioned above, there were five versions for each proposal. The sets of proposals and proposal versions were used to construct 144 lists (i.e., 12 lists per scientific topic area), each of which was composed of two control (White male) proposals and one experimental (non-White-male) proposal (see Fig. S1). We planned for each list to be reviewed by three expert reviewers, which requires a total of 432 reviewers.

Power analysis. Before collecting any data, we conducted a simulation-based power analysis to determine whether our design was adequate to detect scoring gaps between White male and non-White-male PIs. We assumed that our reviewers' Overall Impact scores would be highly similar in distribution to the Priority Scores assigned by the original NIH panels, so we used the Priority Scores to simulate the distribution of Overall Impact scores in our power

analysis. We assumed that the Overall Impact scores assigned by a single reviewer would be correlated at $r = .3$ and the Overall Impact scores received by the same proposal would be correlated at $r = .4$. Further assuming moderate variability in random slopes and a statistical model as described in the data analytic plan, we were able to detect (using $\alpha = .05$) a gap in Impact Scores that is half the size of the gap between our high and moderate quality proposals (1.13 points^c) in 100% of our 1000 simulation runs. When we instead set the gap to a quarter the size (.56 points), we were also able to detect this gap 100% of runs. We conclude that our design yields very high power to detect pragmatically important differences in the scores obtained by White male and non-White-male PIs.

Recruiting reviewers. Our recruitment materials and other communications with reviewers are at <https://osf.io/c5csm/>. We used two primary methods to solicit reviewers for this project. The first relies on the “Similar Projects” function in NIH RePORTER. This function returns 100 projects that have similar topic terms in RePORTER. We used this function to find 100 grant proposal submissions similar to each of our 48 proposals. We scraped the PIs and co-Is from each of these funded proposals and conducted internet searches for each of the emails of these investigators. After filtering out duplicate email addresses and people from whom we had already solicited our stimulus proposals, we sent email invitations to participate in our project. For our second method of recruitment, we asked all participants who completed our study eligibility survey, described below, to recommend people who might be interested in and qualified to conduct grant reviews for our project. In some cases, these two methods were insufficient to obtain our target number of reviewers for a given set. In these cases, we used the “Similar Projects” function to find second-degree similar proposals (i.e., proposals that were highly similar to our target proposals) and used those to recruit our remaining reviewers.

In their initial recruitment email, prospective reviewers were told that they would be asked to review three R01 proposals as the primary reviewer in exchange for \$300. Our first few invited reviewers did not turn in their reviews within a reasonable timeframe, so we set a deadline of one month for subsequent reviewers to complete their reviews. Reviewers were told we would schedule a conference call to discuss the proposals with other reviewers. No conference call would actually occur; we informed the prospective reviewers of this call to better match the actual NIH review process.

We did not want prospective reviewers to recognize the original staff that prepared each of our proposals. We attempted to circumvent recognition by asking all prospective reviewers to complete an “eligibility survey” after the initial recruitment email. As part of the survey, we listed the original PIs of original proposals that we wished the prospective reviewers to review, along with the fictitious PIs of these proposals. This allowed us to assign reviewers only the proposals of PIs with whom the reviewers reported they were unfamiliar. We also asked the reviewers to report if they had served on a past study section, and if so, which section and year, which allowed us to ensure that the reviewers had not encountered our proposals during their past NIH service. We contacted 6,775 prospective reviewers, and 1,135 completed the eligibility survey. Of these, 690 (61%) reported previously serving on an NIH study section.

^c This treats proposals that were not discussed as if their Priority Scores were equal to the worst scores in our pool (5.7).

Once we deemed a reviewer eligible, we sent them an email with links to their assigned proposals, the NIH review form, and resources on the NIH review process. The email also informed the reviewers that the proposals will be a few years old and asked the reviewers to evaluate their proposals in the context of when they were written. Finally, the email reminded the reviewers not to seek outside materials.

Reviewers were sent reminder emails two weeks, one week, and one day from their completion deadline. If, one month after their deadline, they still did either contact us to reschedule their deadline to turn in their reviews, we sent one additional reminder that gave them a new deadline one month after the reminder date. If that additional deadline elapsed with no further contact from the reviewer, we assumed they would not complete their reviews and replaced them with a new reviewer. A total of 446 people turned in reviews. We conducted our preregistered analysis on 412 of these reviewers for reasons described in the “Deviations from preregistration” section.

Once we had received all reviews, we gathered reviewer demographic information by finding pictures of each reviewer via Google searches. If the search resulted in a picture, a coder categorized the PI according to sex (male; female, unsure) and race (White, Black, Hispanic, Asian, other non-White, non-White but cannot be more specific, unsure). Reviewers for whom we could not obtain a picture ($N = 41$) were coded as missing, as were reviewers for whom the coder was uncertain in their categorizations (sex $N = 1$; race $N = 2$). Among the reviewers whose demographics we did not code as missing, the majority were White (59%) and male (76%); a substantial fraction were Asian (28%). Almost half (45%) of our reviewers were both White and male. Based on their responses on the eligibility survey, the majority of our reviewers (58%) had served on a past NIH study section.

Reviewing procedure. We told the participant-reviewers that the proposals they would review were amalgamations and/or alterations of previous, real proposals. Thus, although the participants knew that the proposals had been altered, they did not know the nature of the alterations. We modeled our reviewing procedure closely on the procedure used by NIH. Participants were given one month to complete their three reviews as the primary reviewer and were informed that a conference call would occur with an SRO and other reviewers to discuss the reviews. They received all materials given to NIH reviewers, including a guide for reviewing R01s, confidentiality rules, scoring guidelines, and descriptions of each of the sections of an NIH grant proposal. They were also given a template review form, which we asked they use for all three reviews. To mitigate the possibility of reviewers reading a paper written by a proposal’s original PI and thus discovering the study deception, reviewers were discouraged from using outside resources aside from basic background reading. Reviewers who contacted us to say that they guessed the purpose of the study ($N = 5$) or who guessed the identity of the original grant personnel ($N = 13$) were replaced with new reviewers.

Our review form was modeled after the actual NIH review form, which is divided into five sections: Significance, Investigator, Innovation, Approach, and Environment. In each section, the reviewers were asked to comment on the application’s strengths and weaknesses and to give a score ranging from 1 to 9, with descriptors in Fig. S2.

Overall Impact or Criterion Strength	Score	Descriptor
High	1	Exceptional
	2	Outstanding
	3	Excellent
Medium	4	Very Good
	5	Good
	6	Satisfactory
Low	7	Fair
	8	Marginal
	9	Poor

Fig. S2. NIH scoring criteria.

The reviewers were also asked to evaluate additional special considerations, if applicable, including human subjects considerations, protections for vertebrate animals, biohazards, resource sharing plans for multiple PI proposals, and the budget and period of support. Finally, the reviewers were asked to provide an overall verbal evaluation and Overall Impact score. At NIH, this Overall Impact score is typically given the greatest weight during the discussion of reviews and the assignment of a Priority Score (which is used to determine funding lines).

As they turned in their reviews, reviewers completed a short survey including a yes-or-no question about whether they had used outside resources. If they reported “yes”, they were prompted to elaborate about what resources they used in a free response box. Contrary to their instructions, 139 reviewers mentioned that they used PubMed or read articles relevant to their assigned proposals. We eliminated the 34 reviewers who either mentioned that they learned of our deception or looked up a paper in the PI’s biosketch and therefore were very likely to learn of our deception. The remaining 105 reviewers reported that they looked up a paper from the Research Strategy section, but we retained these reviewers because, unlike in the biosketch, the rate of self-citation in the Research Strategy section was relatively low ($M = 11\%$ across our 48 proposals), hence reading one of these papers is less likely to reveal the study’s central deception. We therefore included these reviewers in our main analysis but investigate how sensitive our results are to this inclusion in our robustness analyses, as detailed in a later section.

After reviewers turned in their reviews, they were paid, debriefed as to the purpose of the study, and informed that, contrary to what they had been led to believe, there would be no conference call.

Deviations from preregistration. Our planned sample size of 432 was based on the desire to recruit three reviewers for each of the 144 lists of proposals. However, some reviewers turned in reviews after we had already replaced them with new reviewers. As a result, 13 of the lists of three proposals were reviewed by four reviewers instead of three.

Close to the end of recruitment, we shortened the amount of time between the submission deadline and our decision to drop a reviewer from the study and recruit a new one in their place from one month to two weeks.

For our final planned participant, three consecutive reviewers were unresponsive two weeks after their submission deadline. The passage of time makes the science presented in the proposals more dated, so after the third dropout we decided to close recruitment rather than spend the time recruiting this last reviewer. This means that one of the lists of three proposals was reviewed by two reviewers instead of three.

Finally, 34 participants turned in reviews without contacting us to say that they noticed the deception, and yet indicated in review submissions that some of the grant personnel were fictitious. We did not specify in our preregistration how to handle these reviewers. We decided to drop these participants from the main analyses because this decision is most consistent with how we handled participants who contacted us during the review process to note that grant personnel were fictitious. However, we also tested how sensitive our results are to the inclusion of these participants in our sensitivity analysis, described in the next section.

Preregistered analysis. Our primary outcome was the Overall Impact scores given to each of the 48 proposals by the 412 reviewers who did not guess that the PI was fictitious. We conducted our analyses using the lme4 package¹⁹ in R. Our fixed effects include quality (represented by the centered priority score received by the proposal when it was originally reviewed), three dummy codes representing the difference between White male PIs and the other three social categories [White male=0, other social category=1], and interactions between quality and the dummy codes.

We used the maximum random effects structure justified by the design²⁰. Each proposal is reviewed multiple times and is assigned to each PI social category, resulting, at the level of proposals, in random intercepts and random slopes, one per PI dummy code. Each reviewer completes multiple reviews, sees proposals of varying quality, and sees varying PI social categories, resulting, at the reviewer level, in random intercepts, random slopes for quality, and random slopes, one per PI dummy code. We computed *p*-values using the Kenward-Rogers approximation from the pbkrtest package²¹ and confidence intervals using bootstrapping.^d

As shown in Fig. 1, we found no evidence that, compared to White male PIs, reviewers gave different Overall Impact scores to White female PIs, $b = -.11$, $F(1, 45.05) = .57$, $p = .45$, 95% CI = [-.40, .15], Black male PIs, $b = .13$, $F(1, 39.98) = .88$, $p = .35$, 95% CI = [-.13, .40], or Black female PIs, $b = -.15$, $F(1, 38.24) = 1.17$, $p = .29$, 95% CI = [-.44, .13].

However, these results do not eliminate the possibility that reviewers gave different scores to White male PIs and non-White-male PIs, but that this gap, while pragmatically important, was simply undetectable in our experiment. We investigated this possibility directly using an equivalence test⁴. An equivalence test requires the user to identify the smallest effect that they consider to be of substantive interest. This value defines a region of equivalence: the set of effects that the user considers to be theoretically or pragmatically uninteresting. In our case,

^d In our preregistration, we specified profile likelihood CIs but ran into convergence issues.

we identified a difference of .5 on NIH's 1-9 rating scale as the social-category-based difference that is pragmatically important. Although the judgment of what is "pragmatically important" is somewhat arbitrary, .5 is only a quarter of the two-point gap in scores between our high and moderate quality grants and represents half the distance between adjacent anchor points on NIH's 1-9 rating scale. Thus, we set the upper and lower bound of our region of equivalence to -.5 and .5.

Once a region of equivalence is defined, the user can conduct two one-sided tests: the first to determine whether the parameter of interest is smaller than the upper bound of the region of equivalence, and the second to determine whether the parameter of interest is larger than the lower bound. If both tests are significant, the user can conclude that the parameter is statistically bounded by the region of equivalence and therefore smaller than the smallest difference that they consider to be of substantive importance.

We conducted this procedure to test whether our observed social-category-based differences were statistically equivalent to the region bounded by -.5 and .5. We used the car package v3.0+²² to conduct the two one-sided tests.^e All of the observed social category-based differences were smaller than the upper bound and larger than the lower bound of the region of equivalence, White female $b + .5 = .39$, $F(1, 45.05) = 7.67$, $p = .004$; Black male, $b - .5 = -.37$, $F(1, 39.98) = 6.62$, $p = .007$; Black female, $b + .5 = .35$, $F(1, 38.24) = 5.89$, $p = .010$.^f Thus, we can conclude that any bias favoring White males over non-White-males is smaller than the smallest difference that we consider to be pragmatically important. This result also contradicts the argument that our findings of no bias are caused by a lack of statistical power; if our design had low power we would have been unable to reject the null hypothesis of non-equivalence to the region bounded by -.5 and .5.

Although not of primary interest, we examined whether any of the coefficients estimating advantage for White males varied by proposal quality. As shown in Fig. S3, they did not, White male vs White female $b = -.09$, $F(1, 42.61) = .66$, $p = .42$, 95% CI = [-.32, .13], White male vs Black male $b = -.09$, $F(1, 43.47) = .69$, $p = .41$, 95% CI = [-.32, .12], White male vs Black female $b = -.19$, $F(1, 38.61) = 2.96$, $p = .09$, 95% CI = [-.42, .00]. In two additional exploratory analyses, we also tested whether the degree of bias in favor of White male PIs varied across the broad topics of science from which we drew the proposals (as defined by their funding institutes) or by whether a proposal's reviewer was a White man. As shown in Fig. S4 and Fig. S5, found no evidence for either proposition, topic $F(9, 52.25) = .67$, $p = .73$, reviewer $F(3, 192.23) = 1.41$, $p = .24$.^g

^e Before v3.0, the car package had a bug in the linearHypothesis() function that made it impossible to conduct tests against values other than 0 in mixed-effects models.

^f As is conventional, we report only the test that yields the largest p -value. Note that these are all one-tailed tests.

^g Given that these are both exploratory analyses, we use omnibus tests here to protect against an inflated rate of false positives. However, when we conduct more specific tests, we find modest evidence that Black female PIs received an advantage from non-White-male PIs that was not present when they were evaluated by White male PIs, $F(1, 118.44) = 4.16$, $p = .044$. We urge extreme caution in interpreting this result.

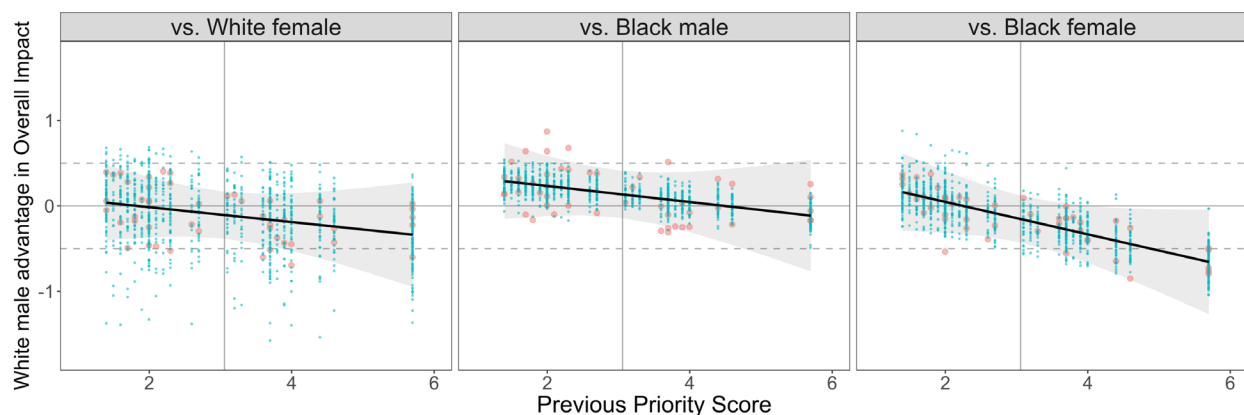


Fig. S3. Relationship between proposal quality and the difference in the Overall Impact scores attained by White male vs non-White-male PIs. The confidence band is a Wald 95% CI; blue dots are reviewer-level random effects, pink dots are grant-level random effects. Quality is operationalized by the Priority Scores given to the original proposals, and the vertical grey line is the mean Priority Score across the 48 proposals. Although descriptively, Black female PIs have an advantage on low quality proposals relative to White male PIs, the overall relationship between Priority Scores and the White male vs. Black female Overall Impact difference is not different from 0.

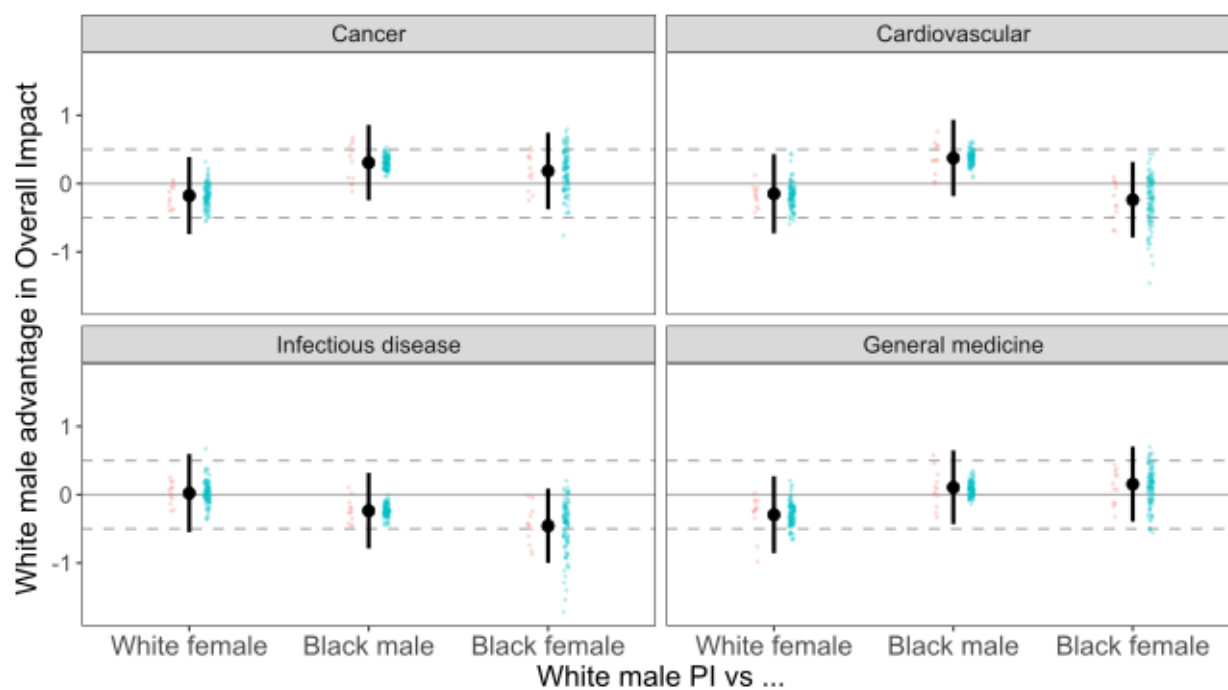


Fig. S4. The difference in the Overall Impact scores attained by White male vs non-White-male PIs in four broad topic areas of science. Topics are defined by the NIH institute that originally funded the proposals assigned to a given reviewer. “Cancer” is the National Cancer Institute; “Cardiovascular” the National Heart, Lungs, and Blood Institute, “General medicine” the National Institute for General Medical Sciences, “Infectious disease” the National Institute for Allergy and Infectious Disease. Dots are the estimated differences from the LMEM, lines are Wald 95% CIs, points to the left and right of each dot are by-proposal and by-reviewer random effects, respectively.

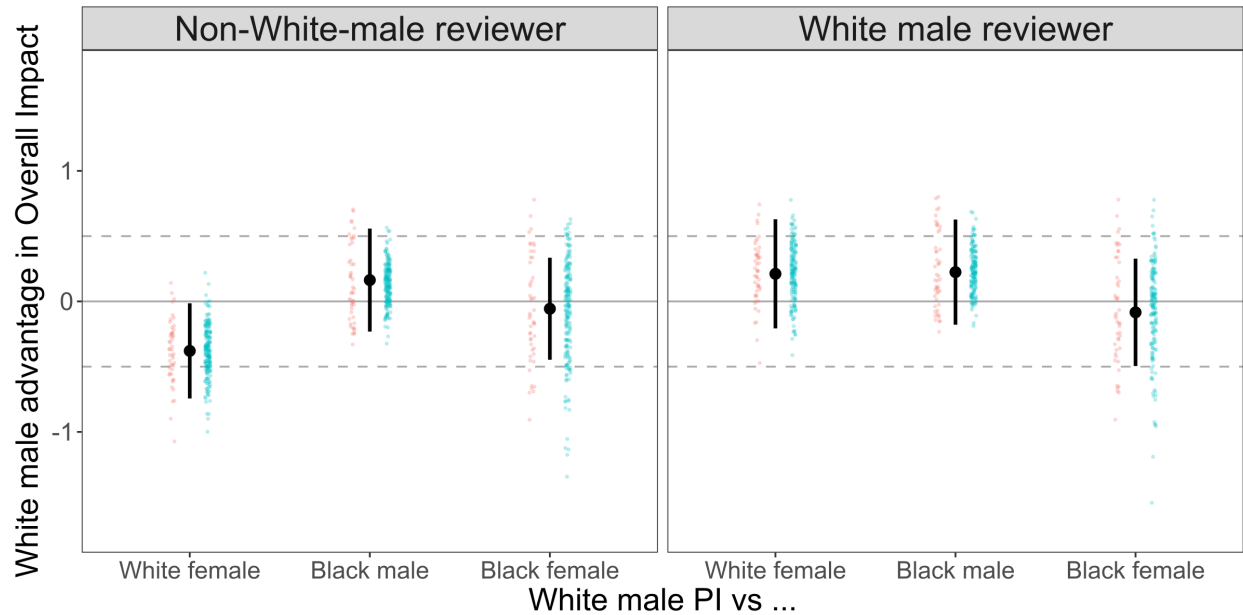


Fig S5. The difference in the Overall Impact scores attained by White male vs non-White-male PIs for non-White-male and White male reviewers. Dots are the estimated differences from the LMEM, lines are Wald 95% CIs, points to the left and right of each dot are by-proposal and by-reviewer random effects, respectively

Fig. 1 suggests that scores received by Black female PIs may be more variable than those of PIs from other social categories. We tested this systematically by using the OpenMx package²³ to fit two sets of multi-group, multi-level Structural Equation Models. In the first set, we allowed either the by-reviewer or by-proposal random intercepts to vary across our four conditions; in the second set, we constrained these random effects to be the same across conditions (see <https://osf.io/mnt8e/>). Although the variability in scores appears to differ by PI social category in Fig. 1, the models where we allowed the variability in scores to differ by PI social category fit no better than the models where they were constrained to be equal across groups; by-reviewer model comparison $\chi^2(3) = 4.50, p = .212$, by-proposal model comparison $\chi^2(3) = 2.12, p = .548$.

Robustness. There are many analytic decisions that are both reasonable and could affect whether we find evidence of a bias in reviews. Table 1 shows many of these points of flexibility, which together yield 4,536 reasonable models that could test for bias in review scores.

The high number of reasonable models to test for bias in review scores raises the possibility that we could obtain different results with models. We assessed this possibility by conducting a sensitivity analysis using a specification curve²⁴. This involves fitting all 4,536 models to test for bias and comparing how this set of models behave compared to their behavior under the null hypothesis. The behavior under the null can be obtained by randomly shuffling the variable for condition to form 500 new datasets and computing the specification curve in these 500 datasets. This process involved a large number of computational resources, so we conducted these analyses using resources provided by the Open Science Grid^{25,26}.

Our results were not very sensitive to alternative model specifications. As shown in Table S3 and Fig. 2, very few of the models resulted in coefficients that were statistically significant

(using $\alpha = .05$). When we examined more closely the coefficients that were significant, permutation tests revealed that the rate of statistically significant results was not substantially different from what one would expect under the null hypothesis (Black male vs White male: 83/2189, $p = .084$; female vs male: 185/2033, $p = .072$; race x gender: 66/2033, $p = .144$).

Model type	Coefficient	N	Median	Positive & significant	Negative & significant	Significantly within [-.5, .5]
Dummy codes	White female vs White male	2189	-0.02	0	0	2189
	Black male vs White male	2189	0.12	83	0	2000
	Black female vs White male	2189	-0.10	0	0	2188
Interaction	Race	2033	0.03	0	0	2033
	Gender	2033	-0.14	0	185	1913
	Race x gender	2033	-0.20	0	66	606

Table S3. Specification curve results. The “N” column gives the number of coefficients of the particular type from the specification curve analysis; “Median” gives the median across those coefficients; the remaining columns denote the number of coefficients with the noted properties. Note that 314 of the models failed to converge, meaning the sum of the number of dummy code (2189) and interaction (2033) models does not equal the number of models tested (4,536).

Moreover, across models, the vast majority of the coefficients comparing White males to White females, White males to Black males, White people to Black people, and men to women stayed significantly within the equivalence bounds of [-.5, .5]. There was one coefficient that did not consistently stay within the equivalence bounds of [-.5, .5], that for the interaction of race and gender. However, this finding seems to reflect the greater uncertainty associated with an interaction term rather than a systematic pattern.

Text analyses. We conducted exploratory analyses assessing the degree to which White male and non-White-male PIs received different written critiques. For each critique, we removed all punctuation except for intra-word dashes, stripped extra whitespace, then created a term-document matrix representing the frequency words from the full corpus that were present in that critique. We neither removed stop words, nor did we stem any words. We then used the term-document matrix to find, for each written critique, the number of words falling into each of 9 categories used by a previous analysis of written NIH critiques⁵. The word categories are shown in Table S4 and include ability, achievement, agentic, research, standout adjectives, the positive and negative evaluation of proposals, and negations. Kaatz and her colleagues⁵ developed and validated 7 of these categories using a modified Delphi method to assess language relevant to the evaluation of proposals; the remaining two categories, negations and pronouns, comes from the Linguistic Inquiry and Word Count (LIWC) software²⁷ and assess whether reviewers use negations at a high rate (e.g., by saying “not enthusiastic”) or use pronouns instead of the names of some PIs (e.g., by saying “she” instead of “Dr. Smith”).

Ability	Achievement	Agentic	Negations	Negative	Positive	Pronouns	Research	Standout
ability	accomplish	achieve	cannot	deficient	acceptable	all	data	amazing
brilliant	diligent	ambition	doesn't	detracts	advances	either	experimental	excellent
flair	improve	boldness	hasn't	fails	convincing	he	grants	outstanding
genius	proficient	initiative	isn't	inappropriate	enthusiasm	nobody	methodology	remarkable
intelligent	solve	leader	neither	limits	impressive	she	published	uniquely
talented	strive	productivity	never	questionable	rigorous	they	research	wonderful

Table S4. The nine categories of words used to test whether critique text differed by PI demographics. Six sample words are shown for each category.

For each category, we assessed whether the proportion of the total number of words differed by PI demographics using a Generalized Linear Mixed-Effects model with a logit link in the binomial family. To ensure our results could be interpreted as proportions, we weighted the response variable by the total word count in each full critique. We used the same random effects structure as we used for our main analysis of the Overall Impact scores.

As shown in Fig. 3, there were no differences in the proportion of words in the critiques of White male and non-White-male PIs. Tables of all model fixed effects are at <https://osf.io/c5csm/>. We also examined whether the finding of no difference in language varied for proposals of different levels of quality, as operationalized by their previous Priority Scores. As shown in Fig. S6, there were few systematic patterns in these analyses. Although there was modest evidence that there was a more positive relationship between Priority Scores and ability words for White women than White men at lower levels of proposal quality, $RR = 1.14$, $\chi^2(1, N = 412) = 6.37$, $p = .012$, 95% CI = [1.03, 1.26], and that there was a more negative relationship between negative evaluation words and Priority Scores for Black women than White men, $RR = .93$, $\chi^2(1, N = 412) = 4.24$, $p = .039$, 95% CI = [.85, .99], the evidence for these differential relationships was weak and the estimated differences in the relationships were slight. Indeed, our model implies that even at the very extreme Priority scores of 1.4 and 5.6, the differences in the percent use of ability and achievement words were tiny. For ability words, White men with a Priority Score of 1.4 received .43% ability words vs the .37% ability words White women received; at a very poor Priority Score of 5.6, these percentages were .34% and .50%, respectively. Similarly, for negative evaluation words, White men with a Priority Score of 1.4 received 1.2% negative words vs the 1.3% Black women received; at a Priority Score of 5.6, these percentages were 1.2% and 1.0%, respectively.

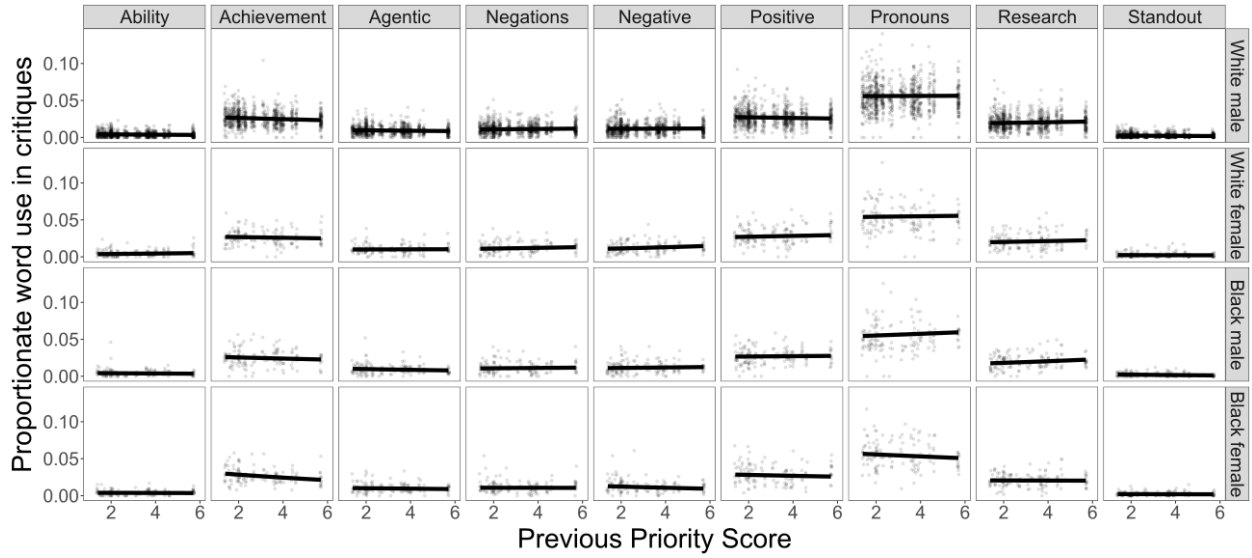


Fig. S6. The relationships between quality, as operationalized by a proposal’s previous Priority Score, and the proportionate word use in each of nine categories. Points are jittered to avoid overplotting.

References:

1. Ginther, D. K. *et al.* Race, ethnicity, and NIH research awards. *Science* **333**, 1015–1019 (2011).
2. Ceci, S. J. & Williams, W. M. Understanding current causes of women's underrepresentation in science. *Proc. Natl. Acad. Sci.* **108**, 3157–3162 (2011).
3. Pohlhaus, J. R., Jiang, H., Wagner, R. M., Schaffer, W. T. & Pinn, V. W. Sex Differences in Application, Success, and Funding Rates for NIH Extramural Programs: *Acad. Med.* **86**, 759–767 (2011).
4. Lakens, D. Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).
5. Kaatz, A., Magua, W., Zimmerman, D. R. & Carnes, M. A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques From Investigators at One Institution: *Acad. Med.* **90**, 69–75 (2015).
6. Crowne, D. P. & Marlowe, D. A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* **24**, 349–354 (1960).
7. Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J. & Handelsman, J. Science faculty's subtle gender biases favor male students. *Proc. Natl. Acad. Sci.* **109**, 16474–16479 (2012).
8. Steinpreis, R. E., Anders, K. A. & 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* **41**, 509–528 (1999).
9. Lerner, J. S. & Tetlock, P. E. Accounting for the effects of accountability. *Psychol. Bull.* **125**, 255–275 (1999).

10. Tomkins, A., Zhang, M. & Heavlin, W. D. Reviewer bias in single- versus double-blind peer review. *Proc. Natl. Acad. Sci.* **114**, 12708–12713 (2017).
11. Peterson, D. A. M. Author Gender and Editorial Outcomes at Political Behavior. *PS Polit. Sci. Polit.* 1–4 (2018). doi:10.1017/S104909651800063X
12. Samuels, D. Gender and Editorial Outcomes at Comparative Political Studies. *PS Polit. Sci. Polit.* 1–5 (2018). doi:10.1017/S1049096518000616
13. König, T. & Ropers, G. Gender and Editorial Outcomes at the American Political Science Review. *PS Polit. Sci. Polit.* 1–5 (2018). doi:10.1017/S1049096518000604
14. Tudor, C. L. & Yashar, D. J. Gender and the Editorial Process: World Politics, 2007–2017. *PS Polit. Sci. Polit.* 1–11 (2018). doi:10.1017/S1049096518000641
15. Nedal, D. K. & Nexon, D. H. Gender in the International Studies Quarterly Review Process. *PS Polit. Sci. Polit.* 1–7 (2018). doi:10.1017/S1049096518000628
16. Williams, W. M. & Ceci, S. J. National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proc. Natl. Acad. Sci.* **112**, 5360–5365 (2015).
17. Link, A. M. US and Non-US Submissions: An Analysis of Reviewer Bias. *JAMA* **280**, 246 (1998).
18. Bertrand, M. & Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
19. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).
20. Brauer, M. & Curtin, J. J. Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables

that Vary Within-Subjects and/or Within-Items. *Psychol. Methods* (2017).

doi:10.1037/met0000159

21. Halekoh, U. & Højsgaard, S. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *J. Stat. Softw.* **59**, (2014).
22. Fox, J. & Weisberg, S. *An R companion to applied regression*. (SAGE Publications, 2011).
23. Neale, M. C. *et al.* OpenMx 2.0: Extended Structural Equation and Statistical Modeling. *Psychometrika* **81**, 535–549 (2016).
24. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electron. J.* (2015).
doi:10.2139/ssrn.2694998
25. Pordes, R. *et al.* The open science grid. *J. Phys. Conf. Ser.* **78**, 12057 (2007).
26. Sfiligoi, I. *et al.* The Pilot Way to Grid Resources Using glideinWMS. in 428–432 (IEEE, 2009). doi:10.1109/CSIE.2009.950
27. Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2009).

Acknowledgments: We acknowledge Ethan Brandt, Kendra Lange, Dahea (Dianne) Lee, Jayme Marsh, Valeria Martinez, Chelsea Mitamura, Neeraja Mohan, Yupheng Lee, Cass Henriques, Rebecca Grzenia, Katharine Scott, Sage Staples, and Dan Statz, and Peter Rienke for their help in conducting this research. We also acknowledge Molly Carnes, Cece Ford, Anna Kaatz, and Josh Raclaw for their help in the design of the research. Finally, acknowledge Jake Westfall for his helpful comments on our analyses and John Fox for his advice on the car package. This research was supported by 5R01GM111002-02, awarded to the last author. Part of this research was conducted using technical resources provided by the Open Science Grid ^{25,26}, which is supported by the National Science Foundation award 1148698 and the U.S. Department of Energy's Office of Science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding: This research was supported by 5R01GM111002-02, awarded to the last author. Part of this research was conducted using technical resources provided by the Open Science Grid ^{25,26}, which is supported by the National Science Foundation award 1148698 and the U.S. Department of Energy's Office of Science.

Author contributions: Conceived research: Devine; Designed research: All authors; Supervised the preparation of materials: Forscher and Cox; Supervised data collection: Cox and Forscher; Prepared preregistration: Forscher; Revised preregistration: All authors; Analyzed data: Forscher and Brauer; Wrote first draft: Forscher; Revised paper: All authors; Principal Investigator of NIH Grant 5R01-GM111002-02, which funded this research: Devine.

Competing interests: Authors declare no competing interests.

Data and materials availability: Our data and materials have been deposited at <https://osf.io/uy7vq/>, which also includes our preregistered protocol. The modified grant proposals have not been deposited for confidentiality reasons. Modified grant proposals can be obtained by contacting the corresponding authors, who will seek permission to share these materials from the research teams that prepared the proposals.