Published in final edited form as:

Nat Hum Behav. 2019 March ; 3(3): 257–264. doi:10.1038/s41562-018-0517-y.

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

Patrick S. Forscher[1,2,*], **William T. L. Cox**[1], **Markus Brauer**[1], **Patricia G. Devine**[1,*]

[1]University of Wisconsin-Madison, Madison, WI, USA.

[2]University of Arkansas, Fayetteville, AR, USA.

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 range 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 3 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.

Reporting Summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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

[*]**Correspondence and requests for materials** should be addressed to P.S.F. or P.G.D. schnarrd@gmail.com; pgdevine@wisc.edu.

Author contributions

P.G.D. conceived the research. All authors designed the research. P.S.F. and W.T.L.C. supervised the preparation of materials and data collection. P.S.F. prepared the preregistration. All authors revised the preregistration. P.S.F. and M.B. analysed the data. P.S.F. wrote the first draft of the manuscript. All authors revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0517-y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Code availability

All code used in this paper can be accessed at our project page at https://osf.io/c5csm/

Data 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.

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. Thus, it is 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 PIs[1], and although initial submissions of male and female PIs are funded at similar rates[2], pro-male gaps emerge in resubmissions[3].

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 ten proposals, one-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 the NIH, proposals are reviewed by study sections and funded by institutes. Our proposals came from 12 study sections that, together, broadly represented the science funded by the 4 largest institutes at the NIH (see Supplementary Table 1). 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 (see Methods for more information). Our 48 stimulus proposals captured a broad range of quality. The mean Priority Score of the high-quality proposals (mean = 1.9, s.d. = 0.65, min = 1.4, max=2.7) was a full 2 points better on the 1 (exceptional) to 9 (poor) Priority Score scale than the moderate-quality proposals (mean = 3.9, s.d. = 0.36, min = 2.7, max=5.7; 4 not discussed and therefore unscored; although Priority Scores are on a 10–90 scale, we use 1–9 throughout for compatibility with the Overall Impact scale). 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 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 pro-posals 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. The NIH does not publicly release demographic information about its reviewers, so we do not know how the demographics of our reviewers compare to the 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, although we did not tell reviewers the nature of these modifications. In exchange for US$300, each reviewer evaluated a set of three proposals as a primary reviewer, a number similar to the number of first-stage primary reviews requested by the 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 the 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 our 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 before 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 males. 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 section, 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 section that the differences in these variances were no greater than one would expect due to sampling error. 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 test[4]. 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 0.5 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 the NIH's 1–9 rating scale. As is shown in more detail in the Methods section, our effects were significantly smaller than 0.5, suggesting that any bias that is present in our data is below this threshold.

Other researchers could reasonably disagree with the decisions that we made when specifying our statistical model. For example, our preregistered analysis treated each race-gender combination as unique (that is, white male versus black male versus white female versus black female), but one could argue that we should look at race (black versus white) and gender (male versus female) as separate variables in our model. In addition, others could argue that any reviewer who used outside resources should be excluded from analysis owing to the possibility that these reviewers discovered that the proposal PIs were fictitious. Table 1 lays out these and other reasonable alternatives for how to analyse 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-analysed 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 significant bias favouring the non-stigma-tized 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 section).

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 scores[5]. To this end, we calculated, for each critique, the proportion of words falling into each of the nine word categories (see Supplementary Table 2) that past research has argued are relevant to grant proposal evaluation[5]. 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.

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 there-fore cannot generalize to it. Our study is not vulnerable to the first criticism. As shown in the Methods section, 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 ones 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 that 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 behaviour[6], 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 the 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 scores[9], 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 preference[10–16], suggesting that reviewers may have little gender bias that influences their reviews. Discovering why initial reviews do not seem to be subject to race or gender bias may help researchers and policy-makers to 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 the NIH even receives a grant proposal, the preparation of proposals requires a 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 the 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 previous audit of the conference submission peer review process[10] suggests that famous authors and authors from prestigious institutions are reviewed more favour-ably than authors without these advantages. A similar dynamic could afflict PIs from the Global South (that is, South and Central America, Africa, South and Southeast Asia, and 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 probably 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 US$300 in compensation. We performed our preregistered analyses on 412 (59% white, 79% male and 58% experienced reviewers) of the recruited participants for reasons outlined in the 'Deviations from preregistration' section. Data were collected blind to the 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 the 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 proposal names from which race and gender can be inferred[7,18]. We chose the names by consulting tables compiled by Bertrand and Mullainathan[18]. Bertrand and Mullainathan compiled the male and female first names that were most commonly associated with black and white babies born in Massachusetts, USA, 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 PI. 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 (that is, >60 odds of being associated with the race in question; categorized as the race in question >90% of the time).

We selected six white male first names (Matthew, Greg, Jay, Brett, Todd and Brad) and three first names for each of the white female (Anne, Laurie and Kristin), black male (Darnell, Jamal and Tyrone) and black female (Latoya, Tanisha and Latonya) categories. We also chose nine white last names (Walsh, Baker, Murray, Murphy, O'Brian, McCarthy, Kelly, Ryan and Sullivan) and three black last names (Jackson, Robinson and 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 (that is, '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, the 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-quality and moderate-quality proposals. Thus, the stimuli seen by participants were always an initial submission that was either funded with relatively high Priority Scores (scores between 1.4 and 2.7, mean = 1.9) or not funded with middling Priority Scores (scores between 2.7 and 5.7, mean = 3.9, 4 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, the National Institute of General Medical Sciences, the National Heart, Lungs, and Blood Institute, and the National Institute of Allergy and Infectious Diseases. However, reviewing 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 Supplementary Table 1), 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.

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 of ≥0.80; see Supplementary Table 3), 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.

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 (for example, 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 of the proposals were high quality and half were moderate quality. 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 and 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 (for example, 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 (for example, graduate school) in their biosketch, we changed these to US equivalents.

We followed a similar process to de-identify 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 (that is, 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 whether asking our reviewers to review more than three proposals would result in an undue burden. Thus, we 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).

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 (that is, 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 Supplementary Fig. 1). 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 $r = 0.3$ and the Overall Impact scores received by the same proposal would be correlated at $r = 0.4$. Further assuming moderate variability in random slopes and a statistical model as described in the data analytic plan, we were able to detect (using $\alpha = 0.05$) a gap in Impact Scores that is half the size of the gap between our high-quality and moderate-quality proposals (1.13 points; this difference treats proposals that were not discussed as if their Priority Scores were equal to the worst scores in our pool) in 100% of our 1,000 simulation runs. When we instead set the gap to one-quarter of the size (0.56 points), we were also able to detect this gap in 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-PIs 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 (that is, 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 US$300. Our first few invited reviewers did not turn in their reviews within a reasonable timeframe, so we set a deadline of 1 month for subsequent reviewers to complete their reviews. Reviewers were told that 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 whether 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.

Once we deemed a reviewer eligible, we sent them an email with links to their 3 proposals to review, which were randomly selected from the 12 lists of possible proposals within their topic area of expertise. The email also contained links to the NIH review form and resources on the NIH review process. In addition, the email 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 2 weeks, 1 week and 1 day from their completion deadline. If 1 month after their deadline they still did not contact us to reschedule their deadline to turn in their reviews, we sent one additional reminder that gave them a new deadline 1 month after the reminder date. If that additional deadline elapsed with no further contact from the reviewer, we assumed that they would not complete their reviews and replaced them with a new reviewer. A total of 446 people turned in reviews; we submitted our timestamped preregistration at participant 445 prior to viewing any of the data. 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 gender (male, female or unsure) and race (white, black, Hispanic, Asian, other non-white, non-white but cannot be more specific or 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 (gender $N = 1$; 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 was 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 that 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 modelled our reviewing procedure closely on the procedure used by the NIH. Participants were given 1 month to complete their three reviews as the primary reviewer and were informed that a conference call would occur with a Scientific Review Officer and other reviewers to discuss the reviews. They received all of the 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 modelled 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 Supplementary Fig 2.

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 the 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 (mean = 11% across our 48 proposals); hence, reading one of these papers is less likely to reveal the study's central deception. Thus, we 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 3 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 3 proposals were reviewed by 4 reviewers instead of 3.

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 reviewer in their place from 1 month to 2 weeks.

For our final planned participant, three consecutive reviewers were unresponsive 2 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 package[19] in R. Our fixed effects include quality (represented by the centred 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 design[20]. 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 package[21] and CIs using bootstrapping (in our preregistration, we specified profile likelihood CIs but ran into convergence issues). In the linear mixed-effects models, distributions were assumed to be normal but were not formally tested.

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, $b = -0.11$, $F(1, 45.05) = 0.57$, $P = 0.45$, 95% CI = −0.40 to 0.15; black male PIs, $b = 0.13$, $F(1, 39.98) = 0.88$, $P = 0.35$, 95% CI = −0.13 to 0.40; or black female PIs, $b = -0.15$, $F(1, 38.24) = 1.17$, $P = 0.29$, 95% CI = −0.44 to 0.13.

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, although pragmatically important, was simply undetectable in our experiment. We investigated this possibility directly using an equivalence test[4]. 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, we identified a difference of 0.5 on the NIH's 1–9 rating scale as the social-category-based difference that is pragmatically important. Although the judgement of what is 'pragmatically important' is somewhat arbitrary, 0.5 is only one-quarter of the two-point gap in scores between our high-quality and moderate-quality grants and represents half of the distance between adjacent anchor points on the NIH's 1–9 rating scale. Thus, we set the lower and upper bound of our region of equivalence to −0.5 and 0.5, respectively.

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 −0.5 and 0.5. We used the car package v3.0+[22] to conduct the two one-sided tests (before v3.0, the car package had a bug in car::linearHypothesis that made it impossible to conduct tests against values other than 0 in mixed-effects models). 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, $b + 0.5 = 0.39$, $F(1, 45.05) = 7.67$, $P = 0.004$; black male, $b - 0.5 = -0.37$, $F(1, 39.98) = 6.62$, $P = 0.007$; black female, $b + 0.5 = 0.35$, $F(1, 38.24) = 5.89$, $P = 0.010$ (as is convention, we report only the test that yields the largest P value). Thus, we can conclude that any bias favouring 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 −0.5 and 0.5.

Although not of primary interest, we examined whether any of the coefficients estimating an advantage for white males varied by proposal quality. As shown in Supplementary Fig. 3, they did not, white male versus white female, $b = -0.09$, $F(1, 42.61) = 0.66$, $P = 0.42$, 95% CI = −0.32 to 0.13; white male versus black male, $b = -0.09$, $F(1, 43.47) = 0.69$, $P = 0.41$, 95% CI = −0.32 to 0.12; white male versus black female, $b = -0.19$, $F(1, 38.61) = 2.96$, $P = 0.09$, 95% CI = −0.42 to 0. In two additional exploratory analyses, we also tested whether the degree of bias in favour 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 Supplementary Figs. 4 and 5, we found no evidence for either proposition, topic $F(9, 52.25) = 0.67$, $P = 0.73$, or reviewer $F(3, 192.23) = 1.41$, $P = 0.24$.

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: $b = 0.60$, $F(1, 118.44) = 4.16$, $P = 0.044$, 95% CI = 0.06–1.14. We urge extreme caution in interpreting this result.

Figure 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 package[23] 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 seems 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, $\chi^2(3) = 4.50$, $P = 0.212$; by-proposal model comparison, $\chi^2(3) = 2.12$, $P = 0.548$.

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 curve[24]. This involves fitting all 4,536 models to test for bias and comparing how this set of models behave compared to their behaviour under the null hypothesis. The behaviour under the null can be obtained by randomly shuffling the variable for condition to form 500 new data sets and computing the specification curve in these 500 data sets. 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 Supplementary Table 4 and Fig. 2, very few of the models resulted in coefficients that were significant (using a = 0.05). When we examined more closely the coefficients that were significant, permutation tests revealed that the rate of significant results was not substantially different from what one would expect under the null hypothesis (black male versus white male: 83 out of 2,189, $P = 0.084$; female versus male: 185 out of 2,033, $P = 0.072$; race × gender: 66 out of 2,033, $P = 0.144$).

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 −0.5 to 0.5. There was one coefficient that did not consistently stay within the equivalence bounds of −0.5 to 0.5, 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 of 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 the nine categories used by a previous analysis of written NIH critiques[5]. The word categories are shown in Supplementary Table 4 and include ability, achievement, agentic, research, standout adjectives, the positive and negative evaluation of proposals, and negations. Kaatz and her colleagues[5] developed and validated seven of these categories using a modified Delphi method to assess language relevant to the evaluation of proposals; the remaining two categories, negations and pronouns, come from the Linguistic Inquiry and Word Count (LIWC) software[27] and assess whether reviewers use negations at a high rate (for example, by saying ‘not enthusiastic’) or use pronouns instead of the names of some PIs (for example, by saying ‘she’ instead of ‘Dr Smith’).

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 that 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 Supplementary Fig. 6, 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 (relative rate = 1.14, $\chi^2(1, N = 412) = 6.37$, $P = 0.012$, 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 (relative rate = 0.93, $\chi^2(1, N=412) = 4.24$, $P = 0.039$, 95% CI = 0.85–0.99), 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 per cent use of ability and achievement words were tiny. For ability words, white men with a Priority Score of 1.4 received 0.43% ability words versus the 0.37% ability words that white women received; at a very poor Priority Score of 5.6, these percentages were 0.34% and 0.50%, respectively. Similarly, for negative evaluation words, white men with a Priority Score of 1.4 received 1.2% negative words versus the 1.3% black women received; at a Priority Score of 5.6, these percentages were 1.2% and 1.0%, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge E. Brandt, K. Lange, D. (Dianne) Lee, J. Marsh, V. Martinez, C. Mitamura, N. Mohan, Y. Lee, C. Henriques, R. Grzenia, K. Scott, S. Staples, D. Statz and P. Rienke for their help in conducting this research. We also acknowledge M. Carnes, C. Ford, A. Kaatz and J. Raclaw for their help in the design of the research. Finally, we acknowledge J. Westfall for his helpful comments on our analyses and J. Fox for his advice on the car package. This research was supported by NIH grant 5R01GM111002-02 to P.G.D. 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 US 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.

References

1. Ginther DK et al. Race, ethnicity, and NIH research awards. Science 333, 1015–1019 (2011). [PubMed: 21852498]
2. Ceci SJ & Williams WM Understanding current causes of women's underrepresentation in science. Proc. Natl Acad. Sci. USA 108, 3157–3162 (2011). [PubMed: 21300892]
3. Pohlhaus JR, Jiang H, Wagner RM, Schaffer WT & Pinn VW Sex differences in application, success, and funding rates for NIH extramural programs. Acad. Med. 86, 759–767 (2011). [PubMed: 21512358]
4. Lakens D Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Personal. Sci. 8, 355–362 (2017). [PubMed: 28736600]
5. Kaatz A, Magua W, Zimmerman DR & Carnes M A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. Acad. Med. 90, 69–75 (2015). [PubMed: 25140529]
6. Crowne DP & Marlowe D A new scale of social desirability independent of psychopathology. J. Consult. Psychol. 24, 349–354 (1960). [PubMed: 13813058]

7. Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ & Handelsman J Science faculty's subtle gender biases favor male students. Proc. Natl Acad. Sci. USA 109, 16474–16479 (2012). [PubMed: 22988126]
8. Steinpreis RE, Anders KA & Ritzke D The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: a national empirical study. Sex Roles 41, 509–528 (1999).
9. Lerner JS & Tetlock PE Accounting for the effects of accountability. Psychol. Bull. 125, 255–275 (1999). [PubMed: 10087938]
10. Tomkins A, Zhang M & Heavlin WD Reviewer bias in single- versus double-blind peer review. Proc. Natl Acad. Sci. USA 114, 12708–12713 (2017). [PubMed: 29138317]
11. Peterson DAM Author gender and editorial outcomes at Political Behavior. PS Polit. Sci. Polit. 51, 866–869 (2018).
12. Samuels D Gender and editorial outcomes at Comparative Political Studies. PS Polit. Sci. Polit. 51, 854–858 (2018).
13. König T & Ropers G Gender and editorial outcomes at the American Political Science Review. PS Polit. Sci. Polit. 51, 849–853 (2018).
14. Tudor CL & Yashar DJ Gender and the editorial process: World Politics, 2007–2017. PS Polit. Sci. Polit. 51, 870–880 (2018).
15. Nedal DK & Nexon DH Gender in the International Studies Quarterly Review process. PS Polit. Sci. Polit. 51, 859–865 (2018).
16. Williams WM & Ceci SJ National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. Proc. Natl Acad. Sci. USA 112, 5360–5365 (2015). [PubMed: 25870272]
17. Link AM US and non-US submissions: an analysis of reviewer bias. JAMA 280, 246–247 (1998). [PubMed: 9676670]
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, 1–48 (2015).
20. Brauer M & Curtin JJ 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. Med. 23, 389–411 (2018).
21. Halekoh U & Hojsgaard S A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models—the R package pbkrtest. J. Stat. Softw. 59, 1–32 (2014). [PubMed: 26917999]
22. Fox J & Weisberg S An R Companion to Applied Regression (SAGE, Los Angeles, CA, USA, 2011).
23. Neale MC et al. OpenMx 2.0: extended structural equation and statistical modeling. Psychometrika 81, 535–549 (2016). [PubMed: 25622929]
24. Simonsohn U, Simmons JP & Nelson LD Specification curve: descriptive and inferential statistics on all reasonable specifications. SSRN Electron. J. 10.2139/ssrn.2694998 (2015).
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 2009 WRI World Congress on Computer Science and Information Engineering (eds Burgin M et al.) 428–432 (IEEE, 2009).
27. Tausczik YR & Pennebaker JW The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29, 24–54 (2009).

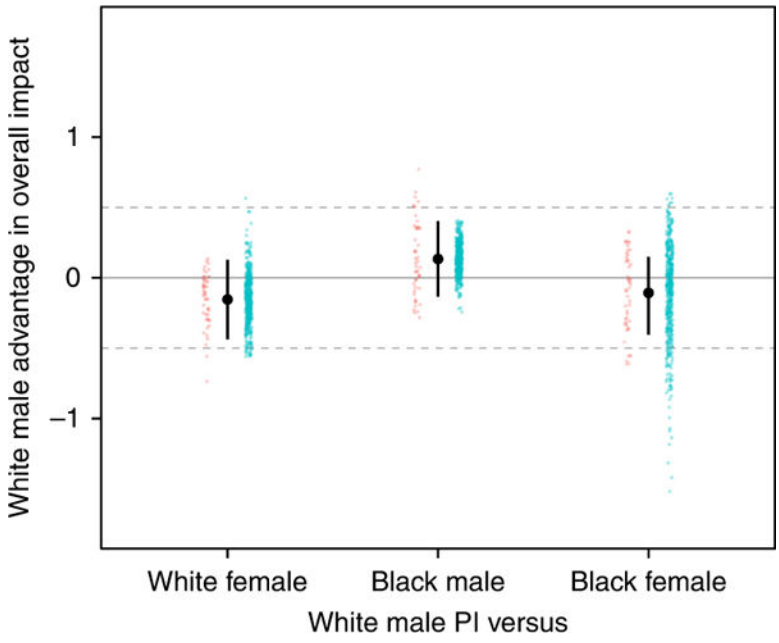

Fig. 1 |. Estimated differences between the Overall Impact scores given to proposals with white male PIs and each of white female, black male and black female PIs.

Black dots represent the mean differences and lines indicate their bootstrapped 95% CIs, all based on the Overall Impact scores provided by 412 participant-reviewers for 3 proposals each. Dotted lines encompass the region bounded by a half-point difference in the 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.

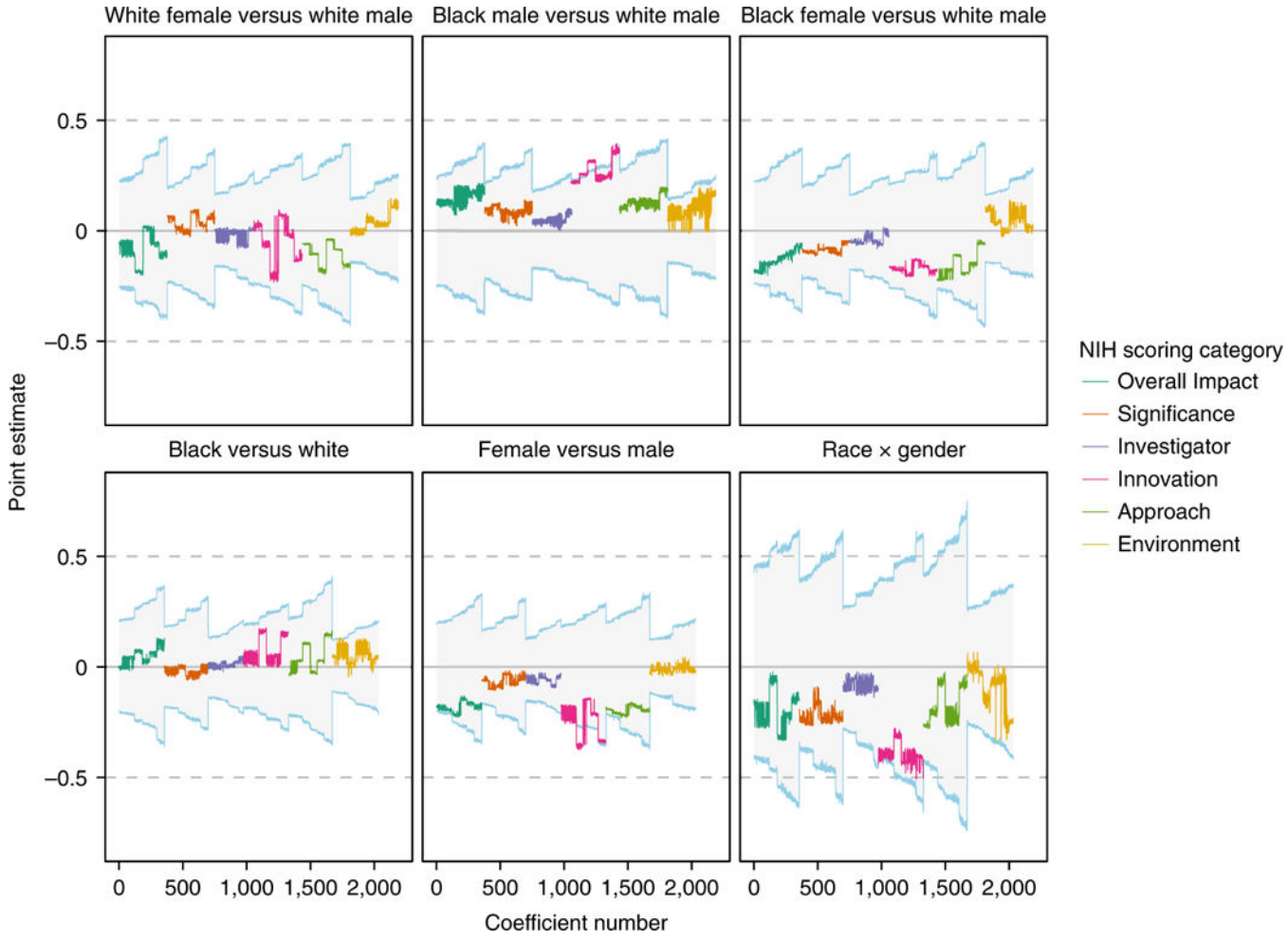

Fig. 2 |. Sensitivity of our results to alternative analytic models.

Each multicoloured 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 (although note that the race× gender interaction cannot be directly 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 a set of point estimates behaves under the null hypothesis. If a given point from the multicoloured lines is within the grey region bounded by the blue lines, this suggests that the point estimate value is not substantially different from what one would obtain under the null hypothesis.

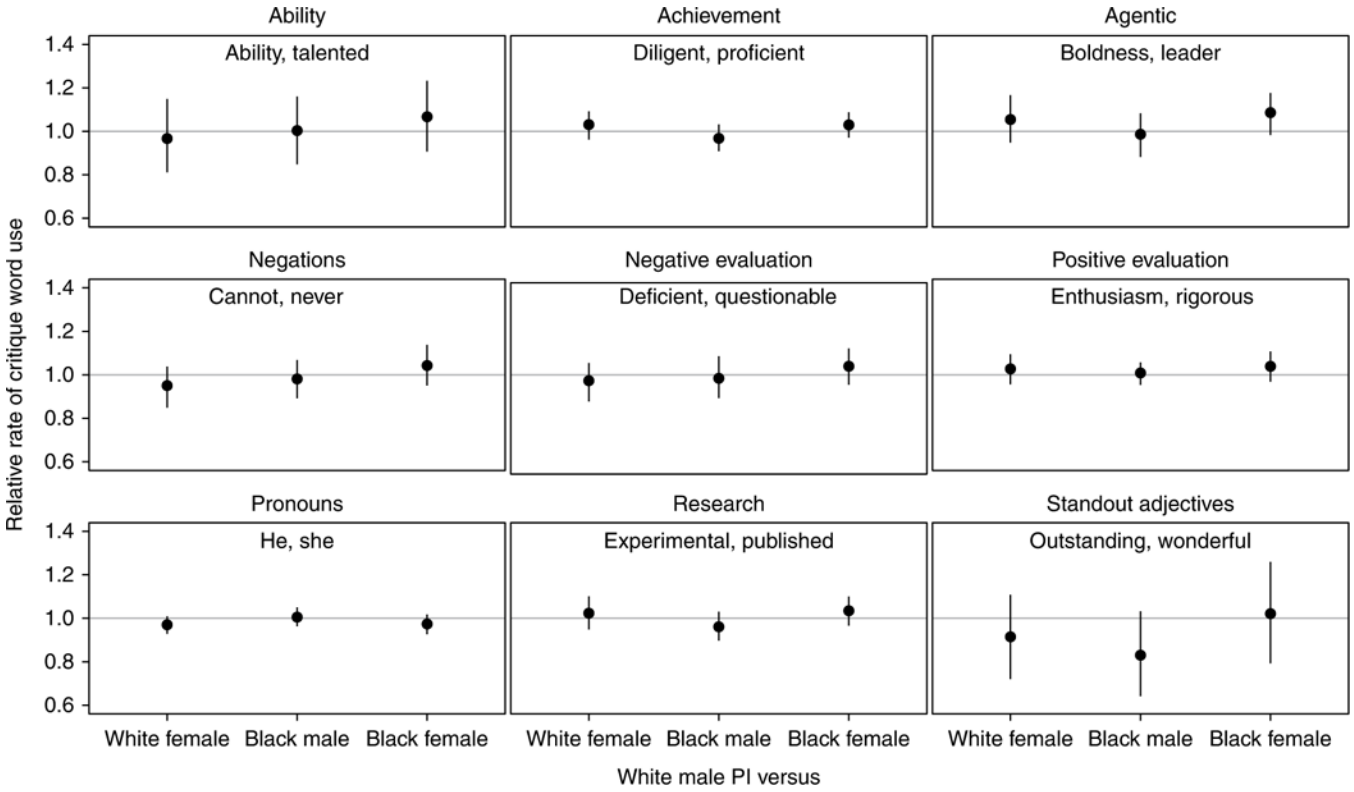

Fig. 3 |. The relative rate of word use in the written critiques given to white male PIs and each of white female, black male and black female PIs.

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 that used more of the word category. Bars represent bootstrapped 95% CIs based on the analysis of the textual responses provided by 412 participant-reviewers for 3 proposals each.

Table 1 |

Reasonable alternatives for how to analyse our data to test for bias in review scores.

	Point of flexibility		Justification
Outcome	(1)	Overall Impact	Most pragmatically important: it has the greatest effect on funding lines
	(2)	Significance	Reviewers may feel that issues of non-white PIs are more ‘niche’
	(3)	Investigator	We manipulated PI race and/or gender, which are investigator characteristics. Bias may therefore show up strongest here
	(4)	Innovation	Judgements 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
Condition	(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
Quality	(1)	Dichotomous	Design assumes proposals in dichotomous quality categories
	(2)	Quantitative, grand mean centred	Dichotomous quality misses variation in the Priority Scores within the ‘moderate’ and ‘high’ categories
	(3)	Quantitative, cluster centred	Proposal means centring conflates between-participant and within-participant variation in proposal quality
Proposal-level random effects	(1)	Intercept only	Proposals probably vary in the average scores they receive
	(2)	Intercept, condition slopes	Proposals probably vary in the 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
Reviewer-level random effects	(1)	Intercept only	Reviewers probably vary in the average scores they assign
	(2)	Intercept, condition slopes	Reviewers probably vary in their susceptibility to condition
	(3)	Intercept, condition and quality slopes	Reviewers probably vary in their susceptibility to proposal quality
	(4)	Intercept, condition, quality, interaction slopes	Reviewers probably 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 and quality slopes, intercept-slope correlations	There may be a relationship between a reviewer’s average scores and their susceptibility to condition and quality
	(7)	Intercept, condition and 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
Observations	(1)	All	Use all reviewers who completed their reviews
	(2)	Remove people who read a biosketch paper	These reviewers very likely realized that some elements of the study were fictitious
	(3)	Remove people who read any proposal paper	These reviewers may have realized that some elements of the study were fictitious
	(4)	Remove ratings of different- quality proposals	Different-quality proposals may have stood out and attracted different reviews

	Point of flexibility	Justification
(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)

In combination, the alternatives yield 4,536 analytic models.