Title: How many reviewers are required to obtain reliable evaluations of NIH R01 grant proposals?

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

Affiliations: [1]University of Wisconsin – Madison; [2]University of Arkansas; Correspondence to: Patrick Forscher at schnarrd@gmail.com.

Abstract: The National Institutes of Health uses small groups of scientists to judge the quality of the grant proposals that they receive, and these quality judgments form the basis of its funding decisions. In order for this system to fund the best science, the subject experts must, at a minimum, agree as to what counts as a "quality" proposal. We investigated the degree of agreement by leveraging data from a recent experiment with 412 scientists. Each of these scientists acted as primary reviewers for three of 48 NIH R01 grant proposals, half of which had been funded and half unfunded. Across all dimensions of NIH's official rubric, we find low agreement among reviewers in their judgments of scientific merit. For judgments of Overall Impact, which has the greatest weight in funding decisions, we estimate that three reviewers yield a reliability .2, and 12 reviewers would be required to bring this reliability up to .5. Supplemental analyses found that reviewers are even less reliable in the language they use to describe proposals.

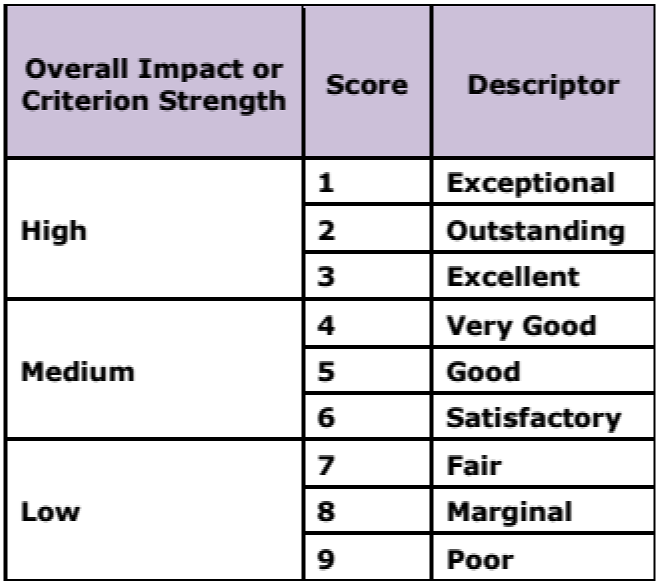
Preprint revised: 2019-04-09

Main Text:

Scientific funding is limited: not all proposals that scientists dream up will receive funding. Determining those proposals that should receive funding is a difficult problem, and as such most funders rely on the peer review judgments of a small group of subject matter experts, who together judge a written description of each project (a grant proposal) for merit according to pre-determined criteria. These judgments are then used to determine which proposals receive funding.

For peer review to serve the purpose of selecting the most meritorious proposals to receive funding, multiple peers must, at a minimum, use similar criteria to judge merit. If reviewers do not agree on their criteria, these different criteria will result in judgments that are essentially arbitrary. In the most extreme case of no agreement, the peer review group will far no better at identifying the best science than a simple dice roll.

In the United States, the primary mechanism through which projects are funded is the National Institutes of Health (NIH) R01 grant. R01 grant applications receive peer review from groups of scientists called Study Sections, which evaluate proposals on a single scientific topic. Each proposal receives primary, secondary, and tertiary reviewers who evaluate each proposal's Overall Impact, Significance, Investigator, Innovation, Approach, and Environment by writing comments and assigning scores on a 1 (exceptional) to 9 (poor) scale for each dimension of merit (see Fig. 1). Proposals typically receive around three reviewers, one primary, one secondary, and one tertiary; primary reviewers have the greatest responsibility for evaluating a given proposal. Discrepancies between reviewers are resolved through discussion, after which the Study Section averages together reviewer Overall Impact scores to yield a Priority Score. NIH uses Priority Scores to determine funding lines.



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. 1. NIH scoring criteria.

We took advantage of a recent experiment of the NIH R01 grant review process [1] to assess the reliability of review judgments. In this experiment, we obtained 48 R01 grant proposals from NIH-funded PIs, which we had re-reviewed by 412 scientists as part of a study of peer review. These proposals came from twelve Study Sections representing a broad swathe of the science funded by the four largest Institutes at NIH. All proposals in our study were eventually funded by NIH, but half the proposals were funded in their first round of review, whereas the half were not (these proposals were revised and funded on resubmission; our study used the original, unfunded proposals). The Priority Scores of our proposals captured a range of scores (*M* = 3.07; *SD* = 1.33; *Min* = 2.7; *Max* = 5.7), though they tended to occupy the "Medium" to "High" ranges of NIH's scoring criteria. Because the experiment on which this analysis is based involved manipulating the identities of the proposal PIs, all proposals had been de-identified by replacing the names of personnel with new, fabricated names prior to review.

We used the NIH RePORTER database to recruit reviewers whose expertise matched the content of the grant proposals. We screened out reviewers who recognized that the PI names were fictitious, leaving 412 reviewers for analysis. Each reviewer evaluated a set of three proposals: One previously funded, one unfunded, and the third either funded or unfunded, depending on condition. Each proposal was evaluated by an average of 25.8 reviewers (*Min* = 21, *Max* = 30, *SD* = 1.6). Because each reviewer evaluated multiple proposal and each proposal was evaluated multiple times, our design allows us to independently estimate the variation in scores due to characteristics of proposals, characteristics of reviewers, and due to other factors. We can then use these estimates to estimate how reliably reviewers assess the scientific merit of the projects described in proposals.

The reliability of review scores

We estimated the amount of variation in scores due to consistent differences between proposals, reviewers, and due to other factors by fitting a series of Linear Mixed Effects Models with reviewer scores as the outcome variables and random by-proposal and by-reviewer intercepts. As shown in Fig. 2, there were very few consistent differences in the average scores obtained by different proposals. There were some consistent differences in the average scores given by different reviewers. By far, the largest source of variation was the observation-level variance, suggesting that other factors aside from reviewer and proposal characteristics were largely responsible for peer review scores.

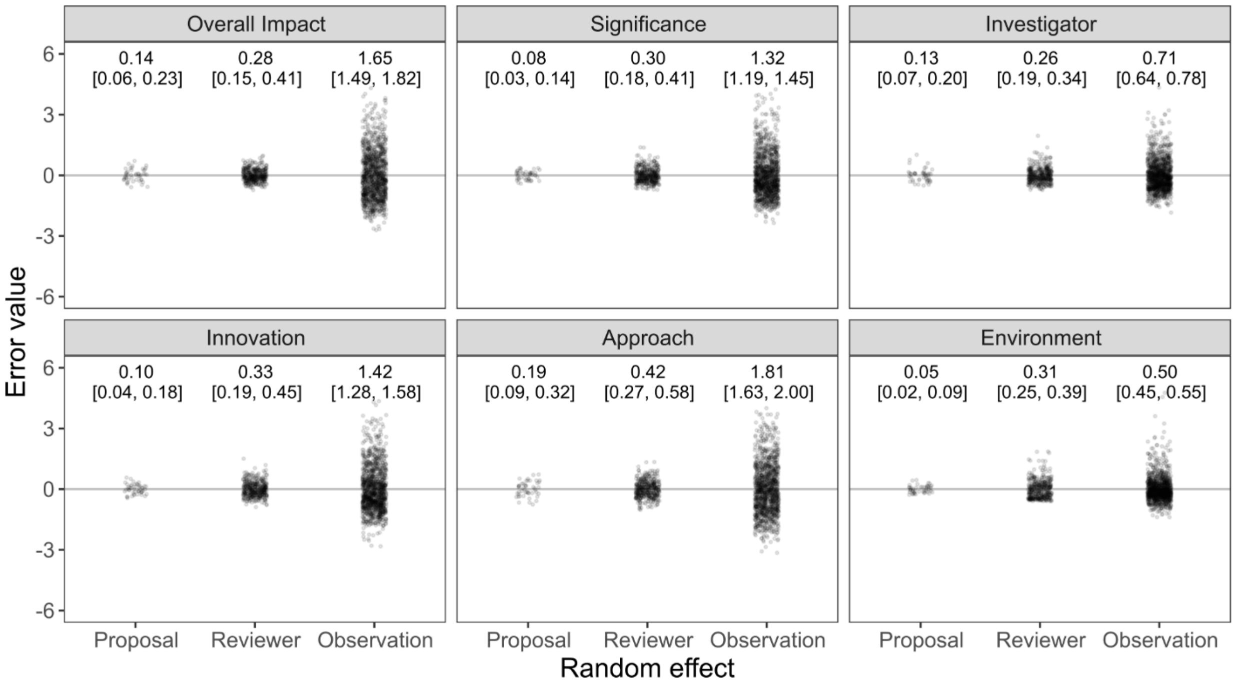


Fig. 2. Proposal-level, participant-level, and observation-level variance components for each of the six NIH R01 dimensions of peer review scores. Numbers represent the overall estimate estimates for the variance component in question and its bootstrapped 95% confidence interval. Points represent the difference between the overall estimate and the average score for each proposal and reviewer, or difference between the overall estimate and each individual observation. Points are jittered to avoid over-plotting.

The small amount of by-proposal variance suggests that very little of the difference in peer review scores is due to consistent differences between proposal content. Even if this is true, however, it is possible to obtain a reliable *aggregate* score by averaging across multiple reviewers who all review the same proposal. Each additional reviewer that goes into the aggregate score for a single proposal reduces the non-proposal variation in the resulting aggregate, increasing reliability [2].

We examined this possibility using Generalizability Theory [2]. Generalizability Theory describes two forms of reliability: absolute, which measures the degree to which an aggregate captures the "true" score, and relative, which measures the degree to which an aggregate captures the relative ranking of the items of interest (in this case, the rank-order of the scientific merit of different grant proposals). Fig. 3 shows how absolute and relative reliability increase with higher numbers of reviewers. NIH grant proposals are typically reviewed by around three reviewers. At this number of reviewers, we estimate that the reliabilities of most dimensions of review are low. The only possible exception is for the Investigator criterion, for which the absolute and relative reliabilities of a three-primary-reviewer aggregate score are .29 and .35, respectively. Speculatively, this dimension may be easier for reviewers to gauge because it is based primarily on a PI's track record of obtaining grants and publishing in journals; counting successful grants and prestigious publications may be easier than judging other dimensions of merit.

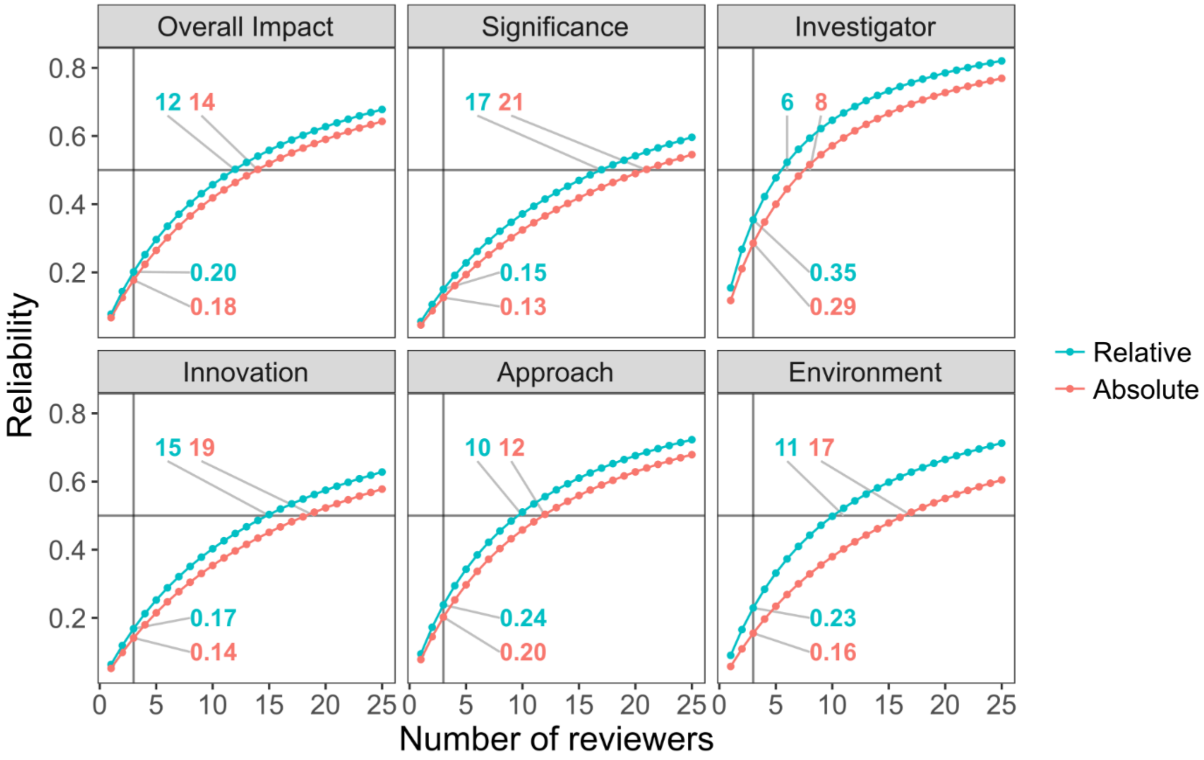


Fig. 3. Predicted absolute and relative reliabilities of peer review scores with different numbers of reviewers. The intersections between the curves and the grey vertical line mark the reliabilities with three reviewers. The intersections between the curves and the grey horizontal line mark the number of reviewers required to obtain a reliability of .5.

We also estimated the number of reviewers required to obtain an absolute and relative reliability of at least .5, which we believe is a reasonable floor for any process designed to select grants to fund. As shown in Fig. 3, for all dimensions except Investigator, obtaining a relative or absolute reliability of .5 requires at least 10 reviewers. For what is arguably the most subjective dimension of review, a grant proposal's Significance, 17 and 21 reviewers are required for, respectively, a relative and absolute reliability of .5.

The reliability of peer review language

Even if peer review scores are relatively unreliable, primary reviewers may converge in the language they apply to the same proposal. To investigate this possibility, we conducted an exploratory analysis on the words used in critiques to describe grant proposals. 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 written NIH critiques [3]. The word categories are shown in Table 1 and include ability, achievement, agentic, research, standout adjectives, the positive and negative evaluation of grants, and negations. Kaatz and colleagues [3] developed and validated 7 of these categories using a modified Delphi method to assess language relevant to the evaluation of grant applications; the remaining two categories, negations and pronouns, comes from the Linguistic Inquiry and Word Count (LIWC) software [4] 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 1. 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 fit a Generalized Linear Mixed Effects Model with a logit link and binomially distributed error, as well as and by-reviewer and by-proposal random intercepts. To control for differences in reviewer verbosity we weighted the response variable by the total word count in each critique. The degree to which proposals varied in the language they elicited from reviewers, as well as the degree to which reviewers varied in the language they gave in their critiques of different proposals, are shown in Fig. 4.

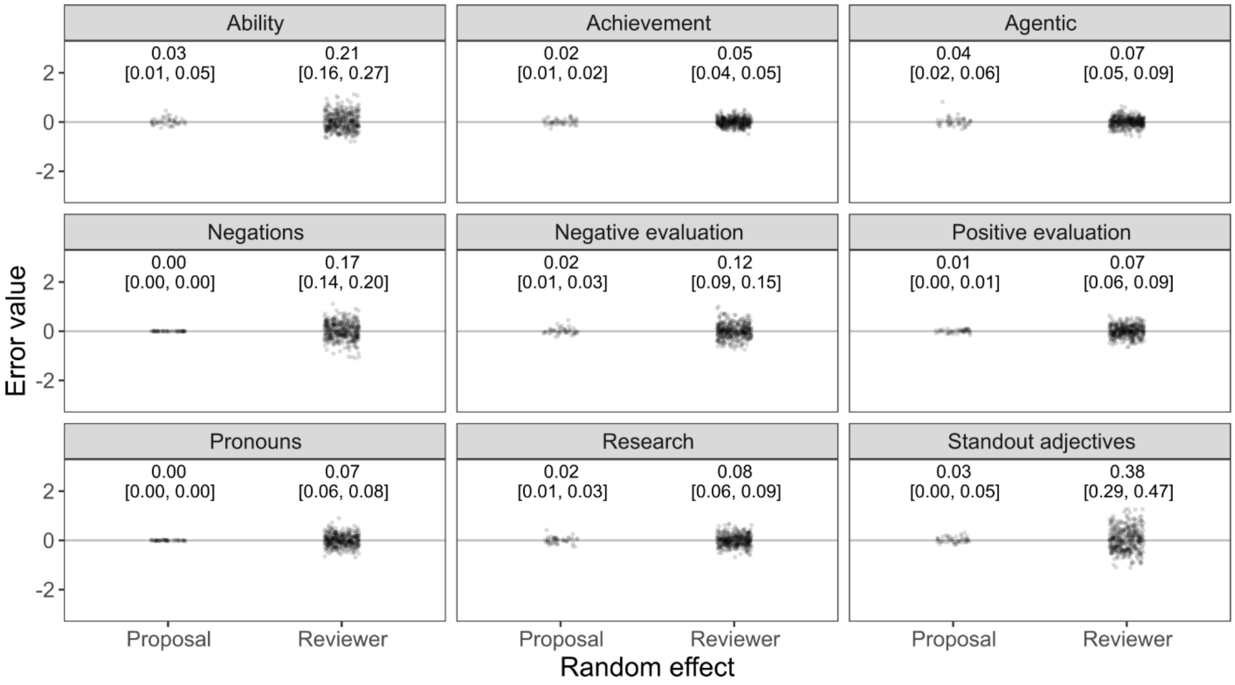


Fig. 4. Proposal-level and reviewer-level variance components for nine categories of peer review language. Observation-level variance is not shown because this variation is fixed to $\frac{\pi^2}{3} \approx 3.29$ in a model with a logistic response [5]. Numbers represent the overall estimate estimates for the variance component in question and its bootstrapped 95% confidence interval. Points represent the difference between the overall estimate and average score for each proposal and reviewer. Points are jittered to avoid over-plotting.

As shown in the figure, the same proposal did not generally receive consistent language across critiques by different reviewers. The same reviewers, however, did use consistent language across different proposals. This suggests that, although the same proposal will not elicit similar language from multiple reviewers, the same reviewer may use a similar style of language to describe multiple proposals.

As in our analysis of peer review scores, we used these variance components to estimate how the reliability of language use, aggregated across primary reviewers, changes with different numbers of reviewer critiques per proposal. As shown in Fig. 5, at the current number of three primary reviewers per proposal, the reliabilities for all nine categories of word use are essentially 0. At least 86 reviewers would be required to bring the reliability of peer review language to a value of .5.

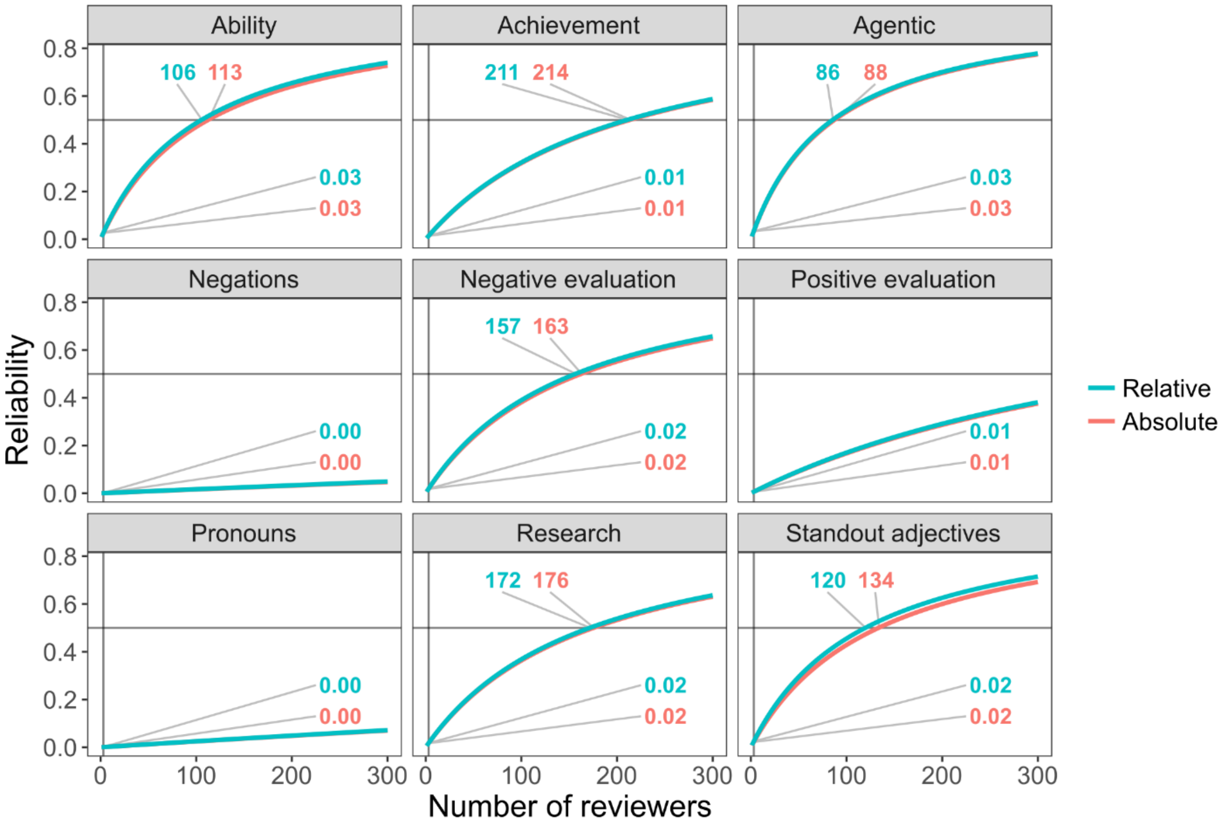


Fig. 5. Predicted absolute and relative reliabilities of nine categories of peer review language with different numbers of reviewers. The intersections between the curves and the grey vertical line mark the reliabilities with three primary reviewers. The intersections between the curves and the grey horizontal line mark the number of reviewers required to obtain a reliability of .5.

Conclusions

Our analysis suggests that, with three reviewers or below, review scores are an unreliable means of judging the merit of a scientific grant proposal. This conclusion held across all the dimensions that NIH uses to judge scientific merit. Primary reviewers were particularly unreliable in their judgments of a proposal's Significance and Innovation. Although one can attain reliability in the judgments of these dimensions by increasing the number of reviewers, the number required to attain even a modest reliability of .5 is high.

Reviewers were fairly consistent on at least one dimension of merit: the Investigator. The exact reason for this finding is unclear, but it may have occurred because the Investigator is one dimension for which it is easy to formulate heuristics, such as counting the number of publications and funded grants a PI has previously received. If this is the strategy reviewers employed, it is one that risks creating a winner-take-all system [6] that advantages PIs who are willing and able to churn out large numbers of grant proposals and publications at the expense of the scientific quality of individual projects [7]. Ironically, although heuristics such as counting publications are not necessarily valid, they are reliable: because they rely on easily-assessed aspects of a proposal, different reviewers who use these heuristics are highly likely to come to similar judgments of scientific merit.

Reviewers did not converge in the language that they used to describe the same proposal. Even for word categories involving rare language, such as standout adjectives or agentic words, reliability was abysmal, and a minimum of 86 reviewers were required to bring the reliability of the aggregate word counts up to an even modestly acceptable level of .5. However, our assessment of similar language was based on a simplistic count of words falling into different categories; it is possible that a more effortful approach that takes into account the context in which words are used would yield higher assessments of reliability. At the very least our analysis suggests that the counts of words used to describe proposals are more specific to particular reviewers than they are to the content of a particular proposal.

The study on which this paper is based uses methods that differ in real ways from the true evaluation process at the National Institutes of Health. Our review process provides a good match to the initial stage of proposal review, when some number (usually three) of scientists serve as either primary, secondary, or tertiary reviewers by providing written critiques and quantitative scores for each grant proposal. However, after this stage, discrepancies between reviewers are resolved through discussion, a process that may improve the reliability of the final Priority Scores assigned to each proposal. We also only had primary reviewers in our study rather than a mix of primary, secondary, and tertiary reviewers. Insofar as primary reviewers put more effort and attention into their reviews, this feature of our study may have even inflated our estimates of the reliability of the true NIH process. What our analysis does show, however, is that a large number of primary reviewers – probably 10 or more – is required to attain even moderate levels of consistency in judgments of scientific merit. To the extent that scientific institutions, such as funders and journals, rely on relatively small panels of experts to make independent quantitative and written assessments to judge the quality of scientific projects, they may be relying on a system that more arbitrary than one might like.

We are not the first to claim that judgments of scientific merit have low levels of consistency [8,9]. However, our results go beyond these past results by showing that large numbers of reviews are necessary to attain even moderate levels of consistency in judgments. If the scientific institutions that rely on quality judgments are willing to invest the resources required to obtain this large number of reviews, they should be able to attain acceptable levels of reliability in these judgments. If they are not, they may need to consider more radical proposals. One particularly intriguing option involves entrusting the allocation of scientific resources to a random number generator [10]: As long as proposals meet a certain threshold of methodological soundness, they are entered into a random drawing for funds. Although this proposal may sound radical, some simulation evidence suggests that in addition to eliminating many of the costs associated with peer review, it may actually accelerate scientific discovery by funding projects that peer reviewers may have overlooked [10].

Whatever changes we do implement to the peer review system, evidence continues to accumulate that the review of grant proposals is less consistent than we might like. We believe the time has come to consider other systems that are more consistent and that therefore might be less a hindrance to scientific progress and the career trajectories of individual scientists.

Method:

For this study, we leveraged a prior dataset of grant proposal reviews collected for a separate experiment on the presence or absence of bias in NIH's initial stages of R01 review [1]. In this experiment, 412 experienced scientists re-reviewed a set of three grant proposals (out of 48 total proposals) in one of 12 general areas of science. The names of all the grant personnel had been stripped and replaced with fictitious names to protect the confidentiality of the original personnel. In the original study, the names of the PIs were changed to names designed to connote White male, Black male, White female, or Black female identities. We ignored this manipulation for the purposes of this study; for more details on this manipulation, see the study from which we drew our dataset [1].

Obtaining grant proposals for review. The original study sought grant proposals that captured a range of quality, from high-quality, funded proposals to moderate-quality, unfunded proposals. However, NIH only provides information about proposals that have been funded, so to obtain stimuli for the original study, we needed to start with proposals that had been eventually funded. In the original study, 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, which presumably was different and lower in quality than the proposal that was eventually funded. The proposals seen by reviewers 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).

In the original study, we wanted the 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. For more details on the process we used to select grant proposals, see the original report from which we obtained our dataset [1].

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.

Our selection process resulted in 48 proposals, four 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/.

Recruiting reviewers. The materials used to recruit reviewers for our original study are at https://osf.io/c5csm/. In our original study, 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.

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.

A total of 446 reviewers turned in reviews. Of these, we filtered out 34 reviewers (as did the authors of the project from which we drew our dataset [1]) because they turned in reviews mentioning that they noticed that some of the original grant personnel were fictitious. We conducted our analyses on the 412 remaining reviewers. Based on their responses in the eligibility survey, the majority (58%) had previously served on an NIH study section.

Reviewing procedure. In the original study, 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 reviewers at NIH. Participants were given one month to complete their three reviews (as 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, reviewers were discouraged from using outside resources aside from basic background reading.

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 as shown in Fig. 1.

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).

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.

The reliability of review scores. To assess the reliability of each dimension on which a proposal receives scores, we estimated the amount of variation in scores due to consistent differences between proposals, reviewers, and other factors with a series of Linear Mixed Effects Models using lme4 [11]. These models used a particular dimension of review as the outcome measure and included by-proposal and by-reviewer random intercepts; the scripts used to conduct these analyses are at https://osf.io/c5csm/.

To examine whether aggregating scores across multiple reviewers could lead to acceptable levels of reliability we used formulas described by Generalizability Theory [2]. Generalizability Theory posits two forms of reliability: absolute, which measures the degree to which an aggregate captures the "true" score, and relative, which measures the degree to which an aggregate captures the relative ranking of the items of interest (in this case, the rank-order of the scientific merit of different grant proposals).

Following Brennan [2] (equation 2.40), given an estimate of the variance due to proposal-specific factors ($\sigma^2_{\text{proposal}}$) and an error variance (i.e., variance due to factors other than the stable characteristics of proposals or reviewers, or what we call "observation-level" variance; $\sigma^2_{\text{observation}}$), the formula for the *relative reliability* of an aggregate score averaged *n* reviewers is:

$$relative\ reliability = \frac{\sigma^2_{proposal}}{\sigma^2_{proposal} + \frac{\sigma^2_{observation}}{n}}$$

A formula for the *absolute reliability* of an aggregate score can be derived from Brennan's [2] equation 2.41. Given the preceding quantities, as well as an estimate of the variance due to reviewer-specific factors ($\sigma^2_{\text{reviewer}}$):

$$absolute\ reliability = \frac{\sigma^2_{proposal}}{\sigma^2_{proposal} + \frac{\sigma^2_{reviewer} + \sigma^2_{observation}}{n}}$$

We used these formulas and our estimates of the difference variances to forecast how the reliability of an aggregate review score would change if we changed the number of reviewers who contributed their scores to the aggregate across each of the dimensions of review. We also estimated the number of reviewers required to obtain an absolute and relative reliability of at least .5, which we believe is a reasonable floor for any process designed to select grants to fund. These results are shown in Fig. 3.

The reliability of critique language. Even if peer review scores are relatively unreliable, reviewers may converge in the language they apply to the same proposal. To investigate this possibility, we conducted an exploratory analysis on the words used in critiques to describe grant proposals. 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 written NIH critiques [3]. The word categories are shown in Table 1 and include ability, achievement, agentic, research, standout adjectives, the positive and negative evaluation of grants, and negations. [3] developed and validated 7 of these categories using a modified Delphi method to assess language relevant to the evaluation of grants; the remaining two categories, negations and pronouns, comes from the Linguistic Inquiry and Word Count (LIWC) software [4] 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").

For each category, used lme4 [11] to fit a Generalized Linear Mixed Effects Model with a logit link and binomially distributed error, with random by-proposal and by-reviewer intercepts. To control for differences in reviewer verbosity we weighted the response variable by the total word count in each critique. The variation in the degree to which proposals elicited words from each category, as well as the variation in the degree to which reviewers used the same words, are shown in Fig. 4.

As in our analysis of peer review scores, we used these variance components to estimate how the reliability of language use, aggregated across reviewers, changes with different numbers of reviewer critiques per proposal. We used similar equations as in our analysis of peer review scores, with the except that our observation-level (i.e., error) variances were fixed to $\frac{\pi^2}{3} \approx 3.29$ rather than estimated from the data [5]. These results are shown in Fig. 5.

References:

1. Forscher, P. S., Cox, W. T. L., Brauer, M. & Devine, P. G. Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nat. Hum. Behav.* **3**, 257–264 (2019).
2. Brennan, R. L. *Generalizability Theory*. (Springer New York, 2001).
3. 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).
4. 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).
5. Wu, S., Crespi, C. M. & Wong, W. K. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemp. Clin. Trials* **33**, 869–880 (2012).
6. Bol, T., de Vaan, M. & van de Rijt, A. The Matthew effect in science funding. *Proc. Natl. Acad. Sci.* **115**, 4887–4890 (2018).
7. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3**, 160384 (2016).
8. Marsh, H. W., Jayasinghe, U. W. & Bond, N. W. Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *Am. Psychol.* **63**, 160–168 (2008).
9. Pier, E. L. *et al.* Low agreement among reviewers evaluating the same NIH grant applications. *Proc. Natl. Acad. Sci.* **115**, 2952–2957 (2018).

10. Avin, S. Funding Science by Lottery. in *Recent Developments in the Philosophy of Science: EPSA13 Helsinki* (eds. Mäki, U., Votsis, I., Ruphy, S. & Schurz, G.) **1**, 111–126 (Springer International Publishing, 2015).

11. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).

Acknowledgments: We would like to 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 would also like to acknowledge Jessica Flake, Jake Westfall, and Jonas Lang for their helpful comments on our analyses.

Funding: This research was supported by 5R01GM111002-02, awarded to the last author.

Author contributions: Conceived research: Brauer, Forscher. Designed research: Brauer, Forscher. Collected the data on which this study is based: All authors. Analyzed data: Forscher, Brauer. Wrote first draft: Forscher. Revised paper: All authors.

Competing interests: Authors declare no competing interests.

Data and materials availability: All data and materials are available at https://osf.io/uy7vq/, with the exception of our modified grant proposals, which we agreed not to share outside of our research team for confidentiality. To gain access to the grant proposals, contact the corresponding authors, who will seek permission from the original authors of the proposals.