Low agreement among reviewers evaluating the same NIH grant applications

Elizabeth L. Pier[a,b,1], Markus Brauer[c], Amarette Filut[a], Anna Kaatz[a], Joshua Raclaw[a,d], Mitchell J. Nathan[b], Cecilia E. Ford[a,e,f], and Molly Carnes[a,g]

[a]Center for Women's Health Research, University of Wisconsin–Madison, Madison, WI 53715; [b]Department of Educational Psychology, University of Wisconsin–Madison, Madison, WI 53706; [c]Department of Psychology, University of Wisconsin–Madison, Madison, WI 53706; [d]Department of English, West Chester University, West Chester, PA 19383; [e]Department of English, University of Wisconsin–Madison, Madison, WI 53706; [f]Department of Sociology, University of Wisconsin–Madison, Madison, WI 53706; and [g]Department of Medicine, University of Wisconsin–Madison, Madison, WI 53792

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved February 5, 2018 (received for review August 23, 2017)

Obtaining grant funding from the National Institutes of Health (NIH) is increasingly competitive, as funding success rates have declined over the past decade. To allocate relatively scarce funds, scientific peer reviewers must differentiate the very best applications from comparatively weaker ones. Despite the importance of this determination, little research has explored how reviewers assign ratings to the applications they review and whether there is consistency in the reviewers' evaluation of the same application. Replicating all aspects of the NIH peer-review process, we examined 43 individual reviewers' ratings and written critiques of the same group of 25 NIH grant applications. Results showed no agreement among reviewers regarding the quality of the applications in either their qualitative or quantitative evaluations. Although all reviewers received the same instructions on how to rate applications and format their written critiques, we also found no agreement in how reviewers "translated" a given number of strengths and weaknesses into a numeric rating. It appeared that the outcome of the grant review depended more on the reviewer to whom the grant was assigned than the research proposed in the grant. This research replicates the NIH peer-review process to examine in detail the qualitative and quantitative judgments of different reviewers examining the same application, and our results have broad relevance for scientific grant peer review.

peer review | social sciences | interrater reliability | linear mixed-effects models

In the past decade, funding at the National Institutes of Health (NIH) has increased at a much slower rate (1) than the number of grant applications (2), and consequently, success rates have steadily declined (3). There are more deserving grant applications than there are available funds, so it is critical to ensure that the process responsible for awarding such funds—grant peer review—reliably differentiates the very best applications from the comparatively weaker ones. Research on grant peer review is inconclusive: Some studies suggest that it is unreliable (4–13) and potentially biased (14–17), whereas others show the validity of review systems and final outcomes (18–20). However, even if peer review effectively discriminates the good applications from the bad, it is now imperative to empirically assess whether, in this culture of decreasing funding rates, it can discriminate the good from the excellent within a pool of high-quality applications. As Chubin and Hackett (21) argue, intensified competition for resources harms peer review because funding decisions rely on an evaluation process that is not designed to distinguish among applications of similar quality—a scenario that they argue is most prevalent at the NIH. Indeed, the findings in the present paper suggest that, in fact, reviewers are unable to differentiate excellent applications (i.e., those funded by the NIH in the first round) from good applications (i.e., those unfunded but later funded by the NIH after subsequent revisions).

Because the grant peer-review process at NIH is confidential, the only way to systematically examine it is to replicate the process outside of the NIH in a highly realistic manner. This is precisely what we did in the research reported in this paper. We recruited 43 oncology researchers from across the United States to participate in one of four peer-review panels (called "study sections" at NIH), each composed of 8–12 reviewers. Fig. 1 presents a deidentified image from one study section meeting. We solicited 25 oncology grant applications submitted to NIH as R01s—the most competitive and highly sought after type of grant at NIH—between 1 and 4 y before our study. Sixteen of these were funded in the first round (i.e., the best applications), whereas 9 of these were funded only after subsequent resubmission (i.e., the good applications).

The NIH uses a two-stage review process. In the first stage, two to five reviewers individually evaluate each grant application by assigning a preliminary rating using the NIH's reverse 9-point scale (1 = exceptional, 9 = poor) and writing a critique describing the application's strengths and weaknesses. Most typically, three reviewers are assigned to an application: a primary, a secondary, and a tertiary reviewer, ranked in order of the relevance of their expertise. Reviewers then convene in study section meetings, where they discuss the applications that received preliminary ratings in the top half of all applications evaluated. After sharing their preliminary ratings and critiques, the two to five assigned reviewers discuss the application with all other study section members, all of whom assign a final rating to the application. This final rating from all members is averaged into a final "priority score." In the second stage, members of NIH's advisory councils use this priority score and the written critiques to make

Significance

Scientific grant peer reviewers must differentiate the very best applications from comparatively weaker ones. Despite the importance of this determination in allocating funding, little research has explored how reviewers derive their assigned ratings for the applications they review or whether this assessment is consistent when the same application is evaluated by different sets of reviewers. We replicated the NIH peer-review process to examine the qualitative and quantitative judgments of different reviewers examining the same grant application. We found no agreement among reviewers in evaluating the same application. These findings highlight the subjectivity in reviewers' evaluations of grant applications and underscore the difficulty in comparing the evaluations of different applications from different reviewers—which is how peer review actually unfolds.

Author contributions: A.K., C.E.F., and M.C. designed research; A.K., J.R., C.E.F., and M.C. performed research; E.L.P. and A.F. coded data with input from M.J.N.; E.L.P. and M.B. analyzed data; and E.L.P. and M.B. wrote the paper with input from all coauthors.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: epier@wisc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1714379115/-/DCSupplemental.

Published online March 5, 2018.

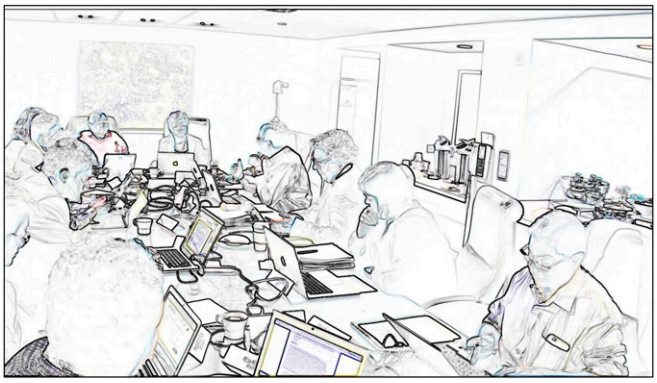

Fig. 1. Deidentified image from one of four peer-review panel meetings.

funding recommendations to the director of the NIH institute or center that awards the funding. Reviewers in study sections are prohibited from discussing or considering issues related to funding and instead are encouraged to rate each application based on its scientific merit alone.

In our study, each reviewer served as the primary reviewer for two deidentified applications. We analyzed only the ratings and critiques from the primary reviewers because their critiques were longer and more detailed than those of the secondary or tertiary reviewers. In total, we obtained 83 ratings and critiques from 43 primary reviewers evaluating 25 grant applications: Each reviewer evaluated two applications, except for three reviewers who evaluated one application, so that every application was evaluated by between two and four reviewers. Our methodology is presented in detail in *SI Appendix*.

We measured agreement among reviewers in terms of the preliminary ratings that they assigned to grant applications before the study section meeting. Our prior research (11) established that discussion during study section meetings worsened rather than improved disagreement among different study sections. The current study aims to examine agreement in the individual ratings before the study section meeting, with a focus on examining the alignment between reviewers' preliminary ratings and their written critiques.

Building off of the approach used by Fiske and Fogg (22) to code the weaknesses in journal manuscript reviews, we coded the critiques, assigning scores for the number of strengths and the number of weaknesses noted by the reviewer. We measured agreement among reviewers in terms of the number of strengths and weaknesses that they noted. We also examined whether different reviewers agreed on how a given number of strengths and weaknesses should translate into a numeric rating.

Results showed that different reviewers assigned different preliminary ratings and listed different numbers of strengths and weaknesses for the same applications. We assessed agreement by computing three different indicators for each outcome variable, and we depict these measures of agreement in Fig. 2.

First, we estimated the intraclass correlation (ICC) for grant applications. The ICC is a statistic that measures how strongly units within a group resemble one another; for our study, the ICC shows the extent to which different reviewers agreed in their evaluations of a single grant application. The ICC turned out to be 0 [$P = 1.0$, 95% CI (0, 0.14)][*] for the ratings, 0 [$P = 1.0$, 95% CI (0, 0.15)] for strengths, and 0.017 [$P = 0.9$, 95% CI (0, 0.18)] for weaknesses, indicating that there was no agreement among reviewers for a given application. Values of 0 for the ICC arise when the variability in the ratings for different applications is smaller than the variability in the ratings for the same application, which was the case in our data. These results show that multiple ratings for the same application were just as similar as ratings for different applications. Thus, although each of the 25 applications was on average evaluated by more than three reviewers, our data had the same structure as if we had used 83 different grant applications.

As another means of assessing agreement, we computed the interrater reliability statistic Krippendorff's alpha (23), which is also an indicator of agreement and can be interpreted similarly to Cronbach's alpha (24): Values above 0.7 are generally considered acceptable. The Krippendorff's alpha values were all near zero, showing that there was no agreement among reviewers in their ratings [$\alpha = 0.024$, 95% CI (−0.047, 0.093)] or in the number of strengths [$\alpha = -0.011$, 95% CI (−0.094, 0.079)] or weaknesses that they listed [$\alpha = 0.004$, 95% CI (−0.063, 0.072)].

As a third means of assessing agreement, we computed an overall similarity score for each of the 25 applications (see *Methods* for computational details). Values larger than 0 on this similarity measure indicate that multiple ratings for a single application were on average more similar to each other than they were to ratings of other applications. We computed a one-sample t test to examine whether similarity scores for our 25 applications were on average reliably different from zero. Results showed nonsignificant results for preliminary ratings [$t(24) = 0.07$, $P = 0.95$, $M = 0.01$,[†] 95% CI (−0.21, 0.22)], for strengths [$t(24) = -0.35$, $P = 0.73$, $M = 0.01$, 95% CI (−0.23, 0.25)], and for weaknesses [$t(24) = 0.29$, $P = 0.77$, $M = 0.01$, 95% CI (−0.21, 0.22)]. In other words, two randomly selected ratings for the same application were on average just as similar to each other as two randomly selected ratings for different applications.

In an additional analysis (see *SI Appendix* for details), we examined whether our reviewers agreed with the original NIH reviewers who decided on each application's outcome when it was first submitted to the NIH. The estimated linear mixed-effects model (LMEM; *SI Appendix*, Table S6) showed that our reviewers rated unfunded applications just as positively as funded applications ($P = 0.58$). Funded and unfunded applications also did not differ in the number of strengths or weaknesses that our reviewers mentioned in their critiques (Ps > 0.25).

Our analyses consistently show low levels of agreement among reviewers in their evaluations of the same grant applications—not only in terms of the preliminary rating that they assign, but also in terms of the number of strengths and weaknesses that they identify. Additionally, our results cannot be explained by differences in reviewers' verbosity: When we repeated all agreement analyses but considered only the major strengths and major weaknesses that were identified in the written critiques, we found similarly low levels of agreement (see *SI Appendix* for computational details). Note, however, that our sample included only high-quality grant applications. The agreement may have been higher if we had included grant applications that were more variable in quality. Thus, our results show that reviewers do not reliably differentiate between good and excellent grant applications. Specific examples of reviewer comments that illustrate the qualitative nature of the disagreement can be found in *SI Appendix*.

We wanted to know whether the lack of agreement stemmed from reviewers' differing opinions about what constitutes the best science or whether they used the rating scale differently. In other words, we assessed whether reviewers' evaluations of an application were simply different or whether they disagreed about how a given number of strengths and weaknesses should be translated into a

[*]Note that because the ICC cannot take values smaller than 0, the lower bound of the 95% CI is 0 for all of the estimates. Confidence intervals were estimated via bootstrapping using the confint function in lme4 within R. See *SI Appendix* for additional clarification about interpreting values of the ICC.

[†]M is the point estimate for the mean difference derived by subtracting (*i*) the average absolute difference among all ratings for one application from (*ii*) the average absolute difference between each rating for that application and the ratings for all other applications (see *Methods* for computational details). The 95% CI is the interval around M.

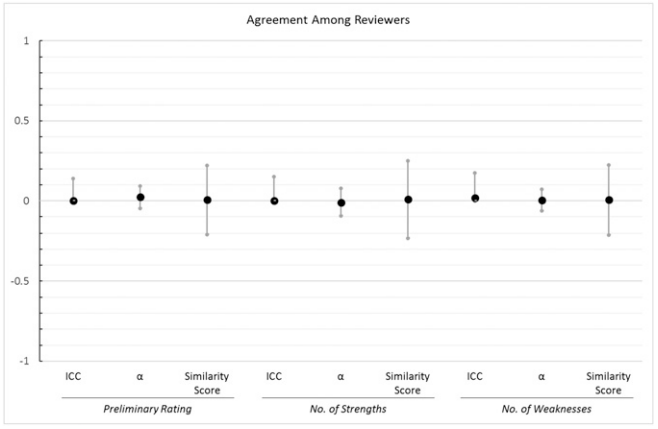

Fig. 2. Visual depiction of the three measures of agreement among reviewers with 95% CIs. Note that only the upper bound of the CI is shown for the ICCs because the lower bound is by definition 0.

numeric rating. To accomplish this goal, we examined whether there is a relationship between the numeric ratings and critiques at three different levels: for individual reviewers examining individual applications, for a single reviewer examining multiple applications, and for multiple reviewers examining a single application.

In an initial analysis (model 1, Table 1), we found no relationship between the number of strengths listed in the written critique and the numeric ratings. This finding suggests that a positive rating (i.e., a low number) did not necessarily indicate a large number of strengths in the critique, but instead reflected an absence of weaknesses. For this reason, we focused only on the relationship between the number of weaknesses and the preliminary ratings in the analyses reported below. For each of the 25 applications, we computed the average number of weaknesses that reviewers identified for that application (the "application cluster means"). For each of the 43 reviewers, we also computed the average number of weaknesses written by that reviewer (the "reviewer cluster means"). We then estimated a LMEM (model 2, Table 1) in which we predicted the preliminary ratings as a function of three fixed-effect variables: the number of weaknesses listed by each reviewer for each of the applications, the application cluster means, and the reviewer cluster means.

As Table 1 shows, the first of these predictors in model 2 was statistically significant: $b_{\text{Weaknesses(Within-Within)}} = 0.13$; $P = 0.003$. This result replicates the result from model 1 showing a significant relationship between preliminary ratings and the number of weaknesses within applications and within reviewers (i.e., for a single reviewer evaluating a single application).

The second predictor, the application cluster means, was also statistically significant, $b_{\text{Weaknesses(App\_Cluster\_Means)}} = 0.17$; $P < 0.001$. This coefficient represents the weakness-rating relationship between applications and within reviewers (i.e., for a single reviewer evaluating multiple applications). This result shows that, when a given reviewer identified more weaknesses for application A than for application B, then the reviewer also tended to give a worse rating to application A than to application B. Conceptually, this finding means that reviewers have internal standards that they apply consistently when rating different applications.

Most importantly for the present paper, the third predictor was not statistically significant: $b_{\text{Weaknesses(Rev\_Cluster\_Means)}} = 0.03$; $P = 0.19$. This coefficient represents the weakness-rating relationship between reviewers and within applications (i.e., across multiple reviewers evaluating a single application): When reviewer A identified more weaknesses for a given application than reviewer B, it was not necessarily the case that reviewer A evaluated that application more negatively than reviewer B. Although null effects should be interpreted with caution, a nonsignificant result here suggests that reviewers do not agree on how a given number of weaknesses should be translated into (or should be related to) a numeric rating.

The importance of this last finding cannot be overstated. If there is a lack of consistency between different reviewers who evaluate the same application, then it is impossible to compare the evaluations of different reviewers who evaluate different applications. However, this is the situation in which members of NIH study sections typically find themselves, as their task is to rate different grant applications that were evaluated by different reviewers. Our analyses suggest that for high-quality applications (i.e., the good or best applications that get discussed), this ranking process has a large random component, since reviewers disagree about how the number of weaknesses and the numeric rating are related to each other (even though they are individually consistent in assigning worse ratings to applications for which they have identified more weaknesses). The criteria considered when assigning a preliminary rating appear to have a large subjective element, which is particularly problematic given that biases against outgroup members (e.g., females and underrepresented racial/ethnic minorities) infiltrate decision-making processes when evaluative criteria are subjective (25).

The findings reported in this paper suggest two fruitful avenues for future research. First, important insight can be gained from studies examining whether it is possible to get reviewers to apply the same standards when translating a given number of weaknesses into a preliminary rating. Reviewers could complete a short online training (26) or receive instructions that explicitly define how the quantity and magnitude of weaknesses aligns with a particular rating, so that reviewers avoid redefining merit by inconsistently weighting certain criteria (27). Second, future studies should examine whether it is possible for reviewers to find common ground on what good science is before they complete their initial evaluation. Prior research examining journal peer review suggests that reviewers identify different, yet "appropriate and accurate" topics in their critiques (ref. 28, p. 591). So, is the problem in grant peer review that reviewers have fundamentally different goals? For example, some choose to focus on weaknesses of the approach, whereas others try to champion research that they believe should be funded (22). Do these goals differ as a function of the reviewer's expertise related to a particular grant application (29)? Or, does the lack of agreement stem from ambiguous, vague evaluative criteria that introduce subjectivity into the way such criteria are applied (25,

Table 1. Parameter estimates from models 1 and 2

Parameters	Model 1	Model 2
Fixed effects	$b(SE)^{Sig}$	$b(SE)^{Sig}$
(Intercept)	3.46 (0.21)***	3.51 (0.15)***
$\text{Strengths}_{\text{(Within-Within)}}$	−0.01 (0.02)	–
$\text{Weaknesses}_{\text{(Within-Within)}}$	0.08 (0.03)*	0.13 (0.02)**
$\text{Weaknesses}_{\text{(App\_Cluster\_Means)}}$	–	0.17 (0.03)***
$\text{Weaknesses}_{\text{(Rev\_Cluster\_Means)}}$	–	0.03 (0.02)
Random effects		
By reviewer		
Intercept	0.97	0.62
$\text{Strengths}_{\text{(Within-Within)}}$	0.00	–
$\text{Weaknesses}_{\text{(Within-Within)}}$	0.00	0.00
By application		
Intercept	0.16	0.00
$\text{Strengths}_{\text{(Within-Within)}}$	0.00	–
$\text{Weaknesses}_{\text{(Within-Within)}}$	0.01	0.00
Residual	0.45	0.69

On the outcome variable (preliminary rating), higher values represent more negative evaluations. The values reported for the random effects are variances. Dashes indicate the predictor was not included in the model. *$P < 0.05$. **$P < 0.01$. ***$P < 0.001$.

27)? Future studies ought to empirically examine whether addressing these issues might help improve agreement among reviewers.

If additional research were to reveal that it is impossible to increase agreement, then a viable solution would be to implement a modified lottery system, in which applications are initially screened by reviewers, and then a given proportion of applications with the best ratings are entered into a lottery (10). Compared with the costly peer-review process that is currently in place, such a lottery would free up financial resources that could be used to fund a larger number of grants. In addition, it would also allow the NIH to assess whether applications with very high ratings from the initial screening really yield more influential results and impactful publications than applications with slightly lower ratings from the initial screening. However, before moving forward with a modified lottery, additional studies with a larger sample of applications covering a wider variety of research areas ought to be conducted, perhaps by the NIH, to replicate the findings of our study.

Our study is not without limitations. First, we examined only evaluations of grant applications that were initially funded or eventually funded after subsequent revision; thus, we cannot say whether these findings would generalize to an entire pool of applications, including those that might never be funded by the NIH. Nonetheless, the results do show that, for grants above a certain quality threshold, the peer-review process is completely random. In addition, evaluating the reliability of grant peer review among strong applications that are considered fundable (i.e., those applications that are discussed during study section meetings) reflects the reality of the current funding climate today, where eligible applications continue to outweigh available funds. Nevertheless, future research should aim to extend the findings in this paper to a pool of applications of more diverse quality.

A second potential limitation stems from the possibility that reviewers in our study may have put less time and effort into their evaluations than real reviewers do when they know there are millions of dollars of research funds at stake. Relatedly, perhaps reviewers were more lenient in their judgments or less committed to their ratings because they knew their decisions would not result in real funding outcomes. However, we have evidence suggesting that the effort our reviewers put in for our study is comparable to the effort they would apply to an actual NIH study section. In a survey administered to reviewers, 81% of them reported that the premeeting process was either very similar or identical to actual NIH study sections in which they had participated. In addition, *SI Appendix* includes transcripts from interviews with our reviewers, statements from the Scientific Review Officer (SRO) in our study, and analyses of the word counts of reviewers' critiques compared with a national sample of more than 18,000 real NIH summary statements, all of which serve to corroborate the claim that reviewers took the task as seriously as if they had been part of a real study section.

One final limitation is that our study has a relatively small sample size, which means that our statistical models are somewhat underpowered. However, our most crucial effects are all estimated to be zero, suggesting that lack of power is not the issue. Furthermore, even if one is willing to accept a much higher type I error rate (e.g., $\alpha = 0.15$), these effects would still be nonsignificant. Nevertheless, a larger-scale study replicating our methods and analyses, and exploring their generalizability to other kinds of grant applications, is a fruitful and exciting arena for future research.

The process of vetting the quality, feasibility, and significance of multimillion dollar research projects is crucial to ensuring that increasingly sparse research funds are spent on the most meritorious applications. In these times of funding austerity, it is as important as ever to subject the current system of NIH peer review to continued empirical scrutiny to assess its efficacy and to evaluate possible interventions to improve the process. Determining additional or alternative practices to maximize reliability while minimizing the burdensome costs of grant peer review is vital for ensuring scientific progress.

Methods

The present study was approved by the Institutional Review Board at the University of Wisconsin-Madison, and informed consent was obtained from all participants (i.e., the reviewers in our study). The peer-review process at NIH is confidential, and all materials used during the peer-review process are destroyed at the end of every study section meeting (30). Thus, the only way to systematically examine the grant peer-review process is to replicate it outside of the NIH in a highly realistic manner. We designed and implemented four "constructed" study section meetings that emulated the NIH peer-review process in every respect possible. Throughout the study design and data collection, our team worked closely with a highly experienced SRO, Jean Sipe, who retired in 2012 after serving as an SRO for the NIH since 1997. Sipe had been the chairperson of the SRO handbook committee and the review policy coordinator since 2003, so she had extensive experience related to conducting and monitoring study sections at the NIH. Sipe guided all of our methodological decisions to ensure that they emulated NIH peer-review practices in every respect possible, and she served as the SRO for all of our constructed study sections.

To assess how closely our constructed study sections emulated real study sections, we administered a survey to our reviewers after the completion of each meeting asking them to rate their experiences on a 7-point Likert scale (1 = completely different, 7 = identical). Eighty-eight percent of the reviewers responded, and of those, 81% reported that the premeeting process was either very similar (6) or identical (7) to actual NIH study sections in which they had participated, and 78% reported that the meetings themselves were very similar or identical to actual NIH study section meetings.

Participants. We recruited 43 reviewers to participate in the study. Sipe used NIH's public database, RePORTER, to identify all investigators who had received an R01 award from the National Cancer Institute between 1 and 4 years prior to our study. Potential reviewers were emailed and invited to serve as a reviewer for an ad hoc study section to evaluate real but deidentified R01 grant applications. They were offered a $500 honorarium for their participation, with all travel and expenses reimbursed. Among interested respondents, Sipe selected the reviewers in the same way she would for a regular NIH study section. *SI Appendix*, Table S1, provides demographic information about our reviewers.

Materials. Our research team utilized RePORTER to identify principal investigators (PIs) who had submitted an R01 application between 1 and 4 years prior to our study to one of two oncology study sections within the NIH's National Cancer Institute: either the Oncology 1 Basic Translational Integrated Review Group or the Oncology 2 Translational Clinical Integrated Review Group. We invited these PIs to donate their funded and any unfunded versions of subsequently funded applications to our study. In consultation with Sipe, our research team selected 25 applications, 16 of which (64%) were initially funded and 9 of which (36%) were funded only after later resubmission. For these latter nine applications, we utilized the initial unfunded application to ensure variability in the quality of the grants in our study.

All applications were deidentified, meaning the names of the PIs, any coinvestigators, and any other research personnel were replaced with pseudonyms. We selected pseudonyms using public databases of names that preserved the original gender, nationality, and relative frequency across national populations of the original names. All identifying information, including institutional addresses, email addresses, phone numbers, and handwritten signatures were similarly anonymized and re-identified as well.

Procedure. In consultation with Sipe and staff from the NIH's Center for Scientific Review, our research team asked each reviewer to evaluate six applications: two as primary reviewer, two as secondary reviewer, and two as tertiary reviewer. This is on the low end of what would be typical in an NIH study section, which was intended to ensure maximal participation in our study. As is the norm for real NIH study sections, Sipe used reviewers' NIH biosketches and curricula vitae to assign applications to reviewers based on their expertise. Based on these assignments, Sipe appointed the reviewers to participate in one of the four study section meetings to ensure that each application was evaluated by three reviewers (one primary, one secondary, and one tertiary) in each study section. Each study section consisted of 8–12 reviewers. Most applications were evaluated by all four study sections, whereas a small subset of applications were evaluated by fewer than four panels. The assignment to applications and to study sections was not entirely random due to the highly specialized nature of the applications under review.

As is typical for NIH study sections, reviewers read the applications assigned to them before the meeting. They prepared a written critique that detailed the perceived strengths and weaknesses in terms of the overall impact and five specific criteria: significance, innovation, investigators, approach, and environment. Reviewers also assigned numeric ratings using a reverse 9-point scale (1 = exceptional, 9 = poor) for the overall impact and for each of the five criteria. The analyses reported in this paper include the critiques and ratings from primary but not from secondary or tertiary reviewers because the primary reviewers are those with the expertise most closely aligned to the application and because the reviewers in our study tended to put more detail and effort into their primary critiques compared with their secondary or tertiary critiques.

The applications were made available to the reviewers 5 wk before their meeting date via an online portal hosted by the institution at which the research took place. In the online portal, reviewers uploaded their written critiques using the same template used by NIH, and they entered their numeric ratings for each application. All reviewers in a given study section meeting were provided access to all of the reviews from other reviewers within their study section 2 d before the meeting, which is in line with real NIH study sections. As is also typical for NIH study sections, our SRO, Jean Sipe, monitored the review submissions and managed communication with reviewers to ensure that their submissions were complete and on time.

Qualitative Analysis. In total, we obtained 83 written critiques and preliminary ratings from the 43 reviewers, since three reviewers evaluated only one application as primary reviewer due to their particular expertise. We devised a coding scheme to analyze the number and types of strengths and weaknesses that primary reviewers pointed out in their critiques of applications. Each critique was coded and assigned two scores: (*i*) the number of strengths mentioned in the critique and (*ii*) the number of weaknesses. *SI Appendix* provides additional details about our coding approach.

Quantitative Analysis.

Agreement among reviewers. We assessed agreement for each of the three key variables: preliminary ratings, number of strengths, and number of weaknesses. We examined agreement with three different approaches, each described in turn below. For complete transparency, and because we wanted to treat both random factors (reviewers and applications) equally, we also examined agreement among applications (i.e., whether the ratings and evaluations across applications were consistent for a single reviewer; see *SI Appendix*), but readers should be aware that the primary focus of this paper is on the indicators for agreement among reviewers (i.e., whether the ratings and evaluations across reviewers were consistent for a single application). In addition, we repeated the analyses below using only the major strengths and major weaknesses that reviewers identified (rather than all strengths and all weaknesses) to ensure that the results were not confounded with reviewers' verbosity, and found similar results (*SI Appendix*).

First, we estimated the ICC, which measures the degree to which observations are clustered by a given random factor (here, application); in other words, the ICC measures the proportion of the total variance in the outcome variable (e.g., rating) that is attributable to the application itself. To compute the ICC, we estimated one model for each of the key variables (ratings, strengths, weaknesses). Each model included an overall fixed intercept and a random intercept for application. We then computed the ICC by dividing the variance of the random intercept by the total variance (i.e., the sum of the variance of the intercept and the variance of the residuals). *SI Appendix*, Table S5, provides the ICC values for ratings, strengths, and weaknesses for grant applications (i.e., examining agreement among reviewers within a grant application). *SI Appendix* also describes alternative specifications of the ICC.

Second, to corroborate the findings from estimating the ICC, we computed agreement among reviewers for each variable using the Krippendorff's alpha statistic (23). Krippendorff's alpha is an interrater reliability statistic that accommodates any kind of data (e.g., nominal, ordinal, interval), allows for missing values, and can be applied to any number of individual raters. It can be interpreted similarly to Cronbach's alpha: Values above 0.7 are generally considered acceptable (31). This set of analyses was carried out on a data file in which reviewers were treated like raters (columns) and applications were treated like targets (rows). We used 1,000 bootstrapped samples to estimate a 95% CI for each estimate. *SI Appendix*, Table S5, displays the values for Krippendorff's alpha.

Third, as an additional means of corroborating the findings from the ICC, we compared the similarity of ratings referring to one application versus the similarity of ratings referring to different applications. We computed two scores for every application: The first score was the average absolute difference between all ratings referring to that application. The second score was the average absolute difference between each of the ratings referring to that application and each of the ratings referring to all other applications. In the next step, we subtracted the first score from the second score to compute an overall similarity score per application. Values above zero on this score indicate that an application's ratings are more similar to each other than to ratings referring to other applications. We then tested whether the 25 overall similarity scores were significantly different from zero. Finally, we repeated this procedure for strengths and weaknesses mentioned in the reviewers' critiques. *SI Appendix*, Table S5, provides the estimates for these similarity tests.

Relationship between ratings and critiques. We next asked whether there is a relationship between the numeric evaluations and the verbal evaluations. No relationship would suggest that individual reviewers struggle to reliably assign similar numeric ratings to applications that they evaluate as having similar numbers of strengths and weaknesses. By comparison, evidence of a relationship would suggest that the lack of agreement among reviewers stems from their having fundamentally different opinions about the quality of the application—and not simply that they used the rating scale differently.

We began by estimating a model using the *lme4* package (32) in R in which we predicted an observation's preliminary rating from the number of strengths and the number of weaknesses (model 1, Table 1). Note that the data contain two random factors—reviewers and applications—that are crossed with each other. The two predictors, strengths and weaknesses, are continuous and vary both within reviewers and within applications. In such a case, it is necessary to "adaptively center" the predictors (33). This approach is similar to what others have referred to as "doubly centering" predictors (34). Failure to do so would lead to regression coefficients that are an "uninterpretable blend" of within-cluster and between-cluster effects (ref. 33, p. 138). Adaptive centering involves subtracting each of the two cluster means from the raw score and then adding the grand mean. For example, we adaptively centered the strength variable by taking the raw score and then (*i*) subtracting the mean number of strengths for a given reviewer (across applications), (*ii*) subtracting the mean number of strengths for a given application (across reviewers), and (*iii*) adding in the grand mean of strengths (the average of all 83 strength values). We adaptively centered both the strength and the weakness scores.

To account for nonindependence in the data, we included the appropriate random effects. We followed the lead of Brauer and Curtin (35) and included, for each of the random factors, one random intercept and one random slope per predictor. In total, we included six random effects—a by-reviewer random intercept, a by-reviewer random slope for strengths, a by-reviewer random slope for weaknesses, a by-application random intercept, a by-application random slope for strengths, and a by-application random slope for weaknesses—plus all possible covariances.

The resulting model was a LMEM with three fixed effects (the intercept and the two predictors) and 12 random effects. The full model did not converge, so we removed all covariances among random effects and reestimated the model, which achieved convergence. The parameter estimates from this model are presented in Table 1.

In model 1, the regression coefficients describe the (partial) relationships between each of the predictors and the outcome variable that are unconfounded with any between-cluster effects. In other words, they describe the within-reviewer/within-application relationships. Note that, when data are clustered by one random factor (e.g., applications nested within reviewers), it is possible to examine the relationship between outcome and predictor variables at two levels: within and between clusters (e.g., within reviewers and between reviewers). In our study, however, the data are clustered by two crossed random factors (i.e., reviewers and applications). In such a case, a given relationship can be examined at three levels: within-within, within-between, and between-within. This is precisely what we did in the following analysis (model 2, Table 1). We decided to focus on weaknesses only, because this predictor was the only one that was significantly related to the outcome variable in model 1.

We adopted a data-analytic strategy by Enders and Tofighi (36) who proposed to include the cluster-mean centered predictor (to examine the within-cluster relationship) and the mean-centered predictor cluster means (to examine the between-cluster relationship). We estimated a LMEM with the preliminary rating as the outcome variable that included the following predictors: the adaptively centered weakness value (to examine the within-reviewer/within-application relationship), the mean-centered reviewer cluster means of the weakness values (to examine the between-reviewer/within-application relationship), and the mean-centered application cluster means of the weakness values (to examine the between-application/within-reviewer relationship). We also included a random intercept and a random slope for the adaptively centered predictor for each of the two random factors (reviewers and applications). The full model with all possible covariances did not converge, but the model without the covariances did. The results of this analysis are shown in Table 1, model 2.

SI Appendix includes additional supplementary analyses, including (*i*) measuring agreement among reviewers for funded versus initially unfunded

applications, (*ii*) replicating analyses with major strengths and major weaknesses only, (*iii*) measuring agreement among applications, (*iv*) exploring the statistical relationship between strengths and weaknesses, (*v*) using alternative model specifications [e.g., the "model selection approach" of Bates and colleagues (37)], (*vi*) computing alternative measures of the ICC, and (*vii*) re-estimating all models with the inclusion of a single outlier in our data.

Data Sharing Plan. Deidentified data can be provided by request from the corresponding author. All code used in statistical analyses is included at the end of *SI Appendix*.

ACKNOWLEDGMENTS. We thank Jennifer Summ, Madeline Jens, Anupama Bhattacharya, Dastagiri Malikireddy, and You-Geon Lee for their assistance and Andrei Cimpian, James Pellegrino, Gerald Pier, and William Klein for their feedback on the manuscript. This work was supported by National Institute of General Medical Sciences of the National Institutes of Health Grant R01GM111002; the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison; and the Department of Medicine at the University of Wisconsin–Madison. The content is solely the responsibility of the authors and does not necessarily represent the official views of the supporting agencies.

1. Office of Budget; National Institutes of Health (2016) Actual total obligations by budget mechanism FY 2000–FY 2016. Available at https://officeofbudget.od.nih.gov/pdfs/FY18/Mechanism%20Detail%20for%20NIH%20FY%202000-FY%202016%20(V).pdf. Accessed August 8, 2017.
2. National Institutes of Health (2016) Research and training grants: Competing applications by mechanism and selected activity codes. Available at https://report.nih.gov/NIHDatabook/Charts/Default.aspx?showm=Y&chartId=200&catId=2. Accessed August 8, 2017.
3. National Institutes of Health (2016) Research and training grants: Success rates by mechanism and selected activity codes. Available at https://report.nih.gov/NIHDatabook/Charts/Default.aspx?showm=Y&chartId=202&catId=2. Accessed August 8, 2017.
4. Cole S, Cole JR, Simon GA (1981) Chance and consensus in peer review. *Science* 214:881–886.
5. Cicchetti DV (1991) The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behav Brain Sci* 14:119–186.
6. Mayo NE, et al. (2006) Peering at peer review revealed high degree of chance associated with funding of grant applications. *J Clin Epidemiol* 59:842–848.
7. Marsh HW, Jayasinghe UW, Bond NW (2008) Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *Am Psychol* 63:160–168.
8. Reinhart M (2009) Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics* 81:789–809.
9. Graves N, Barnett AG, Clarke P (2011) Funding grant proposals for scientific research: Retrospective analysis of scores by members of grant review panel. *BMJ* 343:d4797.
10. Fang FC, Casadevall A (2016) Research funding: The case for a modified lottery. *MBio* 7:e00422-16.
11. Pier EL, et al. (2017) 'Your comments are meaner than your score': Score calibration talk influences intra- and inter-panel variability during scientific grant peer review. *Res Eval* 26:1–14.
12. Obrecht M, Tibelius K, D'Aloisio G (2007) Examining the value added by committee discussion in the review of applications for research awards. *Res Eval* 16:70–91.
13. Fogelholm M, et al. (2012) Panel discussion does not improve reliability of peer review for medical research grant proposals. *J Clin Epidemiol* 65:47–52.
14. Kaatz A, Magua W, Zimmerman DR, Carnes M (2015) A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. *Acad Med* 90:69–75.
15. Ginther DK, et al. (2011) Race, ethnicity, and NIH research awards. *Science* 333:1015–1019.
16. Kotchen TA, et al. (2006) Outcomes of National Institutes of Health peer review of clinical grant applications. *J Investig Med* 54:13–19.
17. Ley TJ, Hamilton BH (2008) Sociology. The gender gap in NIH grant applications. *Science* 322:1472–1474.
18. Gallo SA, et al. (2014) The validation of peer review through research impact measures and the implications for funding strategies. *PLoS One* 9:e106474.
19. Bornmann L, Daniel HD (2004) Reliability, fairness and predictive validity of committee peer review. *BIF Futura* 19:7–19.
20. Wood F, Wessely S (2003) Peer review of grant applications: A systematic review. *Peer Review in Health Sciences*, eds Jefferson T, Godlee F (BMJ Books, London), pp 14–44.
21. Chubin DE, Hackett EJ (1990) *Peerless Science: Peer Review and U.S. Science Policy* (State Univ New York Press, Albany, NY).
22. Fiske DW, Fogg L (1990) But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. *Am Psychol* 45:591–598.
23. Hayes AF, Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 1:77–89.
24. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297–334.
25. Heilman ME, Haynes MC (2008) Subjectivity in the appraisal process: A facilitator of gender bias in work settings. *Beyond Common Sense: Psychological Science in Court*, eds Borgida E, Fiske ST (Blackwell Publishing, Oxford), pp 127–155.
26. Sattler DN, McKnight PE, Naney L, Mathis R (2015) Grant peer review: Improving inter-rater reliability with training. *PLoS One* 10:e0130450.
27. Uhlmann E, Cohen GL (2005) Constructed criteria: Redefining merit to justify discrimination. *Psychol Sci* 16:474–480.
28. Lamont M (2009) *How Professors Think: Inside the Curious World of Academic Judgment* (Harvard Univ Press, Cambridge, MA).
29. Gallo SA, Sullivan JH, Glisson SR (2016) The influence of peer reviewer expertise on the evaluation of research funding applications. *PLoS One* 11:e0165147.
30. National Institutes of Health (2016) NIH reviewer orientation. Available at https://grants.nih.gov/grants/peer/guidelines_general/reviewer_orientation.pdf. Accessed August 8, 2017.
31. Nunnally JC (1994) *Psychometric Theory* (McGraw Hill, New York), 3rd Ed.
32. Bates D, Maechler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67:1–48.
33. Raudenbush SW, Bryk AS (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods* (Sage, Thousand Oaks, CA).
34. Aiken LS, West SG (1991) *Multiple Regression: Testing and Interpreting Interactions* (Sage, Thousand Oaks, CA).
35. Brauer M, Curtin JJ (2017) Linear mixed-effects models and the analysis of non-independent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol Methods*, 10.1037/met0000159.
36. Enders CK, Tofighi D (2007) Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychol Methods* 12:121–138.
37. Bates D, Kliegl R, Vasishth S, Baayen H (2015) Parsimonious mixed models. arXiv: 1506.04967.