Supplementary Methods

Background on NIH Peer Review

At the National Institutes of Health (NIH), peer review takes place in a two-stage process. In the first stage, groups of expert scientists convene to evaluate grant applications submitted in response to a particular funding opportunity announcement. These groups, called Scientific Review Groups (SRGs), are administered by the Center for Scientific Review (CSR) at NIH. In the second stage, the advisory council of the awarding NIH Institute or Center examine reviews and make funding decisions based on the overall impact ratings assigned during the first stage, while taking into consideration the funding priorities of the Institute or Center. The director of the funding Institute or Center makes the final funding decision based on the advisory council's input.

The Scientific Review Officer. Each SRG is overseen by a Scientific Review Officer (SRO). The SRO is an NIH staff scientist tasked with ensuring that the SRG operates according to all relevant laws, procedures, and policies. The responsibilities of the SRO include recruiting qualified expert reviewers to serve in SRGs, assigning applications to reviewers based on their expertise, assigning a chairperson to serve as the moderator of the peer review meeting, managing any conflicts of interest that arise in the SRG, overseeing the SRG peer review meeting, and preparing summary statements to send to the Principal Investigator (PI) for each application reviewed during the SRG meeting.

Study Sections at NIH. At NIH, peer review meetings are referred to as *study sections*. Many study sections are what NIH terms *standing* study sections, in that there are rotating rosters of permanent members (as well as temporary members) who serve on a given study section for multiple years at a time. There are also Special Emphasis Panels that are convened on an ad hoc

basis, which our study aimed to replicate. Both types of study sections generally follow a predetermined procedure overseen by the SRO, described below.

1. Prior to the study section meeting:
    - The SRO assigns each reviewer to evaluate a set of applications, either as primary, secondary, or tertiary reviewer, with the primary reviewer having expertise most closely aligned to the application
    - Individual reviewers read the applications assigned to them, which they access via an online interface (electronic Research Administration, or eRA)
    - Reviewers assign a **preliminary overall impact rating** for each application, ranging from 1 (Exceptional) to 9 (Poor)
    - Reviewers assign **individual criterion ratings** on this nine-point scale for each of five criteria: Significance, Investigator(s), Innovation, Approach, and Environment
    - Reviewers write a critique that summarizes the overall impact of the application, and details the strengths and weaknesses for each of the five criteria. The critiques and ratings are submitted and made available to all reviewers prior to the meeting via the eRA system
    - The SRO calculates an order of review based on the average preliminary overall impact ratings from the three assigned reviewers, starting with the best (i.e., lowest) scoring application. Only the top 50% of applications are discussed during the meeting, and the bottom 50% are "triaged out" from discussion (meaning they do not receive a final overall impact rating and are no longer considered for funding)
2. During the study section meeting:
    - The SRO reviews the meeting procedures

- The chairperson introduces each application, beginning with the best-scoring application
- The three assigned reviewers (preliminary, secondary, tertiary) each announce their preliminary overall impact rating
- The three assigned reviewers each summarize their critique of the application
- The chairperson calls for and moderates discussion to the panel at large
- The chairperson summarizes the discussion for the panel
- The three assigned reviewers each announce their **final overall impact ratings** for that application (which may or may not have changed after discussion)
- The remaining non-reviewing panelists privately write down their final overall impact ratings. Panelists are expected to vote within the range set by the assigned reviewers (e.g., if the assigned reviewers assign ratings of 3, 4, and 5, all panelists are expected to assign ratings between a 3 and 5); however, if they wish to assign a rating outside the range, the chairperson will ask them to identify that they are doing so
- This procedure unfolds for each application discussed in the meeting

3. After the study section meeting:
    - Reviewers have the opportunity to edit their written critiques to reflect any changes to their ratings, if they wish to do so
    - The SRO compiles the reviewers' critiques into a **summary statement** that is provided to the PI, along with the **final overall impact rating** from the panel (i.e., the average rating from all panelists, multiplied by 10; final overall impact ratings range from 10 to 90). Summary statements from standing study sections may also include

the **percentile rating** that indicates the percentage of applications receiving better final overall impact ratings from the study section within the past year

Prior Work on Peer Review

There is prior research examining the grant peer review process. Some studies have assessed agreement among grant reviewers, but these studies investigated the grant peer review process at funding agencies located outside the US, including Canada (6, 12), Australia (7, 9, 38), Finland (13), and Switzerland (8), all of which have a grant peer review process that is rather different from the one at NIH. Furthermore, many of these studies examining reliability (11-13) focus on whether panel discussion improves agreement.

Although some researchers have studied peer review at NIH specifically, they analyzed data that were obtained before 2009 (39, 40), which was when NIH made significant changes to their rating system, in part in order to address the low reliability of the previous rating scale (41). Some of the seminal studies examining agreement in grant peer review even utilize data from the 1980s (4, 5) and 1990s (38). Given that modern research is becoming increasingly complex, broad, large-scale, and interdisciplinary (41), the research on older peer review mechanisms is hardly applicable to today's research landscape.

Note that many studies on pre-2009 NIH reviewer data do not examine the actual level of inter-reviewer agreement. Instead, their goal was to propose procedures for obtaining more reliable rating estimates, such as by increasing the number of reviewers (40) or by applying particular kinds of mathematical models (39), while other researchers focused not on reliability but on investigating NIH funding outcomes for biases based on gender (14, 17), race/ethnicity (15), or type of research (16).

All of the earlier studies on reliability in the grant peer review process focus exclusively on the numeric scores; no study to date has examined agreement in the written evaluations, nor whether reviewers agree on how to "translate" a certain number of strengths and weaknesses into a score. Consequently, no research has yet been conducted to rigorously evaluate the current NIH grant peer review process.

Constructed Study Section Methodology

Our constructed study section methodology is novel, but there is a long history from many different disciplines of relying on simulated or mock groups to examine group processes. For example, in medical education, the use of standardized patients and simulated patients is commonplace (42-44). For decades, studies conducted on mock juries have been used to draw conclusions about the interpersonal processes and group dynamics that unfold during jury deliberations for legal proceedings (45-48). Although the application of such methods to the study of peer review is novel, the use of such simulated methods to draw generalizable conclusions is well established.

Qualitative Analysis

We utilized the qualitative data analytical software program NVivo to engage in open coding (49) of the written critiques. First, three members of the research team collectively open coded several critiques together to establish a baseline coding scheme. Our approach was exhaustive, such that every word in every critique received a code with nothing left uncoded, and the codes were mutually exclusive, meaning that each word received one and only one code. As is common practice in qualitative data analysis, codes are named *in vivo* early in the coding process, meaning the name for the code is derived verbatim from the data. For example, if a

reviewer writes, "The research environment is outstanding," we would create a code called `Environment is outstanding`.

Next, one member of the team applied the coding scheme to a random sample of critiques until she reached *saturation*, i.e., when the codes became repetitive and no new codes are generated. At this point, she engaged in *axial* coding, which means she condensed codes and related codes to one another; for example, `Environment is outstanding` was combined with `Environment is exceptional` and with `Environment is excellent` to become an axial code, `Environment is a strength`. This process of axial coding led to a taxonomy of strengths and weaknesses based on two elements: (i) the *content* (e.g., an evaluation of the scientific approach, versus the qualities of the investigator, versus the innovation of the proposal), and (ii) the relative *magnitude* of the evaluation (e.g., a major weakness versus a minor weakness).

Table S2 lists all of the axial codes our research team derived, along with a brief definition of each code. There are five higher-order categories denoting the magnitude of evaluation: `Strength, Minor Strength, Weakness, Major Weakness`, and `Neither`. Within each category, there are axial codes (e.g., `Application, Approach, Innovation`, etc.) denoting the content of evaluation. Although some codes appear in more than one category (e.g., `Strength–Approach, Weakness–Approach`), note that not every code appears in every category, since some codes did not occur in our data (e.g., there were no major weaknesses related to Environment). Some codes are unique to a category; for example, `Advice` to the PI and `Questions` posed to the PI are each only included in the Weakness category. Table S3 provides an example from our data set for each of the axial codes.

Establishing inter-rater reliability. The next step in the coding process was for two members of the research team to collaboratively code several critiques together using the axial coding scheme in Table S2. Once both team members felt they had achieved a joint understanding of applying the coding scheme, each member independently coded a 20% random subsample of the data that had not yet been coded in order to compute inter-rater reliability. For the 20% subsample, percent agreement was 92.12% and weighted Cohen's kappa was .755.

In some fields in which researchers engage in qualitative research, it is expected or required to report only percentage agreement as a measure of inter-rater reliability. However, a growing body of research (50-53) has established that because percentage agreement does not correct for agreement due simply to chance, it is an upward biased measure of agreement and therefore cannot be trusted. On the other hand, Cohen's kappa (*k*) is considered the gold standard of reliability measures (54). Thus, using percentage agreement on its own is problematic and generally considered not robust enough for most fields and journals, whereas kappa is argued to be a more acceptable index to report.

There are many conflicting guidelines that exist for interpreting acceptable minimum values for kappa (55). The guidelines from Landis and Koch (56) are the most often cited; they label values of .21 - .40 as fair, .41 - .60 as moderate, .61 - .80 as substantial, and .81 - 1.00 as near perfect. Landis and Koch admit that the cutoff values are arbitrary, but argue that they provide useful benchmarks for interpreting kappa values. Other researchers (54, 55, 57, 58) have proposed a cutoff of .75 or larger—or even .70 or larger (59)—as indicative of excellent or substantial agreement, but some recommend providing both values of kappa and the measure of percent agreement (53, 54). Thus, we chose to report both measures here for maximal transparency, and we argue that given 92% agreement and K = .755, we achieved acceptable

agreement between the two coders.

Gold standard coder. The final step in our coding process is for one of the team members to code the remaining critiques using the coding scheme. This is what is known as the use of a *gold standard coder* (54).

Results of qualitative coding. Table S4 provides how frequently each axial code occurred in our data, the average incidence of each axial code per critique, the standard deviation of the axial code across critiques, and the strength of the correlation for each category with our outcome variable of interest, the preliminary overall impact rating. Note that because the preliminary overall impact rating uses a *reverse* nine-point scale (*1* = Exceptional, *9* = Poor), negative correlations indicate a relationship with better-rated applications and positive correlations indicate a relationship with worse-rated applications.

Illustrative Excerpts from the Corpus

In order to illustrate the nature of the disagreement among reviewers evaluating the same proposal, we provide a few examples from the data. The first example illustrates how reviewers perceive the same application to be of fundamentally different quality—that is, they assign starkly different scores. For the "Stavros" application (i.e., the pseudonym we assigned to the PI for our study), the four reviewers assigned scores of 7, 2, 3 and 4 out of a possible 9 (with 9 being the worst possible score). The following excerpts are extracted from the first paragraph of their critique, when asked to summarize the strengths and weaknesses of the application. We omitted additional details that were not coded as a strength or a weakness (i.e., summarizing the aims of the application):

> Reviewer 1: [Score = 7] "PI has experience in DNA technology but PI needs to learn in vivo animal study and design. Proposed radiation dose is inadequate and sample size justification is not described. Significance of the study in respect

> currently available technology/treatment is not well described. Different luciferase imaging is not innovative."
>
> Reviewer 2: [Score =2] "A highly focused application describing a streamlined approach for screening compounds that target glioma stem cells in preclinical models of glioblastoma… Well supported by preliminary data. "
>
> Reviewer 3: [Score = 3] "Strengths of the application are the PI, the solid preliminary data to support the hypothesis, the research team, and the research environment. There are some concerns regarding the approach that weaken the enthusiasm for this study, but overall, the enthusiasm is high and success is likely."
>
> Reviewer 4: [Score = 4] "Though this application has potential to identify druggable target/s, this proposal suffers from several flaws including 1. Open ended; 2. Interdependent aims; 3. lack of definite target/s. These problems reduce the enthusiasm for this otherwise promising proposal.

These excerpts illustrate that reviewers identify different types of weaknesses in this application (e.g., innovation, approach, significance), disagree on specific weaknesses (e.g., the qualifications of the PI), and importantly, assign vastly different scores. The next set of examples come from an application (PI pseudonym of "Rice") that received similar scores from the different reviewers: 4, 3, 3, and 4. However, the excerpts below demonstrate how different reviewers' evaluations are for specific weaknesses:

> Reviewer 1: [Score = 4] "Lack of preliminary data… dampens the enthusiasm of the proposal… Most of the proposed work is already established in other cancel models; hence the innovation of the study appears to be moderate. The experimental design needs more information on the power calculation of the number of patient samples and also for in vivo work. Overall, it is an very interesting proposal with a well-qualified team; however, lack of details diminishes the significance and the impact of the proposal."
>
> Reviewer 2: [Score = 3] "This is an excellent proposal… Dr. Rive is an expert in IGF-1R signaling and its role in breast cancer progression. She has enlisted the help of strong collaborators with expertise… Together this is an outstanding investigative team… The proposal is conceptually innovative, although it is not technologically innovative. The aims are logical and the experiments proposed are well-designed. There is confidence that useful information will be garnered from these studies. The one major weakness lies in the over-expression experiment shown in Figure 3D, where there is an almost undetectable downregulation of

> IGF-1R. The generation of such genetically manipulated cells is critical for the in vivo experiments proposed in Aim 2. Despite this weakness, enthusiasm for this application is high.
>
> Reviewer 3: [Score = 3] "This well-written application by Rice tests the novel hypothesis that Beclin 1 functions as a tumor suppressor…. Overall, this is an excellent proposal that presents a very interesting model that is soundly reasoned. However, at present it is still not clear whether this is truly an autophagy-independent mechanism. Further, other autophagy-independent mechanisms of Beclin 1 that have previously been described such as effects on BCL-2/apoptosis are minimally addressed."
>
> Reviewer 4: [Score = 4] "The hypothesis are sound and the preliminary data is abundant. However, my enthusiasm for this proposal was dampened by the somewhat scattered manner in which it was presented. The approach section was confusing and concepts introduced were not properly addressed in the introduction. Finally, the lack of a sufficient diagrams of models makes it very difficult to track the varying aspects of the experiments to determine whether the expected results were sound."

These excerpts illustrate how reviewers can assign very similar scores while in fact disagreeing about specific weaknesses, such as whether it is a well-written proposal, the study is innovative, or the there are fundamental flaws in the scientific approach.

Supplementary Statistical Analyses

This section contains additional analyses that some readers might want to know about in order to form a more complete impression of the relationships between the variables. The additional models we estimated testify to the stability of the observed effects.

Measuring agreement for funded versus unfunded applications. Given that we found no agreement among reviewers for any of our outcome measures (rating, strengths, or weaknesses), we conducted an exploratory analysis to check whether our reviewers, at the least, (i) assigned better ratings, (ii) listed more strengths, or (iii) listed fewer weaknesses for those applications that were initially funded by NIH compared to those applications that were

unfunded. In other words, we wanted to know whether our reviewers agreed with the original NIH reviewers who decided on each application's outcome when it was first submitted to NIH.

To assess this, we estimated three separate linear mixed-effects models using the *lme4* package in R (32) for three outcome variables (i) preliminary rating, (ii) number of strengths, and (iii) number of weaknesses. In all three models, we included a fixed-effect predictor for funding status, which was coded as a dichotomous variable centered around 0 (i.e., –0.5 for unfunded applications and +0.5 for funded applications). Following the recommendations of Brauer and Curtin (35), we included the maximal random effects structure called for by the design. Here, the design requires us to account for the non-independence introduced by reviewer and by application with random intercepts and random slopes for each. When the maximal random effects structure does not converge, as was the case for these three models, experts recommend removing random effects one by one until the model converges into the maximally converged model (35). After removing the covariances among random effects, all three models converged, so they each contain two random intercepts and two random slopes, but not their covariances.

The result of these exploratory analyses (Table S6) showed that our reviewers rated unfunded applications just as positively as funded applications ($p = .58$). Funded and unfunded applications also did not differ in the number of strengths or weaknesses that our reviewers mentioned in their critiques (ps > .25). Thus, the reviewers in our study did not agree with each other, nor with the original NIH reviewers who evaluated the applications.

Major strengths and major weaknesses only. In order to check the robustness of the agreement analyses and ensure that the low levels of agreement among reviewers was not a function of how many words they wrote, we repeated all agreement analyses with only the major strengths and major weaknesses, rather than all strengths and all weaknesses. These variables

include only those strengths or weaknesses coded as substantial or significant within the context of the critique, rather than a mere mention of a strength or weakness that might be mentioned merely because a reviewer is verbose.

The results of these analyses show similarly low levels of agreement as when we include all strengths and weaknesses. The ICC for major strengths was 0.005 (p = .8, 95% CI [0, 0.19]), and the ICC for major weaknesses was 0.007 (p = 1.0, 95% CI [0, 0.18]). The value of Krippendorff's alpha for major strengths was α = .120 (95% CI [.074, .171]) and for major weaknesses was α = .151 (95% CI [.053 .250]). Finally, the one-sample t-tests of the similarity scores we computed showed non-significant results for both major strengths (t(24) = .156 p = .88, 95% CI [-2.8, 3.3]) and major weaknesses (t(24) = .79, p = .44, 955 CI [-.17, .39]), which indicates that there was not a statistically significant difference between (i) the major strengths (or major weaknesses) listed by different reviewers for the *same* application and (ii) the major strengths (or major weaknesses) listed for *different* applications. Taken together, these supplemental analyses closely mirror the results of the analyses in which we included all strengths and all weaknesses. Since the two indicators analyzed here focus only on substantial strengths and weaknesses that are central to the overall evaluation of the grant application, they are unconfounded with verbosity.

Agreement among applications. In the main article, we measured 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 below). For complete transparency, and because we wanted to treat both random factors equally, we also examined agreement among *applications*, but readers

should be aware that the primary focus of this paper is on the indicators for agreement among reviewers.

By measuring agreement among applications, we answer the question: Are certain reviewers more lenient than others, and does this difference emerge regardless of which application they evaluate? We conducted the same set of analyses as the ones described in the main article, but this time the two random factors traded places. Although not directly related to either of the two main research questions, these analyses nevertheless provide interesting insights. Figure S1 depicts the results of these three analyses for the three outcome variables.

We computed the ICC to estimate the proportion of total variance in the outcome variable that is accounted for by the reviewer random factor. Table S5 provides the values for the ICCs for rating, strengths, and weaknesses. To summarize, the ICC values were small for preliminary ratings, moderate for weaknesses, and substantial for strengths (see Figure S1). This suggests that some reviewers are *slightly* overall harsher on average than others in the ratings they assign to their particular pool of applications, but this difference is not statistically significant in our sample. Some reviewers may write a few more weaknesses on average than other reviewers for their pool of applications, whereas a statistically significant proportion of the variation (59.2%) in the number of strengths listed in a critique can be attributed to the individual reviewer's particular habits. Some reviewers make a greater effort than others to write positive things about the applications they evaluate, and this individual difference accounts for more than half of the variance in the number of strengths listed.

Our second set of analyses to examine agreement among applications was carried out on a data file in which applications were treated like raters (columns) and reviewers were treated like targets (rows). For the preliminary ratings, Krippendorff's alpha was $\alpha = .086$, 95% CI[.007;

.169]. For the number of strengths, α = .601, 95% CI[.564; .636]. For the number of weaknesses, α = .140, 95% CI[.042; .235]. These results show that there is small-to-moderate agreement among applications regarding the level of leniency of each of the reviewers. The fact that some reviewers are more lenient than others is to be expected. What is surprising, however, is the fact that the applications seem to function like interchangeable "raters" (or scale items), suggesting that the characteristics of individual applications hardly play any role in a given reviewer's evaluations, which seem to be driven primarily by his/her level of leniency.

Our third set of analyses to measure agreement among applications involved the same comparison between the similarity of ratings from the same reviewer and the similarity of ratings from different reviewers. Like above, we computed two scores for every reviewer. The first score was the average absolute difference between all ratings from that reviewer. The second score was the average absolute difference between each of the ratings from that reviewer and each of the ratings from all other reviewers. Like before, we subtracted the first score from the second score to compute on overall similarity score per reviewer. Values above zero on this score indicate that a reviewer's ratings are more similar to each other than to ratings from other reviewers. An overall similarity score could only be computed for the 40 reviewers who evaluated two applications (three of the reviewers only evaluated one application as primary reviewer). We tested the 40 overall similarity scores against zero. In total, we performed three one sample *t*-tests: one for ratings, one for strengths, and one for weaknesses. Table S5 displays the results of these three tests. To summarize, the tests yielded non-significant results for rating ($p = .87$) and for weaknesses ($p = .22$), but a significant result for strengths ($p < .001$). This suggests that the number of strengths enumerated in the written critiques are to an important extent determined by reviewer characteristics.

Taken together, these analyses suggest that some reviewers tend to be more lenient than other reviewers in their evaluations and that this difference emerges regardless of the application that they evaluate. Applications function like (nearly interchangeable) scale items that allow us to locate reviewers on the leniency dimension.

Exploring the relationship between strengths and weaknesses. Model 1 in the main article indicated that the number of weaknesses significantly predicts the preliminary ratings, whereas the number of strengths does not. Thus, we wanted to explore whether the number of strengths and the number of weaknesses listed in a critique are inversely related to one another, as we would expect them to be. To examine this question, we estimated an additional model not included in the main article, summarized in Table S7. This model is the same as Model 2 in the main article, except that it includes *strengths* as the outcome variable (rather than the preliminary rating) and the same three weaknesses predictors: the adaptively centered weakness value, the mean-centered reviewer cluster means of the weakness values, and the mean-centered application cluster means of the weakness values.

We found a marginally significant inverse relationship between strengths and weaknesses within reviewers and within applications ($b_{\text{Weaknesses(Within-Within)}} = -0.49$, $p = .07$). This relationship is also marginally significant between-applications-within-reviewers ($b = -0.62$, $p = .06$). When an individual reviewer lists mores weaknesses for application A than for application B, this reviewer also tends to list fewer strengths for application A than for application B. However, this strengths-weaknesses relationship did not hold between-reviewers-within applications ($b = 0.17$, $p = .56$). When reviewer A lists more weaknesses for a particular application than reviewer B, it is not necessarily the case that reviewer A will list fewer strengths

than reviewer B. In other words, the number of weaknesses reviewers identify for a given application won't tell us anything about the number of strengths they will include.

Alternative model specifications. Experts disagree about how to determine the appropriate random effects structure in linear mixed-effects models. Most experts agree that one should start out by attempting to estimate the model with the maximal random effects structure called for by the design, and then, if this model fails to converge, progressively set random effects to zero until convergence is achieved (35). There is disagreement among experts, however, about what comes next. Some experts (60) propose to interpret the model for which convergence has been achieved, i.e., to keep the random effects structure as maximal as possible. Other experts (37) suggest to further simplify the LMEM after convergence by removing random effects that have a near-zero variance.

In the models reported above, we adopted Barr and colleagues' "keep-it-maximal approach" (60) and interpreted the first model that converged. In the following paragraphs, we will adopt the Bates and colleagues' "model selection approach" (37) and delete random effects that have a zero (or near zero) variance. The results of these analyses are provided in Table S8.

We started out by estimating a model (Model 3) that was identical to Model 1, except that we removed the by-reviewer slopes for strengths and weaknesses, because their variance estimate was zero in Model 1. The results remained the same: The partial effect of weaknesses was statistically significant ($b = 0.08$, $p < .02$), whereas the partial effect of strengths was not ($b = -0.01$, $p = .48$).

We also estimated a model (Model 4) that was identical to the previous model (Model 3), but we additionally removed the by-application random slopes that had extremely small variance estimates in the previous models ($s^2 = .002$ and $s^2 = .012$). The results were identical.

We also applied this model trimming approach to Model 2, which was reported in the main article. We reestimated this model, but without the by-reviewer random slope, the by-application random intercept, and the by-application random slope (Model 5). The parameter estimates and the *p*-values were identical to those reported in Table 1 in the main article.

Alternative ICC models. The ICC models reported in the Methods section above included one random factor at a time, but a viable alternative data-analytic strategy would have been to compute the ICCs by estimating models that contain both random factors together. In order to show that our choice of the data-analytic strategy had no influence on the results, we re-estimated the fixed-intercept-random-intercept models with both random factors included (i.e., we regressed the outcome variable on the fixed intercept, the by-reviewer random intercept, and the by-application random intercept). As Table S9 shows, the results from these three models are virtually identical to the ICCs we estimated one random factor at a time. There is no clustering by application— suggesting that there is no agreement among reviewers regarding the relative qualities of the applications—and there is some amount of clustering by reviewer, especially for strengths—suggesting that reviewers differ in leniency.

Readers may be curious as to whether a value of 0 or close to 0 for the ICC is smaller than what one would expect due to random chance alone. This is not the case, however. The ICC is defined as the between-cluster variance divided by the sum of the between-cluster variance and the within-cluster variance: ICC = $\tau^2/(\tau^2+\sigma^2)$. The between-cluster variance is known as the "added variance component" (sometimes written as s_A^2) and is estimated with the following equation: $\tau^2 = (MS_B–MS_W)/n_o$, where MS_B and MS_W are the mean square between clusters and mean square within clusters, and n_o is a measure of sample size. This equation shows that the between-cluster variance (τ^2) is negative when $MS_B < MS_W$. Most statistics programs (including

R and the lme4 package we used) do not allow for negative variances and simply change negative values into 0. In the case of repeated sampling with random data (i.e., from a population in which there is no clustering), MS_B will often be smaller than MS_W and, as a result, the reported value for τ^2 will be zero. Even if $MS_B > MS_W$, the difference between the two will often be quite small for random data, and when this difference is then divided by n_o, the resulting between-cluster variance is so small that statistics programs will return a value of 0 (61, p. 10). With random data, a reasonable number of clusters, and a reasonable sample size, the expected value of ICC is thus zero or very close to zero. Given certain characteristics, the expected value may be slightly larger than zero (due entirely to the small number of samples in which clustering occurred by chance), but the expected modal value for ICCs in data drawn from an unclustered population is zero. Thus, although our ICC estimates for agreement among reviewers are small (regardless of the model specification used), they are not smaller than what would be expected by random chance; instead, they are in line with what one would expect to see with random data. In other words, our reviewers' evaluations of the same application are as similar as their evaluations of different applications.

Re-estimating all models with the outlier. Upon inspection of the data, we realized that one observation's weakness value qualified as an outlier. The observation's value was 83 (i.e., Reviewer #15 listed 83 weaknesses for Application #3). The descriptive statistics of the remaining 82 weakness values were as follows: *M* = 15.57, *Median* = 14.00, *SD* = 9.58, *Min* = 0, *Max* = 41. Following Leys and colleagues (62), we determined the median absolute deviation (MAD) of the weakness values, which turned out to be MAD = 8.90. Observations are considered outliers if they are more than 3 MADs away from the median. The weakness value of 83 was 7.75 MADs above the median and thus clearly qualifies as an outlier.

The analyses reported in the main article and in the supplementary analyses above do not contain the outlier. Below, we repeat all of the analyses from the main article and in the S.I. with the outlier included in order to demonstrate for readers that inclusion or exclusion of the outlier had no bearing on the conclusions reached in this paper. Note that because this is an outlier on the weakness variable, any analysis that only focused on preliminary rating or on strengths is unaffected by the outlier, and thus will not be repeated here.

Agreement among reviewers and among applications. Table S10 provides the estimates for the intraclass correlation (ICC), Krippendorff's alpha, and the similarity score for the weakness variable with the outlier included. Although the point estimates for each of the six statistics increase with the inclusion of the outlier, the substantive conclusions do not. The statistical significance patterns are identical in terms of no agreement among reviewers. The patterns change slightly in terms of agreement among applications, as the ICC ($p = .03$) and the similarity score for agreement among applications ($p = .07$) became statistically significant at $\alpha = .05$ and marginally significant at $\alpha = .10$, respectively. However, this aligns with the overall pattern our results establish: there is some clustering due to reviewer, meaning that some reviewers are more lenient than others in their evaluations. The inclusion of the outlier merely strengthens the degree of clustering by reviewer in our sample.

Relationship between ratings and critiques. The results of the two models reported in the main article (Model 1 and Model 2) are provided in Table S11 and Table S12, respectively, both without and with the outlier, to allow for direct comparison of the estimates. The significance patterns and substantive conclusions are identical in both models with the outlier included or excluded.

Agreement for funded versus unfunded applications. Table S13 provides the estimates for the LMEMs with three separate outcome variables—(i) preliminary rating, (ii) number of strengths, and (iii) number of weaknesses—and a dichotomous funding status predictor (coded -0.5 for unfunded and 0.5 for funded), this time with the outlier included. The point estimates change slightly, but all statistical significance patterns are identical.

Relationship between strengths and weaknesses. Table S7 includes the model with and without the outlier that regresses number of strengths on the three weakness predictors: the adaptively centered weakness value, the mean-centered reviewer cluster means of the weakness values, and the mean-centered application cluster means of the weakness values. The within-within effect of weaknesses on strengths becomes significant with the outlier included ($p = .004$) where it was previously marginally significant ($p = .07$), but this does not change our substantive conclusions, given that this is only the effect within-applications and within-reviewers. The effect at the between-applications-within-reviewers level, which was previously marginally significant ($p = .06$), was no longer significant with the outlier included ($p = .15$), suggesting the relationship between strengths and weaknesses does not hold across multiple applications for an individual reviewer. Most importantly for our conclusions, though, the strengths-weaknesses relationship continues to be non-significant with the outlier included between-reviewers-within-applications ($p = .74$): The number of weaknesses reviewers identify for a given application won't tell us anything about the number of strengths they will include.

Alternative model specifications. Table S14 provides the estimates from the three additional models (Model 3, Model 4, and Model 5) with the outlier included. The only effect that changed was the effect of strengths on rating in Model 4, which became significant with the

outlier included (p = .045). However, this does not change our substantive conclusion that weaknesses are a stronger predictor of preliminary rating than strengths.

Alternative ICC models. Finally, we estimated the alternative models for specifying the ICC due to application and to reviewer with the outlier included for the weaknesses variable, since the ICC for rating and for strengths are unaffected by the outlier. This model estimated the variance due to reviewer as s^2 = 49.17, the variance due to application as s^2 = 0.00, and the residual variance as s^2 = 96.64, resulting in $ICC_{reviewer}$ = (49.17) / (49.17 + 0.00 + 96.64) = 33.72%, (p = .03) with the outlier compared to 15.35% (p = .4) without the outlier. Thus, including the outlier increases the degree to which some reviewers appear to be more or less lenient in the number of weaknesses they enumerate in their critiques. However, the $ICC_{application}$ remains the same at 0: There is no agreement among different reviewers as to the number of weaknesses contained in a given application.

Supplementary Discussion

The analyses reported in the main article establish that our reviewers did not agree with one another in terms of the ratings they assigned to an application, nor in terms of the number of strengths or weaknesses they listed in a critique. They also did not agree with the reviewers who evaluated these applications originally, since they did not evaluate the applications that were funded by NIH more positively than those that were not funded. Supplementary analyses summarized above suggested that there is a slight tendency for some reviewers to be more lenient than others, especially with regard to the number of strengths they identify for an application. Taken together, these results show there is no agreement among reviewers about the relative quality of the applications. A given evaluation of a grant application tells us more about the reviewer's level of leniency than about the scientific merit of the application.

Furthermore, although the three indicators of an application's evaluation—preliminary ratings, number of strengths, and number of weaknesses—tend to be related at certain levels of analysis (within-reviewers-within-applications and within-reviewers-between-applications), they are unrelated to each other at the level that is the most important for the peer-review process: At the between-reviewer-within-application level, there is no relationship between ratings, strengths, and weaknesses. It follows that if there is no relationship between the three indicators when different reviewers evaluate the *same* application, then there cannot be, by definition, a relationship between them when different reviewers evaluate *different* applications, as is the case in real NIH peer review.

Considering all of the analyses reported in the main article and in the Supplementary Information, one can draw a number of conclusions: First, there is no agreement among reviewers regarding the relative quality of the grant applications. Second, reviewers evaluate applications that were funded by NIH just as positively as applications that went unfunded. Third, reviewer's evaluation of an application (preliminary rating, number of strengths and weaknesses mentioned) tells us something about particular characteristics of that reviewer— for example, his/her leniency, his/her particular way of using the 9-point rating scale, the amount of effort s/he is putting into writing the critiques, or his/her kindness by trying to identify as many strengths as possible—but does not say anything about the scientific merit of the application. Fourth, applications function like interchangeable scale items that help us distinguish reviewers along these characteristics. If we want to learn something about a reviewer, it does not really matter which applications we assign to him/her. Finally, although reviewers are internally consistent, there is no consistency when different reviewers evaluate the same application. In

other words, there is no agreement between reviewers on how to "translate" a certain number of strengths and weaknesses into a numerical rating on the 9-point rating scale.

Limitations

Our research is not without limitations. First, our data suffer from a restricted range problem, because only grant applications that were eventually funded are included in our study; we cannot say whether these findings would generalize to an entire pool of applications, including those that might never be funded by NIH. Nonetheless, the results do show that for grants above a certain quality threshold, the peer review process is completely random. Given that at NIH, only the top 50% of proposals are discussed in a study section meeting and considered for funding, it follows that our findings are at least applicable for those applications that move on to the discussion phase of peer review at NIH. Thus, our results suggest that the grants receiving funding may not be more deserving than some grants that are denied funding. Future research should aim to examine the degree to which these patterns of inter-reviewer agreement hold for a pool of applications of more diverse quality than ours.

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 our reviewers put in comparable effort in our study than they would have for an actual NIH study section. In a survey administered to participants, 81% of them reported that the pre-meeting process was either very similar or identical to actual NIH study sections they had participated in. In addition, our research team conducted semi-structured debriefing interviews with participants. During these

interviews, some participants were asked specifically about the process of writing their critiques. The reviewer participants claimed that they put in as much time and effort into their evaluations than for real NIH applications. One participant stated to the interviewer, "You can see our critique, we did it very just like a real NIH review, you know." Three additional excerpts from interviews with participants are shown below, with *I* referring to the interviewer and *P* referring to the participant:

Excerpt #1

I: I asked you about your scores. What about your critiques? Was the way you did your critiques any different do you think than um a typical NIH section?

P: No I don't think so. I mean I think they were pretty, they were about the same.

Excerpt #2

I: How do you compare the way that you reviewed and critiqued your assigned grants compared to an NIH study section?

P: Um overall similar I would say.

Excerpt #3

I: What was the process like for you when you were actually reviewing the grants um initially? Was that at all different from an NIH section?

P: Mm not, not really, no.

Although not exhaustive, these excerpts serve to illustrate our participants' beliefs that the critiques they prepared for our constructed study sections were no different from those that they prepare for a real NIH study section.

In addition, the Scientific Review Officer (SRO) who oversaw the entire constructed study section process, and who herself had presided over NIH study sections for more than 15 years as SRO, wrote the following testimonial to the research team:

> My feeling about the quality of the reviews and discussions [during the meeting] is that they were at least as in depth as the typical review by an NIH study section. The reviewers were experienced and, having volunteered to participate, were committed to the peer review process. In nearly all cases, the critiques were carefully prepared.

It is also worth noting that the reviewers knew they would be participating in a study section led by a highly experienced SRO, that they would be surrounded by real colleagues who would be reading their critiques, that they would need to justify and even defend their preliminary ratings to these colleagues, and that their critiques would be scrutinized by the research team. In fact, while taking a break halfway through each meeting, our reviewers' casual interactions illustrated the very real professional context in which these meetings took place; for example, one reviewer approached another during such a break and asked if he would be attending a conference coming up, and the reviewers proceeded to discuss their upcoming presentations at this conference. Thus, our reviewer participants engaged in this task as seriously as when serving on real NIH study sections.

As a final means of attempting to evaluate the ecological validity of our data, we examined the length of the critiques in our study and compared them to a nationally representative sample of critiques that our research team collected for a different study. Although we did not have access to the raw critiques directly from reviewers, we extracted the primary reviewers' critiques from the summary statements that are sent to the Principal Investigator (PI) after the meeting. In that sample of 18,912 summary statements (13,012 or 68.8% of which were initially funded, and 5,900 or 31.2% of which were not initially funded), the primary reviewers'

critiques were on average 525.16 words long (*SD* = 282.44, *Min* = 14, *Max* = 4207, *Median* = 481). When we selected a random sample of 83 summary statements from the larger national sample (50 or 60.2% of which were funded, and 33 or 39.8% of which were unfunded, which nearly matched the 64% vs. 36% proportion of funded vs. unfunded applications in our data set), the primary reviewers' critiques were on average 488.77 words long (*SD* = 280.56, *Min* = 81, *Max* = 1339, *Median* = 435). In our data set, primary reviewers' critiques were on average 662.89 words long (*SD* = 205.01, *Min* = 260, *Max* = 1994, *Median* = 602). Thus, we believe that although the participants in our study knew they were not participating in real NIH study sections, the quality and depth of the critiques were comparable to those from real NIH study sections.

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 does not alter the ability to detect a small effect—as the effect is zero. Furthermore, since all of our most relevant effects were zero or close to zero, even if we adopted a much higher Type I error rate (e.g., $\alpha = .20$), these effects would still be non-significant. 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.

Supplementary References

38. Jayasinghe UW, Marsh HW, Bond N (2001) Peer review in the funding of research in higher education: the Australian experience. *Educ Eval Policy Anal* 23(4):343-364.
39. Johnson VE (2008) Statistical analysis of the National Institutes of Health peer review system. *Proc Natl Acad Sci USA* 105(32):11076-11080.
40. Kaplan D, Lacetera N, Kaplan C (2008) Sample size and precision in NIH peer review. *PLoS One* 3(7):e2761.
41. National Institutes of Health (2008) 2007-2008 Peer Review Self Study Final Draft. Available at https://enhancing-peer-review.nih.gov/meetings/NIHPeerReviewReportFINALDRAFT.pdf (2008).
42. Helitzer DL, et al. (2011) A randomized controlled trial of communication training with primary care providers to improve patient-centeredness and health risk communication. *Patient Educ Couns* 82:21-29.
43. Dotger BH, Dotger SC, Maher MJ (2010) From medicine to teaching: the evolution of the simulated interaction model. *Innov High Educ* 35(3):129-141.
44. Young OM, Parviainen K (2014) Training obstetrics and gynecology residents to be effective communicators in the era of the 80-hour workweek: a pilot study. *BMC Res Notes* 7:455-460.
45. Foss RD (1976) Group decision processes in the simulated trial jury. *Sociometry* 39(4):305-316.
46. Kerr NL, Nerenz DR, Herrick D (1979) Role playing and the study of jury behavior. *Socio Meth Res* 7(3):337-355.
47. McQuiston-Surrett D, Saks MJ (2009) The testimony of forensic identification science: what expert witnesses say and what factfinders hear. *Law Hum Behav* 33:436-453.

48. Bornstein BH (1999) The ecological validity of jury simulations: is the jury still out? *Law Hum Behav* 23(1)*:*75-91.

49. Strauss A, Corbin J (1990) *Basics of qualitative research: grounded theory procedures and techniques* (Sage, Newbury Park, CA).

50. Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics. *Comput Ling* 34(4):555-596.

51. Craggs R, McGee Wood M (2005) Evaluating discourse and dialogue coding schemes. *Comput Ling* 31(3):289-295.

52. Hallgren KA (2012) Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol* 8(1):23-34.

53. McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochem Med* 22(3):276-282.

54. Syed M, Nelson SC (2015) Guidelines for establishing reliability when coding narrative data. *Emerg Adult* 3(6):375-387.

55. Hruschka DJ, Schwartz D, St. John DC, Picone-Decaro, E, Jenkins RA, Carey JW (2004) Reliability in coding open-ended data: lessons learned from HIV behavioral research. *Field Methods* 16(3):307-331.

56. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1), 159-174.

57. Cichetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 6(4):284-290.

58. Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychol Bull* 76(5):378-382.

59. Bakeman R, Gottman JM (1986) *Observing behavior: an introduction to sequential analysis.* Cambridge University, Cambridge, England).

60. Barr DJ, Levy R, Scheepers C, Tilly HJ (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. *J Mem Lang* 68(3):255-278.

61. Bates DM (2010) *lme4: mixed-effects modeling with R* (Springer: New York).

62. Leys C, Ley C, Klein O, Bernard P, Licata, L (2013) Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* 49(4):764-766.

63. Kenward MG, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53(3):983-997.

R Code for All Models

```
library(lme4)
library(lmSupport)
library(lmerTest)

data = read.csv("data.csv")
hist(data$StrengthsALL)
hist(data$WeaknessesALL)

# Eliminate outlier
data2 <- data
data2$WeaknessesALL <- ifelse(data2$WeaknessesALL==83, NA,
   data2$WeaknessesALL)
hist(data2$StrengthsALL)
hist(data2$WeaknessesALL)
# NOTE: All analyses below were run without the outlier. We re-ran the same
   analyses keeping the outlier, with all syntax identical except that "data
   = data" instead of "data = data2".

########################## RQ #1: AGREEMENT ##########################

#:::::::::::::::::::::::: BY APPLICATION ::::::::::::::::::::::::::::
# RQ 1.1 - Rating (By Application)

# RQ1.1A - ICC(rating) by Application
summary(ICCmodel_app <- lmer(Rating ~ 1 + (1|ApplicationID),data=data2))
rand(ICCmodel_app) #to extract p-value for ICC
confint(ICCmodel_app) #to obtain 95% CI

# RQ1.1B - Krippendorff's Alpha for Ratings by Application
primaryratings <- read.csv("PrimaryRatings.csv")
ratingsmatrix <- (as.matrix(primaryratings))
kripp.alpha(ratingsmatrix, method = "ordinal")
      # Bootstrap conducted in SPSS using syntax from Hayes &
      Krippendorff (2013)

# RQ1.1C – Similarity Score T-tests for Ratings by Application
# See SPSS syntax below


# RQ1.2 - Strengths (By Application)

# RQ1.2A - ICC(Strengths) by Application
summary(ICCStrengths_App <- lmer(StrengthsALL ~ 1 + (1|ApplicationID),
   data=data2))
rand(ICCStrengths_App) #to extract p-value for ICC
confint(ICCStrengths_App) #to obtain 95% CI


# RQ1.2B - Krippendorff's Alpha for Strengths by Application
primarystrengths <- read.csv("PrimaryStrengths.csv")
strengthsmatrix <- (as.matrix(primarystrengths))
kripp.alpha(strengthsmatrix, method = "interval")
      # Bootstrap conducted in SPSS using syntax from Hayes &
      Krippendorff (2013)
```

```
# RQ1.2C – Similarity Score T-tests for Strengths by Application
# See SPSS syntax below

# RQ1.3 - Weaknesses (By Application)
# RQ1.3A - ICC(Weaknesses) by Application
summary(ICCWeak_app <- lmer(WeaknessesALL ~ 1 + (1|ApplicationID),
   data=data2)) # without outlier included
rand(ICCWeak_app) #to extract p-value for ICC
confint(ICCWeak_app) #to obtain 95% CI


# RQ1.3B - Krippendorff's Alpha for Weaknesses by Application
primaryweaknesses <- read.csv("PrimaryWeaknesses.csv") #no outlier
weaknessesmatrix <- as.matrix(primaryweaknesses)
kripp.alpha(weaknessesmatrix, method = "interval")
      # Bootstrap conducted in SPSS using syntax from Hayes &
      Krippendorff (2013)

# RQ1.3C – Similarity Score T-tests for Weaknesses by Application
# See SPSS syntax below


#:::::::::::::::::::::::::: BY REVIEWER ::::::::::::::::::::::::::::

# RQ1.1 - Rating (By Reviewer)

# RQ1.1A - ICC(rating) by Reviewer
summary(ICCmodel_rev <- lmer(Rating ~ 1 + (1|ReviewerID), data=data2))
rand(ICCmodel_rev)
confint(ICCmodel_rev)


# RQ1.1B - Krippendorff's Alpha for Rating by Reviewer
primaryratings <- read.csv("PrimaryRatings.csv")
ratingsmatrix <- (as.matrix(primaryratings))
reviewerratingsmatrix <- t(ratingsmatrix) #transpose the matrix
kripp.alpha(reviewerratingsmatrix, method = "ordinal")

# RQ1.1C - Similarity Score T-tests for Score by Reviewer
# See SPSS syntax below


# RQ1.2 - Strengths (By Reviewer)

# RQ1.2A - ICC(strengths) by Reviewer
summary(ICCStrengths_rev <- lmer(StrengthsALL ~ 1 + (1|ReviewerID),
   data=data2))
rand(ICCStrengths_rev)
confint(ICCStrengths_rev)


# RQ1.2B - Krippendorff's Alpha for Strengths by Reviewer
primarystrengths <- read.csv("PrimaryStrengths.csv")
strengthsmatrix <- (as.matrix(primarystrengths))
reviewerstrengthsmatrix <- t(strengthsmatrix) #transpose matrix
kripp.alpha(reviewerstrengthsmatrix, method = "interval")
```

```
# RQ1.1C - Similarity Score T-tests for Strengths by Reviewer
# See SPSS syntax below


# RQ1.3 - Weaknesses (By Reviewer)

# RQ1.3A - ICC(weaknesses) by Reviewer
summary(ICCWeak_revX <- lmer(WeaknessesALL ~ 1 + (1|ReviewerID), data=data2))
   #without outlier included
rand(ICCWeak_revX)
confint(ICCWeak_revX)


# RQ1.3B - Krippendorff's Alpha for Weaknesses by Reviewer
primaryweaknesses <- read.csv("PrimaryWeaknesses.csv")
weaknessesmatrix <- as.matrix(primaryweaknesses)
reviewerweaknessesmatrix <- t(weaknessesmatrix)#transpose matrix
kripp.alpha(reviewerweaknessesmatrix, method = "interval")

# RQ1.1C - Similarity Score T-tests for Weaknesses by Reviewer
# See SPSS syntax below

############## RQ #2: RELATIONSHIP B/T RATINGS + CRITIQUES #############

# Adaptively centered predictor
data2$StrengthsALLc1M <- ave(data2$StrengthsALL,data2$ReviewerID,
   FUN=function(x)mean(x,na.rm=T))
data2$StrengthsALLc2M <- ave(data2$StrengthsALL,data2$ApplicationID,
   FUN=function(x)mean(x,na.rm=T))
data2$WeaknessesALLc1M <- ave(data2$WeaknessesALL,data2$ReviewerID,
   FUN=function(x)mean(x,na.rm=T))
data2$WeaknessesALLc2M <- ave(data2$WeaknessesALL,data2$ApplicationID,
   FUN=function(x)mean(x,na.rm=T))
data2$StrengthsALLcM <- data2$StrengthsALL - data2$StrengthsALLc1M -
   data2$StrengthsALLc2M + mean(data2$StrengthsALL, na.rm=T)
data2$WeaknessesALLcM <- data2$WeaknessesALL - data2$WeaknessesALLc1M -
   data2$WeaknessesALLc2M + mean(data2$WeaknessesALL, na.rm=T)
data2$StrengthsALLc1M_CCM <- data2$StrengthsALLc1M -
   mean(data2$StrengthsALLc1M, FUN=function(x)mean(x,na.rm=T))
data2$StrengthsALLc2M_CCM <- data2$StrengthsALLc2M -
   mean(data2$StrengthsALLc2M, FUN=function(x)mean(x,na.rm=T))
data2$WeaknessesALLc1M_CCM <- data2$WeaknessesALLc1M -
   mean(data2$WeaknessesALLc1M, FUN=function(x)mean(x,na.rm=T))
data2$WeaknessesALLc2M_CCM <- data2$WeaknessesALLc2M -
   mean(data2$WeaknessesALLc2M, FUN=function(x)mean(x,na.rm=T))

Model1 <- lmer(Rating ~ StrengthsALLcM + WeaknessesALLcM +
                    (1 + StrengthsALLcM + WeaknessesALLcM||ReviewerID)+
                       (1 + StrengthsALLcM +
                              WeaknessesALLcM||ApplicationID),
                  data = data2)
summary(Model1)
Anova(Model1, type = 3, test = "F") #to get p-values for fixed effects

Model2 <- lmer(Rating ~ WeaknessesALLcM + WeaknessesALLc1M_CCM +
                     WeaknessesALLc2M_CCM +
```

```
                    (1 + WeaknessesALLcM||ReviewerID) +
                    (1 + WeaknessesALLcM||ApplicationID),
                  data = data2)
summary(Model2)
Anova(Model2, type = 3, test = "F")



######################## SUPPLEMENTARY ANALYSES ########################

# MODELS BASED ON FUNDING STATUS

data3 = read.csv("data_withfunding.csv")
data3$FundedC <- data3$Funded - .5

data4 <- data3
data4$WeaknessesALL <- ifelse(data4$WeaknessesALL==83, NA,
   data4$WeaknessesALL)

ModelF1 <- lmer(Rating ~ FundedC + (1 + FundedC||ReviewerID) + (1 +
   FundedC||ApplicationID), data = data3)
summary(ModelF1)
Anova(ModelF1, type = 3, test = "F")

ModelF2 <- lmer(StrengthsALL ~ FundedC + (1 + FundedC||ReviewerID) + (1 +
   FundedC||ApplicationID), data = data3)
summary(ModelF2)
Anova(ModelF2, type = 3, test = "F")

ModelF3 <- lmer(WeaknessesALL ~ FundedC + (1 + FundedC||ReviewerID) + (1 +
   FundedC||ApplicationID), data = data3)
summary(ModelF3)
Anova(ModelF3, type = 3, test = "F")


# RELATIONSHIP BETWEEN STRENGTHS AND WEAKNESSES

ModelSW <- lmer(StrengthsALL ~ WeaknessesALLcM + WeaknessesALLc1M_CCM +
   WeaknessesALLc2M_CCM + (1|ReviewerID) +
   (0 + WeaknessesALLcM|ReviewerID) + (1|ApplicationID) +
   (0 + WeaknessesALLcM|ApplicationID), data = data2)
summary(ModelSW)
Anova(ModelSW, type = 3, test = "F")


# MODEL 3 - same as Model 1 but without variance components that were 0
Model3 <- lmer(Rating ~ StrengthsALLcM + WeaknessesALLcM +
                  (1|ReviewerID) + (1|ApplicationID) +
                  (0 + StrengthsALLcM|ApplicationID) +
                  (0 + WeaknessesALLcM|ApplicationID), data = data2)
summary(Model3)
Anova(Model3, type = 3, test = "F")


# MODEL 4 - same as Model 1 but without variance components that were close
   to zero or zero
Model4 <- lmer(Rating ~ StrengthsALLcM + WeaknessesALLcM +
```

```
                  (1|ReviewerID) + (1|ApplicationID), data = data2)
summary(Model4)
Anova(Model4, type = 3, test = "F")


# MODEL 5 - same as Model 2 but without variance components that were zero
Model5 <- lmer(Rating ~ WeaknessesALLcM + WeaknessesALLc1M_CCM +
   WeaknessesALLc2M_CCM + (1|ReviewerID), data = data2)
summary(ModelD5)
Anova(ModelD5, type = 3, test = "F")


# ALTERNATIVE ICC MODELS

ICCmodel <- lmer(Rating ~ 1 + (1|ReviewerID) + (1|ApplicationID), data=data2)
summary(ICCmodel)
rand(ICCmodel)
confint(ICCmodel)

ICCmodel2 <- lmer(StrengthsALL ~ 1 + (1|ReviewerID) + (1|ApplicationID),
   data=data2)
summary(ICCmodel2)
rand(ICCmodel2)
confint(ICCmodel2)


ICCmodel3 <- lmer(WeaknessesALL ~ 1 + (1|ReviewerID) + (1|ApplicationID),
   data=data2)
summary(ICCmodel3)
rand(ICCmodel3)
confint(ICCmodel3)



############################ SPSS SYNTAX ############################

# COMPUTING SIMILARITY SCORES BY APPLICATION

COMPUTE same1 = abs(v1-v2) .
COMPUTE same2 = abs(v1-v3) .
COMPUTE same3 = abs(v1-v4) .
COMPUTE same4 = abs(v2-v3) .
COMPUTE same5 = abs(v2-v4) .
COMPUTE same6 = abs(v3-v4) .
COMPUTE distsame = MEAN (same1 to same6).
VECTOR old = v5 to v85 .
VECTOR newa(81) .
VECTOR newb(81) .
VECTOR newc(81) .
VECTOR newd(81) .
LOOP #I=1 to 81.
+  COMPUTE newa(#I)=abs(v1-(old(#I))).
+  COMPUTE newb(#I)=abs(v2-(old(#I))).
+  COMPUTE newc(#I)=abs(v3-(old(#I))).
+  COMPUTE newd(#I)=abs(v4-(old(#I))).
END LOOP .
EXECUTE .
```

```
COMPUTE distother = MEAN (newa1 to newd81).

# COMPUTING SIMILARITY SCORES BY REVIEWER

COMPUTE same1 = abs(v1-v2) .
COMPUTE distsame = MEAN (same1).
VECTOR old = v3 to v84 .
VECTOR newa(82) .
VECTOR newb(82) .
LOOP #I=1 to 82.
+  COMPUTE newa(#I)=abs(v1-(old(#I))).
+  COMPUTE newb(#I)=abs(v2-(old(#I))).
END LOOP .
EXECUTE .
COMPUTE distother = MEAN (newa1 to newb82).
```

Table S1. Demographic information for participant reviewers.

	CSS1	CSS2	CSS3	CSS4	Total
Number of reviewers	10	12	12	8	42
Gender					
Female	2 (20%)	3 (25%)	3 (25%)	2 (25%)	10 (23.8%)
Male	8 (80%)	9(75%)	9 (75%)	6 (75%)	32 (76.2%)
Race/Ethnicity					
Asian	6 (60%)	7 (58%)	8 (67%)	5 (63%)	26 (61.9%)
Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Hispanic	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
White	4 (40%)	5 (42%)	4 (33%)	3 (37%)	16 (38.1%)
Tenure Status					
Full Professor	5 (50%)	7 (58%)	7 (58%)	3 (38%)	22 (52.4%)
Associate Professor	4 (40%)	3 (25%)	3 (25%)	3 (38%)	13 (31.0%)
Assistant Professor	1 (10%)	2 (17%)	2 (17%)	2 (17%	7 (16.7%)

Table S2. List and description of axial codes derived in qualitative data analysis.

Code	Description
Neither	*Neither a strength nor a weakness of the proposal itself*
Budget	Comment on the application's budget
Cannot be coded	Uninterpretable or related to participating in the study
Description of Aims	Factual summary of the Specific Aims of the application
Description of PI	Factual statement of the name or employer of the PI
Name of Environment	Factual statement of the name of the research environment
Scientific Background	Factual information used to provide scientific background
Strength	*Strengths of the proposal (e.g., excellent, strong)*
Application	Evaluation of application itself (e.g., well written)
Approach	Evaluation of the methodological approach
Enthusiasm High	Comment on the high level of enthusiasm for the proposal
Environment	Evaluation of the research environment
Innovation	Evaluation of the innovation of the proposal
No Major Weaknesses	Statement that there are "no" or "no major" weaknesses
PI	Evaluation of the PI/Co-PI (e.g., expertise, experience)
Preliminary Data	Evaluation of the preliminary data included
Significance	Evaluation of the significance of the proposal's focus
Strengths > Weaknesses	Comment that the strengths outweigh the weaknesses
Team	Evaluation of the team (collaborators, consultants)
Minor Strength	*Moderated strengths of the proposal (e.g., adequate, sufficient)*
Application	Evaluation of application itself
Approach	Evaluation of the methodological approach
Environment	Evaluation of the research environment
Innovation	Evaluation of the innovation of the proposal
PI	Evaluation of the PI/Co-PI (e.g., expertise, experience)
Preliminary Data	Evaluation of the preliminary data included
Significance	Evaluation of the significance of the proposal's focus
Team	Evaluation of the team (collaborators, consultants)
Weakness	*Weaknesses of the proposal (e.g., lacking, insufficient)*
Advice	Advice given to the PI for improvement
Application	Evaluation of application itself
Approach	Evaluation of the methodological approach
Enthusiasm Reduced	Comment on the reduced/mitigated enthusiasm for proposal
Environment	Evaluation of the research environment
Innovation	Evaluation of the innovation of the proposal
Minor Weakness	Qualification that the weakness is "only a minor weakness"
PI	Evaluation of the PI/Co-PI (e.g., expertise, experience)
Preliminary Data	Evaluation of the preliminary data included
Question	Question posed to the PI
Significance	Evaluation of the significance of the proposal's focus
Team	Evaluation of the team (collaborators, consultants)
Weaknesses > Strengths	Comment that the weaknesses outweigh the strengths
Major Weakness	*Major weakness that is explicitly stated (e.g., critical issue, highly problematic)*
Approach	Evaluation of the methodological approach
Enthusiasm Low	Comment on the low enthusiasm for the proposal
Innovation	Evaluation of the innovation of the proposal
No Strengths	Statement that there are "no" strengths
Preliminary Data	Evaluation of the preliminary data included
Significance	Evaluation of the significance of the proposal's focus
Team	Evaluation of the team (collaborators, consultants)

Table S3. Examples from corpus of each axial code.

Code	Example from Corpus
Neither	
Budget	*In budget PI's salary exceeds the NIH cap*
Cannot be coded	*But I guess this grant was submitted much earlier*
Description of Aims	*To test the central hypothesis three Specific Aims are proposed*
Description of PI	*This is a MPI application led by Dr. Susan Albert**
Name of Environment	*The work will be conducted at the University of X**
Scientific Background	*Blockade of CD47 on tumor cells leads to phagocytosis by macrophages*
Strength	
Application	*This is an outstanding-to-exceptional application*
Approach	*The experiments proposed are well-designed*
Enthusiasm High	*The overall project was reviewed with high enthusiasm*
Environment	*The environment at University Y* is superb*
Innovation	*The novel mechanisms of AKT regulation that are proposed are highly innovative*
No Major Weaknesses	*No major weaknesses noted*
PI	*He has over 14 years of experience in brain tumor research*
Preliminary Data	*The hypothesis is supported by strong preliminary data*
Significance	*This application promises to have high translational impact on melanoma treatment*
Strengths > Weaknesses	*Overall, this grant has several merits that outweigh its weaknesses*
Team	*An outstanding supporting team of researchers*
Minor Strength	
Application	*Statistics are presented for all three aims*
Approach	*Aim 3 is technically interesting*
Environment	*The environment is conducive to perform the proposed studies*
Innovation	*It has some innovation in the grant*
PI	*PI has experience in DNA technology*
Preliminary Data	*The proposal has preliminary data suggesting a tumor suppressor function of c-Abl*
Significance	*If successful, application may offer new insights into treating BRAF-driven melanomas*
Team	*The investigative team is appropriately skilled*
Weakness	
Advice	*Therefore, in my opinion, it will be advisable to probe one clinical trial for the proposal*
Application	*In Specific Aim II, the expected outcomes and pitfalls are not well-described*
Approach	*The main weakness lies in the high concentration of the inhibitor to be used in order to achieve the expected outcome*
Enthusiasm Reduced	*These limitations have resulted in dampened enthusiasm*
Environment	*No equipment for UV exposure is mentioned in Facilities or Equipment sections*
Innovation	*The use of paclitaxel or doxorubicin is not novel*
Minor Weakness	*There are some minor weaknesses in research design*
PI	*Principle investigator don't have strong background in area of pancreatic cancer*
Preliminary Data	*Overall, the preliminary data are rather limited*
Question	*Can the strategy to prepare high affinity CD47 agonists be extended to solid tumors?*
Significance	*It's not clear how the aim will lead to further advances in the field*
Team	*This team is lacking experience on PTEN prostate cancer mouse models*
Weaknesses > Strengths	*Overall the weaknesses just seem to out weigh the strengths*
Major Weakness	
Approach	*The lack of consideration of TERT activation is a major weakness of the project*
Enthusiasm Low	*Overall enthusiasm for this proposal is limited*
Innovation	*However, lack of novelty remains a major concern*
No Strengths	*None [listed in the "Strengths" section]*
Preliminary Data	*It is very troublesome that there is no direct preliminary data to support the applicants' expectation that such mechanistically relevant changes might occur*
Significance	*Significance of the knowledge gained from this group is questionable*
Team	*Heavy reliance on multiple outside investigators for specialized techniques (MRI) is concerning*

**Note: These are pseudonyms.*

Table S4. Frequency of axial codes in our data.

Code	Sum	*M*	SD	Correlation with rating*
Neither	**533**	**6.42**	**4.10**	**–.09**
Budget	3	0.04	0.19	.06
Cannot be coded	5	0.06	0.24	.15
Description of Aims	330	3.98	3.40	–.20
Description of PI	47	0.57	0.89	.05
Name of Environment	7	0.08	0.28	.09
Scientific Background	141	1.70	1.84	.11
Strength	**2126**	**25.61**	**13.45**	**–.45**
Application	153	1.84	2.28	–.33
Approach	291	3.51	2.34	–.42
Enthusiasm High	3	0.04	0.19	–.12
Environment	115	1.39	1.10	–.15
Innovation	254	3.06	2.34	–.41
No Major Weaknesses	107	1.29	1.31	–.27
PI	385	4.64	3.88	–.17
Preliminary Data	160	1.93	2.37	–.33
Significance	434	5.23	3.67	–.29
Strengths > Weaknesses	3	0.04	0.19	.01
Team	221	2.66	2.73	–.14
Minor Strength	**379**	**4.57**	**3.98**	**.17**
Application	23	0.28	0.89	–.05
Approach	37	0.45	0.80	.13
Environment	34	0.41	0.73	.22
Innovation	44	0.53	0.93	.22
PI	24	0.29	0.65	.11
Preliminary Data	52	0.63	1.38	.16
Significance	115	1.39	1.64	–.02
Team	50	0.60	1.22	.02
Weakness	**1310**	**15.78**	**11.40**	**0.54**
Advice	91	1.10	1.39	–.01
Application	221	2.66	3.00	.37
Approach	469	5.65	4.37	.39
Enthusiasm Reduced	18	0.04	0.19	.09
Environment	8	0.18	0.42	.11
Innovation	142	0.10	0.43	.32
Minor Weakness	29	1.71	2.21	–.22
PI	37	0.35	0.61	.25
Preliminary Data	90	0.45	1.24	.41
Question	64	1.08	2.45	.30
Significance	112	0.77	1.59	.40
Team	25	1.35	1.80	.16
Weaknesses > Strengths	4	0.30	0.62	.22
Major Weakness	**50**	**0.05**	**0.22**	**.29**
Approach	25	0.30	0.81	.26
Enthusiasm Low	4	0.05	0.22	.30
Innovation	1	0.01	0.11	.11
No Strengths	1	0.01	0.11	.11
Preliminary Data	6	0.07	0.30	.15
Significance	7	0.08	0.32	.13
Team	6	0.07	0.41	.03

**Note.* Rating consists of a *reverse* nine-point scale, so a negative correlation suggests it is associated with a lower (i.e., better) rating, whereas a positive correlation suggests it is associated with a higher (i.e., worse) rating.

Table S5. Estimates of agreement among reviewers and of agreement among applications

		Agreement Among Reviewers	Agreement Among Applications
Rating	ICC	ICC = 0.00% (p = 1.0)	ICC = 3.87% (p = .80)
	Krippendorff's α	α = .024 [-.047, .093]	α = .086 [.007, .169]
	Similarity *t*-test	M = 0.01, SD = 0.75, $t(24)$ = 0.07, p = .95	M = -0.02, SD = 0.97, $t(39)$ = -0.16, p = .87
Strengths	ICC	ICC = 0.00% (p = 1.0)	ICC = 59.22% (p < .001)
	Krippendorff's α	α = -.011, [-.094, .079]	α = .601 [.564, .636]
	Similarity *t*-test	M = -0.50, SD = 7.17, $t(24)$ = -0.35, p = .73	M = 4.36, SD = 6.37, $t(39)$ = 4.33, p < .001
Weaknesses	ICC	ICC = 1.74% (p = .90)	ICC = 15.35% (p = .40)
	Krippendorff's α	α = .004 [-.063, .072]	α = .140 [.042, .235]
	Similarity test	M = 0.27, SD = 4.63, $t(24)$ = 0.29, p = .77	M = 1.42, SD = 7.02, $t(38)$ = 1.26, p = .22

Note. ICCs were estimated via a LMEM with an overall fixed intercept and a random intercept for the cluster variable of interest (application or reviewer). An alternative approach to estimating the ICC can be found in Table S9. *P*-values were estimated via a χ^2 likelihood ratio test on the random intercept from the LMEM. Values of Krippendorff's α above .67 are considered suitable for tentative conclusions about reliability, and above .8 are considered reliable. 95% confidence intervals were estimated using 1000 bootstrapped samples (23). The similarity *t*-tests are one-sample t-tests conducted on the similarity scores; values above zero indicate that an application's ratings are more similar to each other than to ratings referring to other applications (for agreement among reviewers), or that a reviewer's ratings are more similar to each other than to ratings from other reviewers (for agreement among applications.

Table S6. Parameter estimates from estimating the effect of funding status on the rating, number of strengths, and number of weaknesses

	Outcome: Rating	**Outcome: Strengths**	**Outcome: Weaknesses**
Fixed Effects	*b (SE)*^{Sig}	*b (SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	3.48 (.21)***	30.72 (2.25)***	14.90 (1.46)***
Funded	-0.20 (.34)	1.50 (2.76)	-3.73 (3.14)
Random Effects	*Var*	*Var*	*Var*
By-reviewer			
Intercept	0.10	115.77	28.14
$Reviewer_{Funded}$	0.09	0.00	144.83
By-application			
Intercept	0.00	6.41	0.00
$Application_{Funded}$	2.06	0.00	0.00
Residual	2.06	71.79	62.25

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. *P*-values for fixed effects are computed via an approximate F-test with the Kenward-Roger method to compute the degrees of freedom (63).

Table S7. Parameter estimates from a model regressing strengths on weaknesses (without and with the outlier)

	No Outlier	**With Outlier**
Fixed Effects	*b (SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	30.09 (1.93)***	30.09 (1.96)***
$\text{Weaknesses}_{(\text{Within-Within})}$	–.49 (.17)º	–.41 (.13)**
$\text{Weaknesses}_{(\text{App\_Cluster\_Means})}$	–.62 (.26)º	–.32(.22)
$\text{Weaknesses}_{(\text{Rev Cluster Means})}$	.17 (.27)	.07 (.21)
Random Effects	*Var*	*Var*
By-reviewer		
Intercept	123.83	125.82
$\text{Weaknesses}_{(\text{Within-Within})}$	0.05	0.00
By-application		
Intercept	1.22	4.40
$\text{Weaknesses}_{(\text{Within-Within})}$	0.00	0.00
Residual	60.52	60.19

Notes. º $p < .10$ * $p < .05$. ** $p < .01$. *** $p < .001$. *P*-values for fixed effects are computed via an approximate F-test with the Kenward-Roger method to compute the degrees of freedom (63).

Table S8. Parameter estimates from Model 3, Model 4, and Model 5

	Model 3	**Model 4**	**Model 5**
Fixed Effects	*b (SE)*^{Sig}	*b(SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	3.46 (.21)***	3.49 (.26)***	3.52 (.15)***
$\text{Strengths}_{(\text{Within-Within})}$	-0.01 (.12)	-0.03 (.02)	
$\text{Weaknesses}_{(\text{Within-Within})}$	0.08 (.03)*	0.11 (.02)***	0.13 (.02)***
$\text{Weaknesses}_{(\text{App\_Cluster\_Means})}$			0.17 (.03)***
$\text{Weaknesses}_{(\text{Rev\_Cluster\_Means})}$			0.03 (.02)
Random Effects	*Var*	*Var*	*Var*
By-reviewer			
Intercept	.967	.773	.625
$\text{Strengths}_{(\text{Within-Within})}$			
$\text{Weaknesses}_{(\text{Within-Within})}$			
By-application			
Intercept	.164	.969	
$\text{Strengths}_{(\text{Within-Within})}$	.002		
$\text{Weaknesses}_{(\text{Within-Within})}$	.012		
Residual	.446	.636	.689

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. On the outcome variable (preliminary rating), higher values represent more negative evaluations.

Table S9. Parameter estimates from models estimating the ICC by application and by reviewer

	Rating	**Strengths**	**Weaknesses**
Fixed Effects	*b (SE)*^{Sig}	*b(SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	3.55 (.17)***	30.1 (1.94)***	15.6 (1.13)***
Random Effects	*Var*^{Sig}	*Var*^{Sig}	*Var*^{Sig}
By-reviewer intercept	.08	115.17***	14.14
By-application intercept	.00	4.09	0.00
Residual	2.09	72.97	77.96
ICC$_{reviewer}$	3.87%	59.91%	15.35%
ICC$_{application}$	0.00%	2.13%	0.00%

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. *P*-values for fixed effects are computed via an approximate F-test with the Kenward-Roger method to compute the degrees of freedom (63). *P*-values for random effects are computed via a χ^2 likelihood ratio test. The ICC is computed by dividing the variance explained by the random intercept of interest by the sum of the total variance explained by all effects (i.e., the random intercept of interest, the other random intercept, and the residual variance). Here, we express the ICC as a percentage so that it indicates the percent of variance in the outcome variable that is explained by the random factor of interest.

Table S10. Estimates of agreement among reviewers and of agreement among applications) for the weaknesses variable with the outlier included

		Agreement Among Reviewers	Agreement Among Applications
Weaknesses (including outlier)	ICC	ICC = 2.82% (p = .80)	ICC = 33.7% (p = .03)
	Krippendorff's α	α = .034 [-.104, .160]	α = .339 [.253, .422]
	Similarity t-test	M = 0.70, SD = 5.52, $t(24) = 0.64$, p = .53	M = 2.08, SD = 6.95, $t(39) = 1.89$, p = .07

Note. ICCs were estimated via a LMEM with an overall fixed intercept and a random intercept for the cluster variable of interest (application or reviewer). An alternative approach to estimating the ICC can be found in Table S9. *P*-values were estimated via a χ^2 likelihood ratio test on the random intercept from the LMEM. Values of Krippendorff's α above .67 are considered suitable for tentative conclusions about reliability, and above .8 are considered reliable. 95% confidence intervals were estimated using 1000 bootstrapped samples (23). The similarity t-tests are one-sample t-tests conducted on the similarity scores; values above zero indicate that an application's ratings are more similar to each other than to ratings referring to other applications (for agreement among reviewers), or that a reviewer's ratings are more similar to each other than to ratings from other reviewers (for agreement among applications.

Table S11. Parameter estimates from Model 1 without and with the outlier

	No Outlier	**With Outlier**
Fixed Effects	*b (SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	3.46 (.21)***	3.50(.25)***
$\text{Strengths}_{(\text{Within-Within})}$	-.01(.02)	-.03(.02)
$\text{Weaknesses}_{(\text{Within-Within})}$	.08 (.03)*	.09(.03)**
Random Effects	*Var*	*Var*
By-reviewer		
Intercept	0.97	0.89
$\text{Strengths}_{(\text{Within-Within})}$	0.00	0.00
$\text{Weaknesses}_{(\text{Within-Within})}$	0.00	0.00
By-application		
Intercept	0.16	0.70
$\text{Strengths}_{(\text{Within-Within})}$	0.00	0.00
$\text{Weaknesses}_{(\text{Within-Within})}$	0.01	0.01
Residual	0.45	0.47

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. On the outcome variable (preliminary rating), higher values represent more negative evaluations. *P*-values for fixed effects are computed via an approximate F-test with the Kenward-Roger method to compute the degrees of freedom (63).

Table S12. Parameter estimates from Model 2 without and with the outlier

	No Outlier	**With Outlier**
Fixed Effects	*b (SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	3.51 (.15)***	3.53 (.16)***
$Weaknesses_{(Within\text{-}Within)}$	.13 (.02)**	.11 (.02)***
$Weaknesses_{(App\ Cluster\ Means)}$	.17 (.03)***	.15 (.02)***
$Weaknesses_{(Rev\_Cluster\_Means)}$	.03 (.02)	.02 (.02)
Random Effects	*Var*	*Var*
By-reviewer		
Intercept	0.62	0.64
$Weaknesses_{Within\text{-}Within}$	0.00	0.00
By-application		
Intercept	0.00	0.00
$Weaknesses_{Within\text{-}Within}$	0.00	0.00
Residual	0.69	0.66

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. *P*-values for fixed effects are computed via an approximate F-test with the Kenward-Roger method to compute the degrees of freedom (63).

Table S13. Parameter estimates from estimating the effect of funding status on the rating, number of strengths, and number of weaknesses with the outlier included

	Outcome: Rating	Outcome: Strengths	Outcome: Weaknesses
Fixed Effects	*b* (*SE*)^{Sig}	*b* (*SE*)^{Sig}	*b* (*SE*)^{Sig}
(Intercept)	3.48 (.21)***	30.72 (2.25)***	14.94 (1.31)***
Funded	-0.19 (.34)	1.50 (2.76)	-1.46 (2.32)
Random Effects	*Var*	*Var*	*Var*
By-reviewer			
Intercept	.10	115.77	6.22
$Reviewer_{Funded}$	.09	0.00	27.56
By-application			
Intercept	.00	6.41	0.00
$Application_{Funded}$	.00	0.00	0.00
Residual	2.06	71.79	75.309

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. *P*-values for fixed effects are computed via an approximate F-test with the Kenward-Roger method to compute the degrees of freedom (63).

Table S14. Parameter estimates from Model 3, Model 4, and Model 5, with the outlier included.

	Model 3	**Model 4**	**Model 5**
Fixed Effects	*b (SE)*^{Sig}	*b(SE)*^{Sig}	*b (SE)*^{Sig}
(Intercept)	3.50 (.25)***	3.50 (.27)***	3.55 (.15)***
$\text{Strengths}_{(\text{Within-Within})}$	-0.03 (.02)	-0.04 (.02)*	
$\text{Weaknesses}_{(\text{Within-Within})}$	0.09 (.03)**	0.10 (.02)***	0.11 (.01)***
$\text{Weaknesses}_{(\text{App Cluster Means})}$			0.15 (.02)***
$\text{Weaknesses}_{(\text{Rev Cluster Means})}$			0.03 (.02)
Random Effects	*Var*	*Var*	*Var*
By-reviewer			
Intercept	.890	.717	.593
$\text{Strengths}_{(\text{Within-Within})}$			
$\text{Weaknesses}_{(\text{Within-Within})}$			
By-application			
Intercept	.696	1.213	
$\text{Strengths}_{(\text{Within-Within})}$	.001		
$\text{Weaknesses}_{(\text{Within-Within})}$	.001		
Residual	.465	.647	.742

Notes. * $p < .05$. ** $p < .01$. *** $p < .001$. On the outcome variable (preliminary rating), higher values represent more negative evaluations.

Figure S1. Visual depiction of the three measures of agreement among applications with 95% confidence intervals.