Grant reviewer perceptions of panel discussion in face-to-face and virtual formats: lessons from team science?

Stephen A. Gallo[1*], Karen B. Schmaling[2], Lisa A. Thompson[1] and Scott R. Glisson[1]
[1]Scientific Peer Advisory and Review Services, American Institute of Biological Sciences, Herndon, VA
[2] Washington State University, Vancouver, WA
[*]Corresponding Author
Email: sgallo@aibs.org

Abstract

In efforts to increase efficiency and convenience and reduce administrative cost, some granting agencies have been exploring the use of alternate review formats, particularly virtual panels utilizing teleconference-based (Tcon) or Web based (Wb) technologies. However, few studies have compared these formats to standard face-to-face (FTF) reviews; and those that have compared formats have observed subtle differences in scoring patterns and discussion time, as well as perceptions of a decrease in discussion quality in virtual panels. Here we present data from a survey of reviewers focused on their perceptions of the facilitation and effectiveness of panel discussion from their last peer review experience in virtual (Vcon/Tcon/Wb) or FTF panel settings. Reviewers indicated that, in terms of participation, clarifying differing opinions, informing unassigned reviewers and chair leadership, the facilitation of panel discussion was viewed similarly for FTF versus Vcon/Tcon reviewers. However, small but significant differences were found for several of these parameters between FTF and Wb reviews, which may suggest better panel communication (and thus more effective discussion) in FTF panels. Perceptions of discussion facilitation were not affected by our proxy for long-term team membership, frequency of review participation. Surprisingly, no differences were found between any of the reviewers' experiences in virtual or FTF settings in terms of the discussion affecting the outcome, in choosing the best science, or even whether the discussions were fair and balanced. However, those who felt the discussion did not affect the outcome were much more likely to feel negatively about the facilitation of the panel discussion. Small but significant differences were also reported between Wb and FTF reviewers in terms of their perceptions of how well their expertise was utilized on the panel, which may suggest that the level of communication provided in FTF panels allows for better integration of expertise across panel members when evaluating research proposals as a team. Overall, despite clear preferences by reviewers for FTF panels, the lack of differences between FTF and Vcon/Tcon panel facilitation or discussion quality potentially supports the use of this review format by granting agencies, although subtle differences may exist that were not reported by reviewers in this survey. These results also provide some evidence of the perceived limitations in discussion quality in Wb panels, at least in non-recurring panels.

Keywords: Peer Review, Team Science, Communication, Research Funding, Grant Applications, Teleconference, Face-to-Face, Web Based Review, Survey

Introduction

The US National Institutes of Health (NIH), like many major research funders, utilizes a "long standing and time-tested system of peer review to identify the most promising biomedical research [1]." However, peer review is implemented in a variety of ways and meeting formats by different funding agencies and institutes [2]. Even across the NIH, to improve the efficiency, cost-effectiveness, and convenience of the process, panel meetings meet not only face-to-face (FTF), but sometimes via videoteleconference (Vcon) or through a Web based portal [3]. However, these alternate review formats have not completely replaced in-person meetings, in part due to reviewer preferences toward FTF formats [4]. In fact, when the Canadian Institute of Health Research (CIHR) almost completely replaced all face-to-face review meetings with virtual ones, there was a significant backlash from the scientific community, because it was felt the quality of the decision making on virtual panels was much lower [5]. Eventually, ignoring the recommendations from a report from an international working group suggesting that the "asserted benefits of face-to-face peer review are overstated" [6], CIHR relented and abandoned its reforms, focusing again on in-person review meetings [7].

What is striking about these policy shifts is the scant evidence supporting the use of one review format over another. The literature surrounding grant peer review as a whole is very limited, and while some studies in the literature have examined review panel discussion and its effects on scoring [8-12], only four have contrasted traditional and alternate review formats [13-16]. While Gallo et al. (2013) has found no significant differences between face-to-face (FTF) or teleconference (Tcon) panels in terms of the average, breadth or levels of contentiousness in the final scores, both Pier et al (2015) and Carpenter et al. (2015) noted that the effect of discussion on scoring (shifts in scoring by assigned reviewers after discussion) was slightly but significantly muted in Tcon panels as compared to FTF panels (Vcon panels in the case of Pier et al)[13-15]. Consistent with previous findings, these analyses found the magnitude of these scoring shifts after discussion were small and only affected the funding status of a small portion of grants. In addition, these four studies found the average discussion time was reduced for Tcon/Vcon panels, although no correlation between discussion time and the magnitude of shifting scores post-discussion were found in any of these studies. Further, a 2015 NIH survey of reviewers found that the quality of discussions for text-based review was not rated as highly as that of FTF or even Tcon/Vcon reviews [17], while in another study, 43% of reviewers felt that virtual review panels yielded minimal interaction among reviewers [16].

These results point to a slight reduction in reviewer engagement in virtual panels, which is consistent with the literature on distributed teams [18]. Lowered engagement and reduced levels of trust among virtual team members is well documented, more so in text-based communication [19-22]. In addition to the lack of visual cues, opportunities to generate intra-panel trust during panel breaks and meals are also missing in virtual settings [23]. It has also been suggested that virtual teams may have more difficulty developing transactive memory, including an understanding of the location of expertise within the panel [24], which in peer review may reduce productive participation from unassigned panel members. Persuasive tasks, which are crucial to review discussions, have also been shown to be particularly affected by communication setting [19].

It is unclear how precisely the reduced engagement seen in virtual panels manifests itself in terms of review team processes. For a review panel discussion to work effectively, there must be a sense of inclusion across all panel members (such that reviewers feel enabled to lend their expertise to the discussion). Is the decrease in discussion times observed in virtual panels due to lowered engagement of unassigned reviewers, assigned reviewers, or both? Arguments about the quality of research proposals must be clearly communicated to be persuasive, yet it is unclear if virtual communication hinders the clarity of discussions or the persuasiveness of the arguments (as seems may be the case given the reduced levels of scores shifting post-discussion). It is also unclear if team leadership is hindered in a virtual review format, thereby limiting the Chair's ability to facilitate the discussions. Ultimately and most importantly, are the discussions of similar quality across review formats and do they equally promote the best science? These types of questions are not easily answered through the analysis of scores. Pier et al. (2015) suggest that reviewers perceive currently unmeasured benefits of FTF meetings, including "the camaraderie and networking that occurs in person, the thoroughness of discussion, the ease of speaking up or having one's voice heard, the fact that it is more difficult to multi-task or become distracted, reading panelists' facial expressions, and perceived cohesiveness of the panel [15]."

Recently, the American Institute of Biological Sciences (AIBS) developed a survey to address reviewer perceptions of their most recent panel meeting experience and distributed it to biomedical scientists. Two publications have resulted from analysis of the survey responses [4,25]; however, neither addressed discussions quality, despite having included a section in the survey on discussion facilitation and its impact on review outcomes. To examine some of these questions posed above, the AIBS analyzed feedback from the surveyed scientists about the quality and facilitation of their most recent panel discussions and asked respondents to indicate whether their meeting format was FTF, Vcon/Tcon or Wb in the hope to shed some light on the effect of the panel meeting setting on review effectiveness and quality. A better understanding of reviewer perceptions of panel effectiveness could be used, in part, to inform the future implementation of different review formats, which up until this point appear to be largely driven by cost-savings incentives.

Methods

Survey

This study involved human participants who responded to a survey. The survey was reviewed by the Washington State University Office of Research Assurances (Assurance# FWA00002946) and granted an exemption from IRB review (IRB#15268; 45 CFR 46.101(b)(2)). Participants were free to choose whether or not to participate in the survey and consented by their participation. They were fully informed at the beginning of the survey as to the background behind this research, how we acquired their email address, and the importance and intended use of the data. As mentioned, the general survey methodology has been described in two other manuscripts [4,25]. The original survey contained 60 questions and was divided into 5 subsections (the full survey is available in the **S1 File in the Supporting Information**); however, only 3 sections are analyzed in this manuscript to address the issue of discussion quality: 1. Grant Submission and Peer Review Experience, 2. Reviewer Attitudes toward Grant Review and 3. Peer review panel meeting proceedings. The questions regarding discussions quality included here were not analyzed in the previous publications, although other aspects, such as review frequency and reviewer preference were looked at previously.

The questions examined had either nominal (Yes/No) or ordinal (Likert rating) response choices; for example, "on a scale of 1-5 (1 most definitely, 5 not at all), did the grant application discussions promote the best science?"). However, respondents were also given the choice to select "no answer/prefer not to answer." At the end of each section, respondents could reply in free form text to clarify answers. A full copy of the peer review survey is available in the **S1 File**. The raw, anonymized data are available as well (https://doi.org/10.6084/m9.figshare.8132453.v1).

As mentioned in previous publications, the survey was sent out in September of 2016 to 13,091 individual scientists from AIBS's database through the use of Limesurvey, which de-identified the responses from respondents. AIBS's proprietary database has been developed over several years to help AIBS recruit potential reviewers for evaluation of biomedical research applications for a variety for funding agencies, research institutes and non-profit research funders. Most of these reviews are non-recurring and scientists are recruited based on matching expertise to the topic areas of the applications. All individuals participating in this survey were either reviewers for AIBS (36%) or had submitted an application as a PI which was reviewed by AIBS (71%) or both (12%). Respondents were asked to answer questions based on either the most recent peer review or reviews that occurred in the last 3 years (depending on the question); these reviews did not have to be AIBS reviews (it is likely that the majority of reviews reported were not for AIBS).

Statistics

The survey was open for two months; responses were then exported and analyzed through basic statistical software. For this analysis, participant responses were included only if they were fully submitted and included an answer for question 2e, 2.f. and 2.g., which focused on whether they had participated in a peer review panel in the last three years, and if so how often and in what

format. Thus, all questions included in this analysis were focused on reviewer experiences. Reviewers were asked questions related to the qualities of panel discussion and expertise. The data were separated out by reviewers' recent review format (FTF, Vcon/Tcon and Wb) and answers to questions on reviewer experience were compared. Age and review frequency were also included in the analysis. Data were analyzed using Stat Plus software. Mean and percentage comparisons were analyzed using non-parametric tests (e.g. Mann-Whitney, chi-square tests), due to the highly skewed ordinal distributions (most are >1.0; Figure 1). Standard 95% confidence intervals (CI) were calculated for the Likert responses (for proportion data, binomial proportion confidence intervals were calculated). Effect size (d) was calculated via standardized mean difference for all comparisons. Differences between groups were considered significant if there was either no overlap in CI or if there was overlap yet a test for difference indicated a significant result ($p<0.01$).

Results

Response Rate and Demographics

Of the 13,091 individuals contacted for this survey, 1231 responded, giving a 9.4% response rate. Of the 1231 respondents, only 874 of these completed questions 2e, 2f, and 2g, 671 (77%) of whom indicated they had recently reviewed on a panel in the last three years. These 671 reviewer respondents formed the core group upon which the current analysis is based. Demographics were analyzed in detail in previous publications: respondents were 66% male, 80% PhD and 69% in a late career stage (e.g. Tenured Full and Emeritus Professorship) with the majority being age 50 or older (75%; median age 55), Caucasia (76%) and working in Academia (81%). These respondents participated in an average of 4.0 ± 0.08 panel meetings over the last 3 years and, similar to our previous analysis [4], the distribution of their participation was bimodal, with 155 (23%) respondents participating in 7 or more reviews in a 3-year time frame (Rev7 respondents), and the other 516 (77%) reviewer respondents participating in 6 or fewer reviews in this time (non-Rev7).

Review Setting

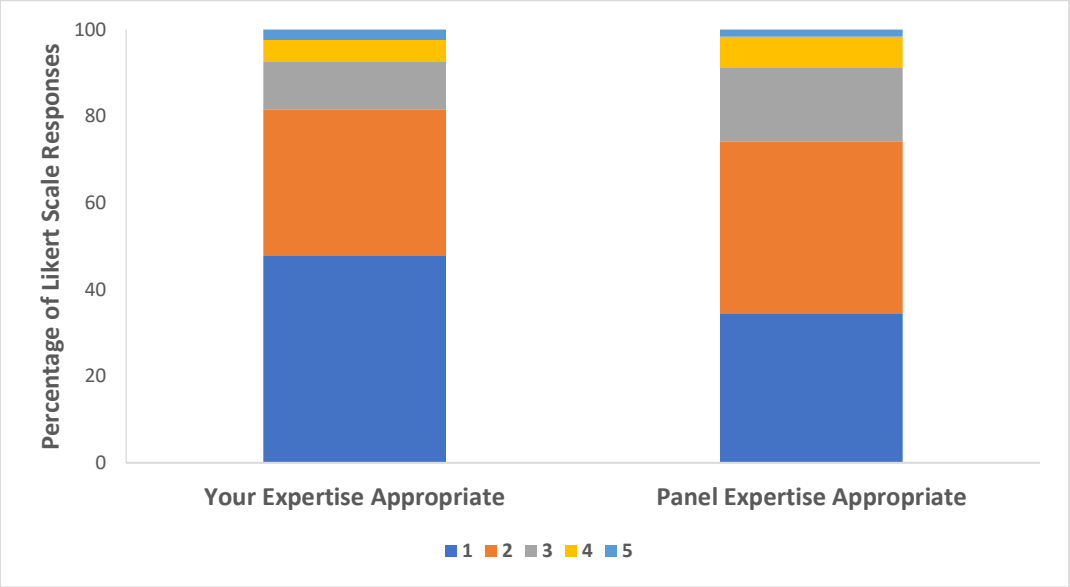
Of the reviewer respondents, 49% (N=331) recently participated in a FTF meeting, 26% (N=172) took part in a Vcon/Tcon meeting, 22% (N=148) took part in a Wb meeting while 3% marked "other" for review setting. When reviewers were asked which review setting they preferred, 80% (N=516) indicated FTF, while only 11% (N=73) indicated Vcon/Tcon and 9% (N=57) indicated Wb (26 indicated no answer). For the clear majority of reviewers that preferred in-person meetings, our previous publication reports that communication was the most influential reason for their preference in review format; for reviewers who preferred virtual meetings, logistical convenience was the most important determinant [4]. It should be noted that reviewer preference for review format was related to reviewer experience; 61% of reviewer respondents who preferred FTF were involved in a recent FTF panel, compared to only 17% of reviewer respondents who preferred virtual formats. The frequency of review participation was also related to review format; 67% (104) of Rev7 reviewers recently participated in a FTF panel compared to 44% (227) of non-Rev7 respondents. As was noted in a previous publication, Rev 7 respondents are likely associated with membership in study sections, which meet several times a year (more often by FTF) and have memberships that may last a couple of years [4]. It should be

noted that, although study section membership is often restricted to more senior scientists [26], the median age was 55 for both Rev7 and non-Rev7 respondents. Furthermore, no differences were found between groups below and above median age in terms of their review setting preferences or experiences; both groups preferred FTF more than other formats, yet both groups experienced other formats more than they would prefer.

Panel Expertise

The majority of reviewers felt their own expertise as well as that of the other panel members was either definitely or most definitely well utilized (82% and 74% for their own expertise versus panel expertise, respectively, for all reviewers). The distributions are shown in Figure 1, where it is shown that reviewers felt more positive about the utilization of their personal expertise (1.80 [1.72-1.88]) as compared to that (2.02 [1.94-2.10]) of other panel members (U[629,622]=224,479; p<0.001, d= 0.22).

Figure 1



As listed in Table 1, FTF reviewers felt more strongly that their scientific expertise was necessary and appropriately used in the review process, more so than reviewers in Wb settings (no difference found with Vcon/Tcon reviewers). However, this difference did not exist when respondents considered the expertise of other panel members.

Table 1 – Panel Expertise

Question	FTF	Vcon/ Tcon	Wb	Mann-Whitney (FTF vs Vcon/Tcon)	Mann-Whitney (FTF vs Wb)
Was your scientific expertise necessary and appropriately used in the review process?	1.67 [1.57 - 1.77]	1.87 [1.62- 2.12]	2.00 [1.82- 2.18]	U[327,167]=29,867 p=0.088, d=0.22	U[327,135]=25,996 p=0.003**, d=0.36
From your perspective was the expertise of the other panel members necessary and appropriately used in the review process?	1.95 [1.85 - 2.05]	2.06 [1.92 - 2.20]	2.16 [1.98- 2.34]	U[329, 164]=29,138 p=0.147, d=0.12	U[329, 129]=24,089 p=0.024, d=0.23

Perceptions of usage of expertise by FTF, Vcon/Tcon or Wb reviewers. Mean values and 95% confidence intervals are displayed on the left and on the right are results from Mann-Whitney tests (U[n1,n2]=value, p=value). The calculated effect size (d) is also provided. **p<0.01

We also analyzed just the Rev7 respondent population to see if long-term panel membership may affect perceptions of expertise utilization. We observed that Rev7 (151) respondents generally find their own expertise better utilized than do non-Rev7 (478) reviewers (Table 2). However, no such difference was found between Rev7 and non-Rev7 respondents regarding panel expertise.

Table 2 – Rev7 Panel Expertise

Question	Rev 7	Non-Rev 7	Significance (Rev7 vs Non-Rev7)
Was your scientific expertise necessary and appropriately used in the review process?	1.54 [1.40-1.68]	1.88 [1.78-1.98]	U[151,478]=28,510; p<0.001**, d=0.31
From your perspective was the expertise of the other panel members necessary and appropriately used in the review process?	1.88 [1.72-2.04]	2.07 [1.97-2.17]	U[150,472]=31,210; p=0.029, d=0.17

Perceptions of Rev7 respondents of usage of expertise in reviews. Mean values and 95% confidence intervals are displayed on the left and on the right are results from Mann-Whitney tests (U[n1,n2]=value, p=value). The calculated effect size (d) is also provided. **p<0.01

Discussion Facilitation

As listed in Table 3, the vast majority (89%-94%) of reviewers felt the panel discussions facilitated reviewer participation and this did not vary significantly across review settings (Table 3). Similarly, 70% of all reviewers felt discussions were most useful or very useful in clarifying opinions and this did not vary significantly across review settings (Table 3). While most reviewers (69%-82%) agreed that the format and duration of the grant application discussions

was sufficient to allow the non-assigned reviewers to cast well informed merit scores, a higher proportion of FTF reviewers as compared to Wb reviewers felt this way (Table 3). No differences were found between FTF and Vcon/Tcon reviewers. In terms of the usefulness of the chair in facilitating the application discussions, 68% of all reviewers reported that the chair's involvement was either extremely useful or very useful. Again, FTF reviewers were more likely than Wb reviewers (but not Vcon/Tcon reviewers) to feel that the chair was useful in facilitating discussions (Table 3).

Age was also influential in how reviewers viewed the facilitation of discussion, specifically in terms of clarifying differing reviewer opinions (S1 Table). Reviewers older than the median age were more positive about the usefulness of discussion in clarifying opinions than their younger counterparts (S1 Table). However, review participation, un-assigned reviewer scoring, and chair facilitation perceptions were not dependent on age (S1 Table).

Interestingly, respondent preference for review format did not influence perceptions of discussion facilitation. For example, of all respondents that recently experienced a virtual meeting (Vcon/Tcon/Wb), 91% [87%-95%] of those who preferred FTF meetings and 87% [80%-94%] of those who preferred virtual (Vcon/Tcon/Wb) meetings felt the discussions facilitated participation ($X^2[1]=1.6$; $p=0.20$, $d=0.14$). Similarly, of respondents that recently experienced a Vcon/Tcon/Wb meeting, 73% [66%-80%] of those who preferred FTF meetings and 78% [70%-86%] of those who preferred Vcon/Tcon/Wb meetings felt the format and duration of the discussions was sufficient to allow the non-assigned reviewers to cast well informed merit scores ($X^2[1]=0.97$; $p=0.33$, $d=0.12$).

Table 3 –Discussion Facilitation

Question	FTF	Vcon/Tcon	Wb	Significance (FTF v Vcon/Tcon)	Significance (FTF v Wb)
Did the grant application discussions facilitate reviewer participation?	Y=94% [91%-97%]	Y=90% [86%-94%]	Y=89% [84%-94%]	X^2[1]=2.9, p=0.090, d=0.17	X^2[1]=3.7, p=0.055, d=0.21
How useful were the grant application discussions in clarifying differing reviewer opinions?	2.06 [1.94-2.18]	2.17 [2.01-2.33]	2.30 [2.10-2.50]	U[326, 167]= 28,977, p=0.241, d=0.10	U[326, 132] = 24,015, p=0.051, d=0.22
Was the format and duration of the grant application discussions sufficient to allow the non-assigned reviewers to cast well informed merit scores?	Y=82% [78%-86%]	Y=80% [74%-86%]	Y=69% [62%-76%]	X^2[1]=0.28, p=0.590 d=0.05	X^2[1]=8.1, p=0.004** d=0.34
How useful was the Chair in facilitating the application discussions?	2.09 ± 0.06 [1.97-2.21]	2.09 ± 0.08 [1.93-2.25]	2.42 ± 0.10 [2.22-2.62]	U[325, 167]= 27,076, p=0.967, d=0.0	U[325, 130]= 24,526, p=0.007** d=0.30

Perceptions of discussion facilitation by FTF, Vcon/Tcon or Wb reviewers. Mean values and 95% confidence intervals are displayed on the left and on the right are results from either Mann-Whitney tests (U[n1,n2]=value, p=value), or chi-square tests (X^2[degree of freedom]=value, p=value). The calculated effect size (d) is also provided.
**p<0.01

There were no differences in perception between Rev7 (142) and non-Rev7 (449) reviewers (Table 4) in terms of the discussion facilitating participation, allowing unassigned reviewers to cast informed scores, and clarifying differing opinions, as well as the usefulness of the chair.

Table 4 – Rev7 Discussion Facilitation

Question	Rev 7	Non-Rev 7	Significance (Rev7 vs Non-Rev7)
Did the grant application discussions facilitate reviewer participation?	Y=94% [90%-98%]	Y=92% [90%-94%]	$X^2[1]=0.69$, p=0.410, d=0.08
How useful were the grant application discussions in clarifying differing reviewer opinions?	2.06 [1.88-2.24]	2.17 [2.07-2.27]	U[153, 489] = 34,469, p=0.142, d=0.10
Was the format and duration of the grant application discussions sufficient to allow the non-assigned reviewers to cast well informed merit scores?	Y=82% [76%-88%]	Y=78% [74%-82%]	$X^2[1]=1.60$, p=0.210, d=0.10
How useful was the Chair in facilitating the application discussions?	2.06 [1.88-2.24]	2.20 [2.10-2.29]	U[151, 488] = 33,673, p=0.110, d=0.13

Perceptions of Rev7 respondents of discussion facilitation. Mean values and 95% confidence intervals are displayed on the left and on the right are results from either Mann-Whitney tests (U[n1,n2]=value, p=value), or chi-square tests (X^2[degree of freedom]=value, p=value). The calculated effect size (d) is also provided. $^{**}p<0.01$

Discussions and Outcome

A total of 71% of reviewers agreed that panel discussion was extremely effective or very effective in influencing the outcome of the grant, although no differences were found across review setting (Table 5). Similarly, 60% of reviewers definitely or most definitely agreed that the grant application discussions promoted the best science; again no differences were found across review setting (Table 5). Interestingly, reviewers were more positive that the discussions were affecting the outcome (2.17 [2.08-2.26]) than they were facilitating the selection of the best science (2.38 [2.30-2.46]; U[638, 640] = 174,842, p<0.001, d=0.19). Moreover, the overwhelming majority (87%-88%) of reviewers felt the discussions were fair and balanced; review format did not affect the perceived fairness of the discussion (Table 5).

Table 5 – Discussion and Outcome

Question	FTF (N=331)	Vcon/Tcon (N=172)	Wb (N=148)	Significance (FTF v Vcon/Tcon)	Significance (FTF v Wb)
Did the grant application discussions affect the outcome?	2.12 [2.00-2.24]	2.18 [2.01-2.35]	2.23 [2.04-2.42]	U[324, 164] = 27,371, p=0.585, d=0.05	U[324, 133] = 22,858, p=0.306, d=0.10
Did the grant application discussions promote the best science?	2.37 [2.26-2.48]	2.29 [2.15-2.43]	2.48 [2.30-2.66]	U[324, 166] = 25,715, p=0.428, d=0.08	U[324, 133] = 22,578, p=0.421, d=0.11
Were the grant application discussions fair and balanced?	88% [84%-92%]	87% [82%-92%]	88% [83%=93%]	X^2[1]=0.16, p=0.69 d = 0.03	X^2[1]=0.002, p=0.97 d = 0.00

Perceptions of review outcomes by FTF, Vcon/Tcon or Wb reviewers. Mean values and 95% confidence intervals are displayed on the left and on the right are results from either Mann-Whitney tests (U[n1,n2]=value, p=value), or chi-square tests (X^2[degree of freedom]=value, p=value). The calculated effect size (d) is also provided. $^{**}p<0.01$

No significant differences were observed between Rev7 and non-Rev7 reviewers (Table 6) in terms of affecting the outcome, promoting the best science or for whether the discussions were fair and balanced.

Table 6 – Rev7 Discussion and Outcome

Question	Rev 7	Non-Rev 7	Significance (Rev7 vs Non-Rev7)
Did the grant application discussions affect the outcome?	2.12 [1.94-2.31]	2.18 [2.08-2.28]	U[153, 485] = 35,469, p=0.411, d=0.05
Did the grant application discussions promote the best science?	2.32 [2.16-2.49]	2.40 [2.31-2.49]	U[154, 486] = 35,208, p=0.268, d=0.07
Were the grant application discussions fair and balanced?	86% [80%-92%]	88% [85%-91%]	X^2[1]=0.54, p=0.462 d = 0.07

Perceptions of Rev7 respondents of review outcomes. Mean values and 95% confidence intervals are displayed on the left and on the right are results from either Mann-Whitney tests (U[n1,n2]=value, p=value), or chi-square tests (X^2[degree of freedom]=value, p=value). The calculated effect size (d) is also provided. $^{**}p<0.01$

Finally, we were interested to explore whether views on panel discussion and outcome were related to those of discussion facilitation. We separated respondents into 2 groups, those who felt the discussions affected the outcome (scoring 1 or 2 on this question; N=450) and those who did not feel the discussions affected the outcome (scoring 3, 4, or 5 on this question; N=184). We then compared the two groups in terms of responses surrounding discussion facilitation (Table 7). Large and significant differences were found between the two groups for all the questions, including views on reviewer participation, clarification of differing opinions, informing unassigned reviewers, and chair facilitation. In all cases, respondents who felt the outcome was affected by panel discussions viewed the discussion facilitation more favorably than those who

felt the outcome was not affected by the discussions. However, similar differences were also seen between these two groups in terms of responses related to the utilization of their expertise (U[444, 181] = 51,303, p<0.001**, d=0.50) as well as that of fellow panel members (U[444, 181] = 53,611, p<0.001**, d=0.58). In both cases, respondents who felt the outcome was affected by panel discussions viewed their expertise and that of the panel more favorably than those who felt the outcome was not affected by the discussions.

Table 7 –Discussion Facilitation versus Discussion Affecting Outcome

Question	Outcome Affected (1 or 2)	Outcome Unaffected (3, 4 or 5)	Significance (Affected vs Unaffected)
Did the grant application discussions facilitate reviewer participation?	Y=96% [94%-98%]	Y=84% [79%-89%]	X^2[1]=30.8, p<0.001**, d=0.33
How useful were the grant application discussions in clarifying differing reviewer opinions?	1.76 [1.69-1.83]	3.09 [2.92-3.26]	U[450, 184] = 67,562, p<0.001**, d=1.15
Was the format and duration of the grant application discussions sufficient to allow the non-assigned reviewers to cast well informed merit scores?	Y=84% [76%-88%]	Y=65% [74%-82%]	X^2[1]=26.2, p<0.001**, d=0.40
How useful was the Chair in facilitating the application discussions?	1.96 [1.87-2.05]	2.71 [2.53-2.90]	U[446, 181] = 54,003, p<0.001**, d=0.59

Respondent perceptions of discussion facilitation in terms of those who felt the discussions affected the outcome of the proposals (scored a 1 or 2 to this question) compared to those felt discussion was relatively unaffected (scored a 3, 4 or 5 to this question). Mean values and 95% confidence intervals are displayed on the left and on the right are results from Mann-Whitney tests (U[n1,n2]=value, p=value) or chi-square tests (X^2[degree of freedom]=value, p=value). The calculated effect size (d) is also provided. **p<0.01

Discussion

Our results indicate a clear preference for FTF panels by respondents, and our previous publication suggests this is largely due to the perceived quality of communication in FTF panels; those who prefer virtual panels suggest logistical convenience as an important motivation [4]. Thus, it is unsurprising that reviewer preference and reviewer experience were found to be related, where respondents who preferred FTF panels were much more likely to have recently participated in a FTF panel as compared to those who prefer virtual panels. It is interesting that these preferences did not seem to have a strong bearing on how reviewers felt about the quality of panel discussion, suggesting that the responses recorded here are more linked to actual reviewer experiences than to any pre-conceived notions of peer review.

We also observed that reviewers generally felt their own expertise as well as that of other panel members was well utilized, although they felt more positive about their own expertise as compared to that of other panel members (Figure 1). Others have reported that individual openness to the diversity of team expertise affects team performance [27]; thus, it may be that the differences found here are related to different degrees of openness amongst reviewers, where perhaps a small proportion of respondents are truly open to the multiplicity of panel expertise. Interestingly, Rev7 respondents (presumably study section members) generally find their own expertise better utilized than non-Rev7 reviewers (Table 2). This is likely the result of long-standing team members having better knowledge how their expertise fits into the decision making process as compared to ad-hoc reviewers.

Overall, a small but significant difference was found between FTF and Wb review settings for the utilization of an individual's expertise (Table 1), but not between FTF and Tcon. It may be this relates to the panel effectiveness of deep knowledge integration of the collective team expertise, which may vary considerably depending on review setting and length of time the team has been together [21]. Indeed, significant differences in expertise utilization are seen between Rev7 respondents and non-Rev7 respondents (Table 2). Poor integration may lead to poorly perceived utilization of an individual's expertise and how this fits into the group. Web-based teams may have difficulty in developing an understanding and trust of where expertise is distributed across the panel, which may negatively influence perceptions of its effective use by the panel [24]. Thus, while it likely can be assumed reviewers are recruited in a similar way for FTF and Wb panels and thus expertise is similarly matched to proposals, it may be that richer communication channels provided in FTF and Tcon panels allow for better knowledge integration across team membership, which leads to a better appreciation of where expertise lies (particularly for ad-hoc teams). It is likely this is compensated for in long standing teams by the strengthening of knowledge integration of team expertise over time. Previous results from a survey of NIH reviewers also found only small differences between FTF (89%) and Vcon (81%)/Tcon (82%) reviewers in terms of the proportion that regarded the adequacy of the panel expertise favorably (no test for significance); unfortunately they did not include Wb meetings in this measure [17].

Interestingly, no differences were found either across review settings (Table 1) or between Rev7/nonRev7 respondents (Table 2) with regards to other panel members' expertise; although again respondents generally felt their own expertise was better utilized than that of other panel members. This may simply be the level of familiarity with other's expertise relative to the

assigned proposals is less than their own, and thus not as sensitive to changes in review parameters.

In general, reviewers felt review discussions were well facilitated (Table 3). However, similar to the results from Table 1, Wb reviewers were also more negative about some aspects of the facilitation of review discussions as compared to FTF reviewers (Table 3). Wb reviewers were less likely than FTF reviewers to find the discussions useful in allowing un-assigned reviewers to make well informed judgements. They were also less likely to find the chair to be a good facilitator of those discussions. Thus, it seems reviewers who recently participated in a Wb meeting are less likely to find the team communication clear and well facilitated. Results from the 2015 NIH survey also indicated smaller proportions of reviewers who had favorable impressions of discussion facilitation with Vcon (70%)/Tcon (76%) and Wb (67%) reviewers compared to FTF (83%), although again no tests for significance were reported [17]. It should be noted that the NIH survey asked only whether "discussions supported the ability of the panel to evaluate the applications being reviewed," which is more general and may have wrapped many of these aspects together.

Again, no differences in opinions of discussion facilitation were found in comparisons between Rev7 and non-Rev7 reviewers (Table 4), suggesting perceptions of discussion quality are not dependent on long-term team membership, despite panel members likely having higher levels of trust and perhaps more established communication among members. However, these results did seem to depend a bit on age, as younger reviewers found clarifying opinions more difficult than reviewers above the median age (S1 Table); this may be related to a level of deference to senior reviewers, who may more often "get the floor" to voice their opinions than younger reviewers, although more research is needed to verify this. Nevertheless, it seems communication setting affects the facilitation of discussion more than long-term team experience, based on the likely assumption that Rev7 reviewers are study section members.

In several areas we have observed differences in how reviewers perceive the facilitation of discussion in FTF panels compared to Wb panels. Moreover, the 2015 NIH survey results suggest much lower reviewer comfort levels with potentially having their own applications reviewed via a Wb panel versus a Vcon/Tcon panel [17]. Taken together, these results are supported by the team science literature that suggests virtual team members in text only communication situations have great difficulty in developing team trust, even when compared to Vcon/Tcon teams, and need richer forms of communication to participate in cooperative tasks [21,22].

Interestingly, we found no significant differences in reported discussion facilitation between FTF and Vcon/Tcon review formats (Table 3). While we and others have previously found subtle differences in scoring and the length of discussion times between Tcon and FTF settings [14,15], this doesn't seem to affect reviewer perceptions of how the discussion was facilitated, although the differences found between FTF and Wb settings underscores the importance of at least audio-facilitated communication and discussion.

The review discussion was generally found to be influential on the outcomes and effective in promoting the best science, although this did not seem to be affected by review setting, including Wb settings (Table 5). This was also the case for the Rev7 comparison (Table 6), suggesting the impact of discussion on review outcomes was not influenced by review setting or team membership. This is in contrast to the differences observed between Wb and FTF reviewers

regarding the facilitation of discussions as well as previous data suggesting that post-discussion shifts in score are reduced in Tcon panels compared to FTF panels [14]. It may be that the some of the effects of review setting on proposal discussion are more subtle than can be detected by reviewers. It may also be that reviewers are overconfident in the effectiveness of panel discussion, potentially because they were directly involved in the discussion [28]. However, respondents that did not feel the discussions influenced the outcome were much more likely to have negative perceptions about the facilitation of discussion (Table 7). Thus, it is likely poorer facilitation limits the ability of the discussion to impact the outcome of the review, which is in agreement with previous studies on post-discussion scoring [14,15]. However, it should be noted that respondents who felt the discussions did not influence the outcome were also more likely to have a negative perception about the utilization of expertise. It may be this group of reviewers is just more negative in its responses than reviewers who felt the discussion impacted the outcome. Future studies could address this by examining actual panel discussions and potential linguistic and stylistic differences in FTF and Vcon/Tcon/Wb panels and comparing them to post-discussion scoring changes [29,30]. It would also be interesting to gather perceptions from outside impartial panel observers, such as scientific review officers who manage panels for funding agencies, which may counter reviewer perceptions.

Interestingly, reviewers are more positive about the discussions affecting the outcome than they are about selecting the best science. This may be related to the natural rater variability in assessing research quality that is inherent in peer review [31,32], which likely exists independent of communication setting. Importantly, the vast majority of reviewers did feel that these panel discussions were fair and balanced, irrespective of review setting, which at least alleviates some of the concern that certain review formats promote bias more than others. However, it should be mentioned that implicit bias may be a very difficult thing for reviewers to detect in a panel discussion, yet it may still have an important impact on panel discussion and scoring. Future work should more rigorously evaluate the relationship between implicit reviewer biases, panel discussion and review format.

One potential limitation to this work is the small practical effects of many of the statistically significant differences between FTF and Wb review settings, although similarly small effects were seen in the NIH survey as well [17]. While small sample sizes are a limitation of this study, our observations somewhat mirror the results of the NIH study, which used much larger sample sizes and still only found small effects in general. Nevertheless, while the effects are subtle, these types of studies can help point the direction for future prospective research.

Another limitation is the relatively low response rate (6.7%), although this rate is similar to those in other recent surveys on journal peer review [33-35]. Furthermore, our demographics are very similar to those of NIH study section members, according to recent reports [26,36]. Additionally, comparing the larger, full sample of incomplete responses (n=1231) to the one used in this manuscript, we find very similar demographics as well as a similar bi-modal distribution of review participation, which shows this sample is representative of the larger population.

Overall, while our reviewer pool indicated a clear preference for FTF panels, perceptions of Vcon/Tcon discussion quality was similar to that of FTF discussion quality; they were viewed as equally clear, inclusive and impactful, and independent of reviewer preference. The previous scoring differences reported aside, it seems our results help bolster the case for Vcon/Tcon panels. It is also clear that reviewers do not feel the same way about the discussion quality of Wb panels and given their low popularity, much more justification should be sought before routinely

implementing this review format. Finally, in terms of review formats that most efficiently avoid bias and promote the best science, from our results, no format seems to be particularly advantageous. Future studies of discussion quality across review formats will need to account for the great variability in reviewer personality and panel leadership. For instance, variability in discussion time may be a function of chair behavior (limit-setting versus allowing discussion). Also, are more persuasive reviewers hindered by review format more than less proactive reviewers? Some have reported the importance of score-calibration comments and even laughter in the effectiveness of panel discussion, although it is unclear if these are affected in any way by review format [30]. And as discussion has traditionally affected the funding status of only a small proportion of proposals [9,10,14], these types of studies should be examined in parallel with those examining the decision making processes that occur at the individual reviewer level.

Figure Legends

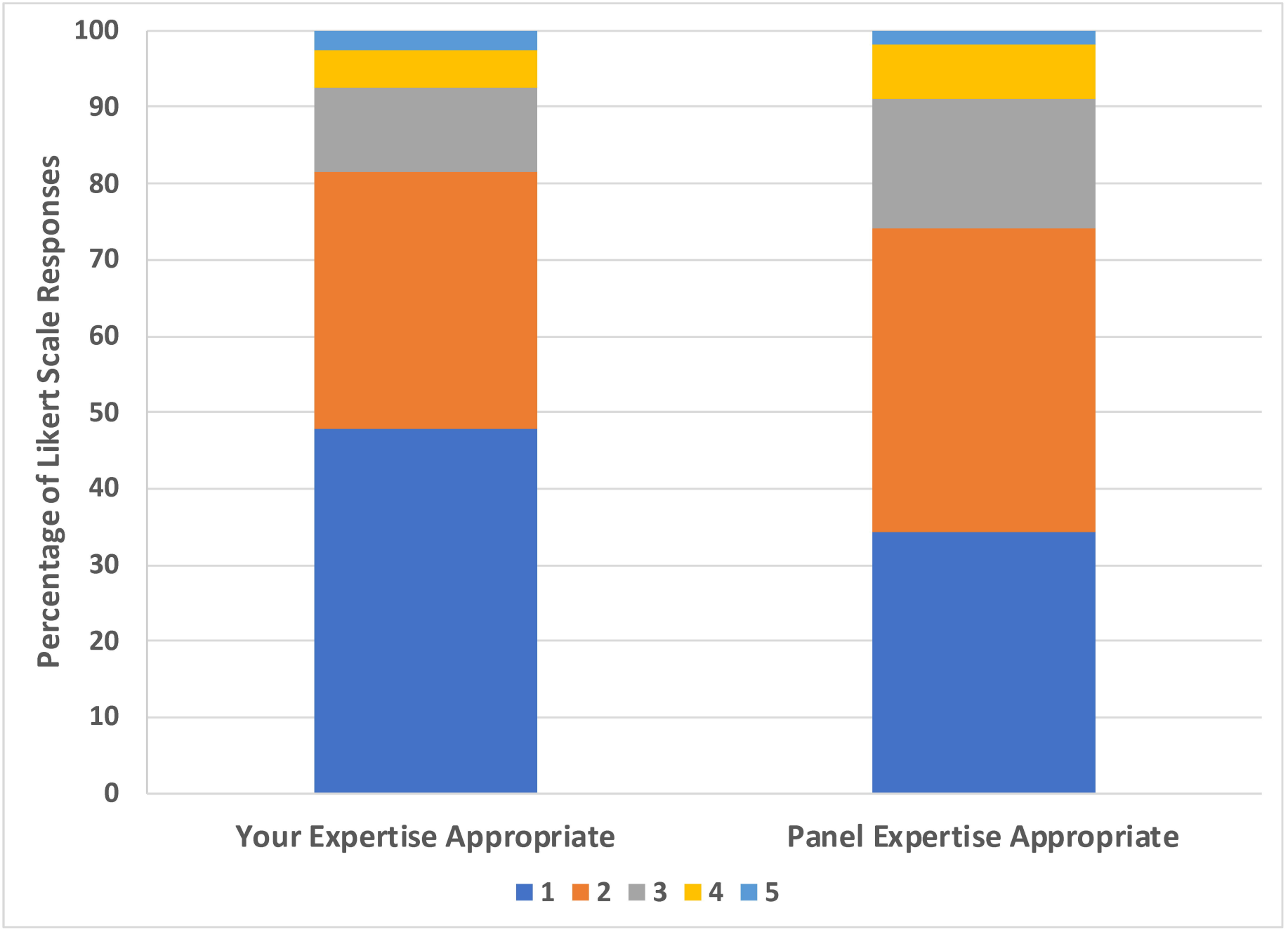
Figure 1 – Likert distribution for individual and panel expertise responses. Responses are for following questions: 1.Was your scientific expertise necessary and appropriately used in the review process? (N=647); and 2. From your perspective was the expertise of the other panel members necessary and appropriately used in the review process (N=640)? Likert scale responses are represented where 1 is most definitely and 5 is not at all.

References

1. NIH. Peer Review. 2018; https://grants.nih.gov/grants/peer-review.htm (Last Accessed January 2019).
2. Liaw L, Freedman JE, Becker LB, Mehta NN, & Liscum L. Peer Review Practices for Evaluating Biomedical Research Grants. Circulation research, 2017; 121(4), e9-e19.
3. NIAID. Serving on a Peer Review Committee. 2018; https://www.niaid.nih.gov/grants-contracts/serving-peer-review-committee (Last Accessed January 2019).
4. Gallo SA, Thompson LA, Schmaling KB, & Glisson SR. Participation and Motivations of Grant Peer Reviewers: A Comprehensive Survey of the Biomedical Research Community. 2018. Preprint. Available from: bioRxiv, 479816.
5. Webster P. CIHR modifies virtual peer review amidst complaints. CMAJ: Canadian Medical Association Journal. 2015; 187(5): E151.
6. Gluckman P, Ferguson M, Glover A, Grant J, Groves T, Lauer M & Ulfendahl M. International Peer Review Expert Panel: A report to the Governing Council of the Canadian Institutes of Health Research. 2017. http://www.cihr-irsc.gc.ca/e/50248.html (last accessed March 2019).
7. Webster P. CIHR's face-to-face about-face. Canadian Medical Association. Journal. 2017; 189(30): E1003.
8. Obrecht M, Tibelius K, D'Aloisio G. Examining the value added by committee discussion in the review of applications for research awards. Res Eval 2007; 16: 79-91. doi:10.3152/095820207X223785
9. Martin MR, Kopstein A, Janice JM. An analysis of preliminary and post-discussion priority scores for grant applications peer reviewed by the Center for Scientific Review at the NIH. PLoS ONE 2010; 5:e13526. doi:10.1371/journal.pone.0013526
10. Fogelholm M, Leppinen S, Auvinen A, et al. Panel discussion does not improve reliability of peer review for medical research grant proposals. J Clin Epidemiol. 2012; 65: 47–52. doi:10.1016/j.jclinepi.2011.05.001
11. Fleurence RL, Forsythe LP, Lauer M, Rotter J, Ioannidis JP, Beal A, Frank L and Selby JV. Engaging patients and stakeholders in research proposal review: the patient-centered outcomes research institute. Ann Intern Med. 2014; 161:122–30.
12. Forsythe LP, Frank LB, Tafari TA, Cohen SS, Lauer M, et al. Unique review criteria and patient and stakeholder reviewers: analysis of PCORI's approach to research funding. Value in Health. 2018; 21(10): 1152-1160
13. Gallo SA, Carpenter AS, Glisson SR. Teleconference versus face-to-face scientific peer review of grant application: effects on review outcomes. PLoS ONE 2013; 8: e71693. doi:10.1371/journal.pone.0071693
14. Carpenter AS, Sullivan JH, Deshmukh A, Glisson SR, & Gallo SA. A retrospective analysis of the effect of discussion in teleconference and face-to-face scientific peer-review panels. BMJ open 2015; 5(9), e009138.
15. Pier EL, Raclaw J, Nathan MJ, Kaatz A, Carnes M, & Ford CE. Studying the study section: How group decision making in person and via videoconferencing affects the grant peer review process. WCER Working Paper No. 2015-6. Wisconsin Center for Education Research. 2015 Oct.
16. Vo NM and Trocki R. Virtual and Peer Reviews of Grant Applications at the Agency for Healthcare Research and Quality. South Med J. 2015; 108(10): 622-6.

17. NIH CSR. Reviewer Quick Feedback Survey Results. 2015; https://public.csr.nih.gov/sites/default/files/2017-10/ReviewerQuickFeedbackSurveyResults.pdf (last accessed January 2019).
18. Rogelberg SG, O'Connor MS, Sederburg M. Using the stepladder technique to facilitate the performance of audioconferencing groups. J Appl Psychol. 2002; 87: 994–1000
19. Driskell JE, Radtke PH, Salas E. Virtual teams: effects of technological mediation on team performance. Group Dyn. 2003; 7: 297–323.
20. Zheng JB, Veinott E, Box N, et al. Trust without touch: jumpstarting long-distance trust with initial social activities. CHI Letters Proceedings of the SIGCHI Conference on Human Factors in Computing System. 2002; 4: 141–6.
21. Cooke NJ. National Research Council. Enhancing the effectiveness of team science. National Academies Press; 2015; Jul 15
22. Bos N, Olson J, Gergle D, Olson G, & Wright Z. Effects of four computer-mediated communications channels on trust development. In Proceedings of the SIGCHI conference on human factors in computing systems. 2002; 135-140. ACM.
23. Blatner A. About nonverbal communications. Part 1: General Considerations. 2009; https://www.blatner.com/adam/level2/nverb1.htm (last accessed February 2019)
24. Kanawattanachai P, & Yoo Y. The impact of knowledge coordination on virtual team performance over time. MIS quarterly, 2007; 31(4).
25. Gallo S, Thompson L, Schmaling K, and Glisson S. Risk evaluation in peer review of grant applications. Environment Systems and Decisions. 2018a; 1-14.
26. National Institutes of Health (NIH) 2007-2008 Peer Review Self-Study Final Draft. 2008. http://enhancing-peer-review.nih.gov/meetings/nihpeerreviewreportfinaldraft.pdf (last accessed November 2018)
27. Homan AC, Hollenbeck JR, Humphrey SE, Knippenberg DV, Ilgen DR, & Van Kleef GA. Facing differences with an open mind: Openness to experience, salience of intragroup differences, and performance of diverse work groups. Academy of Management Journal, 2008; 51(6): 1204-1222.
28. Moore DA, & Healy PJ. The trouble with overconfidence. Psychological review, 2008; 115(2): 502.
29. Raclaw J and Ford CE. Laughter and the management of divergent positions in peer review interactions. Journal of pragmatics. 2017; 113: 1-15.
30. Pier EL, Raclaw J, Carnes M, Ford CE, and Kaatz A. Laughter and the Chair: Social Pressures Influencing Scoring During Grant Peer Review Meetings. Journal of general internal medicine. 2019; Jan2:1-2.
31. Cole S & Simon GA. Chance and consensus in peer review. Science. 1981; 214(4523): 881-886.
32. Pier EL, Brauer M, Filut A, Kaatz A, Raclaw J, Nathan MJ, Ford CE & Carnes M. Low agreement among reviewers evaluating the same NIH grant applications. Proceedings of the National Academy of Sciences. 2018; 115(12): 2952-2957.
33. Ware M, & Monkman M. Peer review in scholarly journals: Perspective of the scholarly community—An international study. London, UK: Publishing Research Consortium. 2008
34. Ware Mark. Peer review: benefits, perceptions and alternatives. Publishing Research Consortium 2008: 4.

35. Sense About Science "Peer Review Survey" 2009 http://archive.senseaboutscience.org/pages/peer-review-survey-2009.html (last accessed May 2019)
36. NIH OER. Enhancing Peer Review Survey Results Report. 2013; https://enhancing-peer-review.nih.gov/docs/Enhancing_Peer_Review_Report_2012.pdf (last accessed May 2019).


100
90
80
70
60
50
40
30
20
10
0
Percentage of Likert Scale Responses
Your Expertise Appropriate
Panel Expertise Appropriate
1
2
3
4
5