BMJ Open

A retrospective analysis of the effect of discussion in teleconference and face-to-face scientific peer-review panels

Afton S Carpenter, Joanne H Sullivan, Arati Deshmukh, Scott R Glisson, Stephen A Gallo

To cite: Carpenter AS, Sullivan JH, Deshmukh A, *et al*. A retrospective analysis of the effect of discussion in teleconference and face-to-face scientific peer-review panels. *BMJ Open* 2015;**5**: e009138. doi:10.1136/bmjopen-2015-009138

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/bmjopen-2015-009138).

Received 19 June 2015
Revised 31 July 2015
Accepted 18 August 2015

Scientific Peer Advisory & Review Services, American Institute of Biological Sciences, Reston, Virginia, USA

Correspondence to
Dr Stephen A Gallo;
sgallo@aibs.org

ABSTRACT

Objective: With the use of teleconferencing for grant peer-review panels increasing, further studies are necessary to determine the efficacy of the teleconference setting compared to the traditional onsite/face-to-face setting. The objective of this analysis was to examine the effects of discussion, namely changes in application scoring premeeting and postdiscussion, in these settings. We also investigated other parameters, including the magnitude of score shifts and application discussion time in face-to-face and teleconference review settings.

Design: The investigation involved a retrospective, quantitative analysis of premeeting and postdiscussion scores and discussion times for teleconference and face-to-face review panels. The analysis included 260 and 212 application score data points and 212 and 171 discussion time data points for the face-to-face and teleconference settings, respectively.

Results: The effect of discussion was found to be small, on average, in both settings. However, discussion was found to be important for at least 10% of applications, regardless of setting, with these applications moving over a potential funding line in either direction (fundable to unfundable or vice versa). Small differences were uncovered relating to the effect of discussion between settings, including a decrease in the magnitude of the effect in the teleconference panels as compared to face-to-face. Discussion time (despite teleconferences having shorter discussions) was observed to have little influence on the magnitude of the effect of discussion. Additionally, panel discussion was found to often result in a poorer score (as opposed to an improvement) when compared to reviewer premeeting scores. This was true regardless of setting or assigned reviewer type (primary or secondary reviewer).

Conclusions: Subtle differences were observed between settings, potentially due to reduced engagement in teleconferences. Overall, further research is required on the psychology of decision-making, team performance and persuasion to better elucidate the group dynamics of telephonic and virtual ad-hoc peer-review panels.

Strengths and limitations of this study

- This is the first study of its kind to investigate and compare the scoring nuances of teleconference and face-to-face peer-review meetings.
- The study examined numerous factors, including the magnitude of shifts in scores, effect of discussion, reviewer contentiousness and application discussion time.
- Only a representative subset of data was utilised.
- The analysis did not examine the psychological processes involved in team decision-making.

INTRODUCTION AND BACKGROUND

Scientific peer review of grant applications is the *de facto* standard in decision-making for most funding bodies, and it plays an important role in guiding research funding and providing scientific direction to projects. With increasing costs, and given advancements in technologies, teleconference (and videoconference) peer-review panels are becoming a desirable forum as compared to the traditional face-to-face setting. The teleconference format reduces costs for funding agencies and offers an increased convenience for reviewers.

The American Institute of Biological Sciences (AIBS) previously explored the efficacy of teleconference reviews compared to face-to-face reviews.[1] Our initial results indicated that there were few differences between the two settings, especially when comparing bulk population parameters, such as average overall scores (OSs), scoring distributions and spreads, and scoring reliability among panel members. However, it was observed that face-to-face panels exhibited longer discussion times than panel reviews conducted by teleconference. Further details on the peer-review process and the study outcomes can be found in our prior publication.[1]

While several studies have investigated the effect of discussion (EOD) on peer-review scoring, there is no consensus in the literature on this topic. Prior studies have found that panel discussion does not improve the reliability or effectiveness of application scoring.[2 3] Conversely, others have found that discussion does significantly impact scoring decisions for at least 13% of applications, in some cases promoting applications into the fundable scoring range.[4] Further, while numerous studies have examined decision-making, teamwork, and the effect of communication setting, no such studies have examined these factors in relation to the scoring and discussion of grant applications in the peer-review process.[5–7]

While the most obvious and major difference between teleconference and face-to-face panels is the communication setting (in person vs virtual), another important difference between the two settings is the level of trust among reviewers. In the face-to-face setting, the panelists are exposed to visual social cues during their discussions and also have the ability to socialise during breaks and meals, allowing for a level of trust to form. For these particular teleconference reviews, visual cues were not available, and there were no instances that allowed for socialising. As prior studies have shown, building trust in distributed settings is difficult, and the type of communication setting can actually play a role in and impact one's commitment to the task at hand.[7 8] Team decision-making is a very critical aspect of peer review; however, these topics are outside the main parameters of this investigation.

To further examine the possible role that review settings may have on peer-review outcomes, we investigated the EOD on the final scoring of applications for teleconference and face-to-face peer-review panels for an anonymous federal programme (PrX), which supports the funding of projects that are biomedical in nature. Specifically, we evaluated the postdiscussion score shift (Δ) between the premeeting merit score (the average of the assigned reviewer scores premeeting) and final overall merit score (the average of voting panel member scores postdiscussion). The relationship of these score shifts with discussion times, premeeting scores and review setting were examined. To the best of our knowledge, this is the first investigation of its kind to explore the EOD based on peer-review setting.

METHODS

PrX programme

This analysis utilised the same PrX data set that was used in our previous publication—namely, scoring data from PrX review panels before and after the programme underwent a transition in peer-review setting in 2011.[1] Thus, this analysis includes data from panels that convened for two funding cycles each via a face-to-face setting (in 2009 and 2010) and a teleconference setting (in 2011 and 2012). PrX, which began receiving funding in 1999, was a programme supported by a federal agency to which submitted applications were from a variety of biomedical research disciplines. Despite the change in review setting in 2011, little else changed procedurally for the review process. In addition, the total amount of funds available remained constant over the period studied, and each year had diverse topic areas. This analysis focused on applications submitted to NIH R01-like basic and translational award mechanisms, with a maximum of $725 000–$2 million in direct costs over a period of 3–4 years available to funded applicants. The majority of applications were submitted to the basic award mechanism. It should be noted that the funding success rate did vary over this time period, being 4.6%, 9.3%, 8.9%, and 10.1% for 2009–2012, respectively. Nevertheless, given the diversity of the different panels and topic areas and the need for programmatic balance, no explicit score or percentile threshold was used to determine application funding status for PrX.

Peer review

As noted in our prior publication, each panel was comprised of 7–12 subject matter experts, including the chairperson, with each panel also having consumer reviewers (individuals that have experience directly with diseases relevant to the panel topic area). PrX did not employ standing panels, so reviewer composition changed from year to year. Each year, approximately 50% of the reviewers were new to the programme. However, the basic reviewer demographics (academic rank and degree) remained fairly similar across all 4 years, regardless of review setting with the face-to-face panel consisting of 25.4% of reviewers at an assistant professor (or equivalent), 27.2% at an associate professor (or equivalent), and 47.4% at a professor or dean level (or equivalent). For the teleconference setting, these values were 22.7%, 42.9%, and 34.5%, respectively. In terms of reviewer degrees, in the face-to-face setting 60.5% of the panel had PhDs (or equivalent), 20.2% had MDs, and 19.3% had MD, PhDs (or equivalent). For the teleconference setting, the makeup was 57.1%, 21.8%, and 21.0%, respectively.

The premeeting scores for this analysis consisted of the primary and secondary reviewer scores (initial assigned reviewer scores determined prior to the panel meeting and discussion), which were entered into the AIBS online peer-review system, SCORES, by a predetermined deadline. The average of these initial primary and secondary reviewer scores is referred to as the average premeeting score (APS) in this manuscript. Premeeting scores are stated at the beginning of application discussion, followed by a verbal summary of the assigned reviewer critiques. Postdiscussion scores were stated by assigned reviewers after the panel discussion and were recorded for all non-conflicted panel members via individual, confidential electronic score sheets within SCORES. The average of all voting members' postdiscussion scores per application

constituted the final OS. Each application was scored using a 1.0–5.0 scale (where 1 is the highest merit, and 5 is the lowest merit). It should be noted that the only differences between primary and secondary reviewers are that primary reviewers lead off the panel discussion by presenting their critique first and are responsible for crafting an overall summary of the panel discussion for the application(s) they are assigned as the primary reviewer.

For PrX, the majority of topic areas would change year to year, while some topic areas would remain the same. For this analysis, we examined four topic areas that remained the same during the 4-year period, with each topic area receiving 2 years of face-to-face and 2 years of teleconference reviews. Four hundred and seventy-two application scores were examined for our analyses related to application scores and associated Δ, with 260 and 212 applications reviewed in the face-to-face and teleconference settings, respectively.

In 2010 and 2012, the programme utilised a preapplication cull. Following the preapplication review, selected investigators were invited to submit full applications. Further, in 2011 and 2012, triage was implemented. Thus, applications receiving scores from both reviewers above a certain threshold were nominated for triage and potentially would not be discussed by the panel. Panel members not in conflict were allowed to remove an application from triage when the panel first convened. If an application was removed from triage, it was added back into discussion to be voted on by all panel members not in conflict. Since there was triage in 2011 and 2012, the lowest score an application received after discussion in our data set was 3.9. To make a fair comparison to 2009 and 2010, any data points that had an OS greater than 3.9 were excluded from this analysis. Only seven data points had to be excluded, which resulted in the 260 face-to-face application data points utilised.

Approach

Differences in application scoring parameters are represented by Δ, whether it is the differential between primary and secondary reviewer scores for a given moment in time (for both prediscussion (Δ_{PS}) and postdiscussion (Δ_{PD}) time points) or the differential in scoring across discussion (for either the primary (Δ_{PRI}) or secondary (Δ_{SEC}) reviewer or the panel (Δ_A=APS—OS)). In this analysis, Δ_{PS} provides an indication of reviewer contentiousness (level of agreement/disagreement between the assigned reviewers' evaluation of an application) prior to the panel discussion, while Δ_A is a measure of the EOD on the scoring of an application. Table 1 provides an overview of the different Δ values used in this analysis.

Discussion times were calculated based on the differences in time between electronic meeting score sheets being locked manually by the designated AIBS staff member, a process done in real-time online via SCORES, ensuring no further access to the score sheet once an application's discussion has ended and each panel member not in conflict has scored. The score sheets are time stamped in SCORES, with times recorded by the system down to the second. Any unrealistic discussion times were removed from the analysis. These included times that were less than 5 min (4 min and 59 s or less) or more than 60 min (60 min and 1 s or more). Additionally, those applications that were the first of the day to be reviewed on a panel were removed from the data set since there is no accurate way to determine when exactly discussion for those applications began. After removing from the data set applications with unrealistic discussion times, applications that were reviewed first, and applications with OS scores greater than 3.9 (as discussed above), 212 face-to-face and 171 teleconference application data points were utilised for the discussion time analysis.

Table 1 Overview of the different Δs used for this analysis

Δ definitions	
Δ_{PRI}	Change in primary reviewer scores (primary reviewer score premeeting—final primary reviewer score postdiscussion)
Δ_{SEC}	Change in secondary reviewer scores (secondary reviewer score premeeting—final secondary reviewer score postdiscussion)
Δ_{PS}	Difference between primary and secondary reviewer scores premeeting (primary reviewer premeeting score—secondary reviewer premeeting score)
Δ_{PD}	Difference between primary and secondary reviewer scores postdiscussion (primary reviewer postdiscussion score—secondary reviewer postdiscussion score)
Δ_A	Difference between average of the assigned reviewer scores premeeting and the final overall score (APS—OS)

APS, average-premeeting score; OS, overall score.

RESULTS

Application score shifts

A regression analysis was performed to compare APS to the final OS (utilising the 260 and 212 face-to-face and teleconference data points, respectively). The analysis showed that there is a strong correlation between the two measures, for the face-to-face (R^2=0.74; p<0.001) and teleconference (R^2=0.82; p<0.001) settings (figures 1A, B). However, differences in variance are apparent in the mean squared errors of the linear fits; we observed a significant difference between face-to-face (0.09±0.01) as compared to teleconference (0.05±0.01) settings. Despite this, running a t test of unequal variance, no statistically significant difference in Δ_{PRI} or Δ_{SEC} was found between the two settings for primary (t(462)=−0.74; p=0.46) or secondary (t(469)=

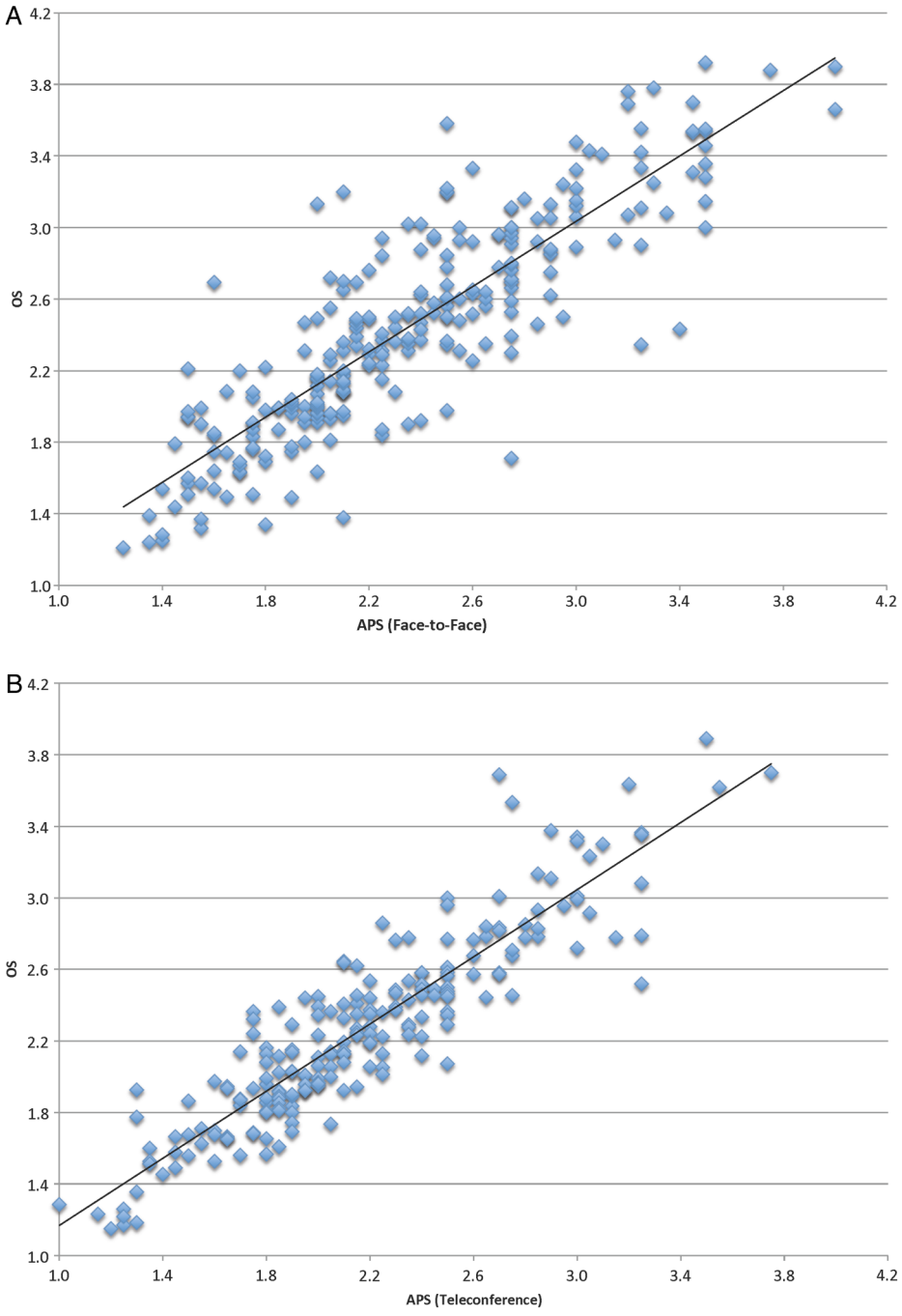

Figure 1 (A) Relationship between APS and OS for face-to-face reviews in 2009 and 2010. (B). Relationship between APS and OS for teleconference reviews in 2011 and 2012. APS, average-premeeting score; OS, overall score.

−0.36; p=0.72) reviewer scores. This was also true for Δ_A (t(464)=0.17; p=0.86). It should be noted that the mean APS and OS were 2.3 and 2.4, respectively, for the face-to-face setting and 2.2 and 2.3, respectively, for the teleconference setting. This indicates that assigned primary/secondary reviewers and voting panel members in 2011/2012 were slightly more generous on average in their initial and final scoring than those in 2009/2010 (this is consistent with our previous publication). When running paired t tests, we found that there were statistically significant differences between premeeting and postdiscussion scores for the primary reviewers during the face-to-face sessions (t(259)=−4.16; p<0.001) as well as the teleconference sessions (t(211)=−4.05; p<0.001). The same was true for the secondary reviewer face-to-face (t(259)=−3.47; p=0.001) and teleconference scores (t(211)=−3.20; p=0.002). Oneway intraclass correlations (ICC) were calculated to determine inter-rater reliability between the two assigned reviewers premeeting and postdiscussion for each year as well as per review setting type. Regardless of setting, it was found that the reliability between reviewers increased postdiscussion (see online supplementary table S1).

Examining Δ_{PRI} showed that 38.8% and 55.7% of primary reviewer scores did not change postdiscussion (table 2), while 18.5% and 13.2% of scores shifted to a better score, and 42.7% and 31.1% of scores shifted to a

Table 2 Magnitude of Δ_{PRI} as compared to magnitude of Δ_{SEC} for face-to-face (A) and teleconference (B) settings as a percentage of the magnitude subgroup and the data set as a whole

(A) Face-to-Face				(B) Teleconference			
Primary	**Secondary**	**Subgroup (%)**	**Whole (%)**	**Primary**	**Secondary**	**Subgroup (%)**	**Whole (%)**
High	High	17.9	1.9	High	High	7.1	0.5
High	Moderate	14.3	1.5	High	Moderate	21.4	1.4
High	Low	14.3	1.5	High	Low	7.1	0.5
High	Zero	53.6	5.8	High	Zero	64.3	4.2
Total		100.0	10.8	Total		100.0	6.6
Moderate	High	13.5	3.8	Moderate	High	5.0	0.9
Moderate	Moderate	16.2	6.9	Moderate	Moderate	27.5	5.2
Moderate	Low	24.3	4.6	Moderate	Low	20.0	3.8
Moderate	Zero	45.9	13.1	Moderate	Zero	47.5	9.0
Total		100.0	28.5	Total		100.0	18.9
Low	High	8.8	1.9	Low	High	5.0	0.9
Low	Moderate	36.8	8.1	Low	Moderate	15.0	2.8
Low	Low	12.3	2.7	Low	Low	25.0	4.7
Low	Zero	42.1	9.2	Low	Zero	55.0	10.4
Total		100.0	21.9	Total		100.0	18.9
Zero	High	19.8	7.7	Zero	High	10.2	5.7
Zero	Moderate	25.7	10.0	Zero	Moderate	19.5	10.8
Zero	Low	19.8	7.7	Zero	Low	19.5	10.8
Zero	Zero	34.7	13.5	Zero	Zero	50.8	28.3
Total		100.0	38.8	Total		100.0	55.7

High, >|0.5|; Low, |0.1| to |0.2|; Moderate, |0.3| to |0.5|; zero is Δ=0.

poorer score for the face-to-face and teleconference settings, respectively (see online supplementary table S2). Table 2 provides an overall summary of the magnitude of change in the assigned primary reviewer score as compared to the magnitude of Δ for the secondary reviewer. For example, 1.5% of the time when a primary reviewer exhibited a high change in score (>|0.5|), the secondary reviewer exhibited a low shift in score (|0.1| to |0.2|). Online supplementary table S2 provides an overview of the occurrence of how often the primary or secondary reviewer improved, worsened (poorer score) or exhibited no change in score. Online supplementary table S3 provides summary information for Δ_{PRI}, Δ_{SEC} and Δ_{A}.

The median value of Δ_{PRI} was 0.0 for both settings. For Δ_{SEC}, 41.5% and 51.9% of scores did not change postdiscussion, while 20.0% and 17.5% shifted to a better score, and 38.5% and 30.7% shifted to a worse score for the face-to-face and teleconference settings, respectively. The median value for Δ_{SEC} was 0.0 for both settings. Thus, primary and secondary scores were more likely to remain the same after discussion in the teleconference setting compared to the face-to-face setting. However, if reviewers did change their score, primary and secondary reviewer scores were more likely to become poorer following discussion rather than better, regardless of setting (see online supplementary table S2). For the majority of applications (53.4% and 50.9% for face-to-face and teleconference settings, respectively), only one of the assigned reviewers shifted their scores postdiscussion. For the face-to-face reviews, 33.1% of applications involved the primary and secondary reviewers shifting scores, compared to only 20.8% for teleconference reviews (table 2 and online supplementary table S2). The occurrence of neither reviewer shifting scores postdiscussion was 13.5% for the face-to-face setting and 28.3% for the teleconference setting.

When examining the magnitude of score change (high=>|0.5|, moderate=|0.5| to |0.3|, low=|0.2| to |0.1|, or 0), table 2A and B demonstrate that there is a decrease in the magnitude shifts observed for assigned reviewers for the teleconference setting as compared to the face-to-face setting. For example, the proportion of applications where at least one assigned reviewer had a moderate to high score shift (greater than |0.1| to |0.2|) was 66.9% versus 45.8% for the face-to-face versus teleconference settings, respectively (table 2).

Observations of Δ_{A} reveal that the overall average opinion of applications often changed as a result of discussion. Just 20.4% and 22.6% of scores did not change postdiscussion, while 26.2% and 20.8% of scores shifted to a better score, and most applications, 53.5% and 56.6%, shifted to a worse score for the face-to-face and teleconference settings, respectively. Thus, $|\Delta_{A}| > 0$ for the majority of applications, regardless of setting (see online supplementary table S3). Further, for both settings, $|\Delta_{A}|$ represented low-magnitude shifts in score (|0.2| to |0.1|) for the majority of the applications (44.6% and 51.9% for the face-to-face and teleconference settings, respectively).

While the overall average Δ_{A} was −0.09±0.02 for the face-to-face setting and −0.10±0.02 for the teleconference

setting, statistically significant differences were observed between settings when the data were separated into subgroups of positive Δ_A (an improvement in score) and negative Δ_A (a worsening in score). The average positive Δ_A was 0.26±0.03 and 0.18±0.02 for the face-to-face and teleconference settings, respectively ($t(110)=2.37$; $p=0.02$). The average negative Δ_A was −0.30±0.02 and −0.24±0.02 for face-to-face and teleconference settings, respectively ($t(254)=2.37$; $p=0.02$). When grouping by common APS, Δ_A was poorly correlated to APS for settings ($R^2=0.15$ ($p=0.06$) and $R^2=0.21$ ($p=0.03$) for face-to-face and teleconference settings, respectively), showing that the EOD on scoring was not altered greatly by the quality of the application initially presented (see online supplementary figures S1A and B).

Score shifts and fundability status

Even though no relationship was found between Δ_A and APS and the majority of Δ_A values were the result of low shifts in application scoring, it is important to explore the impact of discussion-based score shifts on the fundability status of applications for these two review settings. Since PrX did not have a formal funding line, for this investigation we used an assumed funding line of 1.8 (similar to Martin *et al*[1]), meaning those applications scoring worse than 1.8 (1.9 or higher) would theoretically not be considered for funding. Using the assumed funding line, we observed that 10.0% of applications shifted over the funding line in either direction in the face-to-face setting versus 12.7% in the teleconference setting. The rest of the applications remained where their APS had placed them prior to discussion (either within or outside of the fundable range). For the applications reviewed in a face-to-face setting that shifted over the funding line, 34.6% moved to the fundable range following discussion (ie, had an unfundable APS score prior to discussion), and 65.4% moved out of the fundable range. In the teleconference setting, for the applications that shifted over the funding line, 29.6% moved into the fundable range following discussion, while 70.4% moved out of the funding range. Importantly, as noted above, the average APS for the teleconference setting was slightly lower (better score) than that for the face-to-face setting (2.2 vs 2.3, respectively), and the teleconference panels had a larger percentage of applications fall within the presumed funding line (15.4% for face-to-face vs 19.8% for teleconference) following discussion (table 3). When examining the magnitude of score shifts for those applications that moved in either direction over the funding line, 69.2% and 30.8% of these were moderate/high and low score changes, respectively, for the face-to-face setting, as compared to 48.1% and 51.9% for moderate/high and low score changes for the teleconference setting. It should be noted that shift ranges were determined based on rounding Δ_A to the 10ths place.

Contentiousness and EOD

The median Δ_{PS} was 0.0 for both review settings, with Δ_{PS} equal to 0 for 12.3% of applications for the face-to-face setting and 10.4% of applications for the teleconference setting. Conversely, Δ_{PD}, which is a representation of the difference between the primary and secondary reviewer scores postdiscussion, was 0 for 27.7% of applications for the face-to-face setting and 20.8% of applications for the teleconference setting. The median

Table 3 Magnitude of Δ_A as compared to the final overall score ranges for face-to-face (A) and teleconference (B) settings over the entire scoring range

Magnitude of change		1.0–1.2 (%)	1.3–1.5 (%)	1.6–1.8 (%)	1.9–2.0 (%)	2.1–2.5 (%)	2.6–3.0 (%)	3.1–3.5 (%)	3.6–3.9 (%)	Total (%)
(A) Final overall score ranges—face-to-face										
Less than −0.5	High	0.0	0.0	0.0	0.0	0.4	3.5	2.3	0.8	6.9
−0.5 to −0.3	Moderate	0.0	0.0	0.4	2.3	5.8	5.4	3.1	1.9	18.8
−0.2 to −0.1	Low	0.0	0.4	2.3	4.6	10.8	5.0	4.2	0.4	27.7
0	Zero	0.4	1.2	2.3	3.5	6.9	4.6	1.2	0.4	20.4
0.1 to 0.2	Low	0.4	2.7	3.1	2.7	3.1	2.7	1.9	0.4	16.9
0.3 to 0.5	Moderate	0.0	0.8	0.8	1.5	2.3	1.2	0.8	0.4	7.7
Greater than 0.5	High	0.0	0.4	0.4	0.0	0.8	0.0	0.0	0.0	1.5
Total		0.8	5.4	9.2	14.6	30.0	22.3	13.5	4.2	100
(B) Final overall score ranges—teleconference										
Less than −0.5	High	0.0	0.0	0.0	0.5	0.9	0.5	0.5	0.5	2.8
−0.5 to −0.3	Moderate	0.0	0.5	0.9	1.9	8.5	4.2	1.9	0.9	18.9
−0.2 to −0.1	Low	0.5	1.9	4.2	6.6	12.7	6.1	2.4	0.5	34.9
0	Zero	0.5	0.9	3.3	6.6	7.5	3.3	0.0	0.5	22.6
0.1 to 0.2	Low	1.4	0.5	4.7	1.9	5.7	2.4	0.5	0.0	17.0
0.3 to 0.5	Moderate	0.0	0.0	0.5	0.0	1.4	1.4	0.0	0.0	3.3
Greater than 0.5	High	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.5
Total		2.4	3.8	13.7	17.5	37.3	17.9	5.2	2.4	100

Δ_{PD} for both settings was 0.0. Further, examining Δ_{PS} compared to Δ_{PD} provides insight into exactly how discussion altered the assigned reviewer scores. For the face-to-face setting, the distance between the two assigned reviewer scores increased for 10.8% of applications, decreased for 73.2% of applications and remained the same for 16.9% of applications. For the teleconference setting, the distance between the assigned reviewer scores increased for 6.6% of applications, decreased for 61.3% of applications and remained the same for 32.1% of applications. In both settings, this is an indication that discussion resulted in the primary and secondary reviewers coming closer together in score for the majority of applications.

To examine reviewer contentiousness and its relationship to the magnitude of the EOD, we plotted common $|\Delta_{PS}|$ versus average $|\Delta_A|$, which revealed that there was a moderate correlation for the face-to-face setting (R^2=0.35; p=0.002). For the teleconference setting, there was a strong correlation (R^2=0.73; p<0.001) (figures 2A, B). These data demonstrate that, in general, the more contentious applications premeeting (large Δ_{PS}) resulted in larger score shifts following discussion, compared to those applications that were less contentious premeeting.

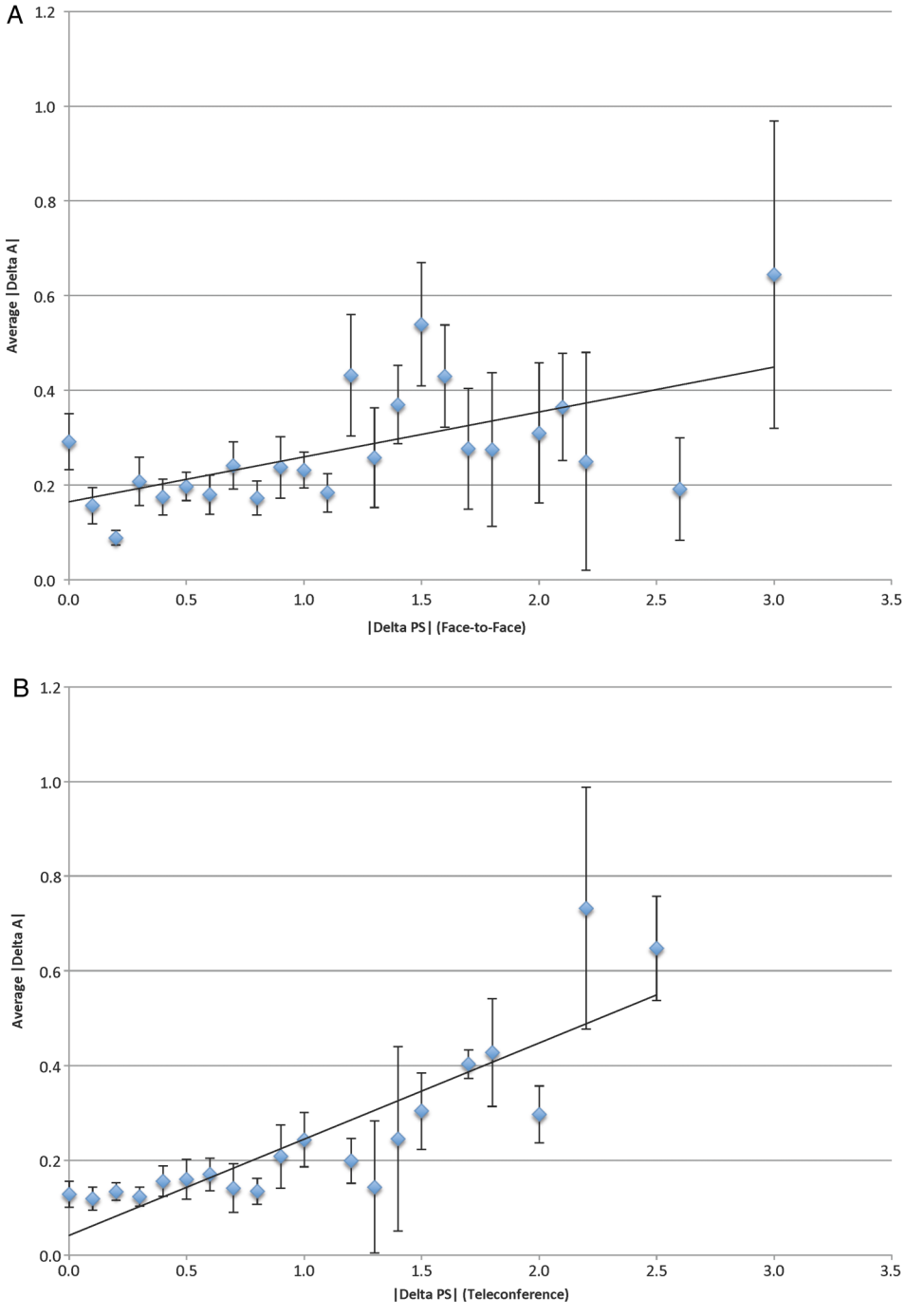

Figure 2 (A) Relationship between common $|\Delta_{PS}|$ and average $|\Delta_A|$ for face-to-face reviews in 2009 and 2010. (B). Relationship between common $|\Delta_{PS}|$ and average $|\Delta_A|$ for teleconference reviews in 2011 and 2012.

Application discussion time

The average panel discussion time per application was 23.9±0.7 min (face-to-face) and 20.0±0.7 min (teleconference). It should be noted that while this is not a large difference in discussion time, it is consistent with our prior results that teleconference panels exhibit shorter discussions. When grouping by common $|\Delta_{PS}|$ and comparing to average discussion time, no correlation was found for either the face-to-face (R^2<0.01; p=0.89) or the teleconference (R^2=0.02; p=0.58) setting. The same was found when plotting average discussion time versus common APS for both settings (R^2<0.01; p=0.81 for face-to-face; R^2<0.01; p=0.77 for teleconference) (data not shown). Grouping by common Δ_A (EOD) and plotting against average discussion time yielded low levels of correlation for both settings (R^2=0.19 (p=0.10) and R^2=0.22 (p=0.11) for face-to-face and teleconference settings, respectively) (figures 3A, B). The same is true for common $|\Delta_A|$ versus average discussion time (R^2=0.06 (p=0.52) and R^2=0.32 (p=0.14) for face-to-face and teleconference settings, respectively). Thus, discussion time neither varies with the contentiousness of reviewers nor is correlated with the EOD. Further, this demonstrates that, though teleconferences have shorter discussion times on average, overall discussion time does not seem to be an important difference between settings

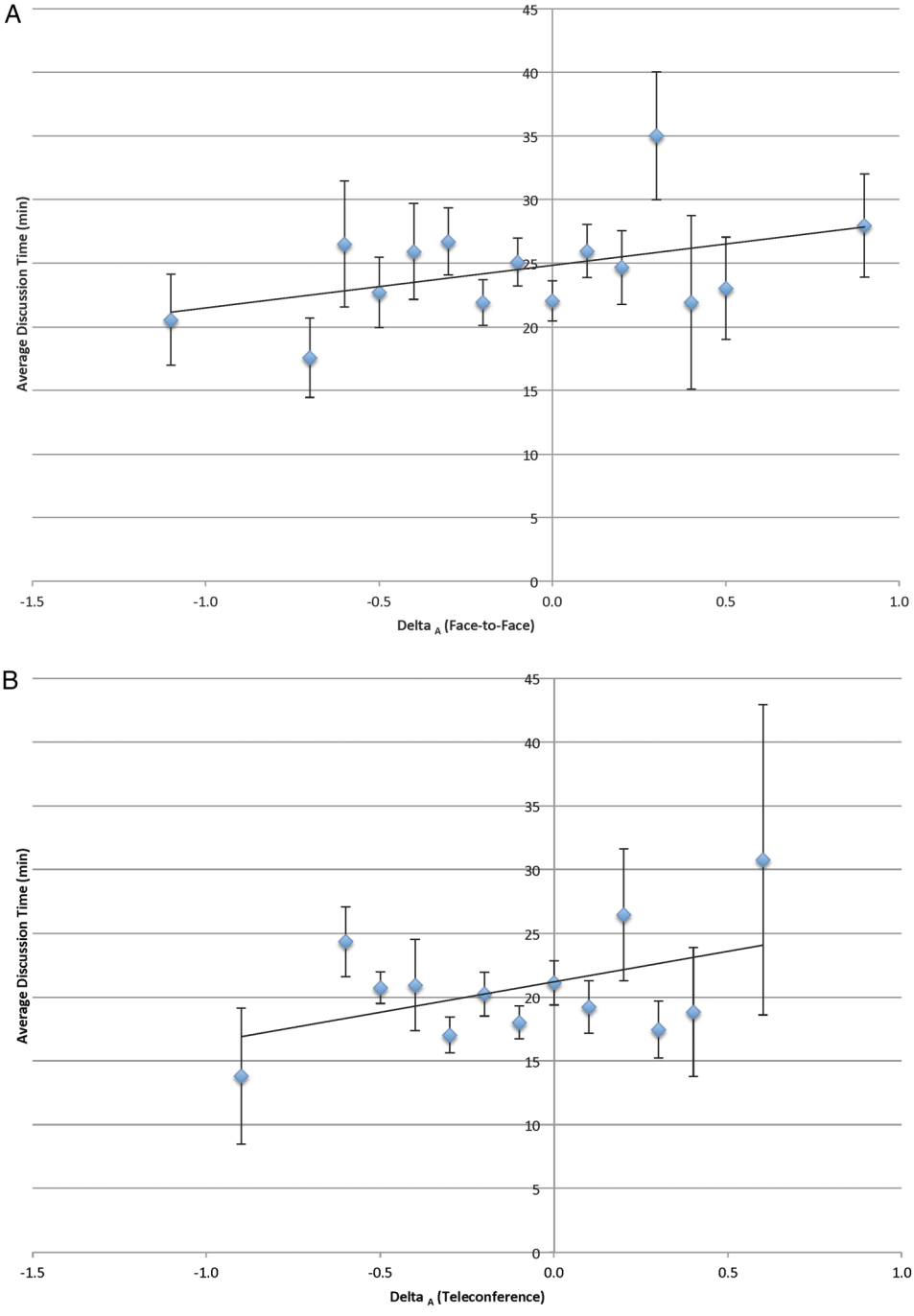

Figure 3 (A) Relationship between common Δ_A and average discussion time for the face-to-face settings. (B). Relationship between common Δ_A and average discussion time for the teleconference settings.

when it comes to application scoring. To uncover whether time of day was an important variable for discussion length (ie, perhaps morning discussions were longer than afternoon) we looked at those applications reviewed earlier in the day than later for review settings and found no real discernable difference (data not shown).

DISCUSSION

Although our studies provide insight on a number of nuances surrounding peer review, the overall analysis resulted in four important findings. First, the EOD, on average, was relatively small, in both settings. Second, there were small but statistically significant differences between the EOD for face-to-face versus teleconference panels. Third, discussion time was observed to have little influence on the magnitude of the EOD. Lastly, panel discussion was observed to more often result in negative, rather than positive, shifts in application scores.

As measured by Δ_A, the majority of applications, 65.0% and 74.5% (face-to-face and teleconference settings, respectively), had either low or no shift in scores postdiscussion, which is similar to what others have found.[4] Despite small adjustments in score being observed for a substantial amount of the applications, using our presumed funding line of 1.8, discussion was found to be of practical importance for 10.0% of applications in the face-to-face setting and 12.7% of applications in the teleconference setting. Further, for applications shifting over the funding line in either direction, we observed low-magnitude shifts in 30.8% of these applications in the face-to-face setting and 51.9% of applications in the teleconference setting. It should be noted that most of these applications moved outside of the fundable range, as opposed to within, following discussion for both settings. Thus, despite relatively low magnitude score shifts overall as a result of discussion, for a subset of applications, discussion played a vital role in determining potential funding status.

The magnitude of the EOD in the teleconference setting was found to be reduced compared to that of the face-to-face setting. This was observed not only in the mean squared error of the APS/OS fits and in the magnitude of negative and positive Δ_A, but also in the proportion of applications where there was a score shift by at least one assigned reviewer. One possible reason for this could potentially be related to the level of engagement between settings, with teleconference reviewers possibly being slightly less engaged than those participating onsite.[9] Some researchers have proposed that there is reduced task commitment, fewer status cues and less expressive behaviours in teleconference settings and that persuasive tasks are particularly susceptible to changes in communication setting.[7] Additionally, the reduced score shifts could possibly be explained by the greater anonymity in teleconference panels. In such settings, perhaps reviewers are more likely to conform to panel norms.[10–12] Further research must be conducted to explore the psychological motivations involved in postdiscussion scoring in onsite and teleconference scenarios.

We also observed that discussion time had little influence on the EOD. We found low to no correlation between average discussion time and APS, Δ_{PS}, and Δ_A, even with teleconference reviews having shorter discussion times on average. These results help to elucidate our prior findings, demonstrating that, even though teleconference reviews exhibit shorter discussion times, this does not seem to be an influencing variable on application scoring.[1] As noted above, one potential reason for this difference could be reduced engagement due to the absence of visual and other cues that are very apparent in face-to-face settings. As discussed, because of the decrease in cues within teleconference settings, there is likely an increase in social distance leading to a decrease in member engagement.[9] Further, the decreased cues could potentially be resulting in the use of more concise reviews in teleconference settings in addition to less side discussions overall. Future areas of exploration should include investigating differences in psychological quality of discussion in these two settings.

Lastly, one clear observation from our analysis was that score changes as a result of discussion were more often negative than positive, which is similar to what others have found.[4 13] Thus, an application's score is more likely to become worse rather than better following panel discussion. This was apparent regardless of setting or assigned reviewer type. It may be that persuading panel members of the potential merits of an application is more difficult than focusing on the less abstract, methodological weaknesses and logical flaws. In addition, studies have demonstrated that ambivalent individuals (presumably like the unassigned reviewers for PrX) are more likely to be persuaded by negative messages.[14] Other studies have even found that the actual level of ambivalence (high or low) someone exhibits also plays a role in whether they can be persuaded (or even resist persuasion) towards the message source (ie, the assigned reviewers).[15] Further research regarding persuasion and team decision-making during the peer-review process are needed.

In contrast with the results of our prior study,[1] this study suggests that there are only subtle differences between the outcomes of face-to-face and teleconference reviews. This seems to be consistent with findings in the literature that performance levels in teleconference settings are similar to face-to-face settings. Overall, as indicated above, further input from psychological research (on persuasion, team performance, etc) is needed to inform and guide future policy decisions regarding peer-review settings for grant application evaluations. This is especially important for ad-hoc panels like those used for PrX, which provide a continuously changing team dynamic from cycle to cycle, as compared to standing panels. Further studies should also be undertaken to evaluate the long-term outcomes of funded applications that were evaluated in face-to-face versus teleconference

settings. While our findings provide insight into the nuances of peer-review scoring and the effects of panel discussion, ultimately there is a need to determine what role the peer-review setting may have on innovative and impactful science.

Twitter Follow Stephen Gallo at @AIBS_SPARS

Acknowledgements The authors would like to thank Caitlin McPartland who aided in the technical editing of the manuscript. In addition, the authors would like to thank the AIBS SPARS staff for their dedication to advancing science and excellent work in implementing PrX reviews for over a decade as well as the managing editor and the two individuals who reviewed our article, as they provided valuable feedback.

Contributors ASC, SRG and SAG conceived and designed the experiments. ASC, JHS, AD and SAG performed the experiments. ASC and SAG analysed the data. ASC, JHS, AD, SRG and SAG wrote the paper.

Funding This research received no grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The anonymised data sets can be found on figshare: http://dx.doi.org/10.6084/m9.figshare.1495503.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Harmon J, Schneer JA, Hoffman LR. Electronic meetings and established decision groups: audioconferencing effects on performance and structural stability. *Organ Behav Hum Decis Process* 1995;61:138–47.
6. Graetz KA, Boyle ES, Kimble CE, *et al*. Information sharing in face-to-face, teleconferencing, and electronic chat groups. *Small Group Res* 1998;29:714–43.
7. Driskell JE, Radtke PH, Salas E. Virtual teams: effects of technological mediation on team performance. *Group Dyn* 2003;7:297–323.
8. 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.
9. 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.
10. Postmes T, Spears R. Deindividuation and antinormative behavior: a meta-analysis. *Psychol Bull* 1998;123:238–59.
11. Postmes T, Spears R, Lea M. Breaching or building social boundaries? SIDE-effects of computer-mediated communications. *Communic Res* 1998;25:689–715.
12. Cooke NJ, Hilton ML. *Enhancing the Effectiveness of Team Science*. Washington, DC: National Academies Press. 2015:152–3.
13. Fleurence RL, Forsythe LP, Lauer M, *et al*. Engaging patients and stakeholders in research proposal review: the patient-centered outcomes research institute. *Ann Intern Med* 2014;161:122–30.
14. Broemer P. Relative effectiveness of differently framed health messages: the influence of ambivalence. *Eur J Soc Psychol* 2002;32:685–703.
15. Cavazza N, Butera F. Bending without breaking: examining the role of attitudinal ambivalence in resisting persuasive communication. *Eur J Soc Psychol* 2008;38:1–15.

Supplementary Figures:

Figure S1a. Relationship between average pre-meeting scores (APS) and average Δ_A for face-to-face reviews in 2009 & 2010.

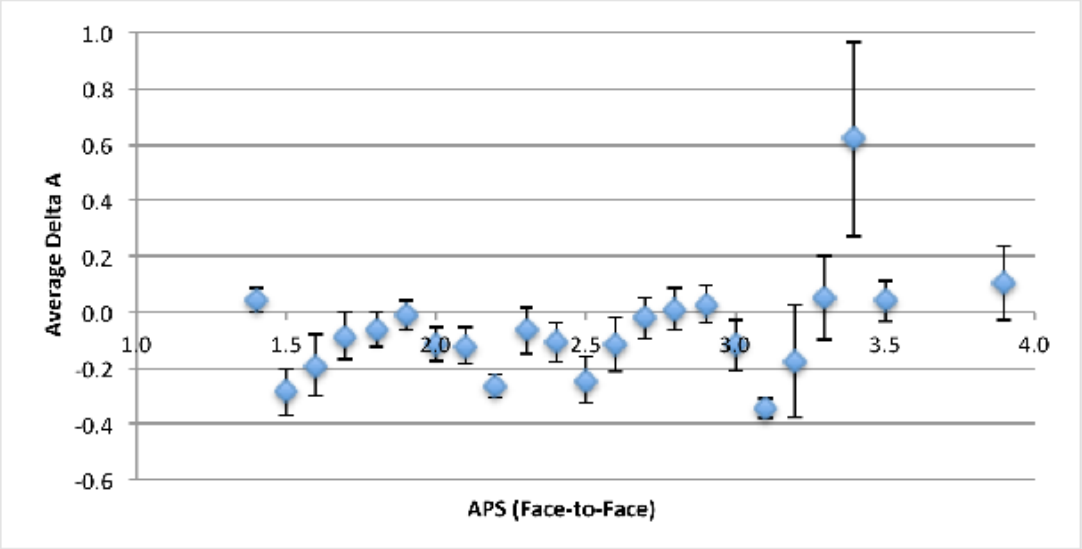

Figure S1b. Relationship between average pre-meeting scores (APS) and average Δ_A for teleconference reviews in 2011 and 2012.

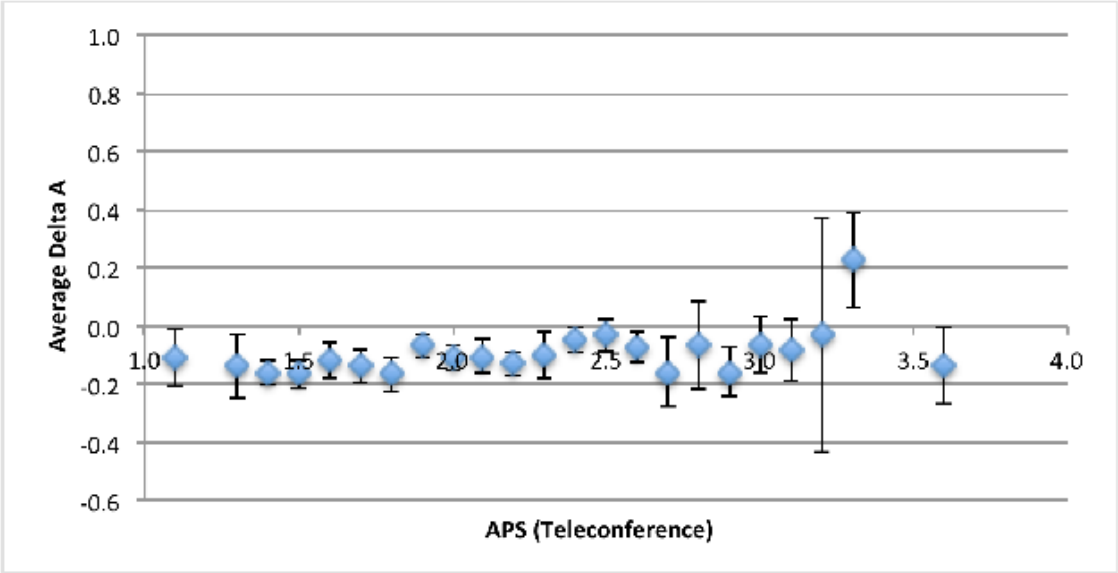

Table S1. Intraclass correlation summary information between primary and secondary reviewers for pre-meeting and post-discussion scores.

Year/Discussion	ICC	95% Confidence Interval	F statistic	P-value
2009 Pre	0.142	(-0.014, 0.291)	F(157,158) = 1.33	p = 0.0369
2009 Post	0.716	(0.631, 0.784)	F(157,158) = 6.04	P<0.001
2010 Pre	0.336	(0.153, 0.497)	F(101,102) = 2.01	P<0.001
2010 Post	0.763	(0.669, 0.834)	F(101,102) = 7.45	P<0.001
2011 Pre	0.190	(0.030, 0.341)	F(145,146) = 1.47	P = 0.0104
2011 Post	0.755	(0.676, 0.817)	F(145,146) = 7.15	P<0.001
2012 Pre	0.411	(0.191, 0.592)	F(65,66) = 2.40	P<0.001
2012 Post	0.650	(0.486, 0.770)	F(65,66) = 4.71	P<0.001
FTF (09&10) Pre	0.221	(0.102, 0.333)	F(259,260) = 1.57	P<0.001
FTF (09&10) Post	0.734	(0.673, 0.786)	F(259,260) = 6.53	P<0.001
TCON (11&12) Pre	0.262	(0.132, 0.382)	F(211,212) = 1.71	P<0.001
TCON (11&12) Post	0.720	(0.649, 0.779)	F(211,212) = 6.15	P<0.001

FTF = face-to-face; TCON = teleconference

Table S2. Changes in primary reviewer scores as compared to the changes of the secondary reviewer scores for the face-to-face (Table S2a) and teleconference (Table S2b) settings.

Table S2a. Face-to-Face

Primary Reviewer	Secondary Reviewer	% of Sub-Group	% of Entire Sample
Improved Score	Improved Score	18.8%	3.5%
Improved Score	Poorer Score	37.5%	6.9%
Improved Score	No Change	43.8%	8.1%
Total	Total	100.0%	18.5%
Poorer Score	Improved Score	19.8%	8.5%
Poorer Score	Poorer Score	33.3%	14.2%
Poorer Score	No Change	46.8%	20.0%
Total	Total	100.0%	42.7%
No Change	Improved Score	20.8%	8.1%
No Change	Poorer Score	44.6%	17.3%
No Change	No Change	34.7%	13.5%
Total	Total	100.0%	38.8%

Table S2b. Teleconference

Primary Reviewer	Secondary Reviewer	% of Sub-Group	% of Entire Sample
Improved Score	Improved Score	17.9%	2.4%
Improved Score	Poorer Score	32.1%	4.2%
Improved Score	No Change	50.0%	6.6%
Total	Total	100.0%	13.2%
Poorer Score	Improved Score	15.2%	4.7%
Poorer Score	Poorer Score	30.3%	9.4%
Poorer Score	No Change	54.5%	17.0%
Total	Total	100.0%	31.1%
No Change	Improved Score	18.6%	10.4%
No Change	Poorer Score	30.5%	17.0%
No Change	No Change	50.8%	28.3%
Total	Total	100.0%	55.7%

Table S3. Summary data for Δ_{PRI}, Δ_{SEC}, Δ_{A}

	Δ_{PRI} FTF	Δ_{PRI} TCON	Δ_{SEC} FTF	Δ_{SEC} TCON	Δ_{A} FTF	Δ_{A} TCON
Average Value	-0.11	-0.08	-0.10	-0.08	-0.09	-0.10
Median Value	0.00	0.00	0.00	0.00	-0.07	-0.07
Range of Values	-1.5 to 2.0	-1.2 to 1.1	1.5 to 2.4	-2.4 to 1.0	-1.1 to 1.0	-1.0 to 0.7
% of scores that improved (+Δ) / average / median	18.5% / 0.4 / 0.3	13.2% / 0.3 / 0.3	20.0% / 0.5 / 0.3	17.5% / 0.3 / 0.3	26.2% / 0.3 / 0.2	20.8% / 0.2 / 0.1
% of scores that stayed the same (Δ = 0)	38.8%	55.7%	41.5%	51.9%	20.4%	22.6%
% of scores that became worse (-Δ) / average / median	42.7% / -0.4 / -0.3	31.1% / -0.4 / -0.3	38.5% / -0.5 / -0.4	30.7% / -0.4 / -0.3	53.5% / -0.3 / -0.2	56.6% / -0.2 / -0.2
% of Δ in high / moderate / low ranges	10.8% / 28.5% / 21.9%	6.6% / 18.9% / 18.9%	15.4% / 26.5% / 16.5%	8.0% / 20.3 % / 19.8%	8.5% / 26.5% / 44.6%	3.3% / 22.2% / 51.9%

high shift = (>|0.5|; moderate shift = |0.3| to |0.5|; low shift = |0.1| to |0.2| Δ = 0 is included in row labeled "% of scores that stayed the same (Δ = 0)"

FTF = face-to-face; TCON = teleconference