Report No. 24-01
June 04, 2024

CSR Analysis of Summer 2023 In-person and Virtual Peer Review Meetings

Executive Summary

This report is one of a series of reports stemming from the Center for Scientific Review's ongoing effort to understand the impact of meeting format on the peer review process. The National Institutes of Health (NIH) depends on the Center for Scientific Review's (CSR) peer review process to ensure that NIH grant applications receive fair, independent, expert, and timely reviews that are free from inappropriate influences. When the COVID-19 pandemic hit in mid-March 2020, CSR shifted its review meetings online, using the Zoom.gov video meeting platform. In Fall 2022, CSR reimplemented in-person meetings and held 1/3 of its standing study sections, small business, and fellowship meetings in-person; the remainder of review meetings were held virtually. It was CSR's first opportunity to hold a substantial number of its regular review meetings both virtually and in-person at the same time. The comparison and analysis of Fall 2022 meetings, by meeting format, can be found here. The present report presents the data collected for the Summer 2023 review meetings (August and October 2023 Advisory Council round), which took place after two rounds of becoming re-acclimated to in-person meetings. During this time, CSR's policy to hold one of the three annual meetings of review panels in person remained in place; the policy applied to standing panels and recurring fellowship and small business panels.

CSR collected two types of data across standing study sections, small business, and fellowship panels to evaluate outcomes: participant survey data and multiple quantitative meeting measures. Survey data were designed to assess whether reviewers' observations of the quality of the review and meeting experiences differed according to the meeting format (i.e., in-person or virtual). Quantitative meeting measures were selected to evaluate roster characteristics and scoring practices.

Reviewer survey highlights

Overall, the survey data indicate that peer review is effective whether the review meeting is held in person or virtually. The vast majority of reviewers, regardless of meeting format, feel that their panel was able to prioritize applications based on scientific and technical merit and rated the overall quality of the review meeting as good to excellent. For some measures, such as effectiveness of the discussion and reviewer engagement, reviewers rated in-person meetings better than virtual meetings. However, all quality of meeting measures were strongly positive regardless of meeting format.. The largest difference between in-person and virtual meetings observed was in terms of reviewer engagement and attention span. Reviewers rated engagement more highly and were better able to sustain attention in in-person meetings; the effect size was moderate.

- Overall review quality and ability to prioritize applications according to their impact and scientific merit was rated highly by over 90% of reviewers in both virtual and in-person meetings. Compared to virtual reviewers, in-person reviewers were more likely to "strongly agree" that their panel was able to prioritize applications, although the magnitude of this effect is small.
- Over 90% of all reviewers thought the overall quality of their meeting was good or excellent. Reviewers who attended in-person meetings rated the meetings significantly more positive and of higher quality than those who attended virtual meetings, but the magnitude of most of these effects are small.

- Reviewers considered reviewer engagement to be less in virtual meetings than in in-person meetings; the differences were statistically significant with a moderate effect size. For other measures of meeting function including the effectiveness of discussion and the ability of the panel to prioritize applications, reviewers also rated in-person meetings better than virtual meetings, although the magnitude of these effects are small.
- Reviewer ratings of their ability to sustain attention were better for in-person vs. virtual meetings; the effect size was moderate.
- Reviewers rated their personal experience of participation better in in-person meetings than in virtual meetings, although the magnitude of these effects are small. For example, the frequency with which they contributed to discussion, and perceived receptivity of others to their opinions were rated more highly for in-person meetings.
- Reviewers in both virtual and in-person meetings prefer in-person over virtual format. Of those who express a preference, reviewers attending in-person meetings prefer to meet in person by a greater than 5:1 margin, while reviewers who attended virtual meetings prefer in-person meetings by a margin of approximately 2:1.

Quantitative meeting measures highlights

- There were no meaningful differences between in-person and virtual meetings in terms of the diversity of the ad-hoc reviewers recruited. Diversity of ad hoc reviewers was evaluated in terms of title (full, associate, assistant professor), geographic location of their place of employment, demographics (representation of underrepresented minorities or women), and their extent of prior service (focusing on new reviewers).

- Meeting format was associated with minor differences in scoring behavior. The statistically significant differences observed had a small to very small effect size.

    - Final overall impact scores were higher (worse) in in-person meetings overall, including in fellowship and small business panels.
    - More out-of-range scores were observed in in-person meetings, including standing study section and small business panels.

In the full report that follows, we present first the survey data followed by the quantitative meeting measures data.

Survey Data and Analyses

See Appendix A for detailed methods.

Survey Results

The survey was administered to 6,800 reviewers, of which 3,223 completed the survey for a response rate of 47%. Of the respondents, 58% attended virtual review meetings (n = 1,864), 37% attended in-person review meetings (n = 1,204), and 5% attended hybrid meetings (n = 155). The following results only include reviewers who attended virtual and in-person meetings. See Table 1 for reviewer characteristics.

Table 1. Reviewer Characteristics of Survey Respondents

Reviewer Characteristics		% Survey Respondents (n = 3,068)
Gender		
	Male	52
	Female	45
	Withheld	3
Race		
	American Indian or Alaskan	< 1
	Asian	23
	Black or African American	4
	More than one race	2
	Native Hawaiian or Pacific Islander	< 1
	White	62
	Withheld	9
Ethnicity		
	Hispanic/Latino	9
	Non-Hispanic	86
	Withheld	5
URM		
	No	78
	Yes	14
	Withheld	8
Career Stage		
	Professor	45
	Associate Professor	33
	Assistant Professor	17
	Other	5

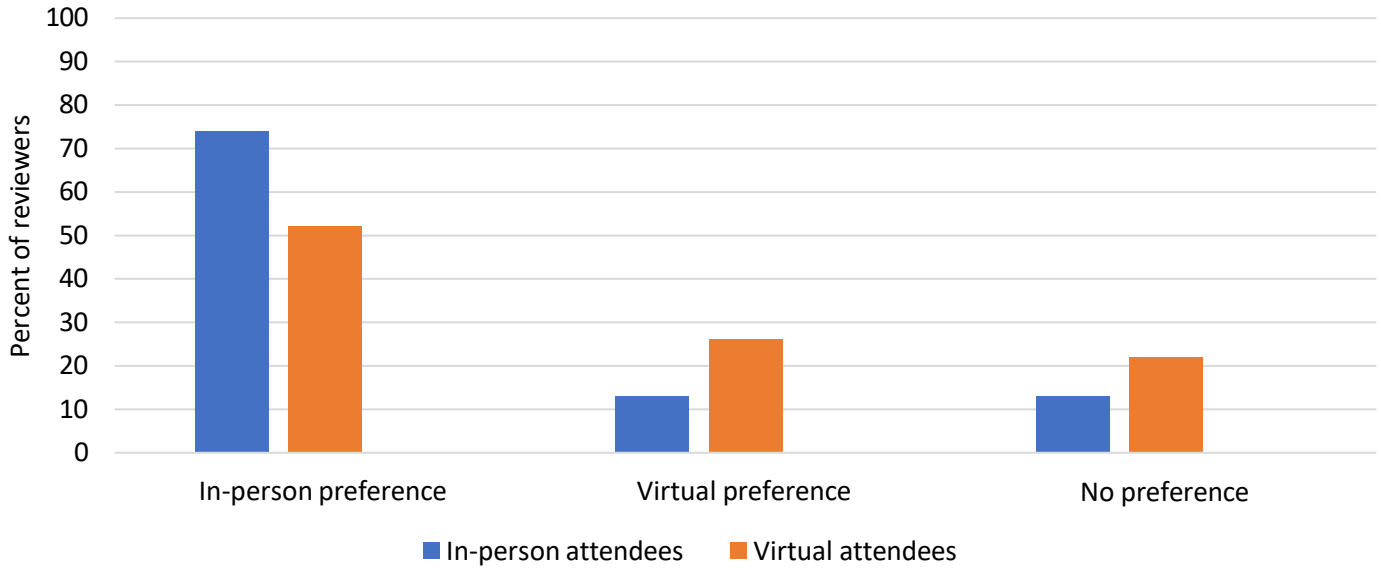
Meeting Format Preferences

The format of the meetings (i.e., in-person or virtual) was not assigned randomly to review meetings. This round of spring/summer meetings was the third and final opportunity for panels to meet in-person and thus to comply with CSR's policy to hold one meeting per year in person.

Figure 1 shows the meeting format preference of reviewers by the format of the meeting they attended.

- Reviewers overall prefer in-person meetings.
- In-person attendees were significantly more likely than virtual attendees to prefer in-person meetings and virtual attendees were significantly more likely to prefer virtual meetings than were in-person attendees.

Figure 1. Meeting Format Preference by Meeting Format



There was a significant association between the format of reviewers' meetings and their preferences ($\chi 2$ (2) = 142.583, p <.001), with a small effect size (ϕ_c = .216, p <.001). The results show that 1) the proportion of in-person attendees who preferred in-person meetings was significantly more than the proportion of virtual attendees who preferred in-person meetings, and 2) the proportion in-person attendees who preferred virtual meetings or had no meeting preference was significantly less than the proportion of virtual attendees who preferred virtual meetings or had no meeting preference.

Quality of Review

Figure 2 shows the distribution of reviewer perceptions of the quality of the review meeting by meeting format. Table 2 contrasts means (and standard errors) of ratings for each format.

- Over 90% of all reviewers thought the overall quality of their meeting was good or excellent.
- Reviewers who attended in-person meetings rated the meetings significantly more positive and of higher quality than those who attended virtual meetings—although the magnitude of most of these effects are small (see Table 2).
- Perceptions of reviewer engagement was the largest difference between formats, with in-person attendees reporting higher engagement at their meetings than reported by virtual attendees. The magnitude of this effect is medium.

Figure 2. Quality of Review by Meeting Format

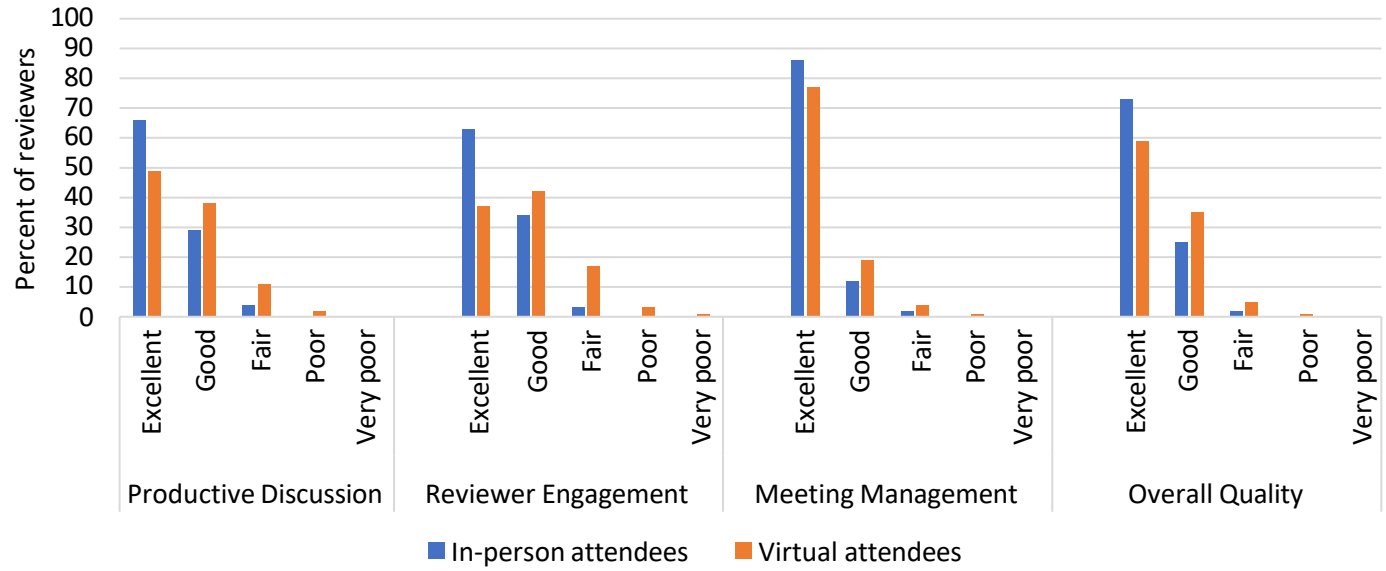


Table 2. Reviewers' Assessment of the Quality of the Review by Meeting Format			
	Virtual Meetings (*M, SE*)	In-person Meetings (*M, SE*)	Independent t-test Statistic and r
Productive Discussions	4.34 (.02)	4.61 (.02)	*t*(2956.07) = 11.275, *p* = .000; *d* = .40
Reviewer Engagement	4.12 (.02)	4.59 (.02)	*t*(3050.96) = 18.168, *p* = .000; *d* = .65
Meeting Management	4.71 (.01)	4.84 (.01)	*t*(3000.399) = 6.892, *p* = .000; *d* = .25
Overall Quality of Review	4.53 (.01)	4.70 (.02)	*t*(2885.305) = 8.441, *p* = .000; *d* = .30

Meeting Experience and Participation

Figure 3 shows reviewer ratings of their own experience and participation at the review meeting by meeting format. Table 3 contrasts means (and standard errors) of ratings for each format.

- In-person reviewers were more likely than virtual reviewers to report contributing to discussion always or often, 59% vs. 43%. Of the five measures of participation shown below, contribution to discussion was the measure for which the largest difference was observed between meeting formats.
- Compared to virtual reviewers, in-person reviewers reported more confidence in expressing their opinions, felt others were responsive to their feedback, and that they communicated clearly. These differences were statistically significant, although the magnitude of these effects are small (see Table 3).

Figure 3. Reviewer Experience and Participation by Meeting Format

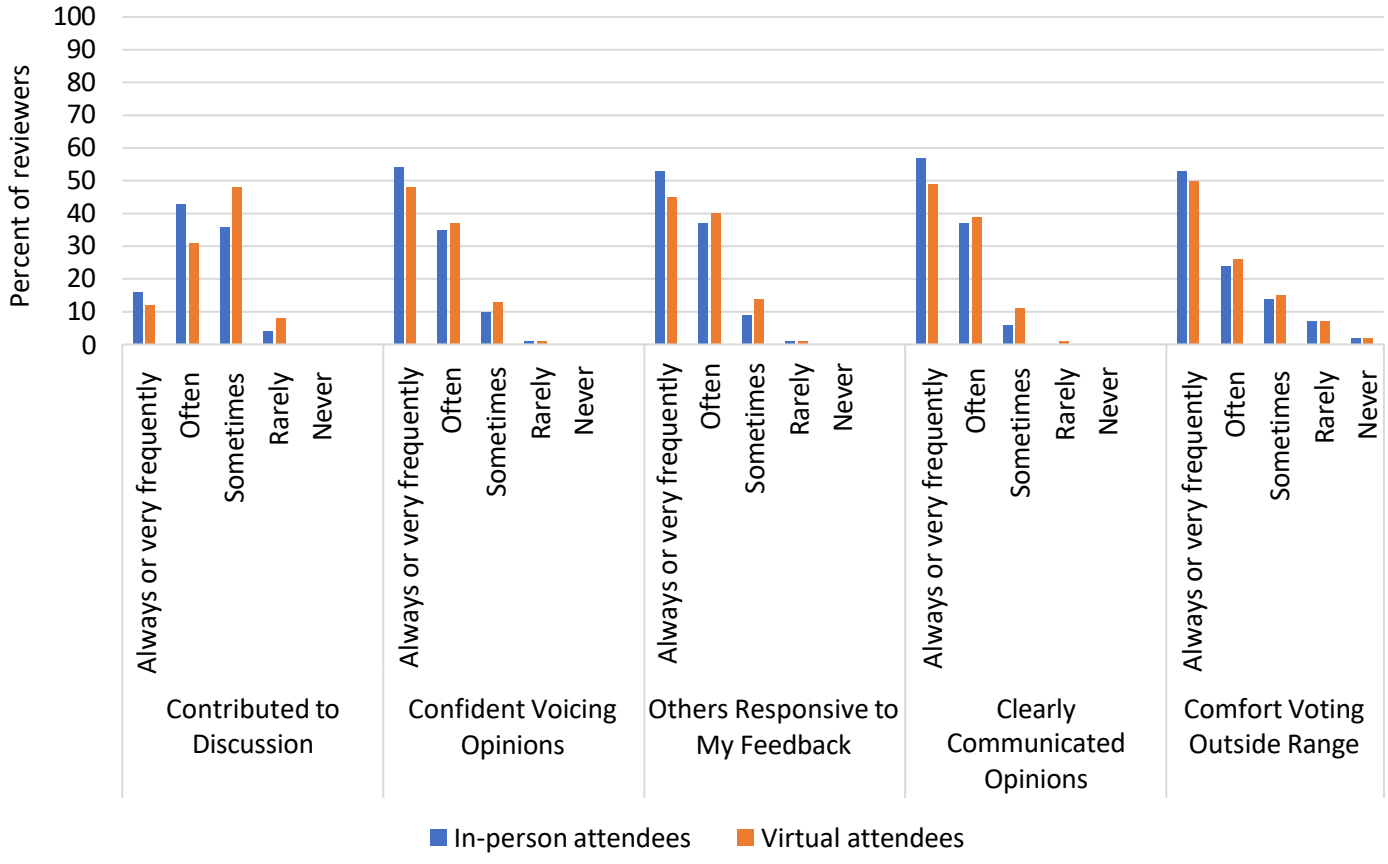


Table 3. Reviewers' Meeting Experience and Participation by Meeting Format			
	Virtual Meetings (*M, SE*)	In-person Meetings (*M, SE*)	Independent t-test Statistic and r
Contributed to Discussion	3.47 (.02)	3.70 (.02)	*t*(2615.39) = 7.639, *p* = .000; *d* = .28
Confident Voicing Opinions	4.33 (.02)	4.41 (.02)	*t*(3049) = 3.073, *p* = .002; *d* = .11
Others Receptive and Responsive to Feedback	4.28 (.02)	4.43 (.02)	*t*(2705.636) = 5.650 *p* = .000; *d* = .21
Clearly Communicated Opinions	4.37 (.02)	4.50 (.02)	*t*(2753.670) = 5.208, *p* = .000; *d* = .20
*Comfortable Voting Outside Range	4.15 (.03)	4.21 (.03)	*t*(2595) = 1.442, *p* = .149

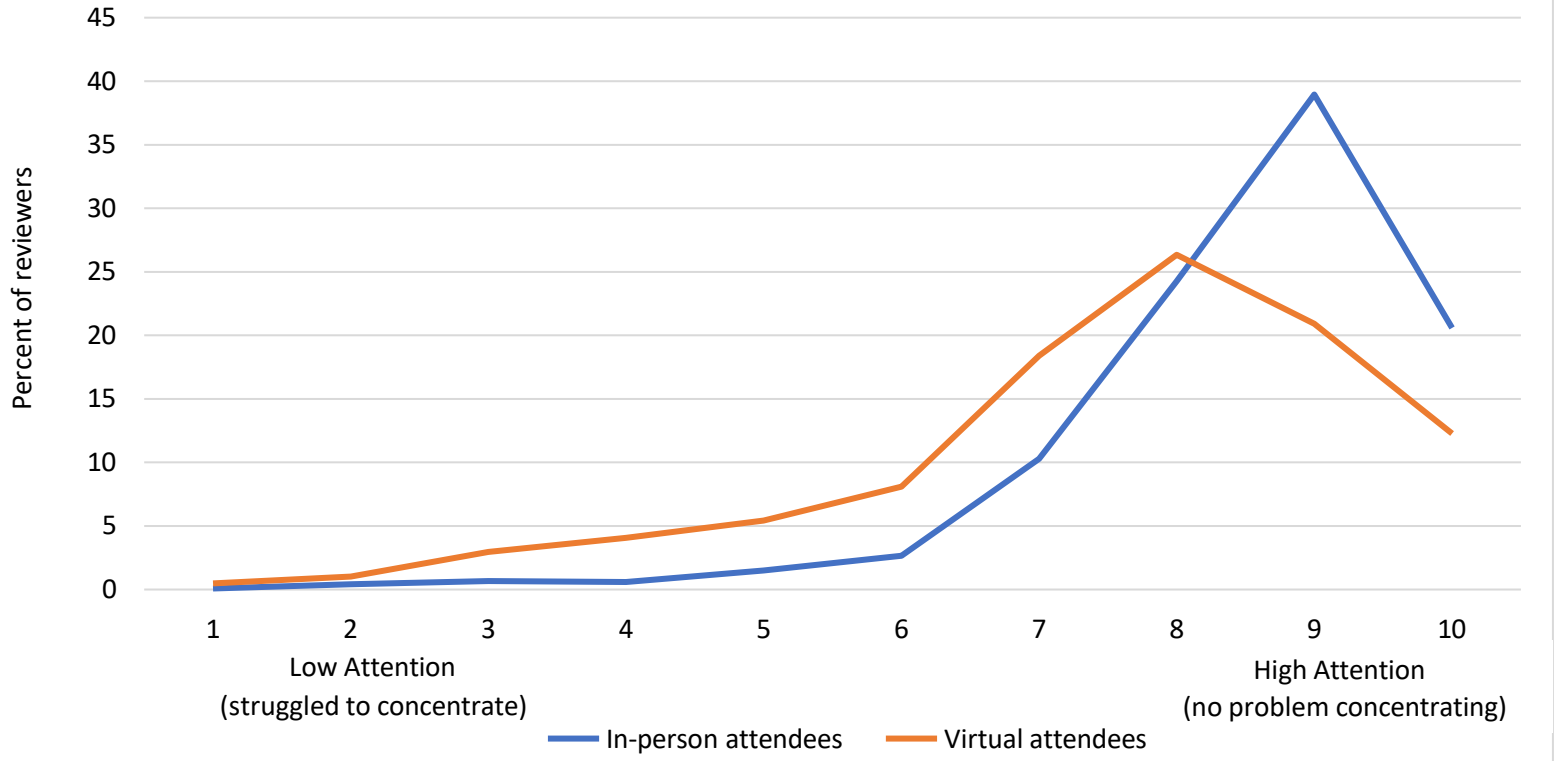
** For this result, significance testing is hampered by low power (32%), and a 68% chance of conducting a Type II error (i.e., a false negative).*

Attention Span

Figure 4 shows reviewers' report of their attention span by meeting format.

- Around 9% of virtual attendees and 2% of in-person attendees had difficulty concentrating at the meeting (i.e., less than 5 on attention scale).
- Around 60% of virtual attendees and 84% of in-person attendees were able to sustain their attention throughout the meeting (i.e., 8 or more on the attention scale).
- Overall, reviewers who attended in-person meetings paid significantly more attention or had significantly longer attention spans than those who attended virtual meetings—the magnitude of this effect is medium.

Figure 4. Reviewers Attention Span by Meeting Format



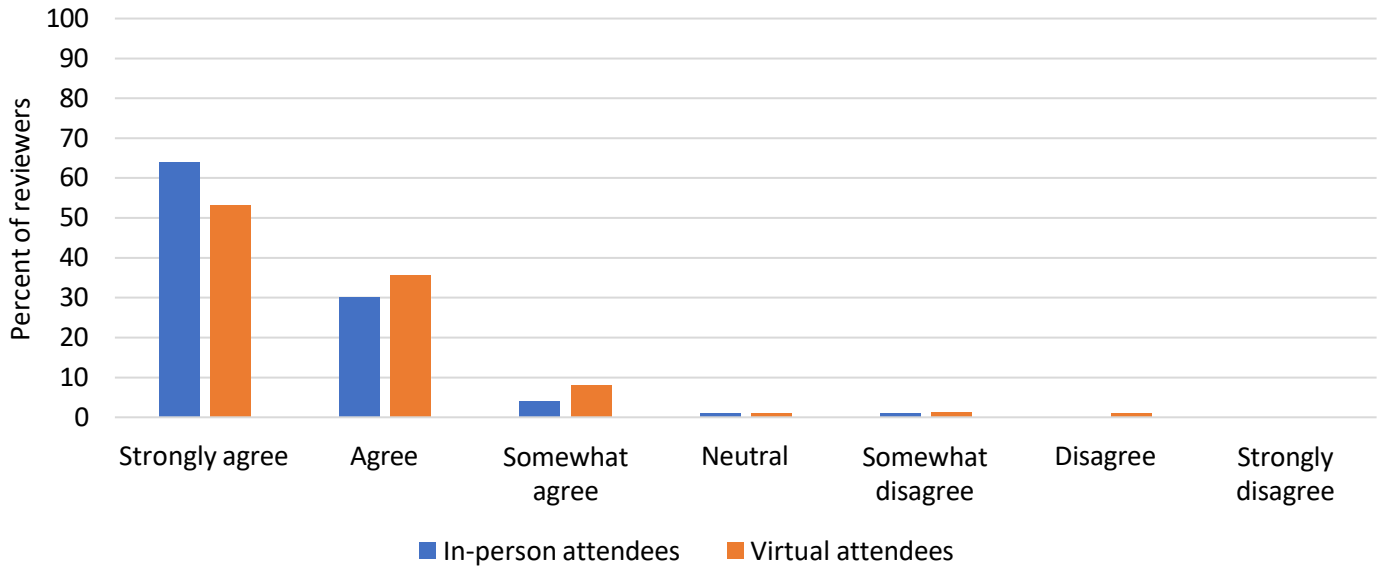
There was a significant difference between reviewers who attended the meeting in-person (M = 8.51, SE = .04) and those who attended virtually (M = 7.54 SE = .04) in their attention span or ability to concentrate at the meeting. This difference (.973, CI [.859, 1.094]) was significant (t(3040.685) = 16.887, p = .000) and represented a medium-sized effect (d = .60).

Prioritizing Applications

Figure 5 shows data on reviewers' perceptions of the panels' ability to prioritize applications by meeting format.

- 91% of all reviewers believed that the panel was able to prioritize applications according to their impact and scientific merit.
- Compared to virtual reviewers, in-person reviewers were more likely to strongly agree that their panel was able to prioritize applications—although the magnitude of this effect is small.

Figure 5. Reviewers' Ability to Prioritize Applications by Meeting Format



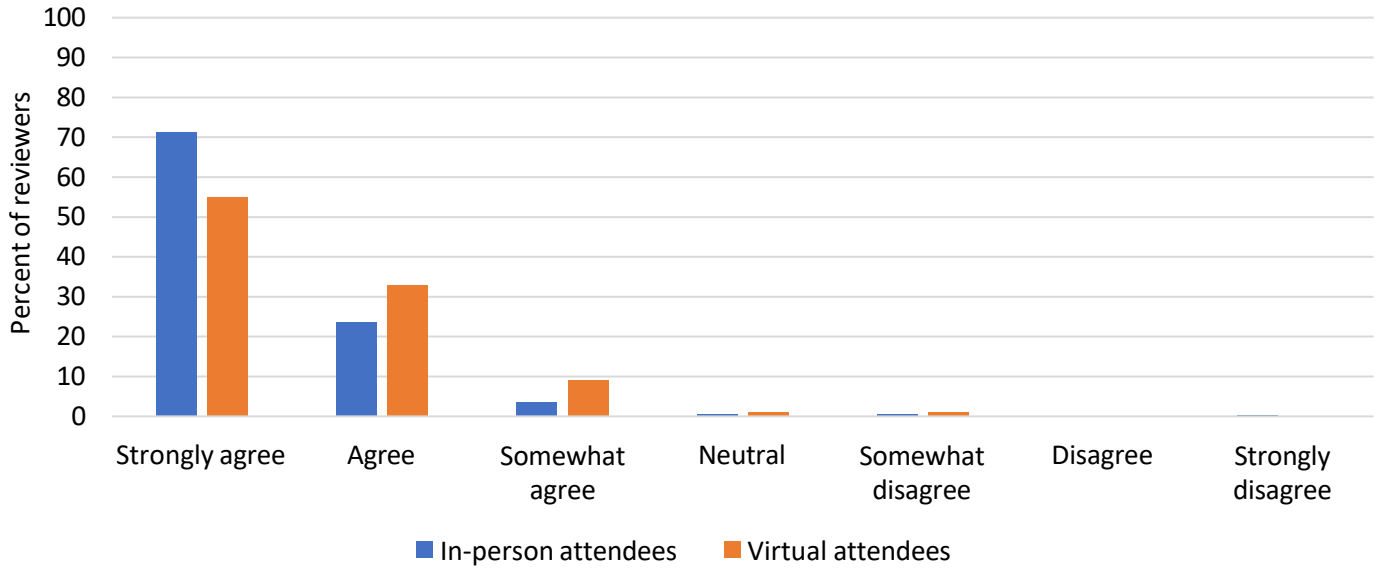
There was a significant difference in ability to prioritize applications between those who attended the meeting in-person (M = 1.45, SE = .02) and those who attended virtually (M = 1.63, SE = .85), This difference (-.174, CI [-.233, -.116]) was significant (t(2727.949) = -5.851, p = .000) with a small effect size (d = .22).

Discussion Quality

Figure 6 shows data capturing reviewers' perceptions of their panel's ability to discuss and evaluate applications by meeting format.

- 90% of all reviewers believed that the scientific discussions helped the panel evaluate the applications being reviewed.
- Compared to virtual reviewers, in-person reviewers were more likely to strongly agree that the discussions were helpful for evaluating the applications; the magnitude of this effect is small.

Figure 6. Discussions Helped with Evaluations by Meeting Format



There was a significant difference between reviewers who attended the meeting in-person (M = 1.37, SE = .02) and those who attended virtually (M = 1.64, SE = .02) in their perceptions on whether the discussions helped the panel evaluate the applications. This difference (-.276, CI [-.333, -.218]) was significant (t(2953.575) = -9.407, p = .000) and represented a small-sized effect (d = .33).

Meeting Measures Data and Analysis

Methods

The purpose of meeting measures data was to evaluate roster characteristics and scoring practices across the same sample as the meetings surveyed in this report. This section focuses on objective measures for the August and October 2023 Advisory Council review meetings, held during the summer of 2023. Data on rosters and scores for all standing study sections, fellowship, and small business panels (n=227) that occurred in these council rounds was collected and compared for meetings held in different formats. (This is departure from previous analyses of meeting format, where meeting measures were confined to a subsample of meetings.) Detailed methods can be viewed in Appendix B.

Approximately 35% of the meetings were held in-person, 61% were held as virtual meetings, and 4% were held as hybrid meetings. Hybrid meetings are excluded from this report due to small sample size and unique considerations. Table 4 shows the breakdown of meetings that were included in this analysis.

Table 4. A sample of meeting in 2023/10 & 08 council round included in the analysis

	In-person Meetings	Virtual Meetings	Total
Standing Study Section	63	105	168
Fellowship	8	17	25
Small Business	11	23	34
Total	82	145	227

Meeting Application Counts, Roster Sizes, and Reviewer Workload Trends

Meeting application counts

The average application count for in-person and virtual meetings were 70.0 and 67.5, respectively. Ninety-six percent of in-person meetings and 90% of virtual meetings were held over two days (vs. one). Table 5 shows the distribution of meetings according to size and meeting format. Table 6 shows the average number of applications by meeting format; there were no statistically significant differences in number of applications per meeting, whether examined by meeting format or by meeting type.

Table 5. Distribution of Number of applications, by meeting format in August and October 2023 Advisory Council

n of applications	In-person Meetings	Virtual Meetings	Grand Total
< 20	0	5	5
30-39	0	5	5
40-49	3	5	8
50-59	17	17	34
60-69	19	40	59
70-79	23	36	59
80-89	20	33	53
90-99	0	4	4
Grand Total	**82**	**145**	**227**

Table 6. Average of application counts in in-person and virtual meetings by meeting type			
Meeting Type	**In-person Meetings *(M, SE)***	**Virtual Meetings *(M, SE)***	**Independent t-test statistic**
All meetings	70.02 (1.89)	67.46 (1.89)	$t(214.9) = -1.36$, $p = 0.912$
Standing Study Section	71.03 (1.84)	70.29 (1.84)	$t(137.9) = -0.41$, $p = 0.657$
Fellowship	69.13 (7.61)	65.24 (7.61)	$t(21.2) = -0.51$, $p = 0.693$
Small Business	64.91 (6.13)	56.17 (6.13)	$t(31.8) = -1.43$, $p = 0.918$

Roster Size & Reviewer Workloads

In-person and virtual meetings did not differ in terms of number of reviewers recruited (roster size) or in the average number of applications assigned to each reviewer. There was also no difference in these two metrics between meetings of standing panels, fellowship panels, or small business panels.

Table 7. Average roster size in in-person and virtual meetings by meeting type			
Meeting Type	**In-person Meetings *(M, SE)***	**Virtual Meetings *(M, SE)***	**Independent t-test statistic**
All Meetings	29.05 (0.74)	28.38 (0.74)	$t(196.2) = -0.90$, $p = 0.816$
Standing Study Section	29.46 (0.74)	29.53 (0.74)	$t(124.3) = 0.10$, $p = 0.461$
Fellowship	26.50 (2.64)	24.18 (2.64)	$t(18.7) = -0.88$, $p = 0.805$
Small Business	28.55 (2.52)	26.22 (2.52)	$t(29.0) = -0.92$, $p = 0.819$

Table 8. Average reviewer workload in in-person and virtual meetings by meeting type			
Meeting Type	**In-person Meetings *(M, SE)***	**Virtual Meetings *(M, SE)***	**Independent t-test statistic**
All Meetings	7.66 (0.13)	7.46 (0.13)	$t(212.6) = -1.55$, $p = 0.938$
Standing Study Section	7.78 (0.12)	7.65 (0.12)	$t(124.5) = -1.04$, $p = 0.851$
Fellowship	7.83 (0.37)	7.91 (0.37)	$t(22.9) = 0.21$, $p = 0.419$
Small Business	6.90 (0.48)	6.25 (0.48)	$t(31.1) = -1.36$, $p = 0.908$

Roster Composition – characteristics of ad-hoc reviewers

Effects of meeting format on roster composition were assessed overall and were examined further by meeting type (standing study section, fellowships, and small business). Only ad hoc reviewers were analyzed in this section. The rationale behind selecting ad hocs for roster composition analyses is that standing members have committed to service throughout their term, and whether a meeting is held in-person or virtually in any given round is unlikely to influence their meeting attendance. Focusing on ad hoc reviewers allows the analyses to determine the effect of meeting formats on reviewer recruitment in relation to title (rank), geographic location, demographics, and their extent of prior service (focusing on new reviewers).

Reviewer Title (Rank), Geographic Location & Demographics

It has been suggested that senior reviewers may have a stronger preference for in-person meetings than do reviewers who are earlier in their career. Table 9 displays the proportion of ad hoc reviewers with different academic titles, by meeting format and meeting type. A Pearson chi-square test revealed no statistically significant differences by meeting format overall (all meetings combined) $\chi 2$ (3) = 1.809, $p = 0.613$. For standing study sections alone, there was a greater proportion of full professors, and a smaller proportion of assistant professors attending in-person review meetings relative to virtual meetings ($\chi 2$ (3) = 8.500, $p = 0.037$, with a small effect size ($\phi_c = 0.059$)). There were also statistically

significant differences in the titles of ad hoc reviewers attending small business panels, however the pattern is opposite of that observed for standing study sections – a greater proportion of full professors, and smaller proportion of assistant professors were recruited for virtual meetings ($\chi 2$ (3) = 7.836, p = 0.0495, with a small effect size (ϕ_c = 0.092)). The proportion of associate professors attending in-person or virtual meetings is comparable for both standing study sections and small business panels. Finally, there were no statistically significant differences observed in the titles of ad hoc reviewers attending in-person vs. virtual fellowship meetings ($\chi 2$ (3) = 2.444, p = 0.486).

Table 9. Distribution of title rank for ad hocs, by meeting format and type

	All Meetings		Standing Study Section		Fellowship		Small Business	
	In-person	Virtual	In-person	Virtual	In-person	Virtual	In-person	Virtual
Professor	28.8%	27.1%	28.2%	23.4%	42.9%	41.4%	20.8%	26.7%
Associate Professor	35.3%	35.1%	36.6%	36.9%	43.4%	42.1%	26.2%	26.0%
Assistant Professor	25.6%	26.9%	30.9%	35.2%	9.9%	13.9%	21.1%	14.8%
Other	10.3%	10.8%	4.3%	4.5%	3.8%	2.7%	31.9%	32.5%

Because virtual meeting formats remove travel as a participation barrier, it has been suggested those who reside far from the meeting location may be more likely to agree to review for virtual meetings. At the time of this report, all in-person meetings are held in the Washington-DC area, so one may expect to see a higher proportion of reviewers from the West or central regions of the U.S. in virtual meetings than in in-person meetings. Table 10 displays the region of the country in which ad hoc reviewers were employed by meeting format. A Pearson chi-square test revealed no statistically significant differences in geographic representation of ad hoc reviewers between the two meeting formats ($\chi 2$ (3) = 0.133, p = 0.988). Further, there were no statistically significant differences in this metric between in-person and virtual meetings when examined by meeting type (standing study sections $\chi 2$ (3) = 1.528, p = 0.676, fellowship $\chi 2$ (3) = 0.978, p = 0.807, small business $\chi 2$ (3) = 4.814, p = 0.186).

Table 10. Region of the U.S. in which ad hoc reviewers were employed, by meeting format and type

	All Meetings		Standing Study Section		Fellowship		Small Business	
	In-person	Virtual	In-person	Virtual	In-person	Virtual	In-person	Virtual
South	33.1%	32.6%	33.2%	31.9%	33.3%	35.9%	32.6%	32.0%
East	25.5%	25.7%	25.1%	27.3%	21.9%	23.3%	29.0%	23.3%
Central	22.5%	22.8%	23.2%	22.4%	24.3%	22.9%	19.4%	23.8%
West	19.0%	18.9%	18.6%	18.4%	20.5%	17.9%	19.0%	20.8%

It has been suggested that alternative review formats, such as virtual or hybrid, may have positive effects on the inclusion of women, persons with health issues or disabilities, and other minorities underrepresented in biomedical sciences (URM). For example, those who traditionally may have had greater caregiving responsibilities, such as women with young children, may have found it difficult to travel to an in-person review meeting.

Table 11 displays the inclusion of URM ad hoc reviewers by meeting format. Those whose URM status is unknown (n=540) were excluded from these analyses. Following the Notice of NIH's Interest in Diversity (OD-20-031), we defined URMs as individuals that identify as: Black or African American, Hispanic or Latino, American Indian or Alaska Native,

Native Hawaiian, and other Pacific Islander. A Pearson chi-square test revealed no statistical differences for the inclusion of URM between the two meeting formats ($\chi 2$ (1) = 1.669, p = 0.196). A further examination of URM representation on the roster by meeting type showed no significant differences between meeting formats (standing study sections $\chi 2$ (1) = 0.108, p = 0.742; fellowships $\chi 2$ (1) = 1.131, p = 0.288; small business $\chi 2$ (1) = 1.778, p = 0.182).

Table 11. Average ad hoc URM representation on meeting rosters		
	In-person Meetings	**Virtual Meetings**
Average URM representation	15.5%	14.0%
Standing Study Sections	14.3%	13.8%
Fellowships	18.4%	15.0%
Small Business	17.1%	13.7%

Table 12 displays the inclusion of women ad hoc reviewers by meeting format. Those ad hoc reviewers whose gender is unknown (n=81) were excluded from these analyses. In terms of the gender of ad hoc reviewers, no differences in meeting rosters were found. The distribution of gender between the two meeting formats was not significantly different ($\chi 2$ (1) = 0.577, p = 0.447). A further examination of gender by meeting type also found no differences between meeting formats (standing study sections $\chi 2$ (1) = 1.040, p = 0.308, fellowships $\chi 2$ (1) = 0.064, p = 0.800, small business $\chi 2$ (1) = 0.009, p = 0.923).

Table 12. Average ad hoc women representation		
	In-person Meetings	**Virtual Meetings**
Average women representation	42.8%	44.1%
Standing Study Sections	44.2%	46.4%
Fellowships	39.3%	40.4%
Small Business	41.3%	40.9%

New Reviewer Recruitment

It has been suggested that virtual meeting formats may remove travel as a participation barrier – incentivizing new reviewers to engage in the peer review process who were unable, opposed to the travel, or ambivalent about participating in in-person meetings. We found no effect of meeting format on the recruitment of ad hoc reviewers with relatively less experience. Tables 13-14 shows the average number of ad-hoc reviewers with little to no review service that were recruited to serve on the panels. A t-test did not reveal any statistical differences on the number of ad hoc reviewers with little (1-2 prior meetings) to no prior review service participating between the two meeting formats. Further, when examining ad hoc new reviewer participation by meeting type, there were no statistically significant findings between meeting formats.

It should be noted that while standing study sections have standing members that are excluded from this analysis, small business panels and fellowships do not have standing members, so the entire panel of ad-hoc reviewers is accounted for those meetings. There are other policy differences between meeting types. CSR has a policy that two Early Career Reviewers (ECRs; those enrolled in the CSR Early Career Reviewer Program) must be recruited for each meeting of a standing panel and, conversely, ECRs are not allowed in small business or fellowship panels. ECRs were included in this analysis and should be considered when interpreting extent of prior review amongst recruited ad hoc reviewers.

Table 13. New Reviewer Recruitment (0 prior reviews)	In-person Meetings *(M, SE)*	Virtual Meetings *(M, SE)*	Independent t-test statistic
All Meetings	3.33 (0.26)	3.21 (0.26)	$t(166.3) = -0.45$, $p = 0.672$
Standing Study Sections	3.49 (0.24)	3.55 (0.24)	$t(136.7) = 0.71$, $p = 0.240$
Fellowships	2.25 (0.97)	1.41 (0.97)	$t(9.2) = -0.86$, $p = 0.796$
Small Business	3.82 (1.06)	3.00 (1.06)	$t(16.7) = -0.77$, $p = 0.775$

Table 14. New Reviewer Recruitment (1-2 prior reviews)	In-person Meetings *(M, SE)*	Virtual Meetings *(M, SE)*	Independent t-test statistic
All Meetings	3.34 (0.33)	3.71 (0.33)	$t(168.6) = 1.12$, $p = 0.131$
Standing Study Sections	2.95 (0.31)	3.23 (0.31)	$t(130.1) = 0.88$, $p = 0.189$
Fellowships	2.88 (1.09)	4.65 (1.09)	$t(22.4) = 1.63$, $p = 0.058$
Small Business	5.91 (1.11)	5.22 (1.11)	$t(15.3) = -0.62$, $p = 0.728$

Scores

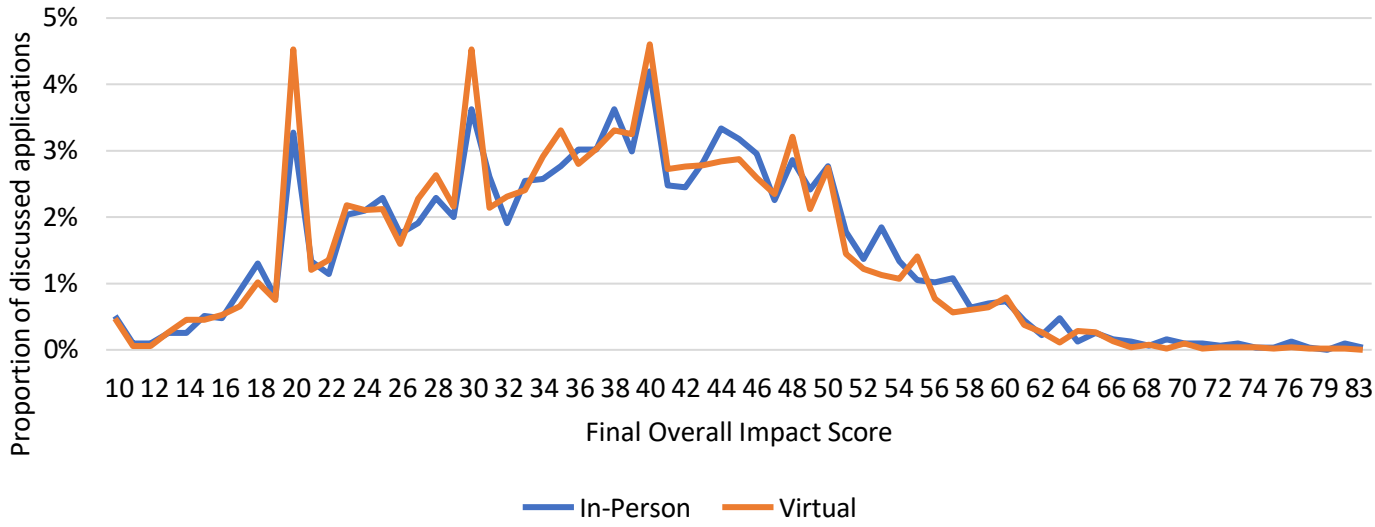
It has been suggested that review quality may suffer in virtual formats – reviewers (especially unassigned reviewers) may be less engaged or distracted in their remote environments, there may be decreased rapport between panel members and less open discussion, and reviewers may be affected by Zoom fatigue. This section assesses scoring behaviors between meeting formats as an indicator of review quality, engagement, and panel function.

Final Score distribution

It has been suggested that virtual meeting may have impact on final overall impact scores. Table 15 shows the descriptive statistics of final overall impact scores for discussed applications only for in-person and virtual meetings, and Figure 7 shows the distribution of final overall impact scores. Differences in the distribution final overall impact score between meeting formats was to be found statistically significant, overall, with a small effect size. A further examination of final overall impact score distribution between in-person and virtual meetings by meeting type were performed (Table 15). Differences in the final overall impact scores between in-person and virtual were statistically significant in fellowship and small business panels with small effect size, but not in standing study sections.

Table 15. Final Overall Impact Score Distribution, discussed applications only, by meeting type and format			
	In-Person *M (SE) [SD]*	**Virtual** *M (SE) [SD]*	**Statistical Testing**
All	38.02 (0.21) [12.0]	37.12 (0.16) [11.5]	$F_{1, 6381.2} = 11.5125$ $p = 0.0007$ with $d = 0.07$
Standing Study Section	38.2 (0.25) [12.2]	37.8 (0.18) [11.5]	$F_{1, 4929.3} = 1.9782$ $p = 0.1596$
Fellowship	33.8 (0.65) [11.6]	31.0 (0.43) [10.8]	$t(586) = -3.55$ $p = 0.0002$ with $d = 0.25$
Small Business	40.2 (0.51) [10.0]	38.9 (0.39) [10.3]	$t(823.2) = -2.04$ $p = 0.0209$ with $d = 0.13$.

Figure 7. Distribution of Final Overall Impact Scores of Discussed Applications only, In-Person vs. Virtual meetings



The absolute value of changes from preliminary to final score entered by assigned reviewers was evaluated by meeting format and type. No differences were found (Table 16).

Table 16. The mean and standard deviation of the absolute value of changes of assigned reviewers' preliminary scores to final scores, by meeting format and type

	In-person *M (SE)*	Virtual *M (SE)*	t-test Statistic
All meetings	0.46 (0.01)	0.45 (0.01)	$t(25488)=-0.801$, $p=0.423$
Standing Study Sections	0.46 (0.01)	0.45 (0.01)	$t(15435.3)=-1.014$, $p=0.311$
Fellowships	0.47 (0.02)	0.47 (0.02)	$t(1869.4)=-0.290$, $p=0.772$
Small Business	0.48 (0.02)	0.49 (0.02)	$t(2364.6)=0.281$, $p=0.779$

Out of Range Scores

It has been suggested that in virtual meetings, fewer unassigned reviewers would score out of the range set by the assigned reviewers, as they may not be engaged enough in discussion to elucidate differences of opinion. Figure 8 shows the percentage of final scores that were out of range for in-person and virtual meetings. There is a significant association between meeting format and scoring outside of the range ($\chi 2$ (1) = 76.578, p <.0001) with a small effect size (ϕc = 0.02).

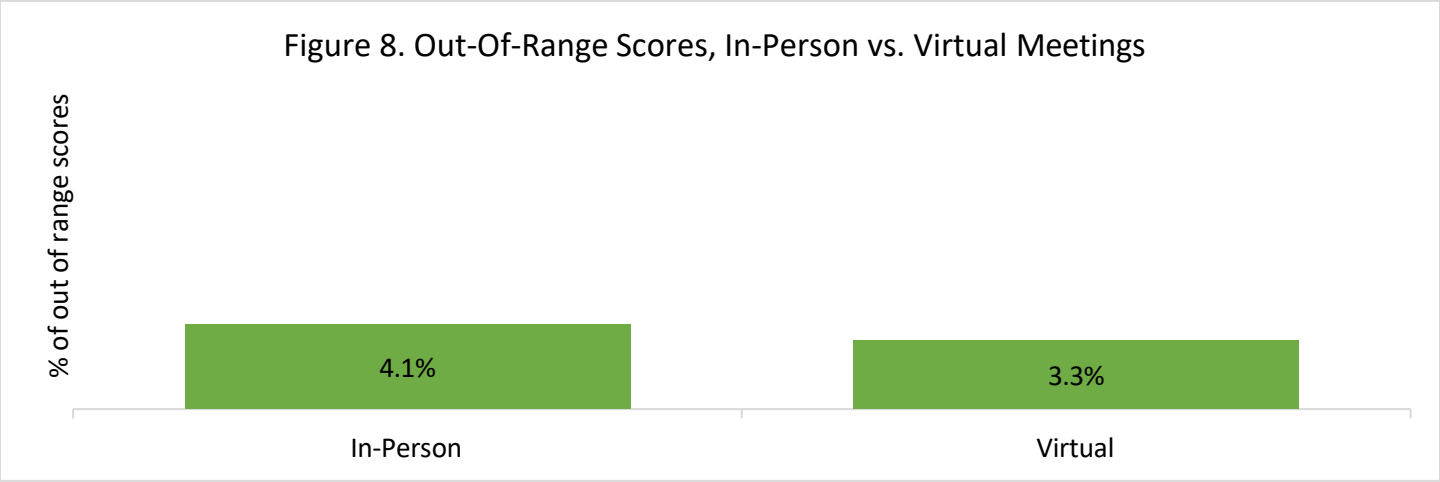


Table 17 shows the number of scores and the percentage of out-of-range scores for in-person and virtual meetings, by meeting format and type. For standing study section meetings, there is significant association between meeting format and scoring outside of the range *$\chi 2$ (1) = 38.102, p <.0001) but the effect size is very small (ϕc = 0.02). In small business panels, there is a significant association between meeting format and scoring outside of the range ($\chi 2$ (1) = 73.970, p <.0001) with a small effect size (ϕc = 0.05). Significant differences in out-of-range scoring were not seen in fellowship panels ($\chi 2$ (1) = 1.128, p =0.2883).

Table 17. Extent of out-of-range scores, in-person vs. virtual by meeting type

	Standing Study Section		Fellowship		Small Business	
Meeting Type	**In-person**	**Virtual**	**In-person**	**Virtual**	**In-person**	**Virtual**
n of scores	59258	96913	7135	13510	9316	16596
% of OOR scores	3.6%	3.0%	4.9%	4.5%	6.6%	4.2%
% of unfavorable scores that were OOR	3.1%	2.7%	4.5%	4.1%	5.8%	3.8%

Conclusions for meeting measures

- It has been suggested that senior reviewers may have a stronger preference for in-person meetings than do junior reviewers. The data provide minimal support for this idea.
    - There were statistically significant differences in the distribution of title rank in standing study section meetings and small business meetings. However, in standing study section meetings, proportionately more full professors and fewer assistant professors participated in in-person meetings than in virtual meetings but opposite was true for small business meetings. Effect sizes for both meeting types were small.
    - There was no evidence that the distribution of title rank in fellowship meetings differed between meeting formats.
- It has been suggested that virtual meeting formats may remove travel as a participation barrier, especially for those who reside far from the meeting location. However, the distributions of reviewers from U.S. geographical regions did not differ between virtual and in-person meetings.
- It has been suggested that a virtual format may have positive effects on the inclusion of women, persons with health/issues/disabilities, and other underrepresented minorities (URM) in peer review. However, the analysis

of rosters from in-person versus virtual meetings showed no differences in the demographics of ad hoc reviewers.

- It has been suggested that virtual meeting format may remove travel as a participation barrier, perhaps incentivizing new reviewers. However, the proportion of new reviewers did not differ by meeting format.
- It has been suggested that review quality suffers in virtual environments. Meeting format was associated with significant differences in scores, albeit with small to very small effect sizes. Specifically:
    - Final overall impact scores were higher (worse) in in-person meetings overall as well as in fellowship and small business panels.
    - Preliminary scores of assigned reviewers were higher (worse) in in-person meetings overall, and in fellowship meetings as well. However, an analysis of the change in score from preliminary to final score entered by assigned reviewers showed no effect of meeting format or type on the absolute value of the change.
    - There was more variability in scores in in-person meetings. This was descriptively true for all meeting types but was not for fellowship meetings.

  These findings may be interpreted as showing greater panel engagement during in-person discussions. More discussion is observationally associated with worse review outcomes. A larger variability in scores is consistent with a more engaged panel in which a wider range of views may be expressed.
- It has been suggested that virtual meetings would have fewer unassigned reviewers scoring out of range, as reviewers may not be engaged enough in discussion to elucidate disagreements with the assigned score range. Differences in out-of-range scores between meeting format were found; effects sizes were very small.
    - For standing study section and small business panels, differences in out-of-range scores were observed but the effect size was very small. Meeting format was not found to affect out-of-range scoring in fellowship panels.

Appendix A. Detailed Methods for Survey Data

Participants

Respondents were reviewers who participated in 226 CSR review meetings (n = 3,068) between May 31st to August 2nd, 2023. The review meetings included standing study section panels, and recurring small business and fellowship special emphasis panels (SEPs).

Survey Administration

Reviewers were asked for their participation in a survey via email on the last day of the review meeting. The email contained a weblink to the survey. Reviewers were told in the email that their responses would be kept confidential and that the survey would take less than five minutes to complete. All surveys returned by August 18th, 2023 were included for analysis.

Measures

Application Evaluation

Two items asked participants to rate on a scale from 1 (strongly agree) to 7 (strongly disagree) the panel's ability to evaluate the applications: 1) the panel was able to prioritize applications according to their impact and scientific merit, and 2) the scientific discussion helped the panel evaluate the applications being reviewed.

Peer Review Quality

Four items asked participants to rate on a scale from 1 (very poor) to 5 (excellent) the following items: 1) overall quality of review, 2) productivity of discussions, 3) level of reviewer engagement, and 4) meeting management.

Reviewer Meeting Experience and Participation

Five items asked participants to rate on a scale from 1 (never) to 5 (always or very frequently) the following items: 1) I contributed to the discussion, 2) I felt confident voicing my opinions, 3) I felt others were receptive and responsive to my feedback, 4) I was able to clearly communicate opinions, and 5) I felt comfortable voting outside the range.

Attention Span

One item asked participants to rate their attention span at the review meeting using a scale from 1-10, with 1 being "really struggled to concentrate" and 10 being "no problem concentrating at all".

Format Preference

One question asked participants if there were no or minimal health risks from COVID-19, would they be more likely to participate in a review meeting if it was held face to face or over Zoom/video? Response options included: face-to face, Zoom/video, and no preference.

Demographic Information

Four questions were used to collect the demographic characteristics of respondents. 1) *Gender*: male, female, I prefer not to respond; 2) *Race:* American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Pacific

Islander, White, I prefer not to respond; 3) *Ethnicity*: Are you Hispanic? Yes, No, I prefer not to respond; 4) *Career stage*: Assistant Professor, Associate Professor, Professor, Other.

Participants' race and ethnicity were used to determine whether they were an underrepresented minority or not. Non-Hispanic Asians and Non-Hispanic Whites were coded as "not URM" and all other participants were coded as "URM". For participants who identified with more than one racial group, if one racial identity was not White or Asian, they were coded as "URM". Participants who identified as both White and Asian were coded as "not URM"[2].

Open-ended Response Options

In an open-ended text box, participants were asked to please share any comments (positive or negative) about their experience or general thoughts on their recent review meeting.

Appendix B: Detailed Methods for Meeting Measures (Quantitative Analyses)

Data Collection

General meeting information was extracted from CSR's internal Meeting Dashboard. Roster data such as information on titles (rank) of the reviewer, review experience, etc. was extracted from the internal NIH IMPAC II database (QVR). Data demographics and scores were provided by the CSR informatics team. Mail reviewers were excluded for all roster analyses, while ECRs were included but not specified.

For roster analyses by gender or URM, those ad hoc members whose gender status is unknown (n=81, 2.0% of sample) or URM status is unknown (n=540, 6.3% of sample) were excluded from these analyses.

Extent of prior review service was measured for ad hoc reviewers only participating in one of the meetings in the sample. CSR Informatic team provided cross-sectional reviewer-level data on extent of prior reviewer, adjusted for one week before the meeting to get the most accurate data at the time of recruitment. Meeting counts for prior extent of service span a 12-year period and include NIH review meetings (both those run by CSR and by other NIH institutes/centers) and Advisory Councils for NIH institutes/centers. Meetings not classified as a meeting of a Federal Advisory Committee, mail reviews, and CSR rump SEPs were excluded from meeting counts.

Limitations

ECRs were not separated out, so all roster analyses metrics will include their metrics as well, including prior review service, where they will disproportionally represent reviewers in standing study sections who have had 0 prior meetings and are "newly engaged in the review system".