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

October 2015

Elizabeth L. Pier
Department of Educational Psychology
University of Wisconsin–Madison
epier@wisc.edu

Joshua Raclaw
Center for Women's Health Research
University of Wisconsin–Madison

Mitchell J. Nathan
Department of Educational Psychology
University of Wisconsin–Madison

Anna Kaatz
Center for Women's Health Research
University of Wisconsin–Madison

Molly Carnes
Center for Women's Health Research
University of Wisconsin–Madison

Cecilia E. Ford
Departments of English and Sociology
University of Wisconsin–Madison

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

Pier, E. L., Raclaw, J., Nathan, M. J., Kaatz, A., Carnes, M., & Ford, C. E., (2015). *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). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Studying the Study Section: How Collaborative Decision Making and Videoconferencing Affects the Grant Peer Review Process

Elizabeth L. Pier, Joshua Raclaw, Mitchell J. Nathan, Anna Kaatz, Molly Carnes, and Cecilia E. Ford

One of the cornerstones of the scientific process is securing funding for one's research. A key mechanism by which funding outcomes are determined is the scientific peer review process. Our focus is on biomedical research funded by the U.S. National Institutes of Health (NIH). NIH spends $30.3 billion on medical research each year, and more than 80% of NIH funding is awarded through competitive grants that go through a peer review process (NIH, 2015). Advancing our understanding of this review process by investigating variability among review panels and the efficiency of different meeting formats has enormous potential to improve scientific research throughout the nation.

NIH's grant review process is a model for federal research foundations, including the National Science Foundation and the U.S. Department of Education's Institute of Education Sciences. It involves panel meetings in which collaborative decision making is an outgrowth of socially mediated cognitive tasks. These tasks include summarization, argumentation, evaluation, and critical discussion of the perceived scientific merit of proposals with other panel members. Investigating how grant review panels function thus allows us not only to better understand processes of collaborative decision making within a group of distributed experts (Brown et al., 1993) that is within a community of practice (Lave & Wenger, 1991), but also to gain insight into the effect of peer review discussions on outcomes for funding scientific research.

Theoretical Framework

A variety of research has investigated how the peer review process influences reviewers' scores, including the degree of inter-rater reliability among reviewers and across panels, and the impact of discussion on changes in reviewers' scores. In addition, educational theories of distributed cognition, communities of practice, and the sociology of science frame the peer review process as a collaborative decision-making task involving multiple, distributed experts. The following sections review each of these bodies of literature.

Scoring in Grant Peer Review

A significant body of research (e.g., Cicchetti, 1991; Fogelholm et al., 2012; Langfeldt, 2001; Marsh, Jayasinghe, & Bond, 2008; Obrecht, Tibelius, & D'Aloisio, 2007; Wessely, 1998) has found that, within the broader scope of grant peer review, inter-reviewer reliability is generally poor, in that there is typically considerable disagreement among reviewers regarding the relative merit of any given grant proposal. While the majority of this research has compared inter-rater reliability among proposals assigned to a single reviewer for independent review, Fogelholm et al. (2012) compared the scoring practices of two separate review panels assigned the same pool

of proposals and found that panel discussion did not significantly improve the reliability of scores among reviewers. Both Langfeldt (2001) and Obrecht et al. (2007) also found low inter-rater reliability among grant panel reviewers assigned to the same proposal, attributing these disparities to different levels of adherence to the institutional guidelines and review criteria provided to panel reviewers. Obrecht et al. (2007) further found that committee review changed the funding outcome of a proposal on only 11% of reviewed proposals, with these shifts evenly split between those grants given initially poorer scores being funded, and grants given initially stronger scores being unfunded following panel review.

Fleurence et al. 2014 investigated the effect of panel meeting discussions on score movement, finding a general trend of agreement (i.e., score convergence) among reviewers following in-person discussions, with closer agreement occurring with grants given very low or very high scores. Among proposals assigned weaker scores during the pre-meeting (or triage) phase of the review process, scores tended to worsen following discussion, though panel discussion of the strongest and weakest proposals resulted in little movement in scores following discussion. Obrecht et al. (2007) also found that panel discussions led to more reviewer agreement on scores.

To summarize, the extant literature on grant peer review has found little inter-reviewer reliability for scores a grant proposal receives within a single review panel and across multiple review panels. Researchers investigating the degree to which scores change after panelist discussion have found the discussion provides only a slight change to a proposal's funding outcome; discussion tends to result in score convergence across reviewers, particularly for proposals with very low or very high scores; and that the greatest change in scores occurred for proposals assigned initially weaker scores (although this change was still quite small).

In addition, the medium of participation is potentially relevant to meeting outcomes. Gallo, Carpenter, & Glisson (2013) compared reviewer performance in panel meetings held in person versus through online videoconference. While the authors noted a small increase in some proposal scores reviewed through videoconference compared to review of these same proposals in face-to-face meetings, they found the medium of review does not significantly affect the inter-rater reliability for panelists scoring the same proposal and therefore concluded meeting medium does not influence the fairness of the review process.

Collaboration and Distributed Cognitive Processes

An important framing of the peer review process involves the acknowledgement of the distributed nature of expertise amongst panel members. As Brown and colleagues (1993) noted, with distributed expertise, some panel members show "ownership" of certain intellectual areas, but no one member can claim it all. Consequently, co-constructed meanings and review criteria are continually being re-negotiated as the members work toward a shared understanding. Schwartz (1995) showed the advantages of collaborative groups engaged in complex problem solving, whereas Barron (2000) examined some of their variability. Barron explained group variability by

noting how groups differentially achieved joint attentional engagement, aligned their goals with one another, and permitted members to contribute to the shared discourse.

Research Questions

Viewing the grant review process through the lens of collaborative decision making via distributed expertise and considering the prior literature on peer review motivates our three research questions:

1. How does the peer review process influence reviewers' scores? Detection of a discernible pattern of change within our data set would constitute a novel finding about the overall impact of collaborative and distributed discourse on individuals' evaluations of grant proposals.

2. How consistently do panels of different participants score the same proposal? This finding would bolster or refute previous findings (e.g. Fogelholm et al., 2012) that there is low reliability in scoring across panels.

3. In what ways does the videoconference format differ from the in-person format for peer review of grant proposals? Specifically, we are interested in how videoconferencing may impact the efficiency of peer review, the inter-reviewer reliability across formats, the final scores that reviewers give to proposals, how scores change as a function of panelist discussion, and whether reviewers prefer one meeting format over another. Given the potential benefits of videoconferencing for peer review panels, investigating how digital technologies might affect the peer review process is crucial.

Method

As the research team did not have access to actual NIH study sections, we organized four "constructed" study sections comprised of experienced NIH reviewers evaluating proposals recently reviewed by NIH study sections. Our goal was to emulate the norms and practices of NIH in all aspects of study design, and our methodological decisions were informed by consultation with staff from NIH's Center for Scientific Review and with a retired NIH scientific review officer who assisted the research team in recruiting grants, reviewers, and chairpersons. This retired officer also served on all of our constructed study sections as the acting scientific review officer for each meeting.

We solicited proposals reviewed from 2012 to 2015 by subsections of the oncology peer review groups for the National Cancer Institute—specifically the Oncology I and Oncology II integrated review groups. Our scientific review officer determined that proposals initially reviewed by these groups would be within her domain of scientific expertise and would provide the research team with a cohesive set of proposals around which to organize the selection of reviewers for the study sections. For each proposal, the original principal investigators, co-principal investigators, and all other research personnel affiliated with a proposal were assigned

pseudonyms. In addition, all identifying information, including original email addresses, phone numbers, and institutional addresses of these individuals, was changed. Letters of support for grant applications were similarly anonymized and reidentified; this process included the replacement of the letter writers' signatures on these letters with new signatures produced using digital signature fonts.

With the scientific review officer's assistance, we used NIH's RePORTER database to identify investigators who had received NIH research project grants (R01) from the National Cancer Institute during review cycles from 2012 to 2015 and would likely have prior review experience. We recruited 12 reviewers for each in-person study section and eight reviewers for the videoconference study section; these numbers are at the low end of what a typical ad-hoc study section might look like. Due to challenges with participant recruitment for this study, the constructed study section for Meeting 1 had 10 reviewers. Additionally, due to one reviewer's inability to travel for Meeting 2, she served as a phone-in reviewer for the entire meeting. As is typical for NIH study sections, the scientific review officer selected the meeting chairperson for each study section based on the potential chair's review experience, chairperson experience, and overall expertise in the science of the proposals the study section reviewed. For the three in-person meetings, the chairpersons also served as an assigned reviewer for six proposals. In the case that those proposals were discussed in the meeting, a pre-selected vice chairperson took over the chairperson duties for that proposal only.

Each study section meeting was organized virtually the same way as an NIH study section. Figures 1 and 2 depict digitally masked screenshots from video of one of the in-person panels (Figure 1) and from the videoconference panel (Figure 2).

Figure 1. **Digitally masked screen shot depicting the layout of panel members in Meeting 1.**

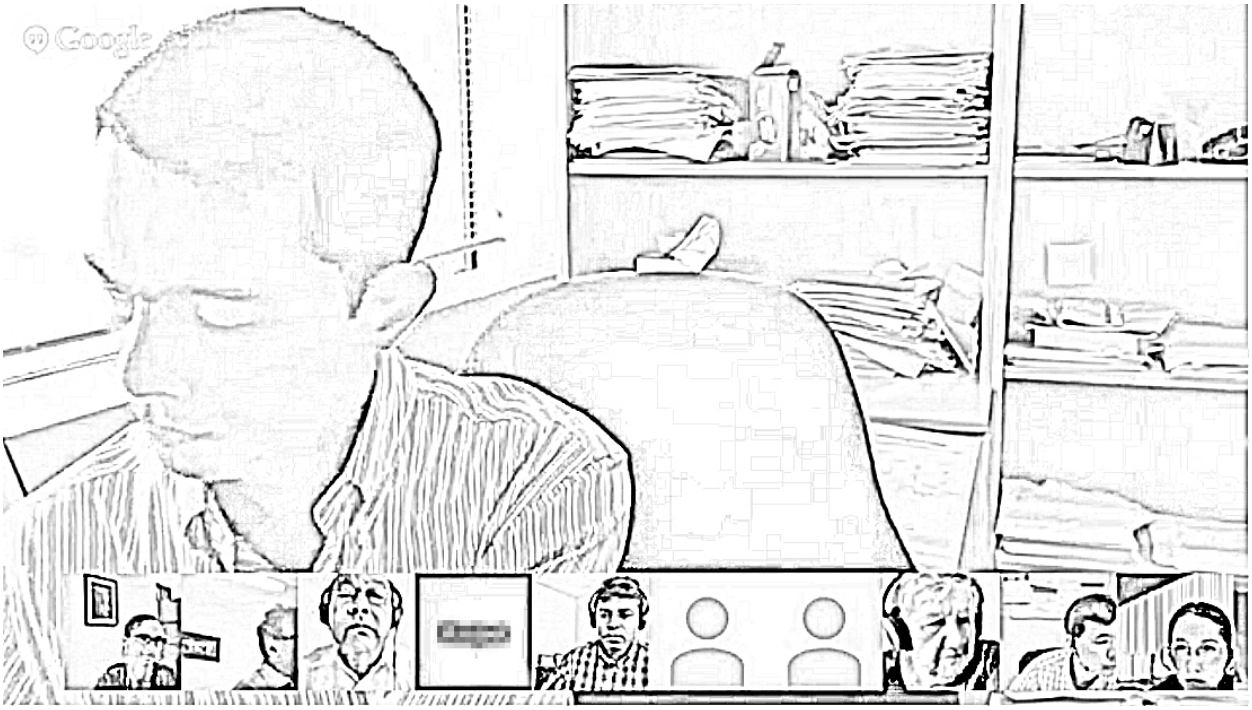

Figure 2. **Digitally masked screenshot depicting the panelists' view of the videoconference meeting (the fourth held). The name of the current speaker, featured in the main window, is displayed in a smaller window along the bottom row (fourth from the left) where he would appear when not speaking, but it has been blurred out here for privacy purposes.**

Figure 3 conveys the overall workflow that occurs for a typical study section meeting. Prior to the meeting, reviewers were assigned to review six proposals: two as primary reviewers, two as secondary reviewers, and two as tertiary reviewers. During NIH peer review generally, the assigned reviewers are responsible for reading all of the proposals assigned to them and writing a thorough critique of the proposal, including a holistic impression of the overall impact of the grant and an evaluation of five criteria: the proposal's significance, quality of the investigators, degree of innovation, methodological approach, and research environment. The scientific review officer monitors the reviewers' written critiques prior to the meeting, ensuring that reviewers adhere to NIH norms for writing a critique and complete all assigned critiques.

In addition to this written evaluation, reviewers provide a preliminary overall impact score for the proposal, as well as a score for each of the five criterion. The NIH scoring system uses a reverse nine-point scale, with 1 corresponding to "Outstanding" and 9 corresponding to "Poor." Individual reviewers' scores fall on this nine-point scale, whereas the *final* impact scores for an application correspond to the average of all panelists' scores (i.e., assigned reviewers and all other panelists who do not have a conflict of interest with the application) multiplied by 10, thus ranging from 10–90. Although the exact funding cutoff score varies depending on multiple factors (e.g., funding availability, number of proposals, NIH institute, etc.), typically, a final impact score of 30 or lower is considered to be a highly impactful project (J. Sipe, personal communication, April 8, 2015).

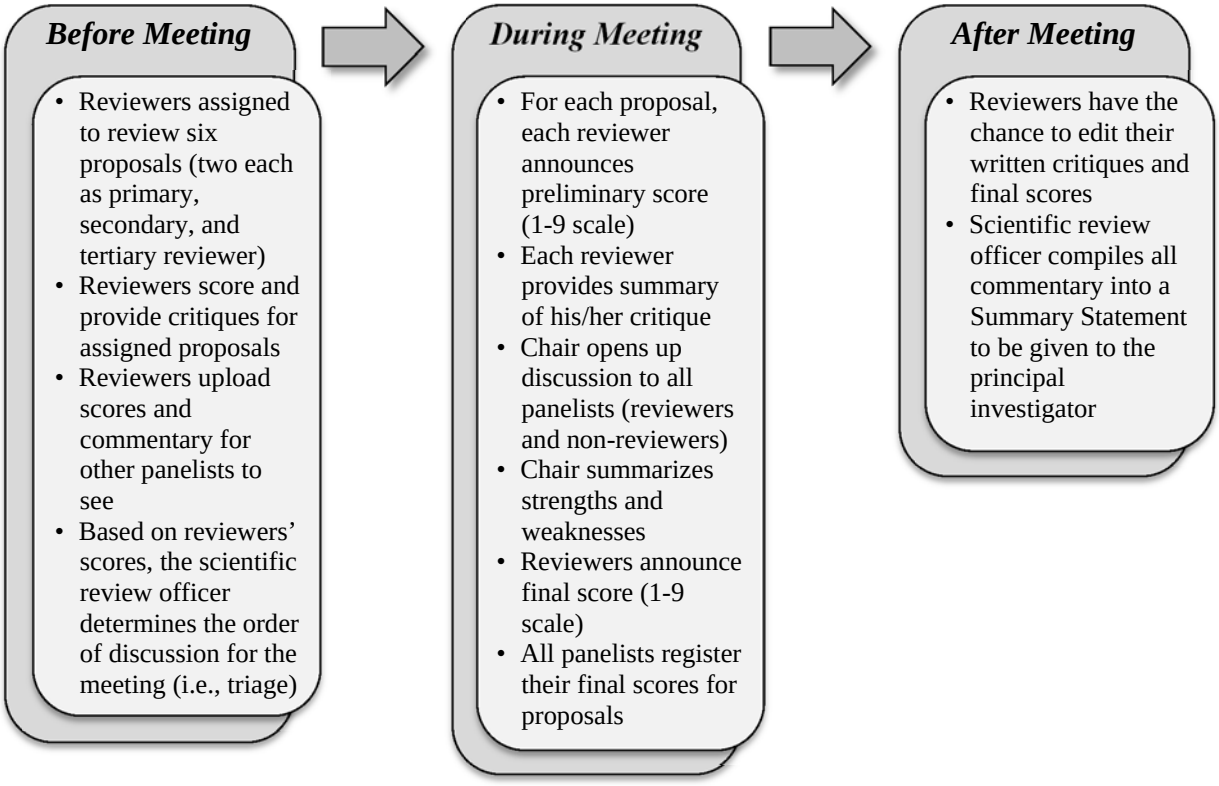

Figure 3. **Typical workflow involved in a NIH study section meeting.**

The order in which the panel discusses proposals is determined via a triage process prior to the meeting; once reviewers provide their preliminary overall impact scores and preliminary scores for each of the five criteria, the scientific review officer determines the top 50% of proposals with the highest overall preliminary impact scores. The order for discussion begins with the best-scored proposal (i.e., the proposal with the lowest preliminary overall impact score) and moves down the list in order of preliminary overall impact score. The proposals receiving preliminary overall impact scores in the bottom 50% of the proposals reviewed do not get discussed during the meeting, and thus, they are not considered for funding. Forty-eight hours prior to the meeting, the submitted written critiques and preliminary scores for an application are made available to all panelists who do not have a conflict of interest to view beforehand, if they wish to access them.

The scientific review officer begins each study section by convening the meeting, providing opening remarks, discussing the scoring system, and announcing the order of review. The officer participates throughout the meeting by monitoring discussion to ensure that NIH review policy is followed, and he or she assists the chair in ensuring there is ample time to discuss all proposals. The chair initiates discussion of individual proposals, beginning with the top-scoring proposal and moving down the list based on each proposal's preliminary overall impact score, by initiating what Raclaw and Ford (2015) refer to as the "score-reporting sequence." Here the chair calls on the three assigned reviewers to announce their preliminary scores and verbally summarize their assessments of the proposal's strengths and weaknesses. The chair then opens the floor for discussion of the proposal from both assigned reviewers and other panel members. Following discussion, the chair summarizes the proposal's strengths and weaknesses, then calls for the three assigned reviewers to announce their final scores for the proposal. All panelists then register their final scores using a paper score sheet or, in the case of videoconference meetings, an electronic document.

Our constructed study sections had 42 reviewers nested within four panels: Meeting 1 had 10 panelists including the chair; Meeting 2 had 12 panelists, one of whom phoned in; Meeting 3 had 12 panelists; and the fourth, conducted through videoconference, had eight panelists. Our sample size is small. Consequently, we take a descriptive approach to these data, as inferential statistics would be severely underpowered. We utilize descriptive statistics and correlational analyses supplemented with qualitative excerpts of discourse from the data to provide an initial, holistic analysis of the processes at play within each of the four constructed study sections.

We compiled transcripts of the verbatim discourse from the four meetings and tabulated multiple outcome measures relevant to our research questions, including: preliminary scores (i.e., scores assigned prior to the meeting) from the primary, secondary, and tertiary reviewers; final scores (i.e., scores assigned after discussion) from these three reviewers; final impact scores (i.e., the average final score from assigned reviewers and other panelists); and the time spent discussing each proposal. We also compare the final impact scores given by our constructed study sections with those given by the actual NIH panel that first reviewed these proposals.

Results

The following sections summarize our preliminary findings for each of our three research questions, including descriptive statistics of scoring and timing data from the four meetings, correlational analyses of the timing data, and evidence from the transcripts of the four meetings and from the debriefing interviews with participants.

Research Question 1: How does the peer review process influence reviewers' scores?

We first examined how peer review discussion affected changes in the scores of the three assigned reviewers. Table 1 lists the average change in individual reviewers' scores for each of the grants discussed across all four study sections, and Figure 4 depicts these changes visually. Change scores are calculated only for the three assigned reviewers, since they are the only panelists to provide a score prior to the study section meeting. The averages were computed by (1) subtracting each reviewer's final score from his or her preliminary score, giving the individual change in score for each reviewer, (2) adding those individual change scores, and (3) dividing by three for the average change in scores.

***Table 1*. Average Changes in Individual Reviewers' Scores Before and After Discussion of Each Proposal**

Proposal	Meeting 1	Meeting 2	Meeting 3	Videoconference
Abel	0	0	–1.667	
Adamsson	–0.667			
Albert	–0.667			–0.333
Amsel	–2	–0.667	+0.333	
Bretz			–0.667	
Edwards		–1.333		
Ferrera			–0.667	
Foster	–1.333	–0.667	0	–1*
Henry	–2.333	–0.333	–0.333	–1
Holzmann				0
Lopez	0	0	+0.333	
McMillan			–1	
Molloy	–2	0		
Phillips	–1	–0.5		
Rice		–1.0	–0.333	
Stavros		–0.667		0
Washington	–0.333	0		+0.333*
Williams	–0.333		–0.333	0*
Wu				+0.667*
Zhang			0	

Notes. Proposals are labeled by the last name of the principal investigator's pseudonym. Blanks indicate a proposal that was not discussed at a given meeting due to triaging. An asterisk (*) indicates these score changes are not exclusively within-subject comparisons, as mail-in reviews were used here due to a reviewer's being unable to participate in the study section. Thus, these are excluded from consideration.

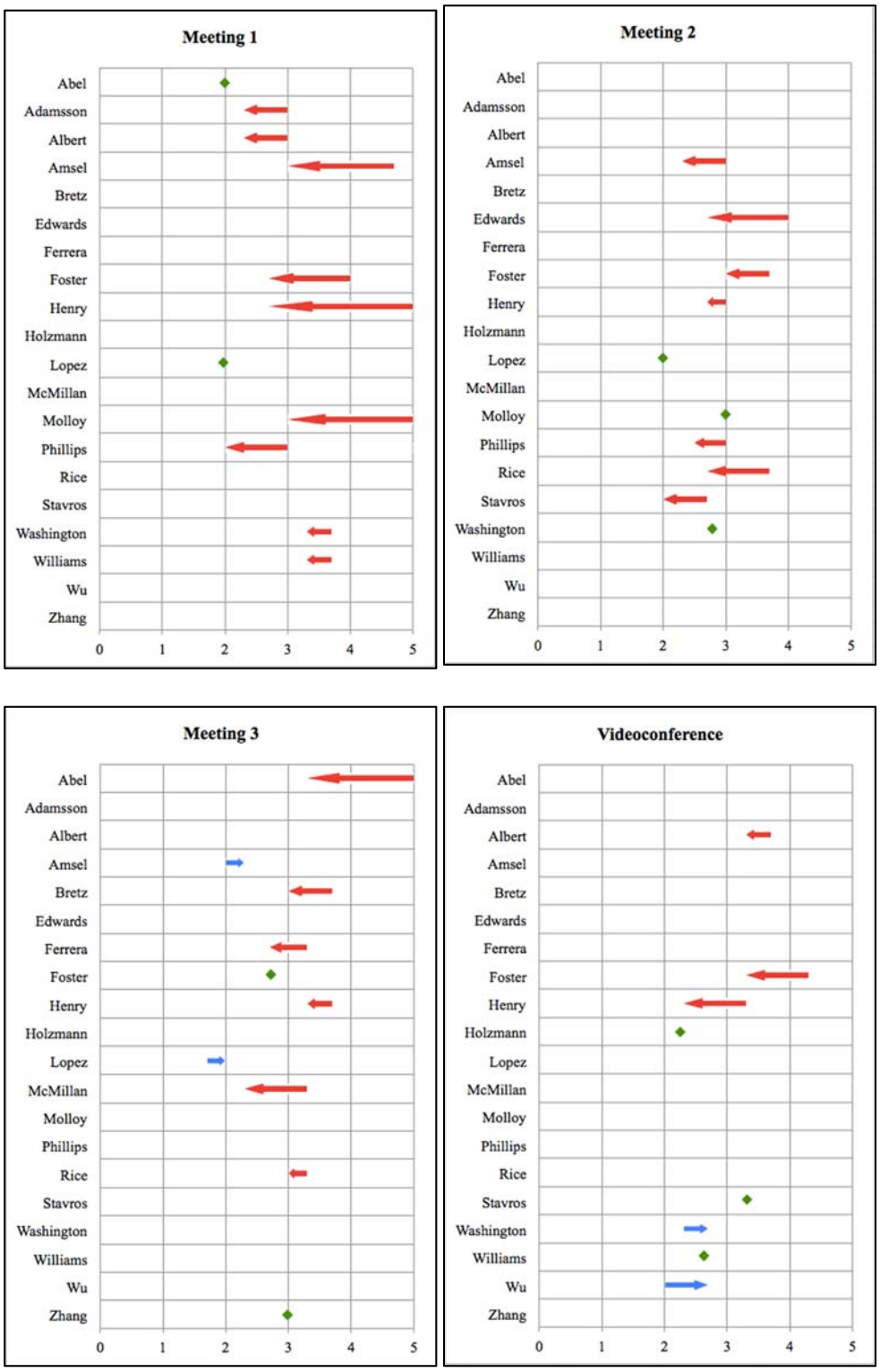

Figure 4. **Visual depiction of grant proposal score changes (from average preliminary score across the three reviewers to the average final impact score across the three reviewers) for each of the four meetings. Any proposal without an arrow or diamond was not reviewed by the panel.**

Overall, a pool of 20 deidentified grant proposals was provided to participants. Across the four study sections, 41 reviews took place (11 proposals in Meeting 1, 11 proposals in Meeting 2, 11 proposals in Meeting 3, and eight proposals in the videoconference). However, in the videoconference meeting, mail-in reviews were used for four proposals in the place of an in-person reviewer. Because these mail-in reviewers do not provide final impact scores following discussion, these are not within-subject comparisons and are therefore excluded from the descriptive statistics we report next. For the remaining 37 proposals, reviewers were far more likely to worsen their scores for an application after discussion (n = 25, 67.57%) than to maintain (n = 10, 27.03%) or improve their scores (n = 2, 5.41%). There were 116 final individual reviewer scores provided across all four meetings (three individual reviewers times 41 proposals, minus seven proposals for which mail-in reviews were used). Of these 116 individual scores, there were n = 56 (48.28%) instances of individual reviewers worsening their scores, n = 47 (40.52%) in which they did not change their scores, and n = 13 (11.21%) in which they improved their scores (see Table 2). Thus, individually and in the aggregate, reviewers tended to give less favorable scores following panel discussion.

***Table 2*. Count of Changes in Individual Reviewers' Scores Before and After Discussion**

Change	Meeting 1	Meeting 2	Meeting 3	Videoconference	Total
Improved *(lower score)*	2 (6.25%)	3 (9.68%)	5 (15.15%)	3 (15.00%)	13 (11.21%)
No change	7 (21.88%)	16 (51.61%)	13 (39.39%)	11 (55.00%)	47 (40.52%)
Worsened *(higher score)*	23 (71.88%)	12 (38.71%)	15 (45.45%)	6 (30.00%)	56 (48.28%)

Although the overall trend was toward scores becoming worse after discussion, there were some differences among panels (see Table 2). In Meeting 1, a majority of reviewers (71.88%) gave worse scores after discussion. In the other two in-person study sections, reviewers were more evenly split between giving worse scores (38.71% in Meeting 2, 45.45% in Meeting 3) and maintaining their scores (51.61% and 39.398%, respectively). In addition, reviewers for the videoconference meeting were more likely to maintain their initial scores (55%) than to worsen (30%) or improve (15%) them (cf. Gallo et al., 2013). This inter-panel variability was evident in other aspects of meeting features as well, as Research Question 2 explores.

Research Question 2: How consistently do panels of different participants score the same proposal?

Our second research question aimed to investigate the degree of scoring variability across panels. Importantly, the set of grant proposals actually reviewed varied across each panel (see Table 3) due to the triaging process prior to the meeting. Thus, variability in terms of which grants were discussed reflects the preliminary scores given by the assigned reviewers prior to the meeting. Out of the 20 grant proposals provided to the panelists, two were discussed in all four study sections, six discussed in three study sections, four discussed in two study sections, and eight discussed in only one study section.

***Table 3*. Final Impact Scores**

Proposal	Meeting 1	Meeting 2	Meeting 3	Video-conference	Average	NIH Score
Abel	20.0	29.1	50.0		33.0	27.0
Adamsson	30.0				30.0	23.0
Albert	35.0			38.6	36.8	39.0
Amsel	50.0	25.5	20.9		32.1	27.0
Bretz			39.2		39.2	20.0
Edwards		37.3			37.3	40.0
Ferrera			33.3		33.3	36.0
Foster	42.0	38.2	29.2	45.0	38.6	23.0
Henry	52.0	35.5	35.0	32.5	38.8	14.0
Holzmann				27.5	27.5	17.0
Lopez	30.0	21.8	16.7		22.8	39.0
McMillan			30.8		30.8	25.0
Molloy	50.0	30.0			40.0	28.0
Phillips	31.1	30.8			31.0	23.0
Rice		39.1	31.7		35.4	ND
Stavros		32.7		33.8	33.3	20.0
Washington	39.0	35.0		26.3	33.4	31.0
Williams	42.0		30.8	38.8	33.9	28.0
Wu				20.0	20.0	44.0
Zhang			29.2		29.2	38.0
Average	38.3	32.3	31.5	31.6	32.8	28.5

Note. Abel and Amsel proposals (shaded) are examples of applications with highly variable final impact scores across constructed study sections.

Variability in final impact scores. In addition to the variability noted above in terms of how individual reviewers change their scores following discussion (see Table 2), we found considerable differences in the final impact scores given to the same grant proposal across study sections (see Table 3). Recall that final impact scores range from 10 to 90.

Importantly, the differences in scores for individual proposals do not reflect a harsh or lenient panel overall, as evidenced by the consistency in their *average* scores across all proposals discussed (see second-to-last column of Table 3). For example, following peer review discussion, the Abel proposal (shaded in grey in Table 3) received a final impact score of 20.0 in Meeting 1, a final impact score of 29.1 in Meeting 2, and a final impact score of 50.0 in Meeting 3. In contrast, the Amsel proposal (also shaded in grey in Table 3) received a final impact score of 50.0 in Meeting 1, 25.5 in Meeting 2, and 20.9 in Meeting 3. Both of these proposals (coincidentally) received a score of 27.0 when reviewed by an actual NIH study section (see last column of Table 3). Thus, the final score for these proposals is highly dependent on the particular study section in which it is discussed.

Overall, despite the fact that the constructed study sections had highly similar average scores across all proposals reviewed, there are notable differences across study sections: reviewers'

tendencies to lower or raise their scores after discussion (as was found for Research Question 1), which proposals are discussed during peer review after triage based on reviewers' preliminary scores, and the final impact scores assigned to a particular proposal.

The process of calibrating scores. Our preliminary analysis of the raw video data reveals that one source of variability among panels stems from panelists' explicit discussion around what constitutes a given score—a process we call *score calibration*. In each of the four study sections, there were numerous instances of panelists who directly addressed the scoring habits of another panelist or of the panel as a whole. For example, toward the end of Meeting 1, one of the non-reviewer panelists raised the issue of what constituted a score of 1. Here, she drew on the institutional authority of NIH to claim that applications with this score must be relatively free of any weaknesses:

> The one thing that I think that is kind of lacking, and I hate to be very critical of the way we're doing things, but you know, every time we go to study section, they give you a sheet where they say the minimal—1 or 2 minimal weaknesses—this is the score. More than, you know, one major weakness is this. I don't think we're following this here.

Similarly, in Meeting 2, during discussion of the first grant, a non-reviewer panelist directly addressed the tertiary reviewer, saying: "So it sounds like a lot of weaknesses given that it's a two [a highly competitive score]. Probably overly ambitious, problems with the model, not necessarily clinically relevant. It's just a long list given that much enthusiasm in the score." The tertiary reviewer responded by saying, "I mean, I was just saying if I had to pick any weakness, that would have been my concern, but you know that that's the only thing and to me, it's a really strong proposal."

Relatedly, toward the beginning of Meeting 3, a non-reviewer panelist commented on the primary reviewer's score after his lengthy initial summary of the grant proposal, telling him, "Your comments are meaner than your score." Cursory comments such as this one, in which a panelist comments on a reviewer's scoring habits, were frequent during our constructed study section meetings.

Finally, in the videoconference panel, during discussion of the first grant, the following conversation transpired between the three reviewers after the chair asked for their final scores following discussion:

> REVIEWER 1: Yes, so, I respect what the other reviewer said, so I will move my score from 2 to 3.
>
> REVIEWER 2: I'm also gonna move from 2 to 3.
>
> REVIEWER 3: Yeah, I'm gonna try and be fair. I mean I think there's a lot of good in it too. I'm gonna say 4. I'm gonna go from 3 to 4.

Here, we see evidence of reviewers calibrating their scores based on the scores and comments of others reviewers. The chair then stepped in, saying:

> Yeah, I would say that it's not unusual for a bunch of folks like us to focus on the weaknesses and take the strengths for granted, and sometimes they don't come out in review, just to be clear.

This professed tendency to focus on the weaknesses provides some insight into our findings that overall, discussion tends to result in worse scores compared to preliminary scores. These examples suggest that how explicit calibration of scoring norms within a study section is an important and common component of the review process. Score calibration appears to directly influence the scoring behaviors of panelists, and is a potential factor for influencing inter-panel reliability of final impact scores.

Research Question 3: In what ways does the videoconference format differ from the in-person format for peer review of grant proposals?

Our final research question compared the videoconference meeting with the three in-person study section meetings. We were interested in investigating how the videoconference format may influence the final scores that reviewers give to proposals, as well as the potential for gained efficiency with this medium.

As the final row in Table 3 shows, the average final impact score across all proposals was virtually identical for the videoconference meeting compared to Meeting 3 and highly similar to the in-person meetings. Thus, the videoconference format does not appear to change the final impact scores of the panel in the aggregate.

Videoconference panels may be more efficient, however: The videoconference reviewed eight proposals for two hours and three minutes, while the three in-person panels each reviewed 11 proposals for 2:53, 3:21, and 3:37, respectively. On average (see Table 4), the videoconference panel spent 42 seconds less per proposal than Meeting 1, but two minutes and 26 seconds less than Meeting 2 and three minutes and 27 seconds less than Meeting 3. An important caveat to these findings of greater efficiency is that the videoconference panel has the fewest panelists in attendance during the study section, so future studies will have to be examine videoconferences more systematically.

***Table 4*. Total Time in Minutes and Seconds Spent on Each Proposal at Each Meeting**

Proposal	Meeting 1	Meeting 2	Meeting 3	Videoconference	Average
Abel	17:39	16:43	14:33		16:18
Adamsson	15:00				15:00
Albert	13:18			13:24	13:21
Amsel	13:08	12:14	20:16		15:13
Bretz			15:20		15:20
Edwards		11:01			11:01
Ferrera			20:22		20:22
Foster	14:46	14:58	15:00	09:29	13:33
Henry	15:41	17:27	14:01	20:17	16:52
Holzmann				15:50	15:50
Lopez	18:58	18:24	17:10		18:11
McMillan			25:47		25:47
Molloy	13:22	09:12			11:17
Phillips	13:37	13:17			13:27
Rice		14:02	13:05		13:33
Stavros		19:52		10:20	15:06
Washington	14:27	31:58		13:30	19:58
Williams	13:33		17:07	15:10	15:17
Wu				12:46	12:46
Zhang			17:41		17:41
Average	14:33	16:17	17:18	13:51	15:48

Correlational analyses. Given this apparent pattern of increased efficiency, we investigated whether a relationship exists between how long a panel spent on an individual proposal and the degree to which the reviewers changed their scores for that proposal. A significant correlation between the time spent discussing a proposal and the degree to which a reviewer changed his or her score following discussion would suggest that the length of time spent discussing a proposal may influence the ultimate funding outcome for a proposal. However, a correlational analysis between review time per proposal and the average change in panel score showed no discernable pattern, indicating that the time spent on each proposal does not strongly predict changes in reviewers' scores (Meeting 1: $r = –0.25$, $p = 0.167$; Meeting 2: $r = –0.42$, $p = 0.204$; Meeting 3: $r = 0.06$, $p = 0.852$; Videoconference: $r = 0.106$, $p = 0.802$). However, our very small sample sizes preclude strong conclusions based on these correlations.

Panelist preferences for panel format. Given that our videoconference panel did not vastly differ from the in-person panels in terms of average final impact scores, but that it took less time to review each proposal, one might conclude that videoconferences provide a cost-effective alternative to the complex, costly, and time-consuming process of implementing the approximately 200 NIH study sections that occur each year (Li & Agha, 2015). However, reviewers do not necessarily see these benefits as outweighing the perceived costs. For example, after the conclusion of Meeting 2, a panelist remained behind and began talking with the meeting chair and the scientific review officer about their experiences doing peer review via

teleconference. During this time, the following exchange occurred between the meeting chair and one of the reviewers (DB):

> Chair: I haven't done video, I've done teleconferences, and they're terrible because here you can see when somebody's interested and they're maybe going to *((slightly raises his hand in the air))*, or you can read the expression on their face. They know that they—and then you can say, so what do you think? *((leans over and points to an imaginary person in front of him, as if calling on them))*. You can do that. You <u>can't</u> do that on a telephone, you're just waiting for people to speak up, they speak up, and if they don't you have no idea who to address.
>
> DB: And you lose, I mean, you lose so much.
>
> Chair: Oh gosh you do.

Later, the conversation continues:

> Chair: I think that the quality of their reviews at the end and the scores are more—better reflect the quality of the science when you have a meeting like this. And so what—the other aspect of it is, I want my grants to be reviewed that way too. So it's not just you know, it's not a good experience for me, I absolutely agree, I would be just, why would you bother? If you're just going to sit on the phone, 'cuz half the time what's happening—
>
> DB: You can't hear it.
>
> Chair: You either can't hear it or worse you know everybody else has put it on mute and they're on their emails, they're writing manuscripts, and they just wait till their grant comes up.
>
> DB: It's true, I mean it would be easier but it would be a waste of time.
>
> Chair: Oh uh uh I would hate if they ever did that. I can't imagine.

Similarly, during a debriefing interview after the videoconference meeting, one panelist who participated in the videoconference said to a member of the research team:

> I would still prefer a face-to-face meeting and that's, you know, that's coming from a person on the West Coast who has to fly a very long way to those face-to-face meetings. But I still think that's by far the best and actually the most fair, too, because I think there's just no barriers for conversation there I think, even with the videoconferencing. It's just not quite the same as it is in an in-person kind of discussion. You know I tend to be candid, but I think when the proposal is contentious or whether there's real issues, then I think it becomes a slightly different process... I think in an in-person discussion, I think all the differences are gonna come out a little more readily than if you're on the

> phone or on a videoconference. I just think that there's more of an inclination on the phone or on videoconference to not get everything said. I think in person you're much more likely to get things said.

Thus, participants mentioned several perceived benefits to the in-person meeting format, 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.

However, not all of the reviewers echoed these panelists' feelings that in-person meetings were preferable to videoconferencing or teleconferencing. Specifically, a few videoconference participants said they preferred this format to in-person meetings. Importantly, all participants selected which date they could participate in the study while knowing whether the meeting was in person or via videoconference, so there is likely some selection bias occurring that would skew the results this way. Nevertheless, panelists mentioned desirable benefits to videoconference panels, such as the chairperson of the videoconference meeting, who said he preferred the online format because:

> Number one, mmm, it is as good as in person, you know hoping that there will be no glitches in terms of video transmission and stuff like that. And the reviewers themselves will be a lot more relaxed you know, because they are doing it from their office and if they need to quickly look for a piece of paper or a reference they can use their computers and look for it. There are a lot of advantages of doing it from your office compared to in a hotel room.

Overall, then, it seems that there are personal preferences at play regarding panelists' feelings toward the format of the study section meeting, and additional research is needed to draw strong conclusions about overall trends.

Discussion

The current study aimed to address three research questions related to the peer review process. Our first research question asked how discussion during grant peer review influences reviewers' scores of grant proposals. Our second research question aimed to evaluate the scoring variability across four constructed study section panels. Our third research question sought to examine the differences between the in-person format and the videoconference format of grant peer review meetings. The follow sections discuss each of our results in turn.

Research Question 1: How does the peer review process influence reviewers' scores?

We found that overall, reviewers tend to assign worse scores to proposals after discussion compared to their preliminary scores. These results provide contrasting evidence to Fleurence and colleagues' (2014) findings that only the weakest applications' scores worsened after discussion and that discussion itself influenced scores very little.

In addition, some inter-panel variability in the proportion of scores became worse. Among the in-person panels, reviewers in Meeting 1 were much more likely to worsen their scores than to maintain or improve their scores, whereas reviewers in the other two in-person meetings were more evenly split between worsening or maintaining scores. The videoconference reviewers were more likely to maintain their initial scores than to change them in either direction. Overall, though, none of the panels were more likely to improve the scores of the grant proposals following peer discussion compared to worsening their scores or leaving them unchanged.

One preliminary hypothesis explaining this phenomenon is that the peer review process may draw out the weaknesses of grant proposals more so than the strengths. In light of research establishing the benefit of collaboration and group discussion for problem solving (e.g., Cohen, 1994; Schwartz, 1995; Webb & Palinscar, 1996), our findings may indicate a heightened capacity for critically evaluating the merits of grant applications in a collaborative team as opposed to panelists' independent grant review. However, we may also be seeing a negative effect of the review process itself. These panelists, by virtue of their expertise, may have a tendency to overemphasize the weaknesses in a grant proposal and take its many strengths for granted, resulting in a preponderance of verbalized weaknesses that may worsen their final impact scores. Yet, while they overemphasize these weaknesses, they may not account for their own overemphasis when setting their final impact scores. This possibility is reminiscent of Schooler's (2002) "verbal overshadowing effect" in which the demands of verbalizing one's views of the overall quality of something that is complex (e.g., an Impressionist painting) can lead one to falsely believe all that they say and to trump their initial, nonverbal impressions. This scenario suggests that one area of improvement of the review process may be ways to regulate or structure the number of positive and negative comments in relation to the overall impressions of the quality of the grant proposal.

Future content analyses of panelists' discussions with score changes in mind and comparisons of the turn-taking frequency in face-to-face versus videoconference meetings will be fruitful for exploring this potential explanation. In addition, because our findings about trends in score changes deviate from those of Fleurence and colleagues (2014), future research—possibly with additional constructed study sections—ought to parse out the components that may lead to significant score changes after discussion, including the behavior of the reviewers, the actions of the chairperson or the scientific review officer, or something intrinsic to the discussion process itself.

Research Question 2: How consistently do panels of different participants score the same proposal?

In addition to considerable variability in the preliminary scores assigned to individual grant proposals (i.e., which proposals were triaged for discussion), there was also substantial variability in the final impact scores given to the same grant proposals across multiple study sections—even though the overall average final impact score was consistent across panels. This aligns with findings reported elsewhere noting vast inter-rater and inter-panel variability (e.g.,

Barron, 2000). Indeed, Obrecht and colleagues (2007), who found significant inter-committee variance in scoring of the same proposals, concluded:

> Intuitively we might expect that group discussion will generate synergy and improve the speed or the quality of decision-making. But this is not borne out by research on group decision-making (Davis, 1992). Nor does bringing individuals together necessarily reduce bias. (p. 81)

Evidence from the transcripts of our constructed study sections revealed that panelists explicitly attempt to calibrate scores with one another during discussion. They directly challenge others' scores (particularly strong scores of 1 or 2), attempting to negotiate a shared understanding of the "meaning" of a score, and acquiesce to other reviewers by changing their final scores from their preliminary scores. This type of calibration or norming is something also noted by Langfeldt (2001), who found that "while there is a certain set of criteria that reviewers pay attention to—more or less explicitly—these criteria are interpreted or operationalized differently by various reviewers" (p. 821).

One potentially fertile area for future research is to determine whether inter-rater reliability is shaped more by differences in panelists' content knowledge or by panelists' adherence to scoring and reviewing norms (including the locally constructed calibration). In terms of the latter, we noted several occasions when panelists held their peers accountable for alignment between their scores, and the strengths and weaknesses they identified. Peer influence suggests researchers also should explore another kind of reliability measure: *intra*-rater reliability—that is, the degree to which a reviewer's words align with her or his scores. Understanding how much (quantitatively) *intra*-rater reliability impacts the *inter*-rater reliability of scores within and across panels may help identify other targeted ways reviewer training could improve the reliability of the review process.

Due to the dynamic, transactional, and contextual nature of discourse (Clark, 1996; Gee, 1996; Levinson, 1983), group discussion in these constructed study sections necessarily involves the local co-construction of meaning (e.g., determining what a numeric score should signify, or which aspects of a proposal should be most highly valued). Though perhaps undesirable from a practical standpoint, such variance is thus to be expected from a sociocultural perspective on collaborative decision-making processes, as Barron (2000) showed. It is therefore imperative for funding agencies such as NIH to determine how to establish confidence in the peer review process given the inter-panel variability that we and others (e.g., Barron, 2000; Langfeldt, 2001; Obrecht et al., 2007) have found. Establishing the degree to which principal investigators submitting grant proposals *are* in fact confident in the peer review process, perhaps through a survey instrument, would be a logical next step. Beyond that, research should aim to investigate how funding agencies may mitigate some of this variability—for example, by piloting a more robust training program for reviewers, examining whether a more stringent rubric for reviewer scoring is superior to a more holistic guide to scoring, or developing professional development programs for the scientific review officers and panelists to ensure stricter adherence to scoring

norms. Importantly, NIH updated its scoring guidelines in 2009, and the Center for Scientific Review has modified its scoring guidelines several times; thus, our recommendations align with what NIH has initiated in response to issues of inter-panel scoring variability.

Research Question 3: In what ways does the videoconference format differ from the in-person format for peer review of grant proposals?

The overall final impact scores across all proposals reviewed were highly similar between the in-person panels and the videoconference panel, which aligns with Gallo and colleagues' (2013) findings. Thus, the videoconference format does not appear to impair or improve panel outcomes in the aggregate. However, we also found that the videoconference panel, while less populated, was more efficient, spending less time discussing each proposal on average than the in-person meetings, which suggests that videoconferencing may serve as a viable, cost-effective replacement for in-person peer review meetings. This finding, of course, is based on data comparing only one videoconference panel to three in-person meetings and will require a more systematic study before strong policy recommendations could be made.

Correlational analyses revealed the lack of a consistent relationship between the time spent reviewing a proposal and the average change in reviewers' scores for a proposal, implying the time spent reviewing a proposal does not affect whether reviewers change their preliminary scores more or less after discussion. Research will need to collect more data from multiple constructed study sections conducted via videoconference to ascertain whether these patterns are consistent with the videoconference format, as opposed to a vestige of this particular videoconference panel.

Although the data support the use of videoconference panels in place of face-to-face formats, some of our reviewers spontaneously expressed their preference for meeting in person to meeting via teleconference or videoconference. While selection bias is at play in this study, since panelists selected into or out of the videoconference format, panelists who participated in the videoconference nonetheless stated a preference for in-person meetings. Thus, although the videoconference format offers a cost-effective and efficient means of reviewing grant proposals without any overall deterioration to the reviewers' final scores, further research is needs to establish the degree to which expert scientists are likely to engage in videoconference peer review meetings, as well as to rigorously investigate the affordances and constraints of utilizing videoconference for peer review.

One notable finding was the recorded exchange after Meeting 2 when the panel chair expressed his belief that in-person scores more accurately reflect the quality of the proposals under review. Our data do not bear this out, and they show little difference in the overall impact scores across the formats. However, these beliefs may influence the quality of the review panels and need to be better understood via more systematic investigation.

Perhaps the onset of more sophisticated videoconferencing technology may alleviate some of the reviewers' concerns with the online medium, but future work needs to investigate how

videoconferencing may be amended or supplemented to address the drawbacks that panelists feel interfere with peer review, so as to improve reviewers' likelihood of participating in and feeling positively about videoconference peer review.

Conclusion

Each year through the study section review process, NIH funds nearly 50,000 grant proposals totaling more than $30 billion. Constructing study sections designed to emulate the scientific review process is a powerful methodological approach to understanding how scientific research is funded and offers valuable insights into the important area of complex group decision making, particularly because the NIH study section review process is a model for other federal funding agencies. Thus, these preliminary findings already contribute to scientific understanding of the review process and to policy recommendations for future review panels.

The vast undertaking of recruiting and convening expert panels of biomedical researchers constrains the number of panelists and grant proposals for this initial study, limiting our sample size and thus power for conducting inferential statistics. Videoconferencing could increase the number of constructed study sections we could investigate in the future. In forthcoming analyses, we plan to examine the relative affordances and constraints of the videoconference format for peer review, as well as the multimodal nature of collaborative discourse during peer review and the developmental trajectory of reviewers as they become socialized into the community of practice of NIH peer review. Ultimately, we believe research of this type can increase the public perception of scientific research activities in the United States and the role scientific research can play for benefiting public policy.

References

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences*, *9*(4), 403–436.

Brown, A. L., Ash, D., Rutherford, M., Nakagawa, K., Gordon, A., & Campione, J. C. (1993). Distributed expertise in the classroom. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 188–228). Cambridge, England: Cambridge University Press.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14,* 119–135.

Clark, H. (1996). *Using language*. New York: Cambridge University Press.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1–35.

Fleurence, R. L., Forsythe L. P., Lauer, M., Rotter, J., Ioannidis, J. P., Beal, A., Frank, L., Selby, J. V. (2014). Engaging patients and stakeholders in research proposal review: The patient-centered outcomes research institute. *Annals of Internal Medicine, 161*(2), 122–130.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., & Väänänen, K. (2012). Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of Clinical Epidemiology, 65*(1), 47–52. doi:10.1016/j.jclinepi.2011.05.001

Gallo, S. A., Carpenter, A. S., & Glisson, S. R. (2013). Teleconference versus face-to-face scientific peer review of grant application: Effects on review outcomes. *PLOS ONE, 8*(8), 1–9.

Gee, J. P. (1996). Sociocultural approaches to language and literacy: An interactionist perspective by Vera John-Steiner; Carolyn P. Panofsky; Larry W. Smith. *Language in Society*, *25*(2), 294–297.

Langfeldt, L. (2001). The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science, 31*(6), 820–841. doi:10.1177/030631201031006002

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, England: Cambridge University Press. doi:10.2307/2804509

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Li, D., & Agha, L. (2015, April 24). Big names or big ideas: Do peer-review panels select the best science proposals? *Science, 348*(6233), 434–438. doi: 10.1126/science.aaa0185

Marsh H. W., Jayasinghe, U. W., Bond N. W. (2008). Improving the peer review process for grant applications: Reliability, validity, bias and generalizability. *American Psychologist, 63*(3), 160–168.

National Institutes of Health (NIH). (2015, January 29). NIH budget [Web page]. Retrieved from http://www.nih.gov/about/budget.htm

Obrecht, M., Tibelius, K., & D'Aloisio, G. (2007). Examining the value added by committee discussion in the review of applications for research awards. *Research Evaluation, 16*(2), 70–91. doi:10.3152/095820207X223785

Raclaw, J., & Ford, C. E. (2015). *Laughter as a resource for managing delicate matters in peer review meetings.* Paper presented at the Conference on Culture, Language, and Social Practice, Boulder, CO.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences, 4*(3), 321–354.

Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences, 6*(8), 339–344.

Webb, N., & Palinscar, A. S. (1996). Group processes in the classroom. In R. Calfee & C. Berliner (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Prentice Hall.

Wessely, S. (1998). Peer review of grant applications: What do we know? *Lancet, 352*, 301–305.

This work was made possible in part by the Avril S. Barr Fellowship to the first author through the School of Education at the University of Wisconsin–Madison. The data collection for this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM111002. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions. The authors extend their deepest appreciation to Dr. Jean Sipe and to Jennifer Clukas for their indispensable contributions to this project.