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

ELIZABETH L. PIER, JOSHUA RACLAW, ANNA KAATZ, MOLLY CARNES, CECILIA E. FORD, & MITCHELL J. NATHAN

UNIVERSITY OF WISCONSIN-MADISON

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

CENTER FOR WOMEN'S HEALTH RESEARCH University of Wisconsin-Madison

ABSTRACT

Grant peer review is a foundational component of scientific research. In the context of grant review meetings, the review process is a collaborative, socially mediated, locally constructed decision-making task. The current study examines how collaborative discussion impacts reviewers' scores of grant proposals, how different review panels score the same proposals, and how the discourse practices of videoconference panels differ from in-person panels. Methodologically, we created and videotaped four "constructed study sections," recruiting biomedical scientists with NIH review experience and an NIH Scientific Review Officer (SRO). These meetings provide a rich medium for investigating the process and outcomes of authentic collaborative decision-making tasks. Implications for research into the peer review process, as well as for the broad enterprise of federally funded scientific research, are discussed.

OBJECTIVES

- The National Institutes of Health (NIH) spends $30.3 billion on biomedical research every year; more than 80% of funding is awarded via competitive grants (NIH, 2015)
- NIH peer review is a model for other federal research foundations, including the National Science Foundation (NSF) and the Institute of Education Sciences (IES)
- Advancing our understanding of peer review has the potential to improve scientific and educational research throughout the nation

THEORETICAL FRAMEWORK

- Peer review consists of a group of **distributed experts** (Brown et al., 1993) within a **community of practice** (Lave & Wenger, 1991) co-constructing meanings and continually negotiating review criteria
- In grant peer review**, inter-reviewer reliability** is typically poor (Cichetti, 1991; Fogelholm et al., 2012; Langfeldt, 2001; Marsh et al., 2008; Olbrecht et al., 2007; Wessely, 1998)
- Gallo et al. (2013) found few differences between **face-to-face (FTF) and teleconference** peer review meetings

RESEARCH QUESTIONS

1. How does collaborative and distributed discussion affect **reviewers' scores** during grant peer review?
2. How **consistent** are different panels of reviewers in scoring the same grant application?
3. In what ways does the **videoconferencing** format differ from the **face-to-face** format for peer review of grant applications?

The research reported in this poster was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM111002. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health.

METHOD & DATA SOURCES

- Four **"constructed" study sections** of experienced NIH reviewers reviewed deidentified applications from 2012 – 2015 submitted to the Oncology groups of the National Cancer Institute
- 42 participants nested in four study section panels
  - Meeting 1 (FTF): *n* = 10
  - Meeting 2 (FTF): *n* = 12
  - Meeting 3 (FTF): *n* = 12
  - Meeting 4 (VC) : *n* = 8
- Participants assigned to review and score 2 applications as primary reviewer, 2 as secondary reviewer, and 2 as tertiary reviewer
- Meetings videotaped and all dialogue transcribed
- Descriptive, case-study approach to the data given small sample size

Figure 1. Anonymized shot from a panel.

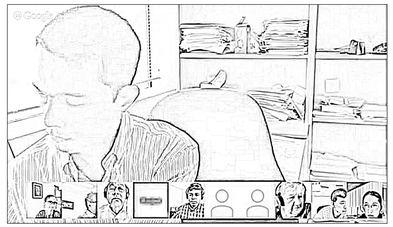

Figure 2. Anonymized screenshot from the videoconference panel.

Overall Impact or Criterion Strength	Score	Descriptor
High	1	Exceptional
	2	Outstanding
	3	Excellent
Medium	4	Very Good
	5	Good
	6	Satisfactory
Low	7	Fair
	8	Marginal
	9	Poor

Figure 4. NIH scoring rubric.

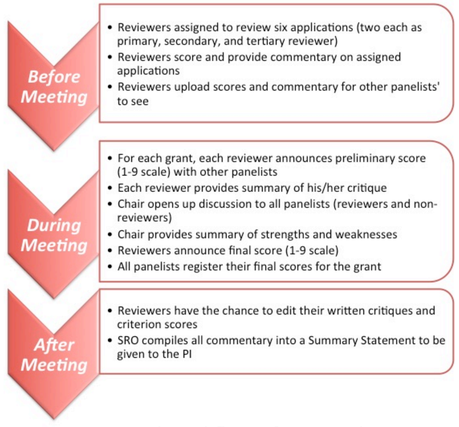

Figure 3. Typical workflow of NIH study section

RESULTS

Table 1. Count of Changes in Individual Reviewers' Scores 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%)	12 (36.36%)	11 (55.00%)	46 (39.66%)
Worsened *(higher score)*	23 (71.88%)	12 (38.71%)	16 (48.48%)	6 (30.00%)	57 (49.14%)

Overall, more common for reviewers to worsen scores after discussion, but panels varied in degree of score change (see Fig. 5)

Table 2. Final Impact Scores

Grant Pseudonym	Meeting 1	Meeting 2	Meeting 3	Video-conference	Average	NIH
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. Grants are labeled by the last name of the PI's pseudonym. The Abel and Amsel grants (shaded) are examples of applications with highly variable final impact scores across constructed study sections.

Scores across panels were highly similar in the aggregate, but highly variable at the grant application level

VC panel did not differ from FTF panel in aggregate final scores

Table 3. Total Time in Minutes and Seconds Spent on Each Application at Each Meeting

Grant	Meeting 1	Meeting 2	Meeting 3	Videoconference	Average
Abel	17:39	16:43	14:33		16:18
Adamsson	15:0				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	14:52	16:17	17:18	13:51	15:48

VC panel was more efficient (>1 minute less per proposal)

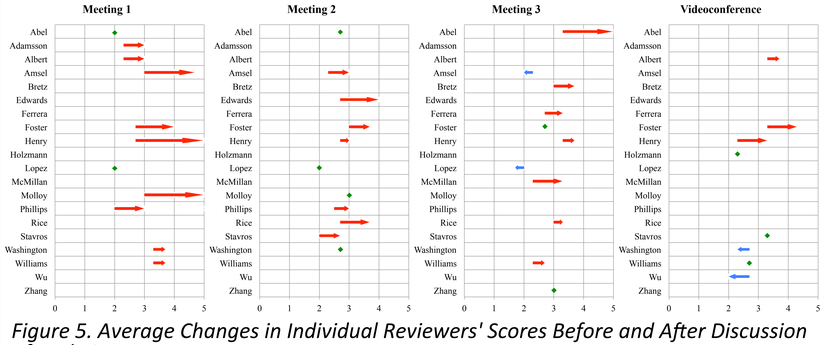

Figure 5. Average Changes in Individual Reviewers' Scores Before and After Discussion of Each Grant

CONCLUSIONS

- This case study aligns with prior research finding **substantial variability** across peer review panels, but is the first study to explore how scores change as a function of collaborative discussion
- Panelists are continually re-negotiating the *meaning* of numerical scores during ***score calibration talk*** (Pier et al., 2015), which may lead to the variability we observe across panels
- Potential for videoconference panels to **increase efficiency—and reduce costs** of convening study sections—without altering final score outcomes
- Future work will examine the nature of the collaborative discussion during peer review panels, explore the relationship between the scores reviewers assign to grant applications and the content of their critiques of the applications, and investigate differences between the process and outcomes of FTF vs. VC panels