Group Communication and Interaction in the Scientific Peer Review Process: How Score Calibration Talk Influences Reviewer Reliability

Abstract

In scientific grant peer review, groups of expert scientists meet to engage in the collaborative decision-making task of assigning scores to grant applications in order to determine the research projects that receive funding. The current study explores the decision-making practices of peer review panels by examining the degree of variability in how multiple panels score the same applications both before and after collaborative discussion. Videotapes of reviewers' interactions are then analyzed for instances of one particular form of communicative behavior—*score calibration talk*—as a potential source of the variability we observe. Results suggest that different individual reviewers score the same grant application differently, that this variability is exacerbated by collaborative group discussion, and that score calibration talk plays an instrumental role in contributing to this inter-panel scoring variability during grant peer review.

The grant peer review process is one of the keystones of scientific research, providing the means by which scientists secure funding for their research programs. The primary funding agency for most clinical, behavioral, and biomedical research in the U.S., the National Institutes of Health (NIH), spends more than 80% of its $30.3 billion annual budget on funding research grants evaluated via peer review (NIH, 2015). The mechanism by which this money is allocated to scientists is via collaborative peer review panels (referred to as "study sections" by NIH), in which expert scientists convene to evaluate grant applications and to assign them numeric scores that are subsequently used for funding decisions. Thus, understanding how peer review functions to serve its intended purpose of identifying and funding the most promising and innovative scientific research is crucial for the scientific community writ large.

Theoretical Framework

The peer review process is structured around study sections that leverage the distributed nature of reviewers' expertise (Brown, Ash, Rutherford, & Gordon, 1993), as reviewers are assigned to evaluate applications based on their particular domain of expertise and are then expected to share their specialized knowledge with others who have related but distinct expertise. Thus, the very structure of study sections facilitates what Brown and colleagues (1993) identify as *mutual appropriation* among groups of distributed experts, whereby participants put forth ideas and knowledge that are then taken up by other participants. In addition, study sections can be thought of as *communities of practice* (Lave & Wenger, 1991) characterized by their particular community norms, values, beliefs, practices, tools, and ways of being—what Gee (1991) refers to as their *big-D Discourse*. In these ways, the format of study sections not only makes the distributed nature of participants' expertise profoundly salient, but also underscores the locally constituted and highly situated nature of peer review as a process embedded in particular study section panels.

It is no surprise, then, that different study sections vary in terms of how reviewers' expertise is brought to bear during peer review meetings, despite the ostensible objectivity of peer review as an evaluative mechanism for awarding funding for the most meritorious scientific research. Indeed, research into peer review has established that, overall, inter-reviewer and inter-panel reliability is typically poor (e.g., Cicchetti, 1991; Langfeldt, 2001; Marsh, Jayasinghe, & Bond, 2008; Obrecht, Tibelius, & D'Aloisio, 2007; Wessely, 1998). Furthermore, several studies challenge the notion that collaborative reviewer discussion remedies this poor inter-reviewer reliability, as panel discussion has not been found to significantly impact grant review outcomes compared to independent peer review (e.g., Fogelholm et al., 2012; Johnson, 2008; Kaplan, Lacetera, & Kaplan, 2008; Obrecht et al., 2007). One study established that upwards of 25 to 40 reviewers may be required to achieve adequate inter-reviewer agreement, unless reviewers' scores differ by only a single standard deviation (Kaplan et al., 2008). These studies align with research in other domains that reveal the variability in decision making that emerges when collaborative groups are engaged in complex problem solving (e.g., Barron, 2000; Forman & Cazden, 1985; Hermann, Rummel, & Spada, 2001; Resnick, Salmon, Zeitz, Wathen, & Holowchak, 1993).

The current study aims to explore the group communication practices of distributed experts within multiple study sections in order to better understanding the variability in their decision making processes and potential sources of this variability. Importantly, NIH made significant changes to its scoring procedures in 2009, and to our knowledge, no study has yet examined inter-reviewer or inter-panel agreement at NIH since these changes were implemented. Additionally, no prior study has been able to peek into the previously obfuscated "black box" of NIH peer review meetings to analyze reviewers' patterns of interaction across multiple study sections. Given the importance of peer review for the field of science broadly, it is crucial to better understand the discourse practices of peer review as experts engage in collaborative decision making, as well as to identify potential sources of variability across review panels in order to evaluate possible steps for improving the reliability of peer review.

Methods

The research team recruited scientists with experience reviewing for NIH to participate in one of four "constructed study sections" (CSS). Participant reviewers evaluated applications reviewed between 2012 and 2015 by study sections within NIH's National Cancer Institute. Applications were donated by R01 Principal Investigators (PIs) listed on NIH's public access database, RePORTER. The CSSs were designed to follow the norms and practices of actual NIH peer review in all aspects of study design, and all methodological decisions were made in consultation with staff from NIH's Center for Scientific Review and with a retired Scientific Review Officer (SRO). As is typical in NIH peer review, the SRO in this study assisted with collecting applications, recruiting reviewers and chairpersons, assigning reviewers to applications, and facilitating each of the four CSSs. The research team additionally drew on ethnographic interviews with experienced NIH reviewers to ensure that the CSSs were in line with actual NIH study sections.

Prior to the review process, each grant application was de-identified, with the names of all research personnel replaced with pseudonyms and all identifying information changed. Each CSS was organized in virtually the same way as an NIH study section. Prior to the meeting, each panelist was assigned to review two grant applications as primary reviewer, two as secondary reviewer, and two as tertiary reviewer (with these designations dictating the order in which reviewers deliver their review of the applications during the meeting). As in NIH peer review, reviewers are responsible for reading each application assigned to them and writing a thorough critique of the application's strengths and weaknesses, including a holistic impression of the overall impact of the grant and an evaluation of five additional criteria: significance, investigator quality, innovation, methodological approach, and research environment. In addition to the written critiques, reviewers provide preliminary scores for each of the five criteria and an overall impact score. The NIH scoring system uses a reverse nine-point scale, with 1 corresponding to "Exceptional" and 9 to "Poor," and the SRO uses the preliminary impact scores to calculate an order of review, with the strongest applications reviewed first. Only the top 50% of applications, based on preliminary overall impact score, are discussed, with the bottom half of applications "triaged" out from discussion.

At study section meetings, the SRO begins by providing opening remarks, discussing the scoring system, and announcing the order of review. The chairperson introduces each application to be discussed, after which the three reviewers state their preliminary overall impact score in a process that Raclaw and Ford (2015) refer to as the "score-reporting sequence." Each reviewer then summarizes his or her critique of the application. Following this, discussion is opened up to the panel at large. After discussion, the chairperson summarizes each critique and the discussion, and then calls for the final overall impact scores from the three assigned reviewers, which then constitute the final "score range" for the panel-at-large for that application. Study section members who are not assigned to review that particular application privately record their final impact scores; scores are expected to fall within the final score range established by the assigned reviewers, and panelists are expected to publically announce if their score falls one or more points outside of the range and provide rationale for divergence.

Our constructed study sections had 42 reviewers nested within four panels: 10 panelists in CSS 1, 12 in CSS 2, 12 in CSS 3, and eight in the fourth CSS, which was conducted via videoconference rather than face-to-face. The research team videotaped the meetings and transcribed dialogue verbatim. We also tabulated the various quantitative outcome measures of each meeting: preliminary criteria scores and preliminary overall impact scores, final impact scores from the three assigned reviewers, and final impact scores from all panelists. Due to the small sample size precluding the use of inferential statistics, we take a case-study approach to these data and utilize descriptive statistics supplemented with qualitative excerpts of discourse from the review panels.

Research Questions

Given the variability that previous researchers have found in studies of grant peer review, this work addresses two research questions:

1. To what extent do we observe variability in how our four CSSs score the same grant applications?
2. What role does *score calibration talk* seem to play in the observed variability across our CSSs?

For Research Question #1, we hypothesize that we will observe substantial variability in how different panels score the same grant applications, descriptively speaking. For Research Question #2, we further hypothesize that one potential source of this variability stems from the distinct ways in which different panels adhere to scoring norms. In previous work (Authors, 2015), we classified discussion amongst panelists of what constitutes a given score as *score calibration talk* (SCT), and our current work (Authors, in preparation) explores how reviewers utilize SCT in order to address and cope with the inherent ambiguity of the review criteria set by NIH. The current study aims to build on this line of research by exploring how SCT may elucidate the sources of the variability we see in panel decision-making practices, thereby informing our understanding of how these groups of scientists communicate and collaborate during the high-stakes process of grant peer review.

Results and Discussion

Each of our four CSSs was assigned nearly the same pool of grant applications for review; however, panels with more reviewers were assigned additional applications to review, and certain applications were not assigned to particular panels due to differing expertise of the panelists or due to conflicts of interest. As a result of the triage process, in which only the top 50% of applications scored by a given panel are discussed, only two applications were discussed by all four CSS panels, five applications were discussed across three of the panels, five applications were discussed by two panels, eight applications were discussed by only one panel, and five applications were not discussed by any panels.

Inter-Panel Variability

To answer Research Question #1, we examined the preliminary and final impact scores given to each grant by its three assigned reviewers in each of the four CSS meetings (see Table 1). In terms of preliminary impact scores (first four columns of Table 1), there was little variability at the panel level: Average application scores (last row of Table 1) ranged from 3.29 to 3.60 (out of a possible range of 1.0–9.0). This suggests that different reviewers assigned to participate in different peer review panels assigned similar scores to the top half of the applications assigned to them, in the aggregate. However, at the level of individual grant applications, there was much more variability in preliminary scores. For example, all three CSSs assigned to review the Lopez application gave it an identical average preliminary impact score of 2.0 (2 = “Outstanding” per the NIH scoring rubric), making it the best scoring application in these three CSSs. By contrast, the Bretz application received scores ranging from 3.0 (3 = “Excellent”) to 6.3 (6 = “Satisfactory”) and thus was only discussed in CSS #3, as it was triaged out of discussion for the other two panels because of poor preliminary scores. Therefore, although the panels were highly similar in how they initially scored applications *overall*, the panels’ scoring practices varied substantially for *specific* grant applications.

This variability within grant applications becomes more pronounced when we investigate the three reviewers’ *final* impact scores (last four columns of Table 1)—that is, the scores assigned after collaborative discussion among all panelists during the meeting. Our data suggest that the process of collaborative discussion exacerbated inter-panel variability instead of facilitating score convergence. For example, for the two applications discussed across all four CSSs (Henry and Foster), the Henry application had a preliminary impact score ranging from 2.3 to 3.3 (range = 1.0) but a final impact score ranging from 3.0 to 5.0 (range = 2.0). Similarly, the Foster application had a preliminary impact score range from 2.7 to 3.3 (range = 0.6) but a final impact score range from 2.7 to 4.3 (range = 1.6). Additionally, when we look at the range of the reviewers’ scores for those applications discussed in each panel meeting (Table 2), the average standard deviation in scores nearly doubles after discussion, increasing from an average *SD*= 0.43 to *SD* = 0.83. This suggests that not only is there substantial variability in how different reviewers initially score the same grant applications, but that this variability *increases* as a function of collaborative discussion, and that the degree of this increase is itself quite variable across panels.

Table 3 depicts how this increased score divergence after collaborative discussion is taken up by the non-reviewing panelists, as well. For example, comparing the Abel application, which received final impact scores of 20.0, 29.1, and 50.0 in the first three CSSs, respectively, with the Amsel application, which received corresponding final impact scores of

50.0, 25.5, and 20.9, highlights the substantial inter-panel variability in final impact scores at the application level, despite having highly similar average final impact scores across applications (final row of Table 3). Therefore, once again we see that scoring patterns are similar across multiple panels in the aggregate, but that the outcome for any one particular grant application is highly dependent upon the particular panel to which it is assigned.

Overall, these findings are in line with other researchers' findings of high inter-reviewer and inter-panel variability of grant rankings (e.g., Langfeldt, 2001; Marsh et al., 2008; Obrecht et al., 2007), and aligns with findings from social psychology that collaborative groups have a tendency to make decisions that are more extreme than the initial judgments of individual members, known as *group polarization* (Moscovici & Zavalloni, 1969; Myers & Lamm, 1976). Given these results that collaborative discussion exacerbates instead of mitigates variability in how different reviewers score the same proposals, a crucial question emerges: What *features* of the collaborative discussion during peer review panel meetings may contribute to this scoring variability?

Score Calibration Talk

To answer Research Question #2, we investigated the extent to which discourse explicitly referencing the scores that reviewers had assigned to applications and the scoring procedures themselves might illuminate potential sources of inter-panel variability during collaborative peer review. In each of the four constructed study sections, we identified multiple instances of what we call *score calibration talk* (Authors, 2015). The following section will provide a few brief examples of score calibration talk to illustrate the sorts of interactions that may contribute to how different panels develop and adhere to locally constructed norms for scoring, despite utilizing the same ostensibly uniform NIH scoring system.

The following transcripts mark the current speaker at the far-left of each new turn at talk. “Ch” is used to indicate the Chairperson, while panelists are identified with the initials of their assigned pseudonym if they are non-reviewing panelists, and with initials followed by a dash and a number to indicate if a person is an assigned reviewer (i.e., -1 for primary, -2 for secondary, and -3 for tertiary).

In Excerpt #1, from CSS 1, the chairperson opens the discussion up to the non-reviewing panelists following the assigned reviewers’ initial summaries and critiques:

01 Ch: Any discussion on the application?
02 LA: I just have a little bit concern about giving a score of one to the grant. I
03 mean, in a regular NIH Study Section, you know, by definition, you
04 give a one to a grant that has no weaknesses, you know not major, not
05 minor. So.
06 GJ: It’s true.
07 LA: Just, just a comment.
08 Ch: So you mentioned that there’s some–Aim 3 is a bit of a weakness.
09 CV-1: Yeah, there are some–I mean, I thought there were minor weaknesses.
10 Uh I guess… you know, yeah, no application is perfect, I agree. I mean,
11 if we decide the top most, the best application–
12 JR: In your pile. Well, my only concern–
13 LA: Well, even with that, you know, even among—excuse me for
14 interrupting—even if it is the best in your pile, if it still has minor
15 weaknesses that you think of, you really cannot score it as a one.
16 Because you’re, you know, globally, you know, there’s so much–
17 JR: Well to begin with one is okay, but after discussion we’ll see. (*laughter*)
18 And I have a concern, that you give one, and both reviewers give one,
19 but the secondary reviewer is really nailing out three ambitious aims,
20 and it is not well connected. All the preliminary data from one cell line,
21 and the concerns about so many kinds of models, and there is not focus
22 at all, so how are you going to defend this grant to get one?

In this example, a non-reviewing panelist, LA, immediately brings up during the discussion phase of the meeting that the primary reviewer assigned this application a score of one—the best possible score—despite having listed several weaknesses for the application. The primary reviewer, CV-1, acknowledges in line 9 that there are “minor weaknesses,” but begins to argue in line 10 that even though “no application is perfect,” this applications was "the top most, the best application." She is interrupted in line 12 by another panelist, JR, who challenges CV-1's assertion by suggesting that this was only the best application in her particular "pile" of assigned applications, though not necessarily the best application overall. Though JR begins to articulate his own concerns with CV-1's scoring of the application, LA interrupts in line 13 by claiming that a score of one cannot be given to an application even if it has minor weaknesses. JR interrupts her in line 17 by saying that "to begin with [a] one is okay," suggesting that a *preliminary* score of one would be acceptable, but that "after discussion, we'll see," indicating his explicit expectation that more weaknesses may be drawn out during the collaborative discussion of the meeting that would motivate the score to be raised (i.e., worsened, given the reverse scale).

Excerpt #2 occurs later in this same meeting (CSS 1). As a secondary reviewer concludes his critique after providing a list of his concerns about the application, the chair interrupts and asks the reviewer to confirm that he had assigned a score of one to the application. The chair's question invites laughter from many of the panelists, and the chair smiles and laughs as he references the issues that LA and other panelists had previously brought up regarding how a score of one should be understood:

01 LZ-2: I don't have any other specific concerns. Overall they use this unique
02 mouse model to clarify what they trying to say in the culture system,
03 and it's–
04 Ch: You had a score of one, right? (*laughter from panelists*) One is uh, as
05 someone said (*points to and gazes at LA*) one is uh–
06 LZ-2: Then I reduce it. (*ongoing laughter*)
07 Ch: Okay.
08 LZ-2: That means I don't have any concern. That's what it is, you know,
09 everybody agrees that you–
10 Ch: Your concerns were somewhat serious. (*laughter*)
11 LZ-2: These are people who publish in *Nature, Science, Cell.* You still have a
12 lot of concern, too.
13 Ch: Well then it shouldn't be a one. (*Ch and JR laughing*).

Amidst the panelists' ongoing laughter, the secondary reviewer, LZ-2, loudly exclaims in line 6 that he will reduce (or worsen) his score. After the chair accepts this proposed score change in line 7, LZ-2 continues by articulating his own understanding of a score of one, that it is an indicator that he doesn't have any concerns with the application. The chair responds in line 10 by rejecting LZ-2's explanation, correcting him and claiming that his "concerns were somewhat serious," which also receives laughter and claims of agreement from some of the other panelists. The discussion continues to focus on reviewers' individual understandings about what constitutes a score of one—for example, that the investigators have previously produced work strong enough to make it into top-tier scientific journals like *Nature*, *Science*, and *Cell*—as the panelists collaboratively negotiate among themselves what the norm is for a score of one.

In addition to illustrating the operation of score calibration talk during the interactive discussion phase of the peer review meeting, the two transcripts above demonstrate how score calibration may directly influence reviewers' assignment of final scores. In the excerpt above, LZ-2 immediately agrees to worsen his score in response to score calibration talk. Another case can be seen in Excerpt #3 below, from CSS 1, as the assigned reviewers stated their final impact scores for the Henry application:

01 Ch: So with that, can we see the scores again?
02 GJ-1: I'll go up to four. After the discussion.
03 CV-2: I'll also go to four.
04 MP-3: Uh, I'm staying at four.
05 Ch: So again, four is, we've heard, four is a pretty good score and one aim
06 is really bad.
07 JR: I hope my grant will be discussed with these three people. (*laughter*)
08 MP-3: Okay, let me–I'll go to five.
09 CV-2: I'll go to five, too
10 GJ-1: Okay, I'll go to five. (*ongoing laughter from panelists*)

After the reviewers announce in lines 2 through 4 that they are each assigning scores of four to the application—a substantially weaker score range compared to the grant's preliminary scores of two, two, and four—the chair responds in line 5 by claiming that four is in fact "a pretty good score" that may not accurately reflect that one aim of the application was deemed "really bad" by reviewers. A non-reviewing panelist, JR, follows this in line 7 by joking that he would want this panel to review his own grants, given their bias towards stronger scores. Amidst the shared laughter that this joke receives, the third reviewer revises his final score to a weaker five (line 8), with the secondary and primary reviewer following suit in lines 9 and 10. It is apparent in this excerpt how episodes of explicit score calibration talk, particularly those involving joking or teasing by the panelists, serves to impact how the reviewers score the application at hand. Excerpt #4 presents another example of this, from CSS 2, during discussion of the Williams application:

01 Ch: Thank you. Uh, any other concerns? None of them? Panel?
02 JG: (*raises hand*) So, based on uh, Jihad [JA-3], what you said about the
03 knockout mice, this is serious stuff, and a score of two, I mean it's like
04 an outstanding grant.
05 JA-3: (*sighs*) Well–
06 JG: You know, this is such a critical point. I mean I do understand that
07 doing that they could have nothing. And if that is the focus–
08 RD: Two is a very good score. (*agreement from various panelists*)
09 JG: That's what I'm trying to say, I mean, two is like a precious score.
10 RD: Two is a very good score!
11 JA-3: I understand about–
12 Ch: Two is not a very good score, twos in an ideal world should be—
13 because one is exceptional. Ones come, according to Sun [another
14 panelist], once in a lifetime, once in a blue moon, so two, two should
15 be– (*raises both hands in the air above head*)
16 JA-3: No, I will revise my score.

Again, here we see non-reviewing panelists (JG and RD) voice their concerns that an assigned reviewer gave an inappropriate score—in this case, assigning a score of two to an application that the reviewer had raised concerns about, which JG frames in line 3 as "serious stuff." In response to this, the tertiary reviewer acquiesces and announces that he will change his final score to be worse in line 16. In this way, score calibration talk not only serves as a collaborative resource for participants to reach a local, normed understanding of the NIH, but also has the potential to directly influence reviewers to change their scores. In our constructed study sections, these changes were invariably from a relatively strong score to a weaker score.

We also see multiple instances of score calibration talk that are initiated by the reviewers themselves about their own assigned scores. For instance, in CSS 2, as the secondary reviewer concludes his critique of the Foster application, he states, "So I gave it a four, which was probably generous." A non-reviewing panelist, RH, displays his agreement by saying, "Yeah, that was generous." The secondary reviewer then continues his turn by saying, "Um, but I can live with it unless somebody wants to talk me to a bigger number." Here, the reviewer explicitly invites other reviewers and panelists to convince him to worsen his score (i.e., assign "a bigger number") based on his critique. Similarly, in Excerpt #5 from CSS 3, the chair calls for reviewers' preliminary scores during the score-reporting sequence (Raclaw & Ford, 2015) in the following exchange:

01 Ch: Score, please.
02 RD-1: Oh score. Sorry, two.
03 Ch: Two. Jack? [JK-2]
04 JK-2: I'm at three, but can be talked down.
05 PM-3: I was at five, but again, on the fence. (*laughter*)

Here, the secondary reviewer similarly invites others to "talk [him] down" to a worse score, whereas the secondary reviewer says she is "on the fence," indicating she is also open to being persuaded by others. In this case, the tertiary reviewer's explicitly stated willingness to change her score may result from her score being misaligned with the primary and secondary reviewers, who assigned the application much stronger scores of two and three. These two excerpts are an example of the “mutual appropriation” that Brown and colleagues (1993) argue is a hallmark of distributed expertise. It is not simply the reviewers who are conveying their knowledge to the non-reviewing panelists, but all members of the panel who are dialogically interacting in order to construct a local, collaborative evaluation of the grant applications under review.

Such mutual appropriation among these scientific experts and dialectical, local norming of scores is often explicitly conducted during the peer review meetings we examine. For example, during Excerpt #6 from CSS 4, when the chairperson calls for the final scores from the assigned reviewers for the Henry grant, the following exchange occurs:

01 Ch: Okay? So are there any, can I have the scores again, from the three
02 reviewers?
03 DJ-1: Yeah, so, I respect what the other reviewer said, so I will move my
04 score from two to three.
05 Ch: Okay. Dr. Rath? [MR-2]
06 MR-2: I’m also going to move from two to three.
07 Ch: Dr. Peters? [RP-3]
08 RP-3: Yeah, I’m gonna try and be fair. I mean, I think there’s a lot of
09 good in it, too. I’m gonna say four. I’m gonna go from three to
10 four.

There is explicit acknowledgement from the primary reviewer, DJ-1, in line 3 that the tertiary reviewer has raised some concerns in his critique that DJ-1 deemed significant enough to encourage him to move his score from a two to a three. The secondary reviewer, MR-2, follows suit in line 6, changing his score from a two to a three. However, the tertiary reviewer, RP-3, further reduces his score from a three to a four, and accounts for his change in score in line 8 by claiming that he is going to “try and be fair” in the scoring process. Such explicit acknowledgement of the role that other panelists and reviewers play in altering or impacting assigned reviewers’ scores is common, and suggests that it is expected that final scores are the result of the local discursive context of the review panel.

Overall, our constructed study sections make frequent use of score calibration talk, often from a non-reviewing panelist directed towards a reviewer, and frequently reoccurring across multiple grant applications throughout the entirety of the panel discussion, as we saw in the above excerpts from CSS 1. We also saw instances where reviewers themselves invite others to persuade them to change their initial scores, acknowledging an appropriate openness to being persuaded by their peers. We further saw examples of reviewers acknowledging the mutual appropriation occurring given the distributed expertise of the panelists—in other words, the reviewers explicitly state that their final scores reflect the collaborative discourse that occurs during discussion. This aligns with the conclusion drawn by Langfeldt (2001), that "while there is a certain set of [scoring] criteria that reviewers pay attention to—more or less explicitly—these criteria are interpreted or operationalized differently by various reviewers" (p. 821).

There is substantial prior literature establishing that in general, collaborative groups tend to outperform individuals on various problem-solving tasks (e.g., Cohen, 1994; Kirschner, Paas, Kirschner, & Janssen, 2011; Webb & Palinscar, 1996). Additionally, within the context of grant peer review specifically, there is evidence that individual reviewers' judgments may be fallible and subject to implicit cognitive biases (e.g., Ginther, Haak, Schaffer, & Kington, 2012; Ginther, et al., 2011 ; Kaatz, Magua, Zimmerman, & Carnes, 2015; Ley & Hamilton, 2008; Polhaus, Jiang, Wagner, Schaffer, & Pinn, 2011). Thus, there is theoretical support for the notion that collaborative peer review may improve the outcome of peer review over and above the contribution of individuals. However, the long-established social psychological phenomenon of *groupthink* (Janis, 1972; 1982), whereby groups of people may make faulty or irrational decisions in order to maintain group rapport and facilitate consensus, would suggest that peer review panels may be subject to social dynamics that would interfere with or hinder effective decision making. Our finding that collaborative discussion exacerbated the initial scoring variability we observed at the grant application level across all four CSSs suggest that groupthink dynamics may be at play during these meetings. Additionally, the immediate and publically announced changes in score that we observed immediately following SCT (Excerpts #2, #3, and #4) further suggest a tendency for panelists to maintain group rapport and cohesiveness; by altering their scores, the reviewers may be attempting to repair the discord brought about by score calibration talk that explicitly challenges a particular reviewers' scores. In this way, score calibration talk may play an instrumental role in the social and group dynamics influencing the increased variability we see in scores following collaborative discussion.

Conclusions and Implications

The current study sought to build upon prior research (e.g., Langfeldt, 2001; Marsh et al., 2008; Obrecht et al., 2007) reporting on low inter-reviewer reliability during peer review by examining how scoring practices may be influenced by the communicative behavior of the group. In our four constructed study sections, we found that although panels had highly similar score ranges across all of the grant applications discussed (i.e., the applications with overall preliminary impact scores in the top 50% of all applications reviewed), the panels varied substantially in how they scored particular grant applications, both in terms of the assigned reviewers and in terms of the panels as a whole. Importantly, to our knowledge, no previous studies other than that of Fogelholm and colleagues (2012) examined how multiple panels scored the same grant applications, so the present study serves to corroborate these findings that inter-panel variability is considerable.

In addition, this study leveraged our unique methodology of conducting constructed study sections that we were able to videotape in order to examine how one particular facet of group communication in peer review panels—*score calibration talk*—may illuminate when and how variability may unfold during the collaborative peer review process. As our data illustrate, during a particular panel meeting, members continually re-negotiate the meaning of particular scores during score calibration talk. We see how score calibration talk is mutually appropriated and invoked by panelists within the local setting of the particular study section meeting. Relying on the differential expertise of the particular scientists serving on a given panel would necessarily produce varying outcomes, but the score calibration talk evident in our constructed study sections suggests that such variability is a function not simply of the differential makeup of peer review panels, but of the locally constructed scoring and participation norms that members negotiate and mutually appropriate during collaborative discussion.

Future research into additional factors influencing scoring variability—including studies that more systematically vary certain aspects of the peer review setting, such as number of meeting participants, variations in application quality, or differential reviewing experience—would serve to corroborate and shed additional light on the findings presented here. Given the descriptive and qualitative nature of this work, we caution against strong conclusions about the reliability of peer review as a policy, and instead offer the present study as an initial foray into how our constructed study section methodology can shed light on the nature of group communication during grant peer review. Overall, though, the findings of this exploratory multiple-case study challenge the assumption that the distributed expertise of scientists collaborating on peer review panels would balance out individual biases or particularities in judgment to result in more objective and more fair outcomes.

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., & Gordon, A. (1993). Distributed expertise in the classroom. In G. Salomon (Ed.), *Distributed Cognitions: Psychological and Educational Considerations* (p. 188–228). Cambridge: 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.

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.

Forman, E. A., & Cazden, C. B. (1985). Exploring Vygotskian perspectives in education: The cognitive value of peer interaction. In J. V. Wertsch (Ed.), *Culture, communication, and cognition: Vygotskian perspectives* (pp. 323–347). New York: Cambridge University Press.

Gee, J. P. (1991). Socio-cultural approaches to literacy (literacies). *Annual Review of Applied Linguistics, 12*, 31–48.

Ginther, D. K., Haak, L. L., Schaffer, W. T., & Kington, R. (2012). Are race, ethnicity, and medical school affiliation associated with NIH R01 type award probability for physician investigators? *Academic Medicine, 87*(11), 1516-1524.

Ginther, D. K., Schaffer, W. T., Schnell, J., Basimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science, 333*(6045), 1015–1019.

Hermann, F., Rummel, N., & Spada, H. (2001). Solving the case together: The challenge of net-based interdisciplinary collaboration. In P. Dillenbourg, A. Eurelings, & K.

Hakkarainen (Eds.), *Proceedings of the first European conference on computer-supported collaborative learning* (pp. 293–300). Maastricht: McLuhan Institute.

Janis, I. L. (1972). *Victims of groupthink*. New York, NY: Houghton Mifflin.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes* (2nd ed.). New York, NY: Houghton Mifflin.

Johnson, V. E. (2008). Statistical analysis of the National Institutes of Health peer review system. *PNAS 105*, 11076–11080.

Kaatz, A., Magua, W., Zimmerman, D. R., & Carnes, M. (2015). A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. *Academic Medicine, 90*(1), 69-75.

Kaplan, D., Lacetera N., & Kaplan, C. (2008). Sample size and precision in NIH peer review. *PLoS ONE 3*(7): e2761. doi:10.1371/journal.pone.0002761

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.

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

Ley T. J. & Hamilton, B. H. (2008). The gender gap in NIH grant applications. *Science, 322*(5907), 1472–1474.

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.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12* (2), 125–135. doi:10.1037/h0027568.

Myers, D. G., & Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin, 83,* 602-627.

National Institutes of Health (NIH). (2015). NIH Budget. 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.

Pohlhaus J. R., Jiang, H., Wagner, R. M., Schaffer, W. T., & Pinn, V. W. (2011). Sex differences in application, success, and funding rates for NIH extramural programs. *Academic Medicine 86*(6), 759–767.

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.*

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and Instruction, 11,* 347–364

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

Tables

Table 1

Preliminary and final impact scores from three assigned reviewers for all applications

	Preliminary Impact Scores				Final Impact Scores			
Panel → **Application ↓**	**CSS 1**	**CSS 2**	**CSS 3**	**CSS 4**	**CSS 1**	**CSS 2**	**CSS 3**	**CSS 4**
Abel	2.0	2.7	3.3	3.7	2.0	2.7	5.0	▨
Adamsson	2.3	4.3	4.7	4.0	3.0	▨	▨	▨
Albert	2.3	5.0	4.3	3.3	3.0	▨	▨	3.7
Amsel	3.0	2.3	2.3	---	4.7	3.0	2.0	---
Bernard	4.3	4.3	---	---	▨	▨	---	---
Bretz	3.7	6.3	3.0	---	▨	▨	3.7	---
Edwards	3.7	2.7	4.7	4.0	▨	4.0	▨	▨
Ferrera	3.7	3.5	2.7	3.7	▨	▨	3.3	▨
Foster	2.7	3.0	2.7	3.3	4.0	3.7	2.7	4.3
Henry	2.7	2.7	3.3	2.3	5.0	3.0	3.7	3.3
Holzmann	3.3	---	4.3	2.3	▨	---	▨	2.3
Kim	4.0	3.3	5.3	---	▨	▨	▨	---
Lopez	2.0	2.0	2.0	---	2.0	2.0	1.7	---
Lukska	4.3	4.3	5.0	4.7	▨	▨	▨	▨
McGuire	---	4.0	4.0	---	---	▨	▨	---
McMillan	---	4.7	2.3	---	---	▨	3.3	---
Molloy	3.0	3.0	4.3	---	5.0	3.0	▨	---
Phillips	2.0	2.5	3.7	---	3.0	3.0	▨	---
Rice	4.0	2.7	3.0	3.7	▨	3.7	3.3	▨
Stavros	4.7	2.0	4.0	3.3	▨	2.7	▨	3.3
Washington	3.3	2.7	3.7	2.7	3.7	2.7	▨	2.3
Wei	3.7	4.3	3.7	4.0	▨	▨	▨	▨
Williams	3.3	3.3	2.3	2.7	3.7	▨	2.7	2.7
Wu	---	4.3	4.7	2.7	---	▨	▨	2.0
Zhang	4.3	4.0	3.0	4.3	▨	▨	3.0	▨
Average	**3.29**	**3.50**	**3.60**	**3.42**	**3.55**	**3.05**	**3.13**	**2.99**

Note. Cells with a dash indicate an application not assigned to a CSS. Cells shaded in grey indicate an application triaged out from panel discussion (i.e., received a preliminary score placing it in the bottom 50% of a CSS's applications).

Table 2

Score range of three assigned reviewers' for discussed applications

	Preliminary Range				**Final Range**			
	CSS 1	**CSS 2**	**CSS 3**	**CSS 4**	**CSS 1**	**CSS 2**	**CSS 3**	**CSS 4**
Lowest (Best) Score	2.00	2.00	2.00	2.33	2.00	2.00	1.67	2.00
Highest (Worst) Score	4.67	6.33	4.67	4.33	5.00	4.00	5.00	4.33
Average Score	3.11	3.35	3.42	3.29	3.55	3.05	3.13	2.99
Standard Deviation	0.51	0.34	0.44	0.43	1.07	0.57	0.89	0.80
Avg. Standard Deviation	0.43				0.83			

Table 3

Final impact scores from all panelists for discussed applications

Grant	**CSS 1**	**CSS 2**	**CSS 3**	**CSS 4**	**Average**
Abel	20.0	29.1	50.0		33.0
Adamsson	30.0				30.0
Albert	35.0			38.6	36.8
Amsel	50.0	25.5	20.9		32.1
Bretz			39.2		39.2
Edwards		37.3			37.3
Ferrera			33.3		33.3
Foster	42.0	38.2	29.2	45.0	38.6
Henry	52.0	35.5	35.0	32.5	38.8
Holzmann				27.5	27.5
Lopez	30.0	21.8	16.7		22.8
McMillan			30.8		30.8
Molloy	50.0	30.0			40.0
Phillips	31.1	30.8			31.0
Rice		39.1	31.7		35.4
Stavros		32.7		33.8	33.3
Washington	39.0	35.0		26.3	33.4
Williams	42.0		30.8	38.8	33.9
Wu				20.0	20.0
Zhang			29.2		29.2
Average	**38.3**	**32.3**	**31.5**	**31.6**	**32.8**