Research Evaluation, 26(1), 2017, 1–14
doi: 10.1093/reseval/rvw025
Article

'Your comments are meaner than your score': score calibration talk influences intra- and inter-panel variability during scientific grant peer review

Elizabeth L. Pier[1,2,*], Joshua Raclaw[1,3], Anna Kaatz[1], Markus Brauer[4], Molly Carnes[1], Mitchell J. Nathan[2] and Cecilia E. Ford[1,5]

[1]Center for Women's Health Research, University of Wisconsin-Madison, 700 Regent Street, Ste. 301, Madison, WI 53715, USA, [2]Department of Educational Psychology, University of Wisconsin-Madison, 1025 West Johnson Street, Madison, WI 53706, USA, [3]Department of English, West Chester University, 700 South High Street, West Chester, PA 19383, USA, [4]Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA and [5]Department of English, University of Wisconsin-Madison, 600 North Park Street, Madison, WI 53706, USA

*Corresponding author. Email: epier@wisc.edu.

Abstract

In scientific grant peer review, groups of expert scientists meet to engage in the collaborative decision-making task of evaluating and scoring grant applications. Prior research on grant peer review has established that inter-reviewer reliability is typically poor. In the current study, experienced reviewers for the National Institutes of Health (NIH) were recruited to participate in one of four constructed peer review panel meetings. Each panel discussed and scored the same pool of recently reviewed NIH grant applications. We examined the degree of intra-panel variability in panels' scores of the applications before versus after collaborative discussion, and the degree of inter-panel variability. We also analyzed videotapes of reviewers' interactions for instances of one particular form of discourse—*Score Calibration Talk*—as one factor influencing the variability we observe. Results suggest that although reviewers within a single panel agree more following collaborative discussion, different panels agree less after discussion, and Score Calibration Talk plays a pivotal role in scoring variability during peer review. We discuss implications of this variability for the scientific peer review process.

Key words: peer review; discourse analysis; decision making; collaboration.

As the primary means by which scientists secure funding for their research programs, grant peer review is a keystone of scientific research. The largest funding agency for biomedical, behavioral, and clinical research in the USA, 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 2016). As part of the mechanism by which this money is allocated to scientists, collaborative peer review panels of expert scientists (referred to as 'study sections' at NIH) convene to evaluate grant applications and assign scores that inform later funding decisions by NIH governance. Thus, deepening our understanding of how peer review ostensibly identifies the most promising, innovative research is crucial for the scientific community writ large. The present study builds upon existing work evaluating the reliability of peer review by examining how the discourse practices of reviewers during study section meetings may contribute to low reliability in peer review outcomes.

The NIH peer review process is structured around study sections that engender distributed expertise (Brown et al. 1993), as reviewers evaluate applications based on their particular domain(s) of expertise but then share their specialized knowledge with others who have related but distinct expertise. The very structure of study sections thus facilitates what Brown and colleagues (1993) identify as *mutual appropriation* among groups of distributed experts, which occurs as participants put forth ideas that are then adopted by others. Although Brown and colleagues devised these constructs in the context of a classroom, they can be fruitfully applied to the context of peer review to understand how panelists' understanding and evaluation of applications becomes co-constructed, as more knowledgeable and less knowledgeable experts collaborate to negotiate a shared understanding of an application's merits—in other words, to construct *common ground* (Clark and Brennan 1991; Baker et al. 1999; Paulus 2009). Participation in peer review panels thus features mutual *appropriation* of knowledge as expertise is distributed yet shared among the group, as well as mutual *construction* of common ground as these groups collaborate to make decisions.

Common ground must be established and maintained not only with regard to the content and perceived merits of grant applications, but also in terms of the evaluative criteria reviewers use. At NIH, reviewers employ a reverse nine-point scale, with each number corresponding to a single descriptive adjective (Fig. 1). What is meant by an 'exceptional' versus an 'outstanding' application, however, is left ambiguous so that reviewers can judiciously apply their expert scientific judgment (NIH 2015). Yet, prior research on evaluative decision making has established that when criteria are ambiguous, the judgments that result are highly susceptible to subjectivity and bias (Bernardin, Hennessey and Peyrefitte 1995; Heilman and Haynes 2008). Peer review is 'expected to operate according to values of fairness and expediency, [and] its product is to be trustworthy, high-quality, innovative knowledge' (Chubin and Hackett 1990: 3), but the ambiguity in review criteria threatens this fairness and trustworthiness. We argue that the collaborative and distributed nature of grant peer review potentially exacerbates this threat, as individual panels locally ground their interpretation of ambiguous evaluative criteria. We examine this claim by investigating how reviewers collaboratively negotiate their understanding of the criteria, as well as how such negotiations affect reviewers' scoring practices.

Overall Impact	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 1. Reverse nine-point scoring rubric used by NIH for the evaluation of R01 grants (adapted from NIH 2015).

The present study builds upon an extensive body of research into the process and outcomes of grant peer review. Such studies have drawn on interviews with various agencies' staff (Cole, Rubin and Cole 1978; Chubin and Hackett 1990) or with reviewers themselves (Hackett 1987; Lamont 2009); surveys of grant applicants (Gillespie, Chubin and Kurzon 1985; McCullough 1989); quantitative analyses of scores and funding outcomes (Klahr 1985; Sigelman and Scioli 1987; Wenneras and Wold 1997); and de-identified written critiques of funded applications (Porter and Rossini 1985). Such studies rely on *post hoc* data, with little attention paid to what goes on *during* review panel meetings; indeed, Olbrecht and Bornmann (2010) identified only five empirical studies that examined the judgment processes of peer review panels. One such example comes from Lamont (2009), who augmented interviews of reviewers with field notes taken from her observations of three interdisciplinary panels evaluating grants and fellowships. The present article aims to extend this work by assigning the *same pool* of grant applications to *multiple* review panels under highly controlled conditions, allowing for the repeated and detailed observation of discourse during the meetings. Furthermore, by videotaping the meetings themselves, we gain unprecedented access to the real-time interactions of different reviewers evaluating the same proposals.

The goals of existing research on grant peer review vary, ranging from examining bias in peer review outcomes (Sigelman and Scioli 1987; Wenneras and Wold 1997; Ley and Hamilton 2008; Ginther et al. 2011, 2012; Pohlhaus et al. 2011 ; Kaatz et al. 2015) to evaluating peer review panel efficiency (Klahr 1985; Gallo, Carpenter and Glisson 2013). Prior research examining the reliability of grant peer review has established that overall, reliability is typically poor (Cicchetti 1991; Wessely 1998). However, such research has either examined the reliability of a few independent reviewers' evaluations of grant applications (Cole and Cole 1981; Cole, Cole and Simon 1981; Marsh, Jayasinghe and Bond 2008) or examined how collaborative panel discussion affects peer review outcomes for a singular or for an isolated set of panels (Langfeldt 2001; Obrecht, Tibelius and D'Aloisio 2007). A notable exception comes from Fogelholm et al. (2012), who randomly assigned 30 reviewers to one of two panels to review the same pool of 65 grant applications to The Academy of Finland's Research Council for Health and found that panel discussion did not improve the reliability of independent reviewers' evaluations.

This study extends this body of work in several ways. We assigned reviewers to one of four different collaborative review panels evaluating the same pool of grant applications. We videotaped the real-time deliberations of experienced reviewers as they evaluated applications recently submitted to NIH. Due to the confidential and highly sensitive nature of NIH peer review meetings, it is not possible to gain access to actual NIH study section meetings.[1] Thus, we organize what we call *Constructed Study Sections* (CSSs; see Methods) to gain access into what both Chubin and Hackett (1990) and Lamont (2009) refer to as the 'black box' of peer review. Our study is the first to conduct fine-grained analyses on the videotaped interactions of multiple panel meetings evaluating the same pool of proposals.

The current article has three objectives: First, we attempt to corroborate prior work establishing low inter-reviewer reliability by examining scoring variability in how multiple reviewers evaluate the same grant applications. We extend this work by further evaluating how scoring variability changes *within* a given panel during collaboration, as well as by comparing the scores *between* multiple panels scoring the same application. Second, we qualitatively mine the discourse of panel meetings to examine how distributed groups of experts collaboratively establish and maintain common ground relating to score criteria. Finally, we connect how these grounding processes may relate to the variability observed between panels. Thus, we pose three research questions:

1. Do different panels of reviewers score the same applications similarly? In particular:
   (1.1)How variable are the preliminary scores of different reviewers who independently evaluate the same application?

(1.2) Do reviewers within a panel agree more (or less) as the peer review process unfolds?
(1.3) Do different panels agree more (or less) as the peer review process unfolds?
2. How do reviewers establish and maintain common ground in interactions during the peer review meeting?
3. Is there a relationship between this grounding process and scoring variability?

For Research Question #1, we hypothesize that we will observe substantial variability in how individual reviewers score the same applications initially (1.1), that this variability within a panel will decrease as a function of collaborative discussion (1.2), and that consequently, the Final Panel Scores of multiple panels evaluating the same application will be less variable than the preliminary scores (1.3). Research Question #2 is exploratory in that we plan to qualitatively mine our data for evidence of discourse related to the grounding of scoring criteria; therefore, we do not pose a specific hypothesis. For Research Question #3, we predict a correlation between instances of grounding discourse and the degree of within-panel score change.

Methods

We leverage a mixed-methods approach in this article that Creswell and Plano Clark (2011) label as a *concurrent embedded correlational mixed-methods design*, in which qualitative and quantitative data are simultaneously collected, with a qualitative strand integrated into an overall quantitative approach. The supplemental strand is added to provide 'a more complete understanding of an experimental trial, such as the process and outcomes' (Creswell and Plano Clark 2011: 73) and 'to enhance planning, understanding, or explaining of [the] primary strand' (p. 75).

Designing CSSs

Our research team employed a novel approach to collecting data on NIH peer review meetings: We recruited scientists with experience reviewing for NIH to participate in one of four CSSs. A key principle guiding our methodological choices was that our CSSs simulate as closely as possible the processes that experienced reviewers are familiar with from their service to NIH. To this end, every methodological decision was made by consulting with staff from NIH's Center for Scientific Review (CSR) and with a highly experienced, retired Scientific Review Officer (SRO; the NIH staff member who oversees a study section). Supplementary Appendix A details at length each methodological decision regarding SRO selection, application selection, deidentification of applications, reviewer selection, and reviewer assignments. The most pertinent of these are briefly described here.

Number of meetings. We consulted with CSR staff and our SRO to determine how many meetings would be ambitious, yet feasible given budgetary and logistical constraints. We conducted three panel meetings in a conference room at a hotel near the Chicago O'Hare airport (a location chosen for maximal convenience for reviewers located across the USA). These meetings were videotaped using three video cameras, and all spoken discourse was transcribed verbatim. Our discussions with CSR staff and the SRO indicated that it is becoming more common for NIH to conduct Video Assisted Meetings; thus, we organized and recorded via screen capture a fourth videoconference meeting to explore potential differences in this format and to serve as a pilot for future studies.

Selection of applications. Using NIH's public access database, RePORTER, our team solicited principal investigators who had submitted an NIH R01-type application between 2012 and 2015 to the Oncology 1 or Oncology 2 Integrated Review Groups within NIH's National Cancer Institute (see Supplementary Appendix A). Because the RePORTER database only enables searches of successfully funded applications, we solicited applications that were funded both as new applications and as resubmissions; for resubmissions, we requested copies of both the original unfunded application and the successfully funded resubmission, but utilized the original unfunded application, thereby providing us with a pool of higher- and lower-quality applications.

The SRO selected 25 applications from this pool based on the complementarity of their methodologies and topics of study to create a cohesive set of applications for review (as a real study section would have). Sixteen of these applications (64%) were initially funded, whereas nine (36%) were not funded when submitted for the first time. We timed each meeting to ensure that the panel spent approximately 15 minutes reviewing each application, creating a level of time pressure similar to what reviewers report experiencing in actual NIH study sections.

Prior to the review process, each application was de-identified, with the names of all research personnel replaced with pseudonyms and all identifying information altered or removed (see Supplementary Appendix A). To reduce the possible effects of prior knowledge of the applicants or applications, participating reviewers were instructed to only refer to the materials our team sent to them (as is the case for real NIH study sections). Reviewers were further asked to disclose any institutional or other conflicts of interest with an application, as is standard NIH practice.

Selection and assignment of reviewers. Our SRO used the NIH RePORTER database to identify investigators who received an R01 award from the NIH National Cancer Institutes between 2012 and 2015 and determine an appropriate pool of potential reviewers given the applications to be reviewed, which is in line with the standard procedures SROs use to identify reviewers for a NIH study section. Reviewers were invited to participate in a CSS, and those who accepted received a $500 honorarium with travel provided to Chicago, Illinois (reviewers for NIH typically receive financial compensation for their time). The SRO reviewed the pool of recruited reviewers, and based on each reviewer's expertise, she assigned them to review a set of applications fit to their scientific background; the reviewers could thus be assigned to the four CSSs so as to ensure an even spread of expertise across the meetings.

We learned from CSR staff and our SRO that the average number of reviewers assigned to a study section could vary widely, with typical rosters for NIH standing study sections within the Oncology 1 and Oncology 2 groups ranging from 18 to 21 reviewers; however, panels of *ad hoc* reviewers and Special Emphasis Panels (SEPs) might have as few as 8–10 reviewers. In consultation with our SRO, we targeted panels of 12 reviewers as being within the 'normal' range of what one might see in an actual NIH study section while allowing for feasible recruitment of both reviewers and applications. For the videoconference meeting, we recruited eight reviewers, given that Video Assisted Meetings at NIH may frequently be smaller in size with lesser reviewer loads. We further learned that although a reviewer may be assigned 9–10 applications for a standing study section, *ad hoc* panels or SEPs can receive assignments as low as 5–6 applications; thus, the SRO assigned each reviewer to evaluate six applications based on their scientific expertise, as we believed a reviewer load on the low end of what is typical would increase the likelihood of study participation. Each reviewer was assigned to serve as the primary, secondary, and tertiary reviewer for two applications; the SRO assigns these designations based on the relevance of reviewers' expertise for the assigned grant, with the primary reviewer having the most relevant expertise. Each application has three assigned reviewers in a given panel meeting.

Our CSSs had 42 reviewers nested within four panels: 10 reviewers in CSS1, 12 in CSS2, 12 in CSS3, and 8 in CSS4 (conducted via videoconference). Reviewer demographic information (gender, race/ethnicity, and tenure status) is available in Supplementary Appendix B (Table B1).

Workflow of a study section meeting. At NIH, 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 criteria: significance, investigator qualifications, innovation, methodological approach, and research environment. The three assigned reviewers provide preliminary scores for each of the five criteria and an overall score based on the reverse nine-point scale in Fig. 1. Prior to the meeting, the SRO calculates the average preliminary score across the three assigned reviewers for each application; the SRO then determines the order of review based on average preliminary score, with the strongest applications reviewed first. Only the top 50% of applications, based on average preliminary score, are discussed during the meeting, with the bottom half 'triaged' from discussion (and thus not considered further for funding).

During panel meetings, the SRO begins by providing opening remarks, reviewing the scoring system, and announcing the order of review. The chairperson introduces the application to be discussed, and the three assigned reviewers state their preliminary scores in a process that Raclaw and Ford (2015) refer to as the 'score-reporting sequence'. Reviewers then summarizes their critique of the application. Following this, the chair opens discussion to the panel-at-large. After discussion, the chairperson summarizes each critique and the discussion, and then calls for the three assigned reviewers to publicly announce their final scores, the lowest and highest of which constitute the 'score range' for that application. Panel members who were not assigned to that application as a reviewer privately record their final scores, which are expected to fall within the score range; if panelists' scores fall outside the range, they are expected to announce that fact and may choose to provide rationale for their divergence. As a result, there are as many final scores as there are panel members, and the final panel score is the average of these individual scores, multiplied by 10 (such that Final Panel Scores range from 10 to 90). These Final Panel Scores are subsequently used by NIH governance to inform their decisions about which applications to fund.

Table 1. List of applications assigned to each CSS and discussed by each CSS

Application	CSS1	CSS2	CSS3	CSS4
Wu		☒	☒	•
Lopez	•	•	•	
Holzmann	☒		☒	•
Zhang	☒	☒	•	☒
Adamsson	•	☒	☒	☒
McMillan		☒	•	
Phillips	•	•	☒	
Amsel	•	•	•	
Abel	•	•	•	☒
Ferrera	☒	☒	•	☒
Stavros	☒	•	☒	•
Washington	•	•	☒	•
Williams	•	☒	•	•
Rice	☒	•	•	☒
Albert	•	☒	☒	•
Edwards	☒	•	☒	☒
Foster	•	•	•	•
Henry	•	•	•	•
Bretz	☒	☒	•	
Molloy	•	•	☒	
Wei	☒	☒	☒	☒
McGuire		☒	☒	
Kim	☒	☒	☒	
Bernard	☒	☒		
Lukska	☒	☒	☒	☒

Note. • designates a CSS discussed the application (i.e. it ranked in the top half of the CSS's applications based on the three assigned reviewers' preliminary scores). ☒ designates a CSS did not discuss an application (i.e. it was triaged out from discussion because it ranked in the bottom half of the CSS's applications). A blank cell designates the application was not assigned to a CSS.

Use of 'constructed' methodology

Our methodology is novel in using CSSs as a means to investigate the process of grant peer review. However, we draw upon similar methodologies used in other domains, including the use of standardized and simulated patient interactions in medical education (Kruijver et al. 2001; Dotger, Dotger and Maher 2010; Helitzer et al. 2011; Young and Parviainen 2014); group simulations in medical and nursing education (Moyer 2016; Stroben et al. 2016) and in clinical psychology (Crosbie-Burnett and Eisen 1992; Romano 1998; Romano and Sullivan 2000; Kane 2003); and the use of simulated juries or 'mock trials' in legal studies (Foss 1976; McQuiston-Surrett and Saks 2009). The literature shows that the outcomes of simulated and constructed juries closely align with those of real juries. For example, Bornstein (1999) conducted a literature review on simulated juries and concluded that the findings 'bode well for the feasibility of generalizing from simulation studies to the behavior of real jurors' (p. 88). Relatedly, a study by Kerr, Nerenz and Herrick (1979) found that when comparing the performance of mock jurors with those tasked with a real evaluation task, both sets of jurors 'delivered nearly identical verdicts, made similar sentence recommendations, deliberated under the same decision scheme for the same length of time, and applied the same criterion of reasonable doubt' (p. 350–1). These findings suggest that our CSSs will likely lead to outcomes on par with actual study sections.

At the end of each CSS, we asked participant reviewers to rate on a seven-point Likert scale how closely their experiences matched that of actual NIH study sections (1 = completely different, 7 = identical). Thirty-seven of the 42 reviewers (88.1%) completed the survey, and of those, 81.08% rated the pre-meeting process as very similar (6) or identical (7) to a real NIH study section ($M = 5.78$, $SD = 1.29$), and 78.38% rated the review meeting itself as very similar or identical to a real NIH study section ($M = 5.70$, $SD = 1.27$). Thus, it is reasonable to infer from our participants' responses that our CSSs strongly resemble study sections conducted by NIH.

Coding of discourse

For the qualitative strand of our mixed-methods approach, two members of the research team exhaustively coded the transcripts from all four CSSs for any instance of discourse related to scores or scoring norms. Members of the research team met to inductively and collaboratively evaluate whether each extracted instance should be included in our collection of discourse relevant to score grounding; once the team determined the final list of instances to include, two members of the research team independently timed the number of seconds spent on each instance. Due to the relatively small number of instances of this genre of discourse and the emergent nature of the qualitative analysis, any discrepancies in the start and stop times for a given instance were discussed and resolved between the two researchers.

Results

Due to the peer review triage process, in which only the top 50% of applications that underwent preliminary review by a given panel are discussed, two grant applications were discussed in all four CSS panels, five applications in three panels, five applications in two panels, and eight applications in one panel; five applications were not discussed by any panel. Table 1 lists all applications in the study (designated by the pseudonym of the application PI), organized from the best (i.e. lowest) average final panel score to the worst (i.e. highest) average final panel score; it pictorially displays which applications were discussed by each panel (a filled circle) compared to which were triaged from discussion (an x-filled box).

Table 2 provides the average preliminary score from the three assigned reviewers ('Preliminary Reviewer Scores'), the average final score from the three assigned reviewers ('Final Reviewer Scores'), and the final panel score from all participating panelists ('Final Panel Scores'). Following NIH standard practice, the Final Panel Score is the average of all panelists' scores (i.e. the three assigned reviewers' plus all non-reviewer panelists) multiplied by 10. The scores from the actual NIH study sections that reviewed each application are available in Supplementary Appendix B (Table B2).

RQ 1: Do different panels of reviewers score the same applications similarly?

RQ 1.1: How variable are the Preliminary Reviewer Scores?

Descriptively, when we average across all applications, we see that the four CSSs have similar *average* Preliminary Reviewer Scores, ranging from 3.29 to 3.60 (final row of Table 2). Thus, there was not one panel that was consistently harsher or more lenient than the others. However, disaggregating the data to the level of the *individual application* reveals much more variability in how reviewers assigned scores to their pool of applications. For example, the Bretz application received scores ranging from 3.0 (3 = Excellent) to 6.3 (6 = Satisfactory) and thus was only discussed in CSS3, as it was triaged out of discussion for the other two panels.

To quantify this variability in the Preliminary Reviewer Scores, we computed Krippendorff's α (Hayes and Krippendorff 2007; Krippendorff 2013), an inter-rater reliability coefficient that accommodates more than two raters and allows for missing values. According to Krippendorff (2004), values of α above 0.80 are considered 'reliable', and values between 0.67 and 0.80 are considered suitable for drawing tentative, but not definitive, conclusions (p. 241). When considering all applications assigned ($n = 25$) across the 12 assigned reviewers (three reviewers per panel), $\alpha = 0.084$, suggesting extremely poor inter-rater reliability among independent reviewers, even before the study sections convene.

When considering *only* the top scoring 50% of the applications discussed within a given panel, $\alpha = -0.088$ for the independent reviewers; a strongly negative value suggests that 'coders consistently agree to disagree…, follow different coding instructions or have conflicting understanding of them' (Krippendorff 2008, p. 329). In addition, we assessed whether panels ended up discussing similar pools of applications by assigning discussed applications a code of 1 and non-discussed applications a code of 0 and computing Krippendorff's α for this binary variable; we found that $\alpha = 0.200$ for the four CSSs.

Summary. Our hypothesis for RQ1.1 was supported, in that we found substantial variability in how reviewers preliminarily scored the same applications. In addition, there was low agreement as to which applications were discussed (i.e. the top 50% of scores) and which were triaged out (i.e. the bottom 50%).

RQ1.2: Do reviewers within a panel agree more (or less) as the peer review process unfolds? We conducted paired-sample *t*-tests for each panel comparing the *range* of the Preliminary Reviewer Scores with the *range* of the Final Reviewer Scores.[2] For all four panels, there was a significant difference between the preliminary score range and the final score range. For CSS1 ($t_{10} = 3.993$, $p = 0.003$), the range decreased from $M = 1.909$ ($SD = 0.944$) to $M = 0.727$ ($SD = 0.905$). For CSS2 ($t_{10} = 4.485$, $p = 0.001$), the range decreased from $M = 1.909$ ($SD = 1.300$) to $M = 0.727$ ($SD = 0.786$). For CSS3 ($t_{10} = 2.803$, $p = 0.019$), the range decreased from $M = 2.091$ ($SD = 1.22$) to $M = 1.091$ ($SD = 0.539$). Finally, for CSS4 ($t_7 = 2.966$, $p = .021$), the range decreased from $M = 1.750$ ($SD = 0.886$) to $M = 0.875$ ($SD = 0.354$). Therefore, in all four meetings, the reviewers within a panel tended to agree more after collaborative discussion than before.

We also computed Krippendorff's α statistics as another measure of intra-panel agreement. Table 3 lists the α values for each CSS for the three sets of scores. There is very low agreement within each panel for the Preliminary Reviewer Scores, ranging from $\alpha = 0.028$ to $\alpha = 0.135$ (Average $\alpha = 0.090$).[3] However, intra-panel agreement substantially improves after collaborative discussion, with Final Reviewer Scores ranging from $\alpha = 0.334$ to $\alpha = 0.717$ (Average $\alpha = 0.553$), and incorporating the additional panelists' scores into the Final Panel Scores improves reliability even further, ranging from $\alpha = 0.477$ to $\alpha = 0.793$ (Average $\alpha = 0.665$). The gray bars in Fig. 2 depict these average intra-panel Krippendorff's α values.

Summary. Our hypothesis for RQ 1.2 was supported. Both the paired-samples *t*-tests and the Krippendorff's α comparisons showed that agreement among reviewers within a panel substantially improves after collaborative discussion for both the Final Reviewer Scores and the Final Panel Scores.

RQ 1.3: Do different panels agree more (or less) as the peer review process unfolds?

We conducted a paired-samples *t*-test between the *range* of the panels' Preliminary Reviewer Scores (i.e. the range of the four average Preliminary Reviewer Scores from each panel) and the *range* of the panels' Final Reviewer Scores (i.e. the range of the four Final Reviewer Scores from each panel) for those applications that were discussed in two or more panels. In other words, this test compared the spread among panels of the Preliminary Reviewer Scores for a given application with the spread among panels of the Final Reviewer Scores for that same application. Despite the small sample size ($n = 12$), there was a significant difference at the $\alpha = 0.10$ level

Table 2. Scores for all applications across all four CSSs

	Preliminary Reviewer Scores					Final Reviewer Scores					Final Panel Scores				
Panel → *Application ↓*	CSS1	CSS2	CSS3	CSS4	Average	CSS1	CSS2	CSS3	CSS4	Average	CSS1	CSS2	CSS3	CSS4	Average
Wu		4.3	4.7	2.7	3.89		ND	ND	2.0	2.00		ND	ND	20.0	20.0
Lopez	2.0	2.0	2.0		2.00	2.0	2.0	1.7		1.89	30.0	21.8	16.7		22.8
Holzmann	3.3		4.3	2.3	3.33	ND		ND	2.3	2.33	ND		ND	27.5	27.5
Zhang	4.3	4.0	3.0	4.3	3.92	ND	ND	3.0	ND	3.00	ND	ND	29.2	ND	29.2
Adamsson	2.3	4.3	4.7	4.0	3.83	3.0	ND	ND	ND	3.00	30.0	ND	ND	ND	30.0
McMillan		4.7	2.3		3.50		ND	3.3		3.33		ND	30.8		30.8
Phillips	2.0	2.5	3.7		2.72	3.0	3.0	ND		3.00	31.1	30.8	ND		31.0
Amsel	3.0	2.3	2.3		2.56	4.7	3.0	2.0		3.22	50.0	25.5	20.9		32.1
Abel	2.0	2.7	3.3	3.7	2.92	2.0	2.7	5.0	ND	3.22	20.0	29.1	50.0	ND	33.0
Ferrera	3.7	3.5	2.7	3.7	3.38	ND	ND	3.3	ND	3.33	ND	ND	33.3	ND	33.3
Stavros	4.7	2.0	4.0	3.3	3.50	ND	2.7	ND	3.3	3.00	ND	32.7	ND	33.8	33.2
Washington	3.3	2.7	3.7	2.7	3.08	3.7	2.7	ND	2.3	2.89	39.0	35	ND	26.3	33.4
Williams	3.3	3.3	2.3	2.7	2.92	3.7	ND	2.7	2.7	3.00	42.0	ND	30.8	38.8	33.9
Rice	4.0	2.7	3.0	3.7	3.33	ND	3.7	3.3	ND	3.50	ND	39.1	31.7	ND	35.4
Albert	2.3	5.0	4.3	3.3	3.75	3.0	ND	ND	3.7	3.33	35.0	ND	ND	38.6	36.8
Edwards	3.7	2.7	4.7	4.0	3.75	ND	4.0	ND	ND	4.00	ND	37.3	ND	ND	37.3
Foster	2.7	3.0	2.7	3.3	2.92	4.0	3.7	2.7	4.3	3.67	42.0	38.2	29.2	45.0	38.6
Henry	2.7	2.7	3.3	2.3	2.75	5.0	3.0	3.7	3.3	3.75	52.0	35.5	35.0	32.5	38.7
Bretz	3.7	6.3	3.0		4.33	ND	ND	3.7		3.67	ND	ND	39.2		39.2
Molloy	3.0	3.0	4.3		3.44	5.0	3.0	ND		4.00	50.0	30.0	ND		40.0
Wei	3.7	4.3	3.7	4.0	3.92	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
McGuire		4.0	4.0		4.00		ND	ND		ND		ND	ND		ND
Kim	4.0	3.3	5.3		4.22	ND	ND	ND		ND	ND	ND	ND		ND
Bernard	4.3	4.3			4.33	ND	ND			ND	ND	ND			ND
Lukska	4.3	4.3	5.0	4.7	4.58	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
Average	**3.29**	**3.50**	**3.60**	**3.42**	**3.47**	**3.55**	**3.05**	**3.13**	**2.99**	**3.16**	**38.3**	**32.3**	**31.5**	**31.6**	**32.8**

Note. ND indicates an application was not discussed (i.e. triaged out) for a CSS. A blank cell indicates an application was not assigned to a CSS. the Preliminary Reviewer Scores and the Final Reviewer Scores range from 1 to 9. The Final Panel Scores are the average of all panelists' scores (i.e. the three assigned reviewers and all non-reviewing panelists) multiplied by 10.

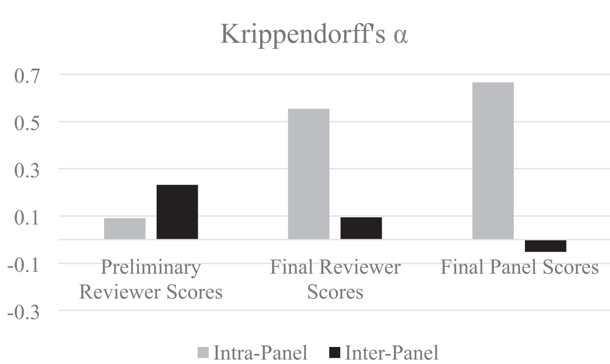

Figure 2. Graph of the Krippendorff's α statistics for the Preliminary Reviewer Scores, Final Reviewer Scores, and Final Panel Scores both within each CSS panel ('intra-panel') and between the four CSS panels ('inter-panel'). A value of $\alpha \geq 0.8$ is considered 'reliable', and values between 0.67 and 0.8 are considered suitable for drawing 'tentative conclusions' (Krippendorff 2004). All applications are included.

($t_{11} = -2.185$, $p = 0.051$), such that the between-panel range of Preliminary Reviewer Scores was smaller ($M = 0.708$, $SD = 0.451$) than the between-panel range of Final Reviewer Scores ($M = 1.306$, $SD = 0.969$). Thus, there is *greater* variability among panels in the score an application receives after collaborative discussion compared to before.

Table 3. Krippendorff's α (intra-panel agreement) for each CSS

	CSS1	CSS2	CSS3	CSS4	Average
Preliminary Reviewer Scores	0.028	0.135	0.132	0.067	0.090
Final Reviewer Scores	0.717	0.334	0.531	0.630	0.553
Final Panel Scores	0.793	0.477	0.707	0.683	0.665

We also computed Krippendorff's α statistics to measure inter-panel variability for the same applications (black bars in Fig. 2). Inter-panel agreement on the Preliminary Reviewer Scores is relatively low ($\alpha = 0.231$).[4] Consistent with the analyses reported in the previous paragraph, inter-panel agreement on the Final Reviewer Scores for a given application *decreases* ($\alpha = 0.095$) and, somewhat surprisingly, inter-panel agreement on the average Final Panel Scores further decreases to $\alpha = -0.052$, suggesting a slight degree of *disagreement* between panels in the Final Panel Score for a given application.

Summary. Taken together, our hypothesis for RQ1 was partially supported. We observed substantial variability in how individual reviewers independently score the same application prior to convening the panel meeting (RQ 1.1). Importantly, this variability is *reduced* within a panel as a function of collaboration during peer review meetings (RQ 1.2). However, contrary to our prediction, the variability *between* panels is not reduced, as the range and the

Table 4. Instances of score calibration talk (SCT) in each CSS

	CSS1	CSS2	CSS3	CSS4	Total
Self-initiated SCT					
Number of instances	15	18	11	12	56
Time (m:s)	3:33	4:36	2:09	2:37	12:55
Other-initiated SCT					
Number instances	7	3	4	1	15
Time (m:s)	6:07	4:28	5:27	1:46	17:48
Total SCT					
Number instances	22	21	15	15	71
Time (m:s)	9:40	9:04	7:36	4:23	30:43

variability of scores between panels increase after discussion (RQ 1.3). We next turn to Research Question 2 as a qualitative and exploratory attempt to better explain these patterns.

RQ2: How do reviewers establish and maintain common ground in interactions?

In prior work (Pier et al. 2015), we identified instances of discourse in which reviewers made explicit references to the scoring habits of fellow panelists, which we refer to as *Score Calibration Talk* (SCT). In this article, we expand upon this line of research by more clearly defining what SCT is, exploring how SCT establishes common ground surrounding the meaning of numeric scores, and examining the relative frequency of this form of discourse in our CSS meetings.

In our meetings, we identified 71 total instances of SCT. As Table 4 shows, we differentiate between what we term *self-initiated* SCT (S-SCT) and *other-initiated* SCT (O-SCT). We define S-SCT as any instance of discourse during which reviewers mention their scores (either the word 'score' or the number itself) *along with* an account (i.e. the rationale or justification) for that score. It is this public accounting of the reviewer's scoring practices that offers the opportunity to establish common ground with other panelists, as the account allows an individual reviewer to provide public access to their own understanding of what numerical scores mean, while also offering a space for other panelists to agree or disagree with this understanding. Thus, we do not consider a reviewer's saying, 'I gave it a four' to be an instance of S-SCT, whereas we do consider a reviewer's saying, 'The reason I gave it a four was mainly due to problems with the approach' to be an instance of S-SCT.

We define O-SCT as any instance of discourse during which one panelist calls on other reviewers to account for their scoring practices by offering a 'challenge' to their score. Here we adopt the analytic understanding of a challenge as it has developed in the field of Conversation Analysis, a qualitative approach to examining the interactional structures and processes of conversational and organizational discourse. Research in this area has examined the various ways that speakers recognizably produce challenges; for example, questions posed in environments where the questioner knows the answer (so-called 'known-answer' questions) are typically heard as offering a challenge (Levinson 1992; Koshik 2002; Schegloff 2007). Thus, we do not consider a question such as, 'What was your score again?' to necessarily be an instance of O-SCT when it is not otherwise followed by a challenge, whereas we do consider a question such as, 'You gave it a score of one, right?'—posed directly following the questioner's clear and repeated visual attention to their scoresheet to verify that the reviewer did, in fact, assign the application a preliminary score of 1—to be an instance of O-SCT. Some instances of O-SCT are extremely brief; for instance, in CSS3, a non-reviewing panelist remarks to the primary reviewer after his summary for the 'McMillan' application, 'Your comments are meaner than your score'. The primary reviewer does not acknowledge this comment and instead moves on to the project of wrapping up his summary, effectively ending that instance of O-SCT. Other instances of O-SCT, however, involve multiple speakers and can thus span multiple turns of talk, as the examples below demonstrate. Details on how we defined, counted, and timed SCT are in Supplementary Appendix C.

To paint a clearer picture of what kinds of discourse were coded as SCT, the following section provides a few illustrative examples. These examples were chosen by the research team as prototypical and representative instances of the kinds of SCT observed in our four CSSs.

Examples of SCT. In the transcripts below, the line number is found to the far left, followed by an identification of current speaker at each new turn at talk. 'Ch' is used to indicate the Chairperson, while panelists are identified with the initials of their assigned pseudonym.

Self-initiated Score Calibration Talk. We provide two typical examples of S-SCT. The first is from a review of the 'Henry' application from CSS1, during the phase of the meeting when the secondary reviewer, whose pseudonym is Camilla di Vitantonio (CV), provides her summary of the application. The transcript (Fig. 3) begins as the Chairperson thanks the primary reviewer for his summary, and then calls for the secondary reviewer to provide her summary in line 1. CV provides an account for her 'high score' in lines 2–3, stating that it was based on her perceived novelty of the application. She extends her turn in line 4 by saying she will likely worsen her score ('come down a little') because of some concerns she has with the application. She then uses a transition marker ('And um' in line 5) to initiate a new (but related) project of listing the relevant weaknesses of the application, which denotes the end of this instance of S-SCT. This example illustrates how this reviewer grounds her understanding of what constitutes a 'high score' in particular elements of an application—specifically here, novelty.

The second example of S-SCT occurred in CSS3 during the evaluation of the 'Zhang' application. The Chairperson begins in Fig. 4 by closing the question and discussion phase of the meeting in line 1 and then initiating the *final score reporting sequence* (Raclaw and Ford 2015), in which the reviewers announce the final scores they are assigning the application and the other panelists privately cast their own scores. The preliminary scores from the three assigned reviewers—Alice Kuntz (AK), Jeff Ma (JM), and Raj Dharanipragada (RD)—were a 1, 5, and 3, respectively. In line 3, the primary reviewer, AK, acknowledges that she is changing her score to a 2 (from a 1) but does not account for why. The secondary reviewer, JM, acknowledges in line 5 that he is also changing his final score, to a 3 (from a 5), and that this constitutes an 'improve[ment]' in line 6. However, JM further *accounts* for this change in line 5 by referencing the primary reviewer 'Alice' (i.e. AK) who had presented multiple strengths of the application in her earlier review. JM then further concedes in line 6 that he had been 'over, too harsh' with his preliminary score. Finally, the tertiary reviewer, RD, simply states his final score of 3 in line 7. The primary and tertiary reviewers thus do not engage in S-SCT, having only announced their final scores for the application, whereas the secondary reviewer additionally provides an account that constitutes S-SCT. This example depicts how a reviewer grounds his

Ch: Okay, thank you very much. Secondary?
CV: Um, I think that's me. I, so my um high score was driven
mainly by the enthusiasm for the novel area of research. Um
I probably will have to come down a little because um there
are concerns with this grant. And um one is the fishing
expedition. Part of the synthetic lethality, I agree with
that, but also looking at it more carefully, um their use
of mouse models, I'm not completely sure how that is going
to pan out...

Figure 3. Example #1 of S-SCT (from CSS1). The underlined portion of the transcript constitutes the S-SCT.

Ch: I think we have to keep on moving. Uh, recap our score,
please.
AK: I'll move to a two.
Ch: Two. Please.
JM: Actually I, I um, after uh, uh Al- Alice. Uh, three. I
improve my, uh, I think I was over, too harsh on this.
RD: Three.
Ch: Two, three, three. Two, three, three is the score.

Figure 4. Example #2 of S-SCT (from CSS3). The underlined portion of the transcript constitutes the S-SCT.

understanding of the scoring rubric in *another* reviewer's (AK's) assessment of the application. Thus, these two examples illustrate multiple means by which reviewers attempt to ground their understanding of the ambiguous review criteria as they engage in collaborative peer review.

Other-initiated Score Calibration Talk. We next present two examples of O-SCT as instances of discourse in which panelists challenge another reviewer's score, thereby grounding their own understanding of a given score's meaning in *another* person's use of that score. The first example (Fig. 5) comes from CSS2, during the question and discussion phase for the 'Foster' application. It begins when a non-reviewing panelist, Debra Burton (DB), addresses the secondary reviewer, 'Doug' Wolfe (DW; he is the chairperson of CSS2 but serves as the secondary reviewer for this application). DB first challenges DW's score by highlighting the misalignment between DW's 'average' summary of the application (line 2) and the score of 4 that DW had assigned, suggesting that it sounded more 'like a five' to her (lines 3–4). DW responds by acknowledging this misalignment (line 5) but proceeds to account for his score by listing some of the relative strengths of the application (lines 6–11). DB responds in lines 18–20 with a follow-up challenge, to which DW in turn responds to in lines 21–29. Then, another non-reviewing panelist, Renmei Huang (RH), presents his assessment of a key weakness of the application (line 30), and DW responds with his agreement, claiming at lines 33–36 that he would improve his score to a 2 in the future if the applicants addressed this weakness in a resubmission. DB then articulates the *upshot* (Heritage and Watson 1979) of DW's description of his scoring practice, asking, 'So you're encouraging them with a four, is that it?' (line 37), to which DW agrees. After a brief pause, the Vice-chairperson (who runs this portion of the meeting while the Chairperson serves as the secondary reviewer, marked on the transcript as VCh) begins a new project of opening the floor for more comments, thus ending the episode of O-SCT in our coding scheme. By calling on DW to account for his score (and producing an upshot of the rationale behind his score that illustrates her shared understanding of his account), DB works to establish common ground with DW (and potentially other panelists) based on the score that DW assigned and his accompanying evaluation of the application.

The second example of O-SCT (Fig. 6) occurred in CSS1 during the final score reporting sequence for the 'Molloy' application. The three assigned reviewers—Manish Patel (MP), Camilla di Vitantonio (CV), and Gopal Joshi (GJ)—each announce their final scores to be a 4 (lines 3–7). The Chairperson then begins to ask if any of the panelists will score outside this range (line 8), but halts this action prior to its completion to comment that 'these are pretty serious concerns that were raised' during the prior discussion of the application, considering 'four is a very high score' (line 9). This challenge to the reviewers' scores is met with agreement by a non-reviewing panelist, Jayan Ramachandran (JR) (line 10). The secondary reviewer, CV, responds to this challenge by worsening her score from a 4 to a 5. The tertiary reviewer immediately follows suit in line 14, as does the primary reviewer in line 15. The Chairperson calls for the final scores again in line 16 (erroneously referring to them as preliminary scores), and the three reviewers announce their new final scores of 5, 5, and 5. In this case, O-SCT begins with the Chairperson's challenge in line 8 and continues until the Chairperson begins the new project of re-initiating the final score reporting sequence in line 16, illustrating how brief yet influential SCT can be during study section meetings.

Summary. Reviewers engage in S-SCT as a means to account for the scores that they assign to applications. Such S-SCT can be used to ground their understanding of a particular score relative to certain aspects of an application (such as novelty) or in other reviewers' use of the scoring rubric. Reviewers engage in O-SCT as a means to challenge the scores that other reviewers assigned to applications. Such O-SCT can be used to establish common ground among participating panelists surrounding their shared understanding of the scoring rubric. We next turn to our third and final research question to explore the relationship between this genre of discourse and the patterns of scoring variability we observed in RQ1.

RQ3: Is there a relationship between grounding discourse and scoring variability?

RQ 1 established that reviewers *within* a given panel tend to agree more after collaborative discussion (RQ 1.2), whereas agreement *between* panels in the final scores is not improved as a function of collaboration (RQ 1.3). RQ 2 established that one way in which reviewers attempt to ground their understanding of a particular

DB: Doug, you make it sound, as the breast cancer expert, you make it sound like a very average application. That's why I'm thinking it's—you were talking, it sounded like a five to me.
DW: Yeah, and that's probably not fair on my part. I think, I mean there are strengths to it. I mean really, the preliminary data are, are, uh quite convincing. Um it's not hard to believe that the Cdk2-Cyclin-A connection probably is relevant. Does it change the way that we think about AKT signaling? Does it change the way in which we would approach clinical treatment? There, I'm not so sure.
DB: Um—
DW: And so, it may be, if they do this work, they will find something that we don't—I don't know what it would—I don't think they know what it would be, that it would change the way we would look at things. I just didn't see it.
DB: What about the disconnect that I heard between the triple negative preliminary data and, are they not using triple negative cells in the appro—*(DH shakes his head)* they are.
DW: They are. They use everything but they, but you know they have a little bit of Aim 1, they say, we'll do this in the tri—in these lines, and they're all triple negative, and then, they, they go at this other part and they tell you the cell lines, and they're mostly luminal, and there isn't a clear explanation and there isn't a clear descr— evidence from their preliminary data, why some parts of the pathway might be more relevant in luminal and some parts might be more relevant in triple negative.
RH: So it sounds like they don't quite understand the disease, you know.
DW: I think they just didn't think it through. You know, and this is the problem, you know a lot of my issues, they could probably come back and address and I'd be, okay go for it, you know here's a two. Ehh, but as it stands, I- I I I can't cut them any more slack than-
DB: So you're encouraging them with a four, is that it?
DW: Yes. That's kind of what I'm thinking.
VCh: *(Pause)* Any other comments?

Figure 5. Example #1 of O-SCT (from CSS2). Everything up until the Vice-chairperson begins a new project in line 39 is included in this instance of O-SCT (see Supplementary Appendix B).

Ch: Other comments? *(pause)* So with that, let's hear our new scores?
MP: So I'll move to a four.
Ch: Secondary?
CV: Uh, I'll move to a four also.
Ch: Dr. Joshi?
GJ: I had four to begin with and I'll stay there.
Ch: Anyone outside that ra—these are pretty serious concerns that were raised. Four is a very high score.
JR: Yeah.
CV: Yeah mine, actually go to a five.
AA: *(laughter)*
Ch: Okay.
GJ: I'll go to five.
MP: I'll go to five.
Ch: Let's go again. The preliminary—um new scores are? *(group laughter)*. Preliminary? Dr. Patil?
MP: Five.
CV: Five.
GJ: Five.
Ch: Okay. Anyone outside that range?

Figure 6. Example #2 of O-SCT (from CSS1). The underlined portion of the transcript constitutes the O-SCT.

score and of the scoring rubric is via S-SCT and O-SCT. RQ 3 bridges these two research questions by exploring a relationship between SCT and scoring variability. Although we do not have the means to examine whether SCT influences or causes reviewers to change their scores, we wish to instead preliminarily establish whether these phenomena might co-occur in our meetings.

To this end, we calculated several correlations to explore how SCT relates to reviewers' scoring behaviors. First, we computed the correlation between the number of times SCT occurred in a panel and the number of times reviewers changed their score during the meeting. The correlation between (i) Total SCT and score change was $r = 0.717$, (ii) S-SCT and score change was $r = 0.107$, and (iii) O-SCT and score change was $r = 0.978$. Given that each correlation is based on only eight data points (two for each CSS), we caution against strong inferences. However, it does suggest a robust relationship specifically between O-SCT and changes in reviewers' scores. We repeated these correlation analyses using the average percent of times a reviewer changed their score out of the total number of scores (e.g. if any reviewer changed his/her score 20 times out of a possible 33 scores, the average percent is $20/33 = 60.61\%$), and the pattern was strikingly similar: for Total SCT, $r = 0.608$, for S-SCT $r = -0.001$, and for O-SCT $r = 0.970$. Thus, this finding suggests that O-SCT is related to reviewers' score changes.

Next, we computed the correlation between the amount of time (in seconds) spent on SCT and the number of times reviewers changed their scores, in an attempt to use a more nuanced measure of SCT. The overall pattern was again the same: the correlation between the number of times reviewers changed their score and time spent on (i) Total SCT was $r = 0.809$, (ii) S-SCT was $r = 0.067$, and (iii) O-SCT was $r = 0.961$. Again, repeating these analyses with the average percentage of times a reviewer changed his/her score revealed a similar pattern: $r = 0.653$ for Total SCT, $r = -0.032$ for S-SCT, and $r = 0.824$ for O-SCT. This indicates a positive relationship between the amount of time a panel spends on O-SCT and the number of times reviewers alter their scores.

Finally, in an attempt to measure whether SCT is associated with score *convergence* within a panel, we calculated the range in Preliminary Reviewer Scores within each panel ('$\text{range}_{\text{prelim}}$') and the range in Final Reviewer Scores within each panel ('$\text{range}_{\text{final}}$'). We then computed the correlation between the difference in these ranges ($\text{range}_{\text{final}} - \text{range}_{\text{prelim}}$) and the (i) count of SCT and (ii) time spent on SCT. For the raw count, the relationship between Total SCT and the difference in score range was $r = 0.980$, for S-SCT $r = 0.682$, and for O-SCT $r = 0.858$. For the time spent, the relationship between Total SCT and the difference in score range was $r = 0.936$, for S-SCT $r = 0.657$, and for O-SCT $r = 0.788$. For both sets of correlations, the pattern mimics the ones found for the number/percentage of times reviewers changed their score above, but S-SCT seems to have a stronger relationship with score *convergence* than it does with how *often* reviewers change their score. Taken together, these correlations suggest a relationship between what we have deemed as SCT and the degree to which reviewers change their scores to converge toward a narrower score range.

Using only four data points precludes strong conclusions about these correlations; however, it is not appropriate to calculate these correlations at the level of the individual application for multiple reasons. First, there are relatively few instances of SCT for any given application (between 0 and 4, see Table B3), which restricts the variability observed in the SCT variable. Second, because there are no cases in which a panel engaged in more than one instance of O-SCT for a given application, the standard deviation of the count of O-SCT for each application is 0, preventing the calculation of a correlation coefficient. Relatedly, because CSS4 only had one instance of O-SCT, the standard deviation is 0 and the correlation cannot be calculated. Finally, because SCT can span across applications (e.g. be initiated when discussing one application and be later re-invoked during another application), discussions of applications are not independent of one another, and so it becomes an arbitrary distinction to make when considering how much SCT occurs during a meeting.

Relating within- and between-panel variability. We found a relationship between SCT and the number of times reviewers changed their scores, as well as between SCT and the degree to which reviewers' scores within a panel converged. In addition, RQ 1.2 found that reviewers *within* a panel tend to agree more following discussion than they did initially, which might suggest that collaborative peer review would serve to decrease the variability in how individual reviewers preliminarily evaluate grant applications (RQ 1.1). However, RQ 1.3 found that there is still substantial variability *between* panels following collaborative discussion. That is, although the *within-panel* score range significantly decreased from preliminary to final scores, the *between-panel* score range significantly increased from preliminary to final scores. This begs the question: Does score convergence within panels lead to greater between-panel variability? As a first pass at answering this question, we calculated the correlation between the difference in the range of the three reviewers' scores within panels ($\text{range}_{\text{final}} - \text{range}_{\text{prelim}}$) and the difference in the range of the average score between panels ($\text{range}_{\text{PanelFinal}} - \text{range}_{\text{PanelPrelim}}$). The correlation was $r = -0.606$, and with 20 data points (10 applications that were discussed by at least one panel), this is statistically significant ($p = 0.005$). Given the small sample size and the moderate size of the correlation, it is wise to interpret this with caution; nevertheless, it suggests that for a given application, the more the reviewers *within* a panel converge in the scores, there is a greater tendency for the scores *between* panels to diverge. Because SCT appears to be related to the number of times a reviewer changed his/her score, as well as to the degree to which reviewers converge in their scores, this implicates SCT as a potential factor influencing the high degree of variability—and therefore the low inter-panel reliability—that we observed in how our four CSS panels evaluated the same applications.

Summary. Our hypothesis for RQ3 was supported. We found that there is a strong correlation between the number of times (or the percentage of times) reviewers change their scores during a meeting and the number of instances of (or the duration of time spent on) O-SCT. We did not find the same robust relationship for S-SCT. We further found that O-SCT is strongly associated with score convergence within a panel, with a weaker relationship for S-SCT. Finally, we established that there is a moderately strong relationship between score convergence within a panel and score divergence between panels.

Discussion

This investigation found that there is little agreement among individual reviewers in how they independently score scientific grant applications, but that collaboration during peer review facilitates score convergence. Reviewers belonging to the same panel agree more with one another about an application's score after collaborating together, though separate panels reviewing the same proposals do not converge with one another.

Investigation of our first research question revealed that inter-rater reliability of individual reviewers' preliminary scores was extremely low. This aligns with others' work establishing low reliability in grant peer review (Cicchetti 1991; Wessely 1998; Langfeldt 2001; ; Obrecht et al. 2007; Marsh et al. 2008), as well as prior research demonstrating that ambiguous review criteria can lead to subjective or biased judgments (Bernardin et al. 1995; Heilman and Haynes 2008).

We next found that the range of scores significantly narrowed within each panel, indicating score convergence among the three assigned reviewers. Olbrecht and Bornmann (2010) detail how many well-established social psychological phenomena, including groupthink, motivation losses, and group polarization, may operate during panel peer review. Of particular relevance is the pressure toward *conformity* in groups (Asch 1951, 1955; see Cialdini and Goldstein 2004 for a comprehensive review). Individuals may be motivated by a desire to maintain an accurate interpretation of a situation, as well as by a desire to gain social approval from other people, which Deutsch and Gerard (1955) defined as *informational conformity* and *normative conformity*, respectively. In peer review meetings, when expert scientists are collaborating with their professional colleagues to make ostensibly objective decisions, both informational and normative motivations may be at play. Furthermore, when individuals are held accountable for their decisions and must explain them to others (as in peer review), they tend to conform to the majority's position more so than for private judgments (Lerner and Tetlock 1999, Pennington and Schlenker 1999).

However, this conformity does not necessarily translate into agreement or reliability across different groups engaged in the same decision-making task. We established that the range of scores between panels was greater following discussion compared to the preliminary score range, which aligns with prior work establishing variability among different groups engaged in decision making during complex tasks (Resnick et al. 1993; Barron 2000), as well as with research in peer review specifically establishing that panel discussion does not significantly impact outcomes compared to independent evaluations (Obrecht et al. 2007; Fogelholm et al. 2012).

Research Question 2 explored how reviewers who possess distributed expertise (Brown et al. 1993) work to establish and maintain common ground (Clark and Brennan 1991) relating to the scores they assign during peer review meetings. We found SCT to be one of the means by which such grounding occurs in our meetings. We defined, identified, and measured the number and duration of instances of SCT—both discourse involving a reviewers' accounts for their own scoring practices (self-initiated) and discourse involving a panelist challenging another's scoring practices so as to invite an account (other-initiated). Our examples illustrate the range of ways in which reviewers ground their understanding of their own and of others' scoring practices. As reviewers grapple with the ambiguity inherent in the NIH review criteria, there are multiple instances in each meeting when panelists are faced with conflicting interpretations of what constitutes a given score, and it is through SCT that they attempt to resolve such conflicts. In this way, SCT facilitates the mutual appropriation (Brown et al. 1993) of locally constructed scoring norms, which theories of social constructivism (Vygotsky 1978) would predict are publicly negotiated via SCT and subsequently internalized by an individual reviewer, as manifested by a reviewer's changing his or her score to align with socially constructed scoring norms.

Our final research question aimed to explore this relationship between SCT and the patterns of score change we observed within panels. We found that both the raw O-SCT count and the time spent on O-SCT were highly correlated with the number or proportion of times reviewers changed their score, but the correlations were near zero for S-SCT. Thus, when one reviewer challenges another's score, this is associated with the target of that challenge changing their score; however, merely accounting for or justifying one's own scoring behaviors does not have the same effect. Our data suggest being held accountable by another person may exert pressure toward conformity, even if it is only one other individual making this accountability salient (Quinn and Schlenker 2002).

Our results show that SCT is related to reviewers' attempts to achieve consensus with others—but that the point of convergence differs across panels for any given application. Reviewers grapple with the ambiguity of the review criteria in part through SCT, which facilitates agreement among the participating reviewers, while also undermining the external validity of the scores that a given panel assigns to an application. Thus, although telling another reviewer, 'Your comments are meaner than your score' may help the reviewers within a particular panel achieve agreement about the relative merits of their pool of applications, such SCT in fact reinforces collective subjectivity in the peer review process.

Conclusion

In a 2012 survey of NIH grant applications, NIH found that applicants overall expressed 'concern about the reviewers' tendency to assign scores unevenly across the range of available scores' and noted that 'reviewers expressed a desire for more scoring guidance, coaching and direct scoring instruction' (NIH 2013: 1). The current study makes a unique contribution to the literature establishing the subjectivity present in peer review processes by identifying SCT as a key mechanism by which reviewers address, grapple with, and attempt to overcome the ambiguity inherent in the rating scale. By showing that instances of and time spent on SCT are strongly related to reviewers changing their scores to reach consensus—and that there is a moderate relationship between such score convergence within a panel and score divergence between different panels—this study takes an important first step in determining possible places for interventions that could attempt to enhance the reliability of peer review. For example, one might examine the efficacy of a jigsaw technique (Aronson et al. 1978) in which groups of reviewers evaluate a set of grant applications, and then mix in with other groups of reviewers for a different set, to allow for 'osmosis' of SCT and of score norming. Another potential intervention could include a set of 'standard' applications that all NIH study sections within a given specialty evaluate, to serve as a benchmark for that panel's scoring practices against which other applications could be measured. These and other interventions could leverage our initial findings on SCT to evaluate practical ways to mitigate the variability introduced during peer review meetings.

The constructed nature of our study sections limits the ecological validity of our study, since the reviewers knew that they were participating in a research study and that their decisions regarding applications would not have funding implications. However, our reviewers rated their experiences as highly similar or identical to real NIH study sections they had served on, and it is clear from the videos of the meetings that the reviewers who participated took the task seriously—a view shared by the SRO who served on each of the four CSSs (J. Sipe, Personal Communication). It is reasonable to believe that the interactions we observed in our meetings were comparable to those that occur in real ones. Second, due to the voluntary nature of our sampling design, our sample of applications consisted of initially funded applications (64% of our sample) or of applications that were funded after resubmission (36% of our sample), leading to sampling bias slightly skewed toward higher-scoring applications. Third, the relatively small number of panels, applications, and reviewers is a limitation of our study design, and it precludes us from making strong inferences or generalizations from our observations. Instead, we intend for the current exploratory study to be descriptive enough to guide future larger-scale experimental studies. Such experimental studies would enable us to draw causal conclusions about the role of SCT in promoting score convergence, as well as to leverage linear mixed-effects models to systematically examine whether increased agreement within panels causes increased disagreement across panels. We are also interested in examining how the race/ethnicity and gender of applicants and of reviewers affect the peer review process during these meetings, to explore potential factors influencing the race- and gender-based biases that others have found to occur in NIH peer review (Ley and Hamilton 2008; Ginther et al. 2011, 2012; Pohlhaus et al. 2011; Kaatz et al. 2015).

Overall, this mixed-methods study builds upon a strong foundation of prior work establishing low levels of inter-rater reliability in grant peer review, while leveraging our access to videotaped discussions of CSS meetings to explore the interactions that contribute to poor reliability. Being able to closely examine and analyze such interactions is a crucial step toward improving the peer review process—at NIH and beyond. We do not intend to suggest with this line of work that peer review is a flawed or broken process beyond repair. Indeed, we believe peer review is a crucial pillar of the enterprise of science. Therefore, as we continue to shine a brighter light into the 'black box' of peer review, we must be cautious not to thrust it into the shadows.

Notes

1. Indeed, Chubin and Hackett (1990) describe the drawn-out litigation process involved in gaining access to a list of names of NIH applicants in 1980–1, and they discuss in detail 'how difficult it is to examine the peer review system from the outside' (p. 53).
2. Note that these analyses only included those applications that were discussed in a meeting (i.e. the top scoring 50%) because those are the only applications that receive final scores.
3. When including only discussed applications, agreement within each panel is even lower overall (CSS1 α –0.179, CSS2 α –0.304, CSS3 α –0.246, CSS4 α –0.167, Average α –0.224). Agreement for final reviewer and final panel scores are unchanged.
4. When including only discussed applications, agreement between panels is even lower overall (α –0.012). Agreement for final reviewer and final panel scores are unchanged.

Supplementary data

Supplementary data is available at *Research Evaluation* Journal online.

Acknowledgements

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health [R01GM111002] and in part by the Arvil S. Barr Fellowship through the School of Education at the University of Wisconsin-Madison to E.L.P. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest statement. None declared.

References

Aronson, E. et al. (1978) *The Jigsaw Classroom*. Beverly Hills, CA: Sage Publishing Company.

Asch, S. E. (1951) 'Effects of Group Pressure on the Modification and Distortion of Judgments'. In: Guetzkow H (ed.) *Groups, Leadership and Men*, pp. 177–90. Pittsburgh, PA: Carnegie Press.

Asch, S. E. (1955) 'Opinions and Social Pressure', *Scientific American*, 193: 33–5.

Baker, M. et al. (1999) 'The Role of Grounding in Collaborative Learning Tasks.' In: Dillenbourg P (ed.) *Collaborative Learning: Cognitive and Computational Approaches*, pp. 31–63. Oxford: Pergamon.

Barron, B. (2000) 'Achieving Coordination in Collaborative Problem-Solving Groups', *The Journal of the Learning Sciences*, 9/4: 403–36.

Bernardin, H. J., Hennessey, H. W., Jr, and Peyrefitte, J. S. (1995) 'Age, Racial, and Gender Bias as a Function of Criterion Specificity: A Test of Expert Testimony', *Human Resource Management Review*, 5: 63–77.

Bornstein, B. H. (1999) 'The Ecological Validity of Jury Simulations: Is the Jury Still Out?', *Law and Human Behavior*, 23/1: 75–91.

Brown, A. L. et al. (1993) 'Distributed Expertise in the Classroom'. In: Salomon G (ed.) *Distributed Cognitions: Psychological and Educational Considerations*, pp. 188–228. Cambridge: Cambridge University Press.

Chubin, D. E., and Hackett, E. J. (1990) *Peerless Science: Peer Review and U. S. Science Policy*. Albany, NY: State University of New York Press.

Cialdini, R. B., and Goldstein, N. J. (2004) 'Social Influence: Compliance and Conformity', *Annual Review of Psychology*, 55: 591–621.

Cicchetti, D. V. (1991) 'The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation', *Behavioral and Brain Sciences*, 14: 119–35.

Clark, H. H., and Brennan, S. A. (1991) 'Grounding in Communication'. In: Levine JM, and Levine LB, Teasley SD (eds.) *Perspectives on Socially Shared Cognition*, pp. 127–49. Washington, DC: American Psychological Association.

Cole, J., and Cole, S. (1981) *Peer Review in the National Science Foundation: Phase II of a Study*. Washington, DC: National Academy of Sciences.

Cole, S., Cole, J. R., and Simon, G. A. (1981) 'Chance and Consensus in Peer Review', *Science*, 214/4523: 881–6.

Cole, S., Rubin, L., and Cole, J. R. (1978) *Peer Review in the National Science Foundation: Phase I of a Study*. Washington, DC: National Academy of Sciences.

Creswell, J. W., and Plano Clark, V. L. (2011) *Designing and Conducting Mixed Methods Research* 2nd edn. Thousand Oaks, CA: SAGE Publications, Inc.

Crosbie-Burnett, M., and Eisen, M. (1992) 'Simulated Divorced and Remarried Families: An Experiential Teaching Technique', *Family Relations*, 41/1: 54–8.

Deutsch, M., and Gerard, H. B. (1955) 'A Study of Normative and Informative Social Influences upon Individual Judgment', *Journal of Abnormal Social Psychology*, 51: 629–36.

Dotger, B. H., Dotger, S. C., and Maher, M. J. (2010) 'From Medicine to Teaching: The Evolution of the Simulated Interaction Model', *Innovative Higher Education*, 35: 129–41.

Fogelholm, M. et al. (2012) 'Panel Discussion Does Not Improve Reliability of Peer Review for Medical Research Grant Proposals', *Journal of Clinical Epidemiology*, 65/1: 47–52.

Foss, R. D. (1976) 'Group Decision Processes in the Simulated Trial Jury', *Sociometry*, 39/4: 305–16.

Gallo, S. A., Carpenter, A. S., and Glisson, S. R. (2013) 'Teleconferenec versus Face-to-Face Scientific Peer Review of Grant Application: Effects on Review Outcomes', *PLoS ONE*, 8/8: e71693.

Gillespie, G. W., Jr, Chubin, D. E., and Kurzon, G. M. (1985) 'Experience with NIH Peer Review: Researchers' Cynicism and Desire for Change', *Science, Technology, & Human Values*, 10/3: 44–54.

Ginther, D. K. et al. (2011) 'Race, Ethnicity, and NIH Research Awards', *Science*, 333/6045: 1015–9.

Ginther, D. K. et al. (2012) 'Are Race, Ethnicity, and Medical School Affiliation Associated with NIH R01 Type Award Probability for Physician Investigators?', *Academic Medicine*, 87/11: 1516–24.

Hackett, E. J. (1987) 'Funding and Academic Research in the Life Sciences: Results of an Exploratory Study', *Science & Technology Studies*, 5/3-4: 134–47.

Hayes, A. F., and Krippendorff, K. (2007) 'Answering the Call for a Standard Reliability Measure for Coding Data', *Communication Methods and Measures*, 1/1: 77–89.

Heilman, M. E., and Haynes, M. C. (2008) 'Subjectivity in the Appraisal Process: A Facilitator of Gender Bias in Work Settings'. In: Borgida E and Fiske ST (eds.) *Beyond Common Sense: Psychological Science in Court*. Oxford: Blackwell Publishing, Ltd.

Helitzer, D. L. et al. (2011) 'A Randomized Controlled Trial of Communication Training with Primary Care Providers to Improve Patient-Centeredness and Health Risk Communication', *Patient Education and Counseling*, 82: 21–9.

Heritage, J. C., and Watson, D. R. (1979) 'Formulations as Conversational Objects'. In: Psathas G (ed.) *Everyday Language: Studies in Ethnomethodology*, pp. 123–62. New York: Irvington.

Kaatz, A. et al. (2015) 'A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques from Investigators at One Institution', *Academic Medicine*, 90/1: 69–75.

Kane, M. N. (2003) 'Teaching Direct Practice Techniques for Work with Elders with Alzheimer's Disease: A Simulated Group Experience', *Educational Gerontology*, 29: 777–94.

Kerr, N. L., Nerenz, D. R., and Herrick, D. (1979) 'Role Playing and the Study of Jury Behavior', *Sociological Methods & Research*, 7/3: 337–5.

Klahr, D. (1985) 'Insiders, Outsiders, and Efficiency in a National Science Foundation Panel', *American Psychologist*, 40/2: 148–54.

Koshik, I. (2002) 'A Conversation-Analytic Study of Yes/No questions Which Convey Reversed Polarity Assertions', *Journal of Pragmatics*, 34: 1851–77.

Krippendorff, K. (2004) *Content Analysis: An Introduction to its Methodology*, 2nd edn. Thousand Oaks, CA: Sage Publishing.

Krippendorff, K. (2008) 'Systematic and Random Disagreement and the Reliability of Nominal Data', *Communication Methods and Measure*, 2/4: 323–38.

Krippendorff, K. (2013) *Content Analysis: An Introduction to its Methodology*, 3rd edn. Thousand Oaks, CA: Sage Publishing.

Kruijver, I. P. M. et al. (2001) 'Communication Between Nurses and Simulated Patients with Cancer: Evaluation of a Communication Training Programme', *European Journal of Oncology Nursing*, 5/3: 140–50.

Lamont, M. (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press.

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: 820–41.

Lerner, J. S., and Tetlock, P. E. (1999) 'Accounting for the Effects of Accountability', *Psychological Bulletin*, 125: 255–75.

Levinson, S. (1992) 'Activity Types and Language'. In: Drew P and Heritage J (eds.) *Talk at Work: Interaction in Institutional Settings*, pp. 66–100. Cambridge: Cambridge University Press.

Ley, T. J., and Hamilton, B. H. (2008) 'The Gender Gap in NIH Grant Applications', *Science*, 322/5907: 1472–4.

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008) 'Improving the Peer Review Process for Grant Applications: Reliability, Validity, Bias and Generalizability', *American Psychologist*, 63/3: 160–8.

McCullough, J. (1989) 'First Comprehensive Survey of NSF Applicants Focuses on Their Concerns About Proposal Review', *Science, Technology, & Human Values*, 14/1: 78–88.

McQuiston-Surrett, D., and Saks, M. J. (2009) 'The Testimony of Forensic Identification Science: What Expert Witnesses Say and What Factfinders Hear', *Law and Human Behavior*, 33/5: 436–53.

Moyer, S. M. (2016) 'Large Group Simulation: Using Combined Teaching Strategies to Connect Classroom and Clinical Learning', *Teaching and Learning in Nursing*, 11: 67–73.

National Institutes of Health (NIH). (2013) *Enhancing Peer Review Survey Results Report*, <https://grants.nih.gov/grants/peer/guidelines_general/scoring_system_and_procedure.pdf> accessed 28 July 2016.

National Institutes of Health (NIH). (2015) *Scoring System and Procedure*, <https://grants.nih.gov/grants/peer/guidelines_general/scoring_system_and_procedure.pdf> accessed 27 April 2016.

National Institutes of Health (NIH). (2016) *NIH Budget*, <http://www.nih.gov/about/budget.htm> accessed 24 March 2016.

Obrecht, M., Tibelius, K., and 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.

Olbrecht, M., and Bornmann, L. (2010) 'Panel Peer Review of Grant Applications: What Do We Know From Research in Social Psychology on Judgment and Decision-Making in Groups?', *Research Evaluation*, 19/4: 293–304.

Paulus, T. M. (2009) 'Online but Off-Topic: Negotiating Common Ground in Small Learning Groups', *Instructional Science*, 37: 227–45.

Pennington, J., and Schlenker, B. R. (1999) 'Accountability for Consequential Decisions: Justifying Ethical Judgments to Audiences', *Personality and Social Psychology Bulletin*, 25: 1067–81.

Pier, E. L. et al. (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>

Pohlhaus, J. R. et al. (2011) 'Sex Differences in Application, Success, and Funding Rates for NIH Extramural Programs', *Academic Medicine*, 86/6: 759–67.

Porter, A. L., and Rossini, F. A. (1985) 'Peer Review of Interdisciplinary Research Proposals', *Science, Technology, & Human Values*, 10/3: 33–8.

Quinn, A., and Schlenker, B. R. (2002) 'Can Accountability Produce Independence? Goals as Determinants of the Impact of Accountability on Conformity', *Personality and Social Psychology Bulletin*, 28: 472–83.

Raclaw, J., and 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. et al. (1993) 'Reasoning in Conversation', *Cognition and Instruction*, 11: 347–64.

Romano, J. L. (1998) 'Simulated Group Counseling: An Experiential Training Model for Group Work', *The Journal for Specialists in Group Work*, 23/2: 119–32.

Romano, J. L., and Sullivan, B. A. (2000) 'Simulated Group Counseling for Group Work Training: A Four-Year Research Study of Group Development', *The Journal for Specialists in Group Work*, 25/4: 366–75.

Schegloff, E. A. (2007) *Sequence Organization in Interaction: A Primer in Conversation Analysis*, Vol 1. Cambridge: Cambridge University Press.

Sigelman, L., and Scioli, F. P. Jr (1987) 'Retreading Familiar Terrain. Bias, Peer Review, and the NSF Political Science Program', *Political Science & Politics*, 20/1: 62–9.

Stroben, F. et al. (2016) 'A Simulated Night Shift in the Emergency Room Increases Students' Self-Efficacy Independent of Role Taking Over During Simulation', *BMC Medical Education*, 16: 177–83.

Vygotsky, L. (1978) *Mind in Society*. London: Harvard University Press.

Wenneras, C., and Wold, A. (1997) 'Nepotism and Sexism in Peer-Review', *Nature*, 387: 341–3.

Wessely, S. (1998) 'Peer Review of Grant Applications: What Do We Know?', *Lancet*, 352: 301–5.

Young, O. M., and Parviainen, K. (2014) 'Training Obstetrics and Gynecology Residents to be Effective Communicators in the Era of the 80-Hour Workweek: A Pilot Study', *BMC Research Notes*, 7: 455–60.