ALMA Memo 616

Analysis of the ALMA Cycle 7 Supplemental Call

John Carpenter (JAO),
Jennifer Donovan Meyer (NRAO),
Andrea Corvillón (JAO),
Violette Impellizzeri (JAO),
Robert Kurowski (ESO),
and
Alex Chalevin (ESO)

November 5, 2020

Executive Summary

The scientific merit of proposals submitted to the Cycle 7 ACA supplemental call was evaluated using distributed peer review (DPR), where each proposal team designates one person to participate in the review process. This memo analyzes the results of the supplemental call to determine if the implementation of DPR is mature operationally to consider for future cycles and to assess the level of community acceptance of the process. The assessment is based upon analysis of the (i) proposal rankings, (ii) help-desk tickets submitted by users, and (iii) surveys of the reviewers and Principal Investigators (PIs) in both the Cycle 7 main and supplemental calls. The main findings are as follows.

- Most reviewers ($\geq$ 78%) rated the documentation and the reviewer tool positively. Help-desk tickets from users typically addressed benign issues, although a technical issue in the reviewer tool that affected $\sim$ 2% of users needs attention.
- PIs rated 75% of the individual reviewer comments as "very" or "somewhat" helpful. Younger PIs typically found the comments more helpful than senior PIs.
- The helpfulness of the reviewer comments, as judged by the PIs, is not correlated with the career status or expertise of the reviewer. Students, postdocs, and senior reviewers had a similar percentage of comments viewed positively by the PIs, while experts and non-experts wrote equally helpful comments.
- PIs give similar marks to the overall clarity, accuracy, and helpfulness of the reviewer comments in the supplemental and main calls.
- The systematics with regional affiliation, PI experience, and gender follow similar trends as in the Cycle 7 main call. While women and PIs from East Asia showed improvement in their rankings relative to the main call, it is not possible to conclude that the nature of the systematics changed materially because of different demographics among the reviewers and the low number of proposals.
- 46% of the reviewers believe DPR is either suitable for regular proposals or equally effective as panel reviews; 27% felt DPR should not be used in the main call, while the remaining reviewers are unsure.
- The majority of PIs indicated DPR would be suitable to review "small" proposals (61%; defined as less than 25 h on the 12-m array) or an ACA supplemental call (71%). Most PIs do not support using DPR to review "medium" (41%; 25-50 h) and especially "large" (13%; $>$ 50 h) proposals, and 20% of PIs indicate DPR should be not used to review any proposals.
- Most PIs feel confidentiality and bias issues are either the same between DPR and review panels or have no strong opinion. However, more PIs expressed concerns about biases and especially confidentiality in DPR than in panel reviews.

Based on the results of the supplemental call, the DPR process is mature from an operational perspective. The main technical risks in implementing DPR for the main call are scaling the tools for $\sim$5$\times$ more concurrent users and running DPR in parallel with panel reviews. In terms of the quality of the reviewer feedback, the individual comments in DPR and the consensus reports from the review panels received similar ratings by the PIs, with no clear reason to prefer one process over the other. Most PIs in the supplemental call are receptive to using DPR to review small and future supplemental call proposals, but do not want to see DPR applied to the largest proposals.

Contents

1 Introduction **6**

2 Overview of the Cycle 7 Supplemental Call **6**

2.1 Timeline 7

2.2 Documentation 8

2.3 Surveys 9

2.4 Help-desk tickets 9

3 Demographics **10**

3.1 Proposal submissions 10

3.2 Designated reviewers 12

3.3 Review assignments 12

3.4 Review submissions 15

3.5 Reviewer ranks 15

3.6 Correlation of ranks with reviewer expertise and career status 17

3.7 Systematics in the rankings 19

4 Analysis of the reviewer survey **28**

4.1 Documentation and reviewer tools 28

4.2 Proposal reviews 28

4.3 Self-assessment of expertise 30

4.4 Viability of DPR in the main call 33

5 Analysis of the PI survey **33**

5.1 General comments on reviewer feedback 33

5.2 PI assessment of individual comments 36

5.3 Suitability of DPR in future calls 44

6 Lessons learned from the Cycle 7 Supplemental Call **51**

6.1 Dispersion in the individual ranks 51

6.2 Relative ranks versus absolute scores 53

6.3 Creating the global ranked list 53

6.4 Improvement in the tools . . . 54
6.5 Evaluating the appropriateness of the comments . . . 55
6.6 Improving the proposal assignment process . . . 55
6.7 Confidentiality and conflicts of interest . . . 56
6.8 Improving the quality of the reviewer feedback . . . 56
6.9 Providing feedback to the reviewers . . . 56
6.10 High-risk / high-reward proposals . . . 57
6.11 Risks in using DPR in the main call . . . 57

Appendix A Reviewer survey results **59**

A.1 How helpful were the guidelines on writing comments to the PI? . . . 59
A.2 How relevant were the review criteria to evaluate the proposals? . . . 60
A.3 How much time did you spend, on average, reviewing each proposal (including writing comments)? . . . 61
A.4 How satisfactorily were you able to evaluate the proposals for which you were a non-expert? . . . 62
A.5 How easy was it to navigate the interface to review the proposals? . . . 63
A.6 Would you submit ALMA proposals in future cycles if you were required to review 10 proposals for every proposal submitted? . . . 64
A.7 How many years has it been since you obtained your PhD? . . . 65
A.8 Have you ever participated in the ALMA review panels as a Science Assessor or Chair for the main proposal call? . . . 66
A.9 Proposals submitted to the Cycle 7 Main Call for Proposals were reviewed by one of 25 review panels, while the Cycle 7 Supplemental Call used distributed peer review, requiring each proposal team to designate a person to review 10 proposals. ALMA is considering using distributed peer review in an upcoming Main Call for regular proposals, while continuing to have a face-to-face review panel for Large Programs. Do you think distributed peer review would be appropriate for a Main Call? . . . 67
A.10 How extensively did you consult with the mentor on the science evaluation? [If you do not have a PhD.] . . . 68
A.11 How extensively did you consult with the mentor on writing the comments to the PI? [If you do not have a PhD.] . . . 69
A.12 Please rate your level of expertise on each of your review assignments, using the text of your reviews for reference. . . . 70

Appendix B PI survey results **71**

B.1 Are the individual comments on your proposal clear and understandable? . . 71

B.2 Are the comments scientifically accurate? 72

B.3 Will the comments help you to improve future ALMA proposals? 73

B.4 Were the comments written in a respectful and professional manner? 74

B.5 If you have ever submitted a proposal to an ALMA Main Call, how do you rate the general quality of the comments you have received in the Supplemental Call versus consensus reports you have received in a Main Call? 75

B.6 For which types of proposals do you think Distributed Peer Review would be beneficial? (check all that apply) 76

B.7 Are you concerned about confidentiality in ALMA review processes? 77

B.8 Are you concerned about the robustness of ALMA review processes against any biases? 78

B.9 Would you submit ALMA proposals in future cycles if you were required to review 10 proposals for every proposal submitted? 79

B.10 How many years has it been since you obtained your PhD? 80

B.11 Please take a few minutes to rate the helpfulness of each review that you received, indicating the extent to which this comment will help to improve your proposal in the future. Positive comments like "best proposal I ever read" can be ranked as not helpful as it does not improve the proposal further. 81

Appendix C Reviewer comments marked by the PI as inappropriate or unprofessional **82**

1 Introduction

ALMA issued a Supplemental Call for Proposals in Cycle 7 to allocate additional time on the Atacama Compact Array (ACA). The main goal of the supplemental call is to maximize the scientific output of the ACA by making sure the observing queue is full and to allow more timely science to be proposed since the supplemental call follows the main call by five months.

This is the second time ALMA has issued an ACA supplemental call. The first call was issued in Cycle 4, when the proposals were reviewed by ALMA staff and Chilean representatives. The number of proposals overwhelmed the limited number of reviewers, and further supplemental calls were put on hold until Cycle 7 when a more operationally feasible review process, distributed peer review (DPR), could be implemented. In this review process, each proposal team selects one member from the investigator list to review a subset of proposals. The review process thereby naturally scales with the number of the proposals.

In the April 2019 ALMA Board meeting, the Board stated their intention to implement DPR for the main call starting in Cycle 9 pending a successful pilot run in the Cycle 7 supplemental call. A key aspect in measuring the success of the review process is to determine if the community is satisfied with the process and the results. Accordingly, ALMA surveyed both the reviewers and the Principal Investigators (PIs) in the main and supplemental calls to gather community input on the review process.

This memo analyzes the results of the supplemental call and the various surveys to gauge community input on the viability of distributed peer review for ALMA. Section 2 describes the information that was provided to the community regarding the proposal call and the review process. Section 3 presents the demographics of the supplemental call, including the proposal submissions, the designated reviewers, the review assignments, and systematics in the proposal rankings. The results of the reviewer and PI surveys are discussed in Sections 4 and 5, respectively. Section 6 summarizes the lessons learned and potential changes in the review process that should be considered if DPR is used in future cycles. The appendices present the detailed results of the reviewer and PI surveys from the supplemental call.

2 Overview of the Cycle 7 Supplemental Call

Recognizing that the Cycle 7 supplemental call was a pilot for implementing DPR in future cycles, extensive preparation went into issuing the call so that the feedback from reviewers and PIs would not be colored by improper mechanics. The resulting documentation, reviewer tools, and ARC support equaled, and arguably surpassed in some respects, that offered in the main call. This section summarizes the review process and the documentation that was provided to the community, and analyzes the help-desk tickets submitted to the ARCs to understand the typical issues faced by the community.

Table 1: Cycle 7 Supplemental Call Timeline

Date	Milestone
19 December 2018	Cycle 7 pre-announcements (Main Call and Supplemental Calls)
3 September 2019	Supplemental Call for Proposals released
1 October 2019	Deadline to submit Supplemental Call proposals
15 October 2019	Proposals released to reviewers
12 November 2019	Deadline to submit reviews and ranks
3 December 2019	Notification emails sent to PIs

2.1 Timeline

Table 1 summarizes the timeline of the supplemental call. ALMA first notified the community of the supplemental call in the December 2018 Cycle 7 pre-announcement. Proposals were formally solicited in September 2019 and the review process took place in October/November 2019. The JAO notified the community of the results in December 2019. The basic rules of the process, provided to PIs in the supplemental call documentation, are as follows.

- All participants in the review process are expected to behave in an ethical manner. If it is found that a reviewer has not behaved in an ethical manner, the proposal(s) associated with the reviewer may be rejected.
- Each proposal must designate one reviewer to participate in the review process. The designated reviewer may be the PI of the proposal or one of the co-Investigators (co-Is).
- The reviewer must be specified in the Observing Tool (OT) at the time of proposal submission and cannot be changed after the proposal deadline.
- Reviewers must declare any major conflicts of interest with their assigned proposals. Any assignment with a major conflict of interest will be replaced by another proposal.
- Each designated reviewer is responsible for writing comments and scientific ranks for ten proposals. If a person is the designated reviewer on multiple proposals, they will receive ten unique review assignments per submitted proposal.
- If a designated reviewer does not submit their reviews and ranks by the review deadline (12 November 2019 15:00 UT), the proposal for which they were identified as the reviewer will be rejected.
- All participants in the review process agree to keep the materials confidential and will not use the materials for any other means other than the proposal review. Participants will delete any proposals after they have completed their assessments.
- PIs who do not have a PhD may be selected as the designated reviewer. In such cases, a mentor must be specified who will assist the PI in the review process. The mentor must have a PhD and be specified in the OT at the time of proposal submission.

Each reviewer receives 10 proposals, and therefore each submitted proposal will have 10 reviews. Reviewers were asked to assess the scientific merit proposals to the best of their ability using the following criteria:

- The overall scientific merit of the proposed investigation and its potential contribution to the advancement of scientific knowledge.
  - Does the proposal clearly indicate which important, outstanding questions will be addressed?
  - ll the proposed observations have a high scientific impact on this particular field and address the specific science goals of the proposal? ALMA encourages Reviewers to give full consideration to well-designed high-risk/high-impact proposals even if there is no guarantee of a positive outcome or definite detection.
  - Does the proposal present a clear and appropriate data analysis plan?
- The suitability of the observations to achieve the scientific goals.
  - Is the choice of target (or targets) clearly described and well justified?
  - Are the requested signal-to-noise ratio, angular resolution, spectral setup, and $u - v$ coverage provided by the ACA sufficient to achieve the science goals?

A reviewer assigns rank=1 to the top rated proposal, rank=2 to the second best proposal, and continuing to rank=10 to the lowest rated proposal. Each rank must be unique and only integer values are allowed. The reviewers also provide written comments that summarize the strengths and weaknesses of the proposals, which are then sent to the PIs verbatim[1].

2.2 Documentation

Full documentation on the supplemental call is available on the ALMA Science Portal[2]. The documentation includes explicit review criteria, criteria to declare conflicts of interest, guidelines for writing comments, a discussion of unconscious bias, a page for frequently asked questions (FAQ), instructions for mentors, and instructions on how to use the reviewer tool.

The documentation was released in stages. A description of DPR and a FAQ page were posted on the ALMA Science Portal in December 2018 with the release of the Cycle 7 pre-announcement in order to introduce the community to the basic process. The review criteria, conflict criteria, discussion of unconscious bias, guidelines for writing comments, guidelines for mentors, a description of the reviewer tools and an updated FAQ page were released with the Call for Proposals in September 2019. This informed the community of the expectations of the review process before PIs designated the reviewers.

[1] In practice, the JAO reviewed the comments before sending them to PIs. The JAO made editorial changes to the comments for one reviewer since it was felt the comments were unduly harsh. The Lead of the Proposal Handling Team had a conversation with the reviewer to explain why the comments were modified.
[2] https://almascience.org/proposing/7m-array-supplemental-call

2.3 Surveys

Surveys were conducted for the reviewers and the PIs for both the main call and the supplemental call. The questions and responses for the supplemental call surveys are presented in Appendix A (reviewer survey) and Appendix B (PI survey). Table 2 lists the response rate for the various surveys. The response rate for both the main and supplemental call reviewer surveys and the supplemental call PI survey were excellent, with 70-95% of the people responding. These surveys should accurately reflect the consensus opinions of these groups. While a large number of PIs (406) responded to the Cycle 7 main call survey, it was a minority (23%) of the overall number of PIs.

Table 2: Response rate for the Cycle 7 Surveys

Survey	Number people	Number responses	Response rate
Main call: reviewer	158	137	87%
Main call: PI	1773	406	23%
Supplemental call: reviewer	225	213	95%
Supplemental call: PI	226	158	70%

2.4 Help-desk tickets

The Joint ALMA Observatory (JAO) Proposal Handling Team (PHT) tracked the requests for assistance submitted to the ALMA help-desk from proposers and reviewers starting with the release of the call for proposals and through the delivery of the PI notification letters. Users submitted 42 tickets in total. Eighteen tickets addressed user errors (e.g., withdrawing duplicate proposals, changing the designated reviewer, contacting users about late submission of conflicts of interest). The remaining 24 tickets clarified questions about the proposal submission or review processes, or reported errors in the tools. The tickets do not include an issue in the PI notification letters when the JAO proactively notified PIs that the reviewer feedback was not visible in SnooPI; PIs were notified within two days that the issue had been fixed.

Table 3 lists the topics in which more than one help-desk tickets was filed. The largest category of tickets was withdrawing duplicate proposals with 6 tickets. The most significant issue is that five people reported errors in submitting conflicts of interests or proposals ranks, which mostly likely are related to the autosave feature in the reviewer tool. Given that this was the first time DPR was used at ALMA and the tools and documentation had to be created anew, the low number of tickets and their mostly benign nature shows that logistically the process from proposal submission through proposal review went smoothly. This is reflected in the reviewer surveys as well (see Section 4.1).

Table 3: Most common helpdesk tickets in Cycle 7 Supplemental Call

Number of tickets	Category of tickets
6	Withdraw duplicate proposal
5	Error when submitting conflicts of interest or ranks by reviewers
4	New conflicts identified after conflict submission
2	Problems to submit the reviewer survey
2	Questions about instructions: emails were sent to reviewer's spam folder
2	Questions about review: how to handle a proposal requesting only total power
2	Email apologizing for late submission of conflicts
2	Users contacted to ask them to do the final submission of their reviews/ranks
2	Reviewers requested to include small edits in one of their review reports

3 Demographics

This section analyzes the demographics of the supplemental call proposals, including the submission statistics, the scientific rankings of the reviewers, and systematics in the rankings.

3.1 Proposal submissions

Tables 4 and 5 summarize the number of proposals submitted and accepted by region and scientific category, respectively. ALMA received 249 proposals requesting 8199 h on the 7-m array. By comparison, in the Cycle 7 main call[3], PIs submitted 80 ACA standalone proposals that requested 3541 h on the 7-m array . Thus the number of submitted proposals tripled and the requested time more than doubled compared to the main call. Categories 2 and 3 had the largest number of submitted proposals and requested time, while categories 4 and 5 had the fewest number of proposals. These trends are broadly consistent with the main call. Relative to the main call, Europe and East Asia have a larger share of submitted proposals while North America had a lower share.

The supplemental call proposals could either be new proposals or declined proposals re-submitted from the main call. Proposals were identified as a resubmission if the PI proposed to observe the same sources in a declined Cycle 7 main call proposal, or if the supplemental proposal had the same title as in the main call but with a different PI. Figure 1 shows the origin of the supplemental call proposals. The majority (63%) are new proposals that were not submitted to the Cycle 7 main call. Thus the supplemental call successfully sparked the community to submit new ideas on how to use the ACA. The remaining originated from a Cycle 7 main call proposal, either as a declined 12-m+ACA proposal, a 12-m proposal recast to use only the ACA, or an ACA proposal. In many cases the details of the proposal changed in these resubmissions (e.g., time requested, requested spectral lines, number of sources) but the proposal was considered resubmitted nonetheless.

[3] https://almascience.eso.org/news/alma-cycle-7-proposal-review-detailed-report

Table 4: Cycle 7 supplemental call proposal statistics by region

	Chile	East Asia	Europe	North America	Other	Total
Submitted proposals						
Number of proposals	9	75	101	53	11	249
7-m Array time (hours)	259	2547	3085	1986	322	8199
Total Power Array time (hours)	81	1550	974	973	117	3695
Accepted proposals						
Number of proposals	9	22	38	26	4	99
7-m Array time (hours)	259	652	1091	995	73	3069
Total Power Array time (hours)	81	307	366	452	88	1294

Table 5: Cycle 7 supplemental call proposal statistics by category

	Category 1	Category 2	Category 3	Category 4	Category 5	Total
Submitted proposals						
Number of proposals	41	79	90	18	21	249
7-m Array time (hours)	1818	2872	2382	594	532	8199
Total Power Array time (hours)	0	1118	2531	6	40	3695
Accepted proposals						
Number of proposals	17	33	33	6	10	99
7-m Array time (hours)	713	1156	870	33	297	3069
Total Power Array time (hours)	0	540	741	6	7	1294

Category 1: Cosmology and the high redshift universe
Category 2: Galaxies and galactic nuclei
Category 3: Interstellar medium, star formation, and astrochemistry
Category 4: Circumstellar disks and the solar system
Category 5: Stellar evolution and the Sun

Table 6: Cycle 7 supplemental call reviewer statistics

	Chile	East Asia	Europe	North America	Other	Total
Number of proposals	9	75	101	53	11	249
Number of reviewers	9	74	101	53	12	249
Number of non-PhD reviewers	0	10	9	6	0	25
Number of delegated reviewers	2	13	18	5	3	41

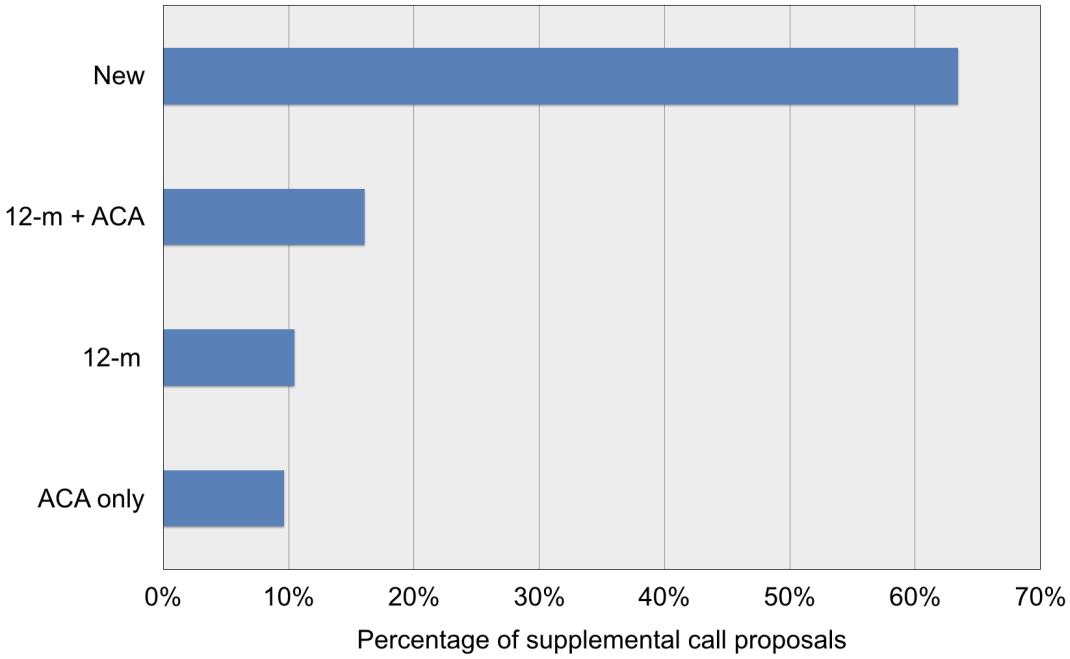

Figure 1: The source of proposals submitted to the Cycle 7 supplemental call. The majority of submissions (63%) were new proposals not submitted in the Cycle 7 main call. The remaining proposals were declined in the main call and resubmitted to the supplemental call, usually after making some changes to the source list, spectral setup, or requested sensitivity. These include 12m+ACA where the ACA portion was resubmitted, a 12-m array proposal recast as an ACA proposal, or an ACA standalone proposal.

3.2 Designated reviewers

Table 6 presents the data on the designated proposal reviewers. Of the 249 proposals, there were 225 unique reviewers: 202 people reviewed one set of 10 proposals, 22 reviewed two sets, and one person reviewed three sets. The PI served as the designated reviewer in 84% of the proposals, which was roughly the same in each of the regions. Figure 2 shows the regional distribution of PIs and reviewers. The regional distributions are virtually identical, indicating that the designated reviewers tend to have the same regional affiliation as the PI. PIs that do not have a PhD (typically students, but could also be staff) were the designated reviewer for 25 proposals (10% of all proposals). The fraction of non-PhD reviewers was similar across regions.

3.3 Review assignments

The PHT prioritized assigning proposals to reviewers whose research aligns with the category and keywords of the submitted proposal, with the underlying assumption that a reviewer had general knowledge if not expertise in the subject area of their own proposal. In practice, it was not possible to assign all proposals following this guideline due to conflicts

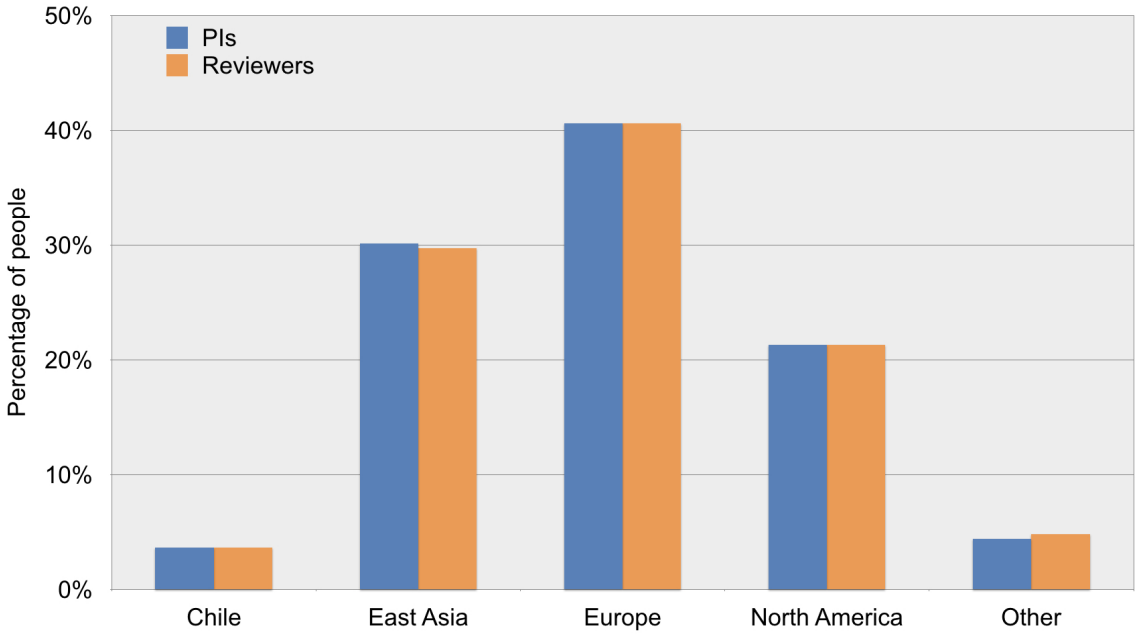

Figure 2: The regional distribution of PIs (blue) and reviewers (orange) in the Cycle 7 supplemental call.

of interest and the limited number of proposals in some categories (especially categories 4 and 5). The PHT developed a prioritized list of rules to assign the proposals based on the scientific category, keywords, and major and minor conflicts of interest among the PI, reviewers, and mentors. The JAO identified major conflicts in the supplemental call as follows:

- The PI, designated reviewer, or mentor of the submitted proposal is a PI or coI on the proposal to be reviewed.
- The PI, designated reviewer, or mentor of the submitted proposal is a coI on another proposal - in the Cycle 7 main call or supplemental call - led by the PI of the proposal to be reviewed.
- The PI, designated reviewer, or mentor of the submitted proposal has led a proposal - in the Cycle 7 main call or supplemental call - where the PI of the proposal to be reviewed has been a coI.
- The PI, designated reviewer, or mentor of the submitted proposal is at the same institution as the PI on the proposal to be reviewed.

A minor conflicts was defined as a PI, designated reviewer, or mentor of the submitted proposal that is at the same institution as any one of the coIs on the proposal to be reviewed. Reviewers identified any conflicts that were not identified with these automated criteria.

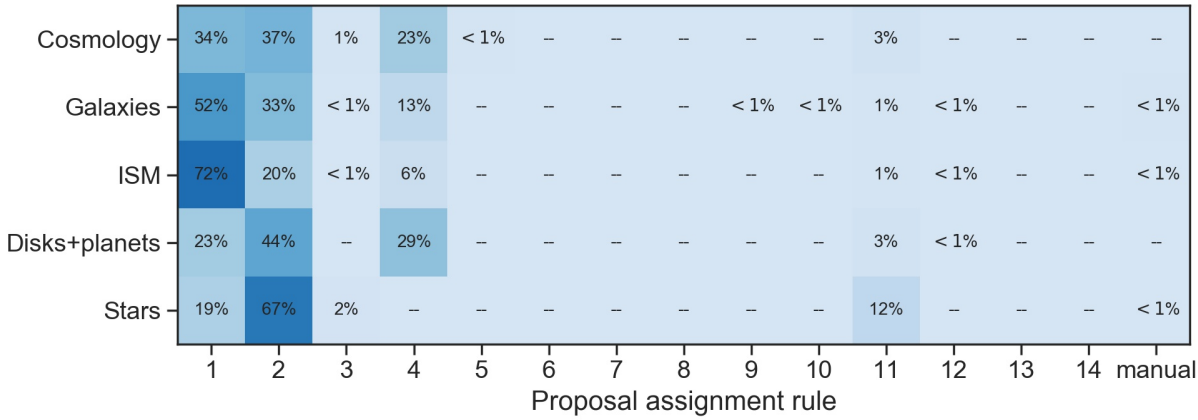

Figure 3: Percentage of proposal assignments by the assignment rules shown in Table 7. The results are normalized separately for each proposal category. A dash indicates no proposals were assigned under that rule.

Table 7 lists the prioritized rules. Proposal assignments were first attempted using Rule 1, which assigns proposals within the same category as the reviewer's submitted proposal that share one or more common keywords and have no major or minor conflicts with the PI, reviewer, or mentor. If not all proposals could be assigned with Rule 1, then Rule 2 was considered, which is similar to Rule 1 except that the assignment was not required to have common keywords. Rule 3 dropped the requirement to avoid minor conflicts with the PI, reviewer, or mentor. In Rules 4 and 5, similar proposal categories were considered, where Category 1 and Category 2 are defined to be similar, and Category 3 is defined as similar to Category 4. No similar categories were designated for Category 5. The last column in Table 7 indicates the percentage of assignments made under that particular rule, including a small number of manual assignments (which mostly followed Rules 1-4).

Figure 3 shows the percentage of proposals assigned by rule for each proposal category. In all categories, the majority of the proposal assignments were in the same category, and for categories 2 and 3, with common keywords. Category 5 is anomalous in that 12% of the proposal assignments in this category came from Rule 11. This is mainly because no proposal category was formally designated as "similar" to implement Rule 4 assignments. Of the 2490 review assignments, 84% of the proposal assignments were in the same category as the submitted proposal and 96.7% were in the same or similar science categories as the submitted proposal. Given that the proposal categories generally contain a broad range of topics, a reviewer assigned a proposal from within the same category as their own submitted proposal may not necessarily be an expert on the proposal. The reviewer self-assessment of their expertise is discussed in Section 4.

Table 7: Proposal assignment rules

Rule	Same category	Similar category[1]	Common keyword(s)	No major conflicts[2]		No minor conflicts[2]		Percentage of assignments[3]
				PI	R/M	PI	R/M	
1	✓		✓	✓	✓	✓	✓	51%
2	✓			✓	✓	✓	✓	33%
3	✓			✓	✓			< 1%
4		✓		✓	✓	✓	✓	12%
5		✓		✓	✓			< 1%
6	✓		✓		✓		✓	–
7	✓				✓		✓	–
8	✓				✓			–
9		✓			✓		✓	< 1%
10		✓			✓			< 1%
11				✓	✓	✓	✓	3%
12				✓	✓			< 1%
13					✓		✓	–
14					✓			–
manual[4]								< 1%

[1] In assigning proposals, categories 1 and 2 are defined as similar, as are categories 3 and 4.
[2] PI: conflicts for Principal Investigator; R/M: conflicts for the reviewer or mentor.
[3] Percentage of all proposals assigned based on that rule.
[4] A few manual assignments were made to override the automatic selections, primarily in dealing with conflicts submitted by reviewers. Most of the manual assignments were equivalent to Rules 1-4.

3.4 Review submissions

The left panel in Figure 4 shows the number of days before the review deadline in which the reviewers completed their assignments. The submission date refers to the time when the last review was submitted. Reviewers could have entered the reviews over an extended period of time and saved their work. Not surprisingly, 55% of the reviewers completed their assignments in the last two days before the review deadline. A potential concern is if reviewers rushed to finish the reviews in order to meet the deadline and compromised the quality of the review. While that is difficult to assess directly, the length of the comments may signify if comments were written in haste. The right panel in Figure 4 shows the median length of the review comments in a proposal set versus the submission time. Some reviewers submitted short reviews in the last couple of days, although this occurred well before the deadline as well. While some reviews were likely written in a rush to meet the deadline (which could be masked by using the median length of the comments), most reviews appeared to have been completed carefully at least as measured by the comment length.

3.5 Reviewer ranks

Since all reviewers completed their review assignments by the deadline, each proposal received 10 individual rankings. The JAO used these ranks to generate a global ranked

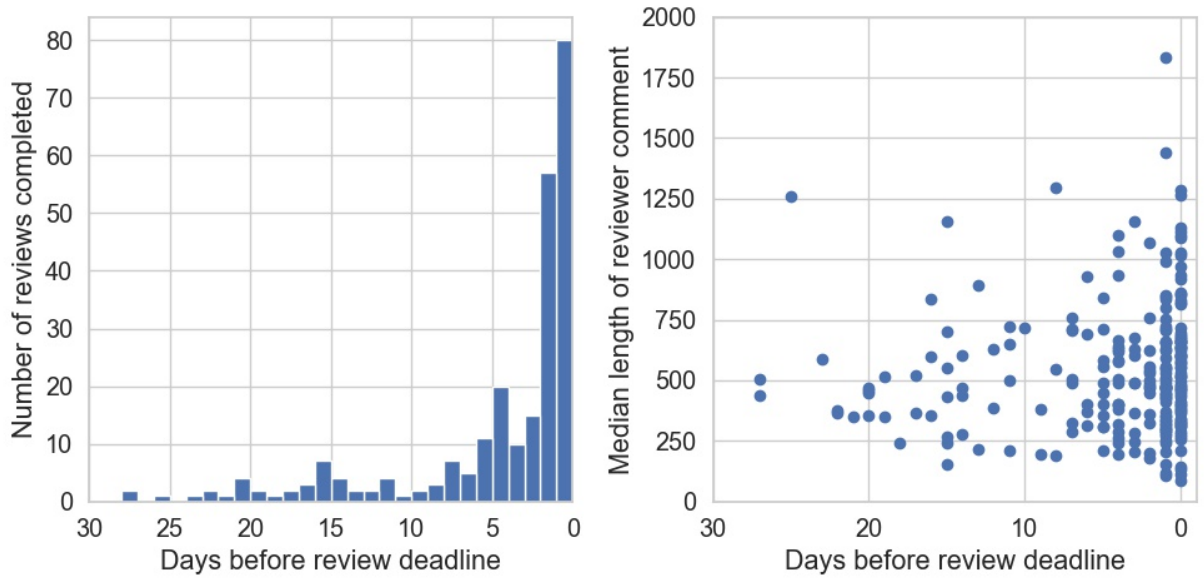

Figure 4: Left: Number of proposal sets completed as a function of the days before the review deadline. Right: Median number of characters in the reviewer comments versus the submission time before the deadline.

list by averaging the individual ranks. Proposals were then added to the observing queue based primarily on the scientific rank, but also the proposal pressure in the requested right ascension, the weather conditions for the observed frequency, the balance of time across executives, and the amount of time available.

Figure 5 shows the mean and standard deviation of the individual proposal ranks (blue) as a function of the global rank. The poorest (red) and best (green) individual rank for each proposal are also shown. The dispersion in the ranks is approximately the same across the proposals. Even the best-rated proposals often had at least one reviewer who gave a poor rank, and the poorest ranked proposals often had one reviewer who gave a very good rank. The dispersion in the individual ranks is attributed to two factors. First, reviewers receive only ten proposals per set and are required to assign relative ranks one through ten regardless of the perceived intrinsic merit. In an extreme case in which a reviewer received the ten most meritorious proposals, one proposal must be assigned a rank of 10 even though it is an excellent proposal. The second factor contributing to the scatter in the individual ranks is the difference in opinions amongst the reviewers on the scientific merit of the proposals.

To assess the dominant cause of the scatter in the individual ranks, Monte Carlo simulations were run for two limiting cases. The simulations assume that there is an intrinsic relative scientific merit to the proposals such that the proposals can be ordered from 1 to N, where N=249 for the supplement call. In the first limiting case, reviewers can perfectly judge the relative scientific merits of the proposals and rank their proposals 1 to 10 in accordance with the intrinsic scientific merit. Therefore, any differences in ranks between reviewers is attributed only to which set of 10 proposals are assigned. In the second case, reviewers have

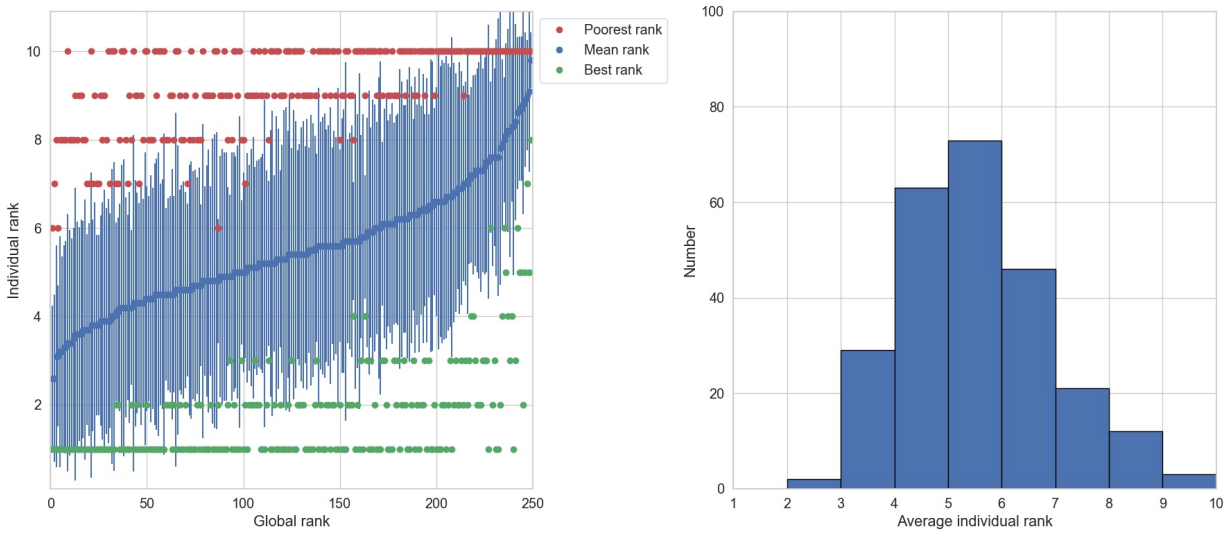

Figure 5: Left: The mean and rms of the individual proposal ranks (blue) as a function of the global rank for the 249 proposals submitted to the Cycle 7 supplemental call. The top-ranked proposal has rank=1. The red circles indicate the poorest individual rank given for each proposal, and the green circle indicates the best individual rank. Right: Histogram of the average of the individual ranks for the 249 proposals.

no correlation in their assessment of the scientific merit. Figure 6 shows one instance of the ideal case where reviewers have the perfect ability to rank the proposals according to the intrinsic scientific merit. The average ranks are uniformly distributed and there is relatively small dispersion in the individual ranks for each proposal. By contrast, a simulation with the uncorrelated scientific ranks (Figure 7) shows a Gaussian-like distribution of average ranks with a larger dispersion of the individual ranks per proposal. A qualitative comparison of the two simulations with the actual data in Figure 5 clearly indicates that the scatter in the individual ranks is dominated by differences in the scientific opinions and not by having access to only 10 proposals. Nonetheless, close comparison of Figures 5 and 7 indicates the actual average ranks have a broader distribution than the randomized case, indicating that there is some correlation in the ranks between reviewers.

3.6 Correlation of ranks with reviewer expertise and career status

Reviewers may be assigned proposals covering a wide range of topics. This could introduce bias in the rankings if reviewers tend to favor or disfavor proposals outside their area of expertise. Figure 8 presents histograms of the individual proposal ranks, separated by the self-declared expertise of the reviewers in the reviewer survey (see Appendix A.12). The expertise categories are “expert”, “some knowledge”, and “little or no knowledge”. Each histogram should have a flat distribution if no bias is present in the individual ranks. The results suggest that reviewers with little or no knowledge on a proposal were somewhat hesitant to assign a top-rating (rank=1) to a proposal and to a lesser extent proposals in the top half of

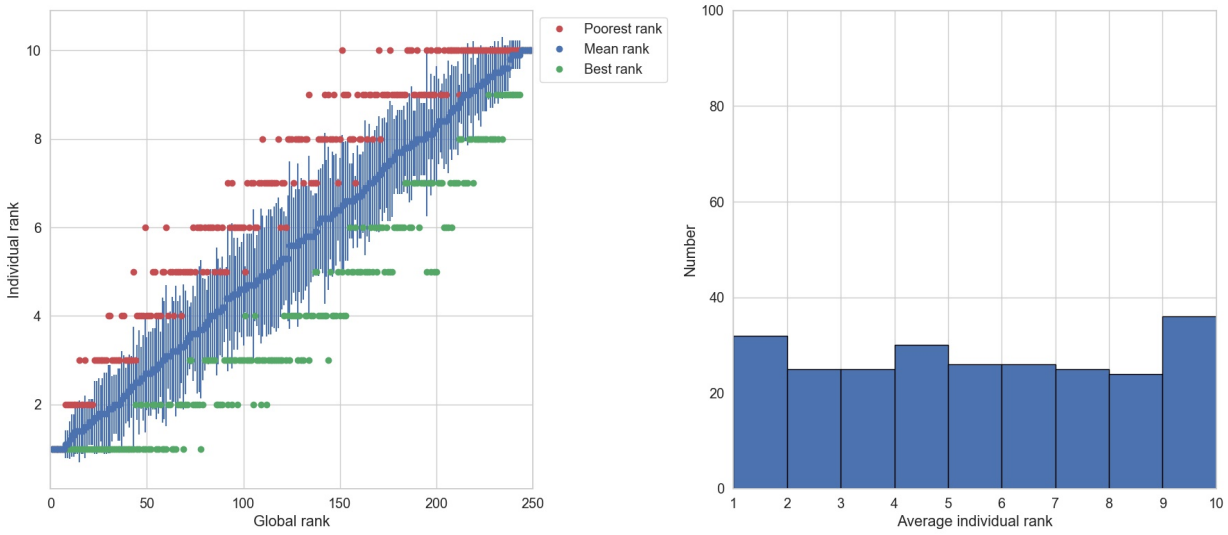

Figure 6: Same as in Figure 5, except for a Monte Carlo simulation in the ideal case where reviewers have the perfect ability to assess the relative scientific merit of the proposals. The dispersion in the ranks then reflects which set of 10 proposals are assigned to a reviewer.

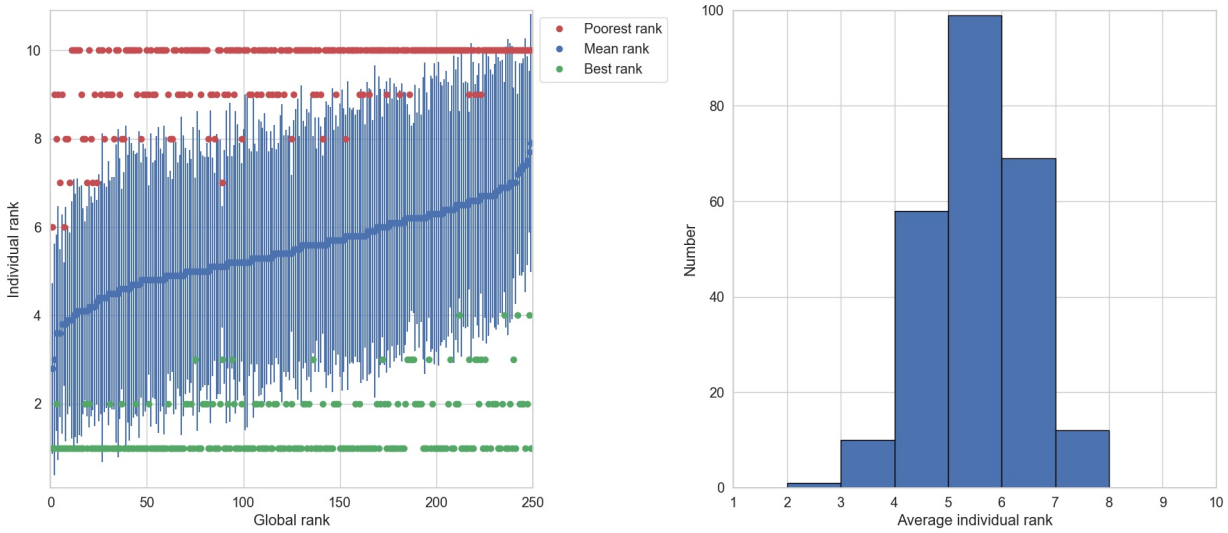

Figure 7: Same as in Figure 5, except for a Monte Carlo simulation for the case where reviewers have completely uncorrelated opinions on the relative scientific merit of the proposals.

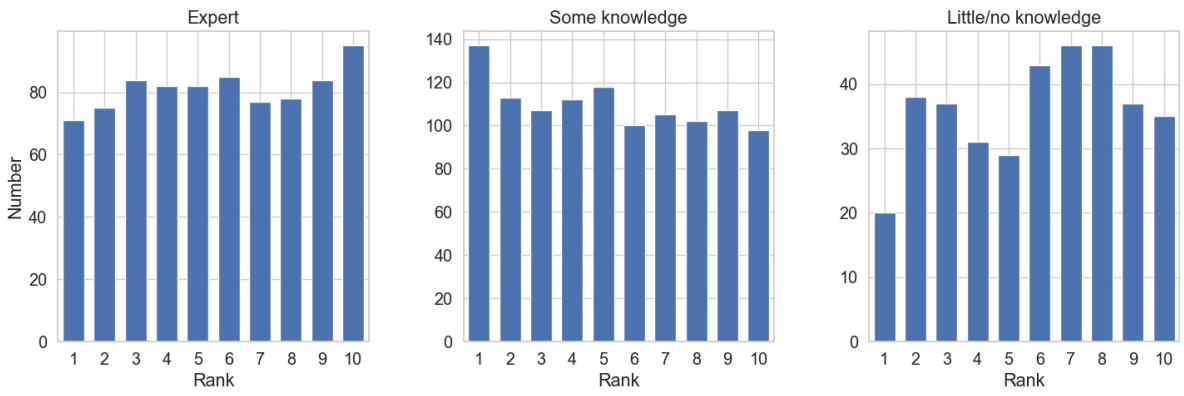

Figure 8: Histogram of the individual ranks assigned to a proposal grouped by the self-declared expertise from the reviewers.

the rankings. Reviewers with some knowledge were more willing to assign the top ranking, and the expert reviewers were more likely to assign the poorest ranking. A chi-squared test indicates the probability that the histograms are consistent with a uniform distribution is p=0.85, 0.31, and 0.06 for expert, some knowledge, and little/no knowledge, respectively. Thus with marginal significance, reviewers did not assign uniform ranks for proposals in which they had little or no knowledge. Given only 16% of review assignments were classified with little or no knowledge, the net impact of the final rankings is marginal. Nonetheless, any future implementations of DPR need to be mindful of this apparent systematic.

Another potential impact of having non-experts as reviewers is that the dispersion in the scores would increase if they cannot reasonably judge the scientific merit of a proposal. This might lead to an increased dispersion in the ranks and a less precise measure of the proposal rank. Figure 9 shows histograms of the dispersion in the individual reviewer ranks grouped by the expertise of the reviewers. These results show that the dispersion in the ranks from different reviewers does not depend strongly on the expertise level. Expert reviewers have an equally wide range of opinions on the scientific merits of the proposals as reviewers with moderate expertise and with little or no expertise.

3.7 Systematics in the rankings

In the main call, a global ranked list of all proposals is created by normalizing the ranked list within a panel by the number of proposals in the panel. This allows the fraction of proposals accepted in a category to be proportional to the fraction of proposals submitted in that category. No explicit normalization is done in the supplemental call when forming the ranks. Given that most of proposals are reviewed against other proposals in the same category and the distributions of ranks are identical for all reviewers, no systematics should appear. To verify that this is the case, Figure 10 shows the cumulative distribution of proposal ranks by scientific category. The probability that the distributions are drawn from the same parent population is 0.85, confirming that there is no significant difference in the proposal rankings by category as anticipated.

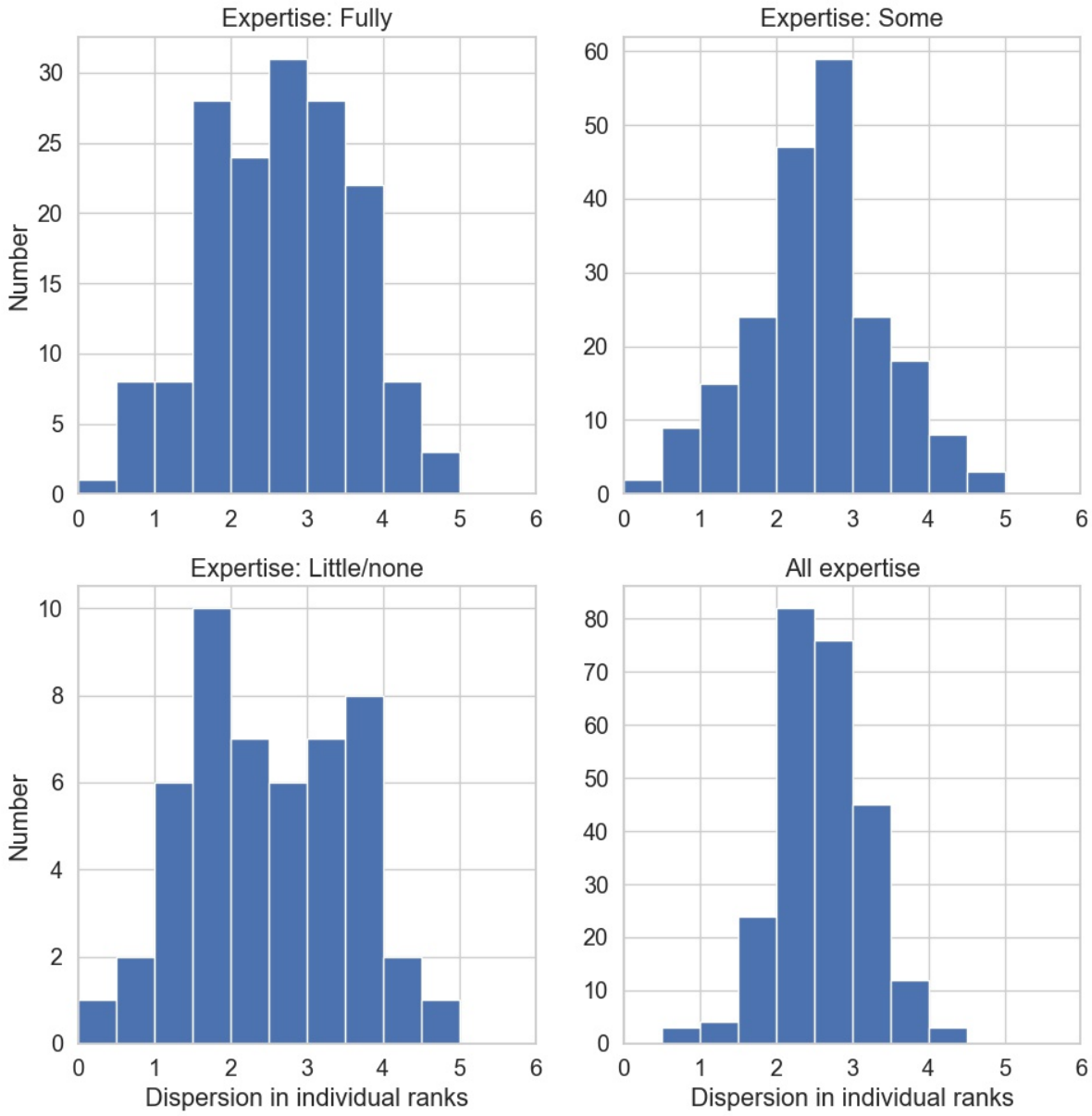

Figure 9: Histogram of the dispersion of individual ranks grouped by the self-declared expertise from the reviewers. Only proposals which have 3 or more reviewers with the relevant expertise are included. The bottom right panel shows the dispersion when including all reviewers.

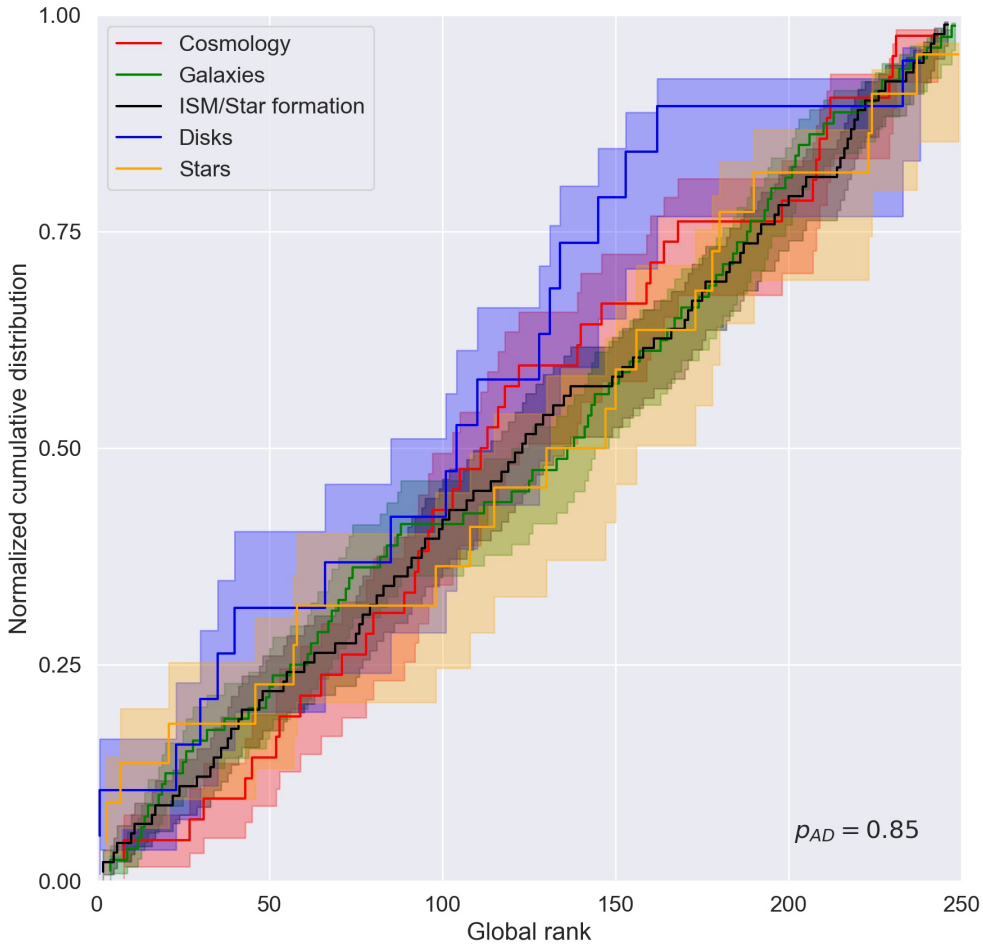

Figure 10: Cumulative distribution of proposal ranks in the supplemental call for the five ALMA categories. The ranks vary between 1 (best) to 249 (poorest). The shaded region indicates the 68.3% confidence interval computed using the beta function. The probability from the Anderson-Darling k-sample test that the distributions are drawn from the same population is indicated in the lower right corner of each panel.

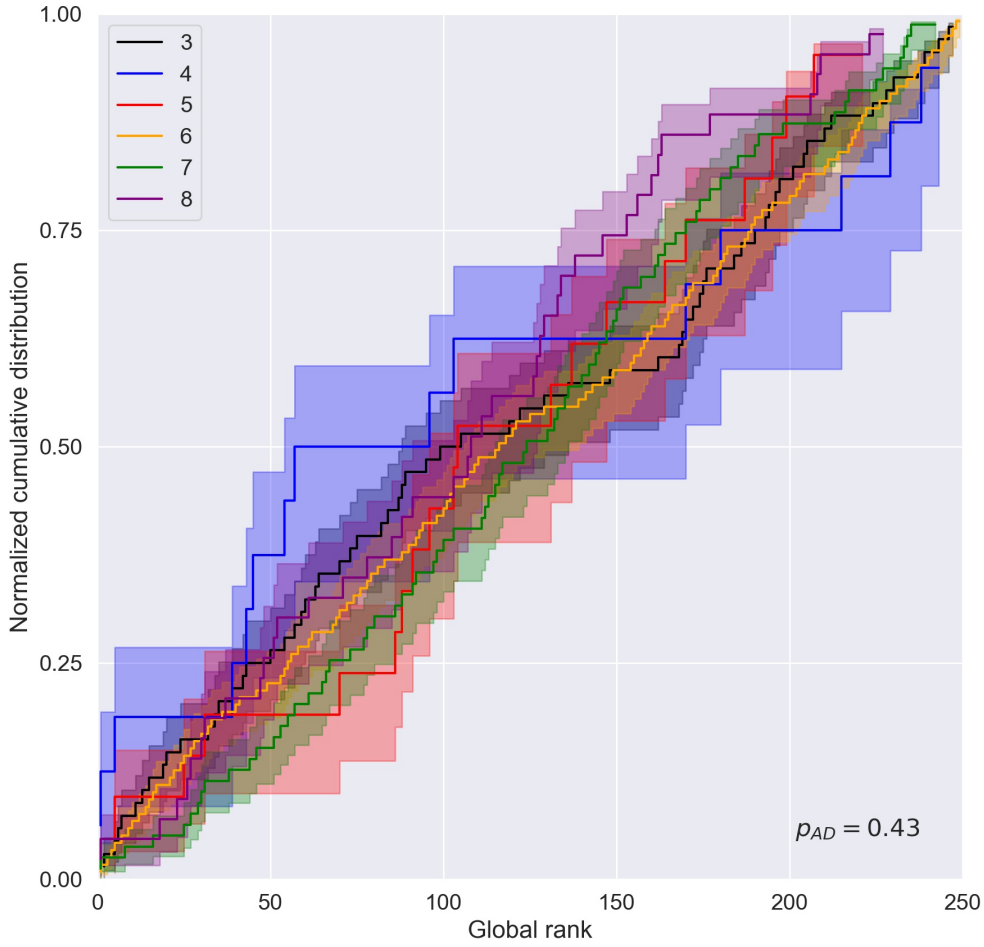

Figure 11: Cumulative distribution of proposal ranks in the supplemental call for the ALMA receiver bands.

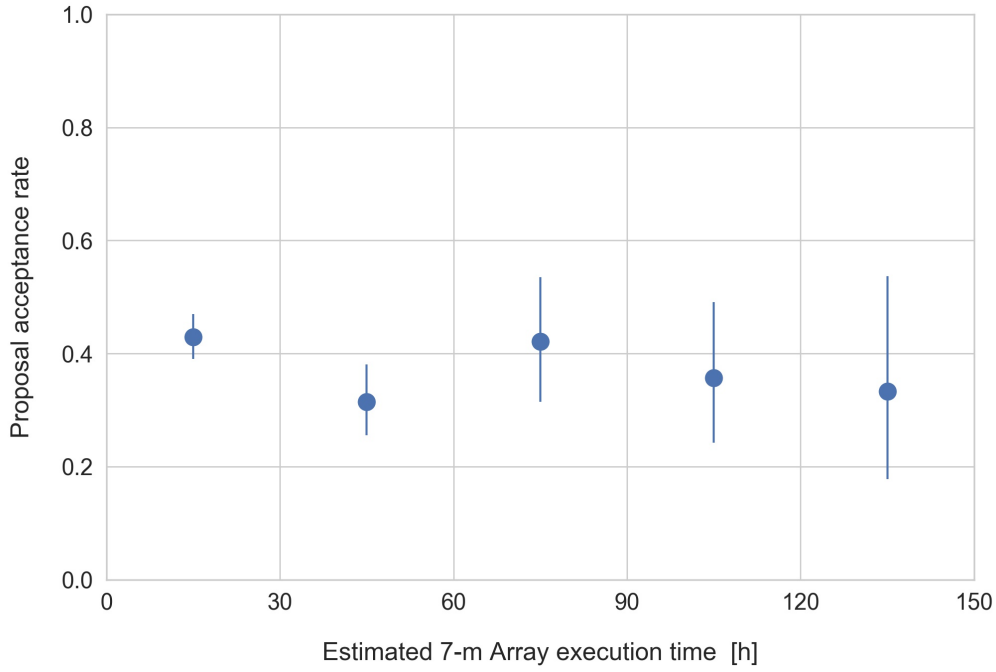

Figure 12: Fraction of the proposals accepted in the supplemental call as a function of the estimated execution time on the 7-m array. The uncertainties indicate the 1σ uncertainties based on binomial statistics.

A concern that has been expressed about DPR is that "risky" proposals would be penalized since review committees are thought to be inherently conservative, and in DPR there is no opportunity for a reviewer to persuade others on the merits of a proposal. Since "risky" proposals are difficult to define and identify objectively, two different characteristics were examined: the relative ranks of high frequency proposals (Band 8 in particular since Band 9 and 10 were not offered in the supplemental call) and the acceptance rate of proposals that request a large amount of observing time. Figure 11 shows the cumulative proposal ranks by receiver band: no significant systematics in the proposal rankings are present with the probability of $p = 0.43$ that the ranks are drawn from a common parent population. Band 8 had the third best proposal ranks out of the six bands offered as measured at the median in the cumulative distribution. Figure 12 shows the acceptance rate of proposals as a function of the estimated execution time on the 7-m array. The acceptance rate is approximately uniform with requested 7-m array time up to the maximum allowed time of 150 h per proposal, indicating that the reviewers were not unduly ranking large time requests harshly.

Carpenter (2019) identified systematics in the proposal rankings, especially with region affiliation and experience of the PI but also gender, that may signify bias in the review process. While DPR is not expected to remove any such biases, it is important that the biases are not exacerbated. Figure 13 shows the cumulative distribution of ranks by region. As in the main call, proposals from East Asian PIs have poorer overall ranks than PIs from Europe and North America. The differences between regions are not statistically significant though, with a probability of $p = 0.21$ that the distributions are drawn from the same

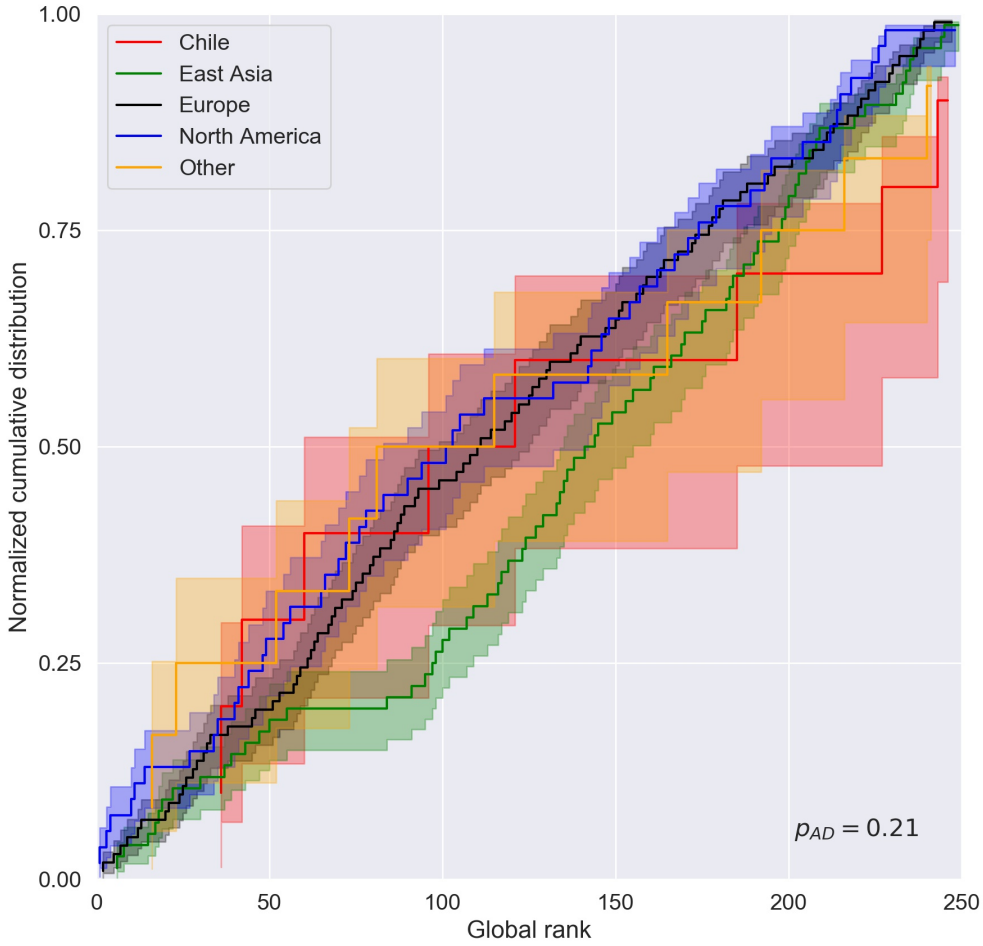

Figure 13: Cumulative distribution of proposal ranks in the supplemental call by region.

Table 8: Median normalized ranks by region in Cycle 7

Region	**Main call (Stage 1)**	**Supplemental call**
Chile	$0.52^{+0.05}_{-0.06}$	$0.49^{+0.26}_{-0.10}$
East Asia	$0.65^{+0.02}_{-0.02}$	$0.58^{+0.04}_{-0.04}$
Europe	$0.44^{+0.02}_{-0.01}$	$0.45^{+0.06}_{-0.05}$
North American	$0.48^{+0.03}_{-0.02}$	$0.41^{+0.16}_{-0.08}$
Other	$0.57^{+0.03}_{-0.05}$	$0.46^{+0.31}_{-0.17}$

distribution. Table 8 compares the normalized median proposal ranks in the Cycle 7 main call (the Stage 1 ranks) and the supplemental call, where the top-rated proposal has a normalized rank of 0 and the lowest-rated proposal has a normalized rank of 1. In the main call, the median proposal from East Asia had a normalized rank of 0.65 ± 0.02, where the expected value is 0.5 if no systematics in the review process are present. The uncertainty in the median rank was estimated by running 10,000 bootstrap simulations with replacement. In the supplemental call, East Asian PIs had a normalized rank of 0.58 ± 0.04, and therefore showed an improvement relative to the main call. However, it should not be interpreted that the regional systematics are reduced with DPR. Besides the smaller number of proposals, a larger fraction of the reviewers in the supplemental call are from East Asia (30%) relative to the main call (20%). Without further analysis, it cannot be ruled out that the differences in the demographics of the reviewers contributed to the apparent improvement in the East Asian ranks.

Figure 14 shows the cumulative rankings grouped by experience, where experience reflects the number of cycles the PI has submitted a proposal in the *main* call. The results show that the least experienced PIs (who have submitted a proposal in at most one main cycle) have the poorest overall scores, which has been true in all cycles to date. The most experienced PIs had the best overall proposal ranks up until Cycle 7, when the investigator lists were randomized on the coversheet. After randomizing the names, the most experienced PIs had average rankings. In the supplemental call, which also randomized the investigator names, the most experienced PIs also had average rankings. While the tendencies in the supplemental call are consistent with Cycle 7, the overall differences in the ranks with experience level are not statistically significant with $p = 0.13$.

Figure 15 shows the cumulative distribution of ranks by gender. In the supplemental call, proposals led by women had better proposal ranks than proposals from men, although the probability that the two distributions are drawn from the same parent population is 0.09 and is not statistically significant. Nonetheless, women compared more favorably to men in the supplemental call than in any main cycle.

In summary, the reduced number of proposals submitted to the supplemental call compared to the main call makes it more difficult to identify systematics conclusively. While not statistically significant, the same general tendencies are present in the supplemental call as in the Cycle 7 main call; e.g., East Asian proposals have poorer ranks (but with some improvement relative to the main call), the least experienced PIs have poorer ranks, and the most experienced PIs have average grades after randomizing the investigator list on the proposal cover sheet. On the other hand, women did better than men in any previous cycle. As indicated previously, it is premature to conclude that any of the systematics identified in the main call have been fundamentally changed with DPR. It does seem apparent though that any systematics have not be exacerbated by the DPR process.

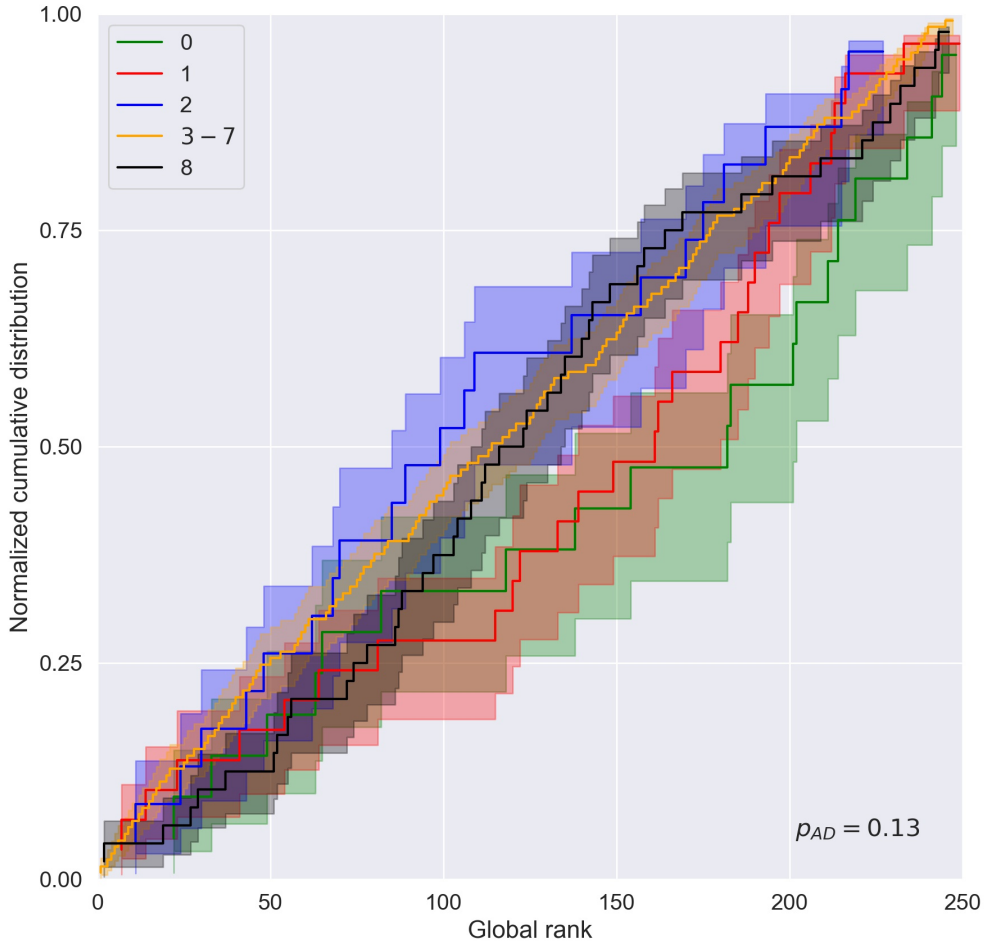

Figure 14: Cumulative distribution of proposal ranks in the Cycle 7 supplemental call grouped by the number of cycles in which a PI has a submitted proposal to the main call. PIs who have submitted to the main call between 3 and 7 cycles are grouped together for clarity.

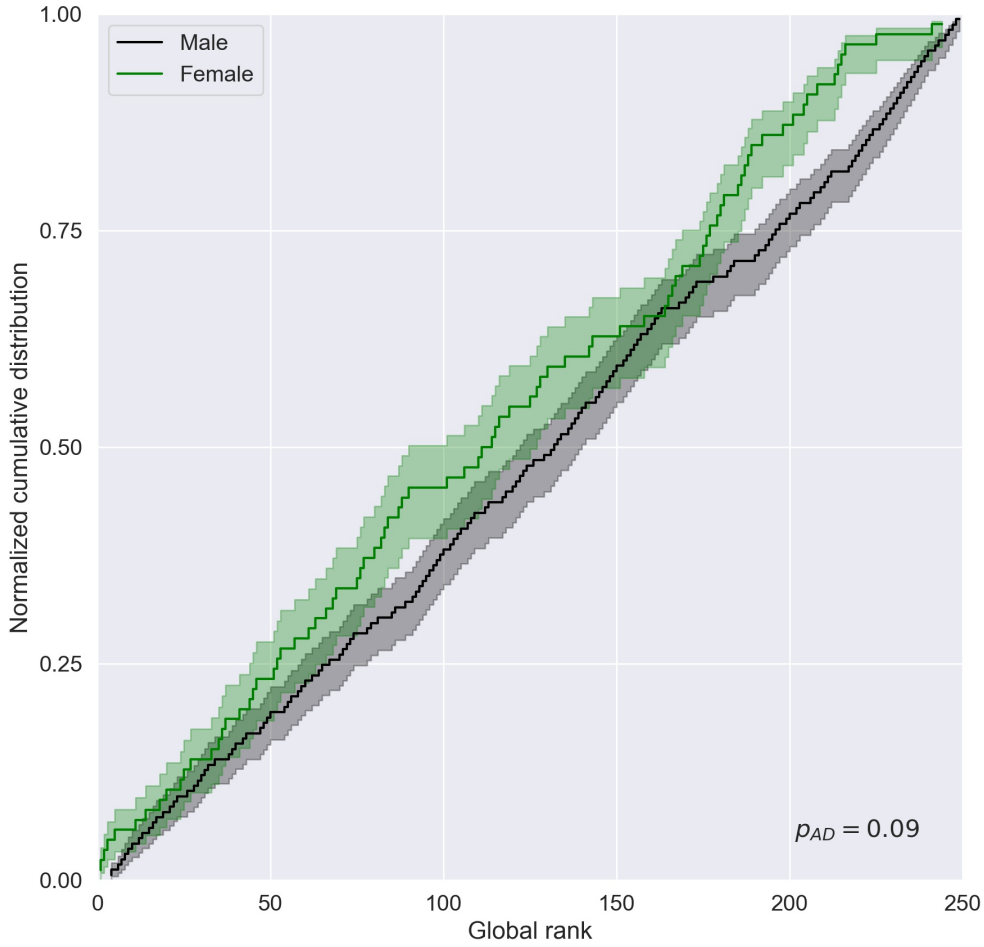

Figure 15: Cumulative distribution of proposal ranks in the supplemental call by gender.

4 Analysis of the reviewer survey

After reviewers completed their assignments, they were directed to a survey designed to assess their experience with the review process itself as well as a self-assessment of their expertise on each of the assigned proposals. In addition to specific questions, reviewers could provide free-form feedback on the documentation and the overall review process. The full text of the questions, the responses to the survey questions, and the free-form feedback are presented in Appendix A. This section summarizes the results and correlates the responses with the reviewer expertise and career status. Since 95% of the reviewers responded to the survey, the results are considered very representative of the thoughts of the reviewer population.

4.1 Documentation and reviewer tools

Reviewers rated the helpfulness of the guidelines to writing comments to the PI (see Appendix A.1), the relevancy of the review criteria (see Appendix A.2), and the ease-of-use of the reviewer tool (see Appendix A.5). The responses from the reviewers were positive on all three aspects: 78% of the reviewers felt that the guidelines to write comments are clear and appropriate, 88% indicated the reviewer criteria are fully or mostly relevant, and 98% rated the reviewer tool as easy or mostly easy to use. While not shown here, the satisfaction levels are similar across all career levels of the reviewers, as defined by the number of years since the PhD was obtained. The low number of help-desk tickets from users (Section 2.4) and the high positive responses with respect to the documentation and tools indicates that logistically, the proposal submission and review processes were successful.

4.2 Proposal reviews

Reviewers were asked how long they spent on average to review a proposal (see Appendix A.3). Figure 16 shows the distribution of time spent reviewing proposals by career level. Reviewers without a PhD or who obtained their PhD within the past 3 years spent more time on average reviewing proposals than more senior reviewers, with 57% of the non-PhD reviewers spending more than 45 minutes per proposal. A number of factors could contribute to this tendency. Younger reviewers almost certainly require more time to study a proposal due to inexperience. They also potentially may have more time available to evaluate proposals than senior astronomers. Given that younger reviewers likely have served on few if any committees previously, there may also be self-pressure to be particularly diligent in performing the evaluations.

Figure 17 compares the average time spent by reviewers to evaluate proposals in the main call and the supplemental call. For the supplemental call, only reviewers that received their PhD 4 years ago or more are shown since younger researchers typically do not serve on the panels. The results show that reviewers typically spent 30-45 minutes reviewing a proposal in the supplemental call compared to 15-30 minutes in the main call. Presumably this is because there are fewer proposals to review in the supplemental call.

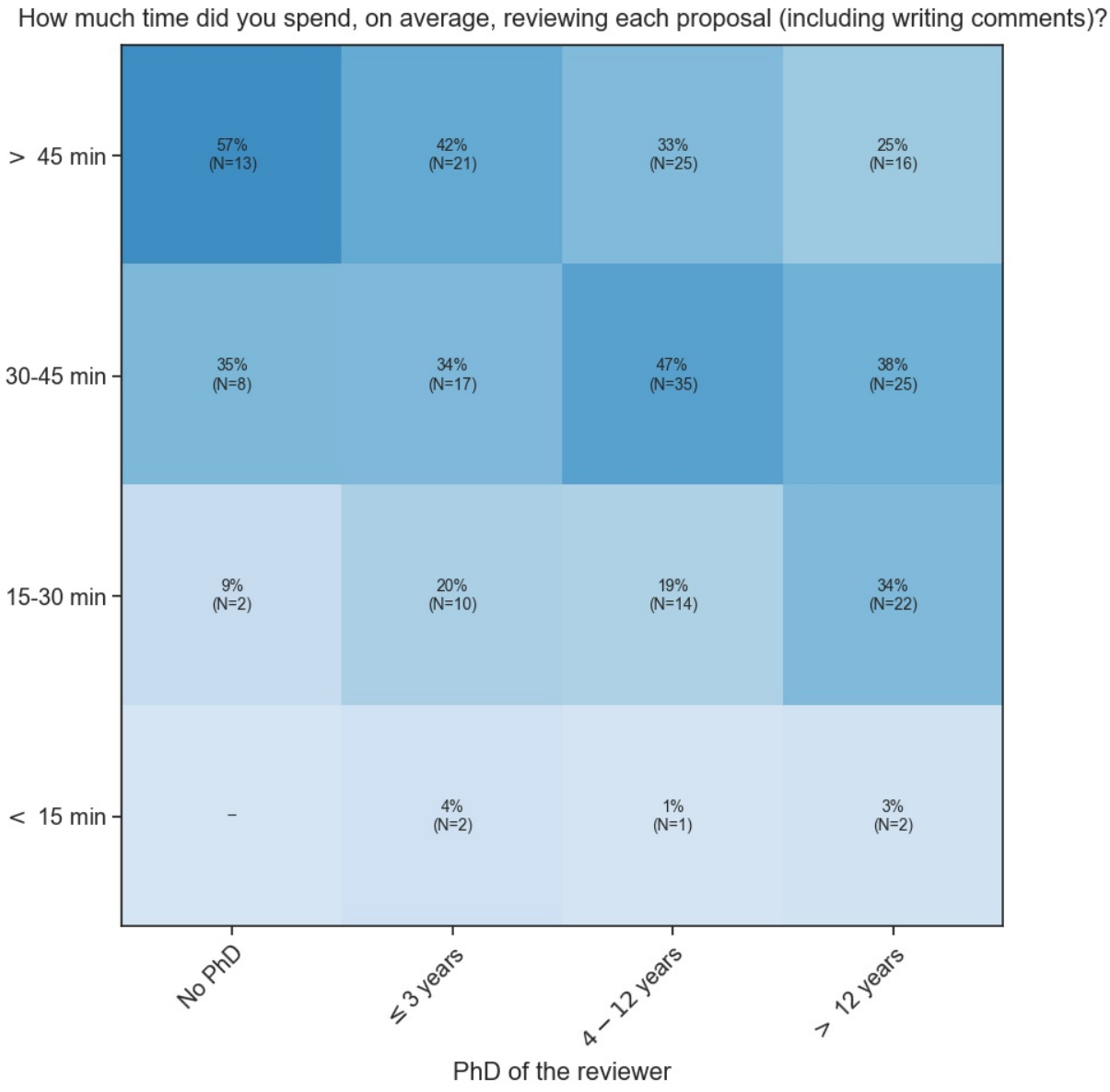

Figure 16: Distribution of the average time spent reviewing the supplemental call proposals by career level. The responses are normalized separately for each career level.

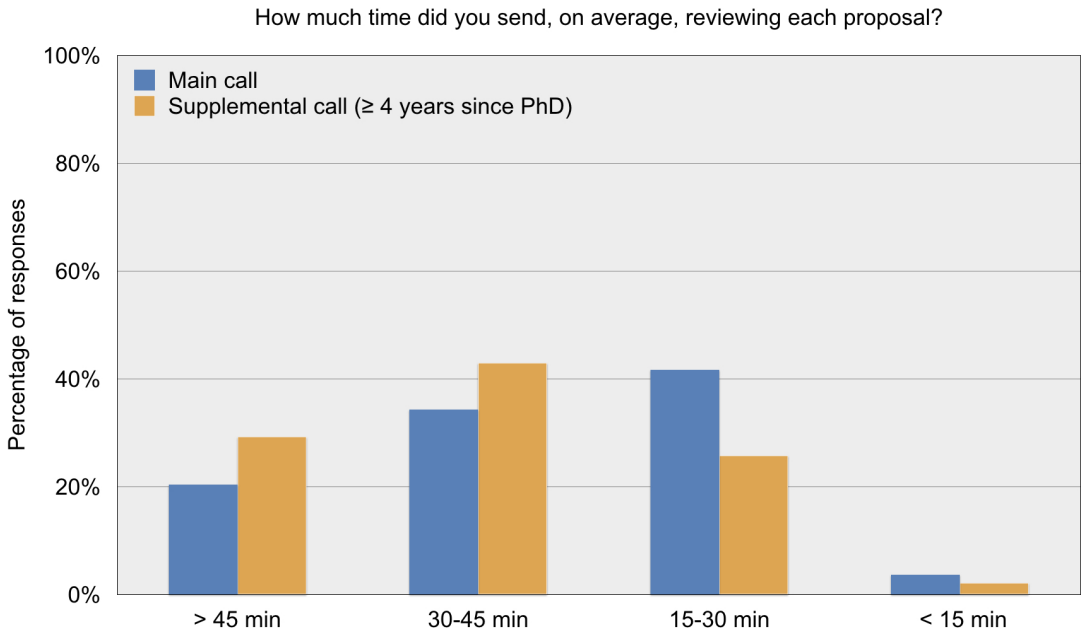

Figure 17: Average time spent reviewing a proposal in the Cycle 7 main call and supplemental call based on responses in the reviewer surveys. The results for the supplemental call are shown only for reviewers that received their PhD 4 years ago or more since younger researchers do not typically serve on the review panels.

Reviewers were given review assignments in the same category and with the same keywords as their submitted proposal to the extent possible. However, the topics within a category can be broad, and in some cases reviewers were assigned proposals from other categories. Nonetheless, 59% of the reviewers indicated that they were able to fully or mostly provide a fair assessment; only 8% indicated they were not able to provide a satisfactory evaluation. Figure 18 shows the distribution of responses between year of PhD and the self-assessment. Each career level shows a similar distribution of responses.

4.3 Self-assessment of expertise

The reviewers provided a self-assessment of their expertise on each assigned proposal, with the options of "This is my field of expertise", "I have some general knowledge of this field", and "I have little or no knowledge of this field". Figure 19 shows the responses separated by career level. Not surprisingly, younger reviewers declared themselves experts on a lower percentage of proposals than senior reviewers. Figure 20 shows the distribution of responses between reviewer expertise and rule assignment. Even for assignments in the same category and with at least one common keyword as the reviewer's submitted proposal (Rule 1; see Table 7), more often than not reviewers indicated that they had some or little/no knowledge. The trend remains even if only expert reviewers are considered. This may indicate that reviewers are conservative in declaring their expertise or the keywords cover a broad range of topics.

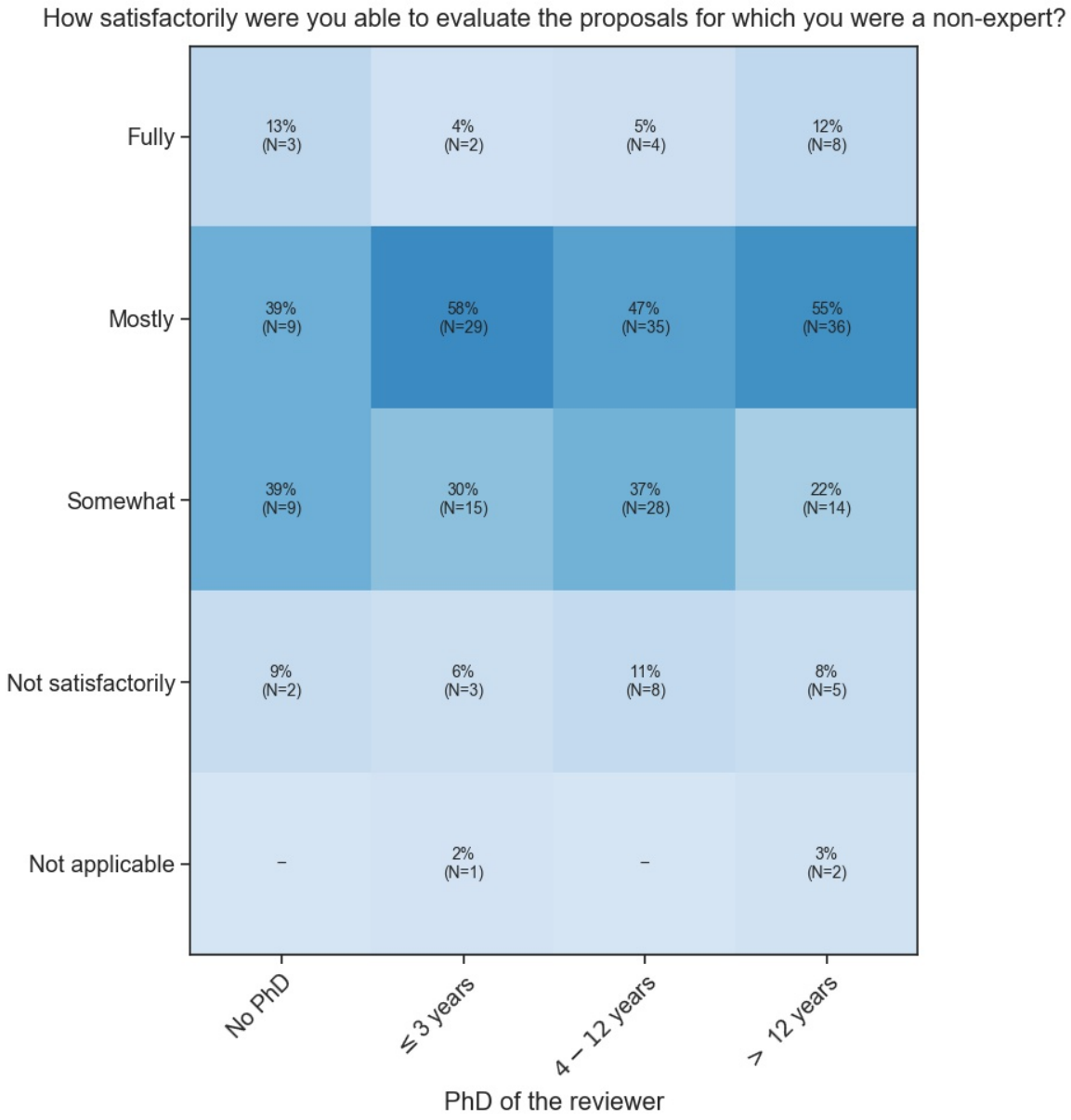

Figure 18: Distribution of the reviewer self-assessment in how fairly they were able to assess proposals in which they were a non-expert. The responses are normalized separately for each career level.

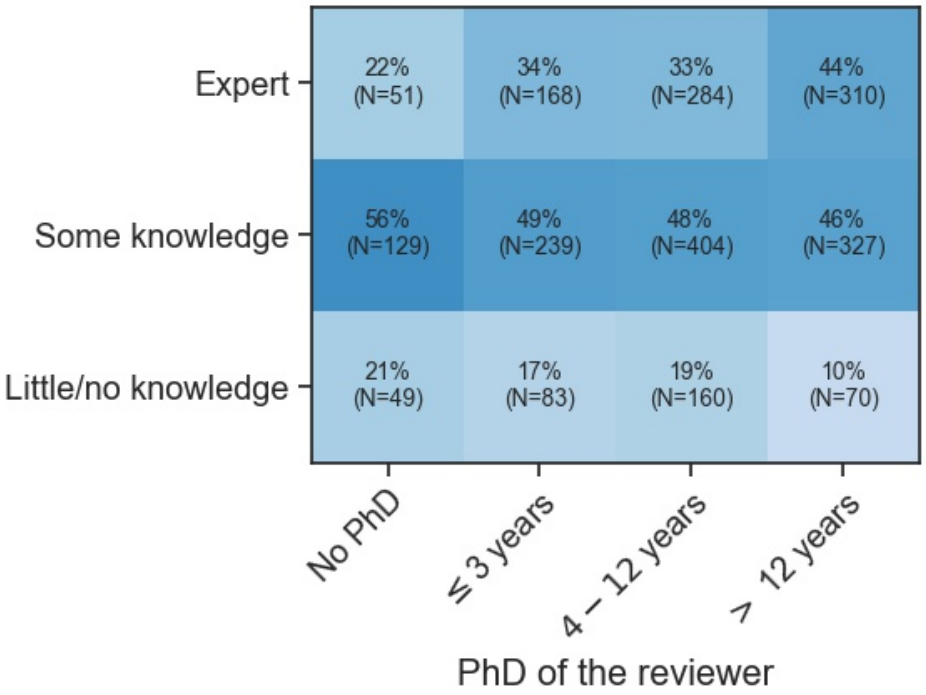

Figure 19: Distribution of the reviewer self-assessment of their expertise on individual assigned proposals across career levels. The responses are normalized separately for each career level.

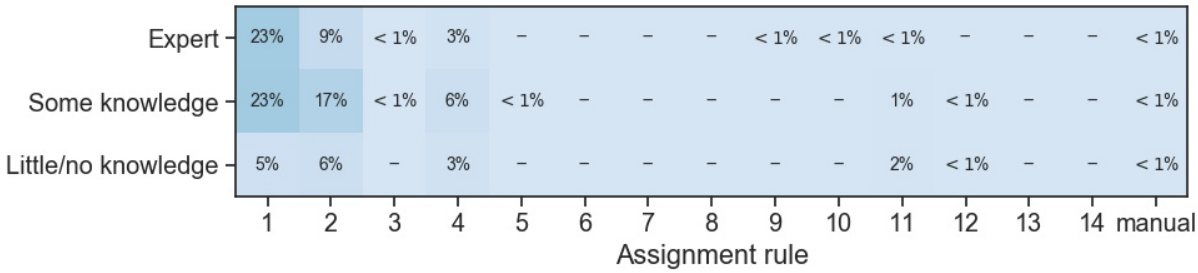

Figure 20: Distribution of the reviewer self-assessment of their expertise on individual assigned proposals across the rule assignment (see Section 3.3). The responses are normalized by the total number of responses.

4.4 Viability of DPR in the main call

The survey asked the reviewers whether distributed peer review would be appropriate as a means to review regular proposals in the main call while continuing with a panel review for Large Programs. Besides "yes" (36% of the responses) or "no" (27%), reviewers could indicate that both review models are equally effective (9%) or if they are unsure (27%). Treating "yes" and "equally" as supportive responses, 46% of the reviewers found DPR to be suitable for the main call. Figure 21 shows the correlation of the responses with career status. The responses are similar across reviewers with a PhD, although non-PhD reviewers are less supportive in that 43% do not feel DPR is suitable. Excluding the reviewers that are unsure, the probability that the non-PhD reviewers have a different opinion on the viability of DPR than the reviewers with a PhD is $p = 0.12$, where the sum of the "yes" and "equally effective" responses are compared against "no". Therefore the difference in responses between the non-PhD and PhD reviewers is not statistically significant.

The supplemental call reviewers included 39 people who had previously served on a review panel in the main call. Figure 22 shows the distribution of responses for these reviewers along with supplemental call reviewers who have not previously served and received their PhD 4 years ago or more. Overall, main-call reviewers are somewhat less favorably disposed to DPR compared to those who have not served, but the difference is not statistically significant ($p = 0.24$). Nonetheless, it is notable that previous panel reviewers are not overwhelmingly in favor of DPR, suggesting that they place value in the face-to-face panel discussions that they have experienced.

5 Analysis of the PI survey

A key metric to evaluate DPR is PI satisfaction with the process in general as well as how helpful they found the reviewer comments to be. The PI survey gauged the quality of the comments overall and garnered a rating of the helpfulness of individual comments. Appendix B provides the results of the PI survey. This section summarizes the results and examines the correlation of the PI assessment of the quality of the comments versus the career status of the PI and the reviewer that wrote the comments. Since 70% of the PIs completed the survey, the results should reflect the consensus opinions reasonable well.

5.1 General comments on reviewer feedback

PIs rated the overall quality of the reviewer comments in terms of clarity, accuracy, helpfulness in improving future proposals, and professionalism of the comments. Figures 23–26 show the responses grouped by the PhD year of the PI. The main results are as follows.

- The vast majority of PIs (85%) indicated that the comments are mostly or fully clear overall even if they disagreed with the comments scientifically (Figure 23). This sentiment is expressed across all career levels.

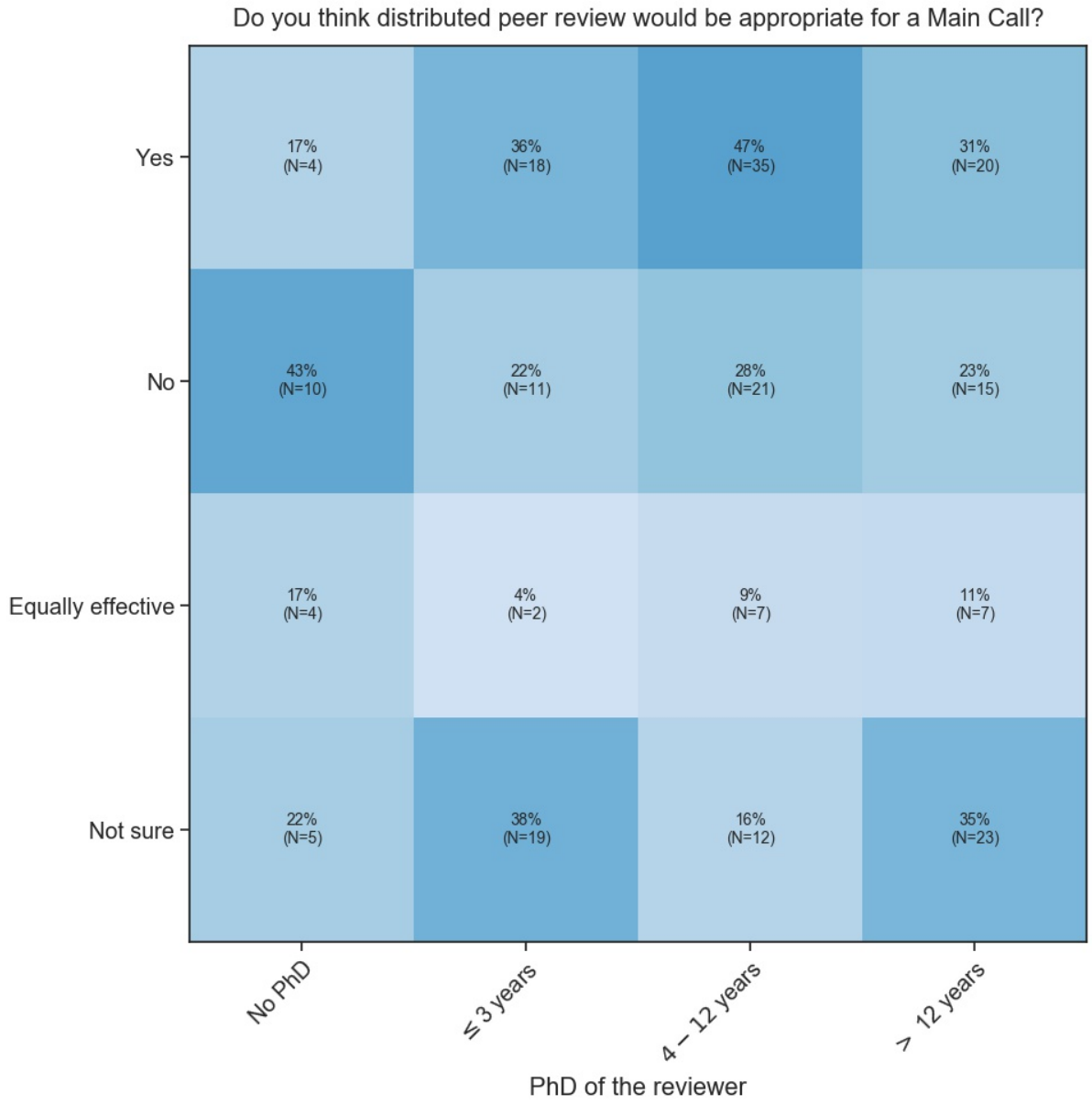

Figure 21: Distribution of the reviewer opinions on the viability of DPR for regular proposals in the main call, grouped by the career status of the reviewer. The results are normalized for each career level.

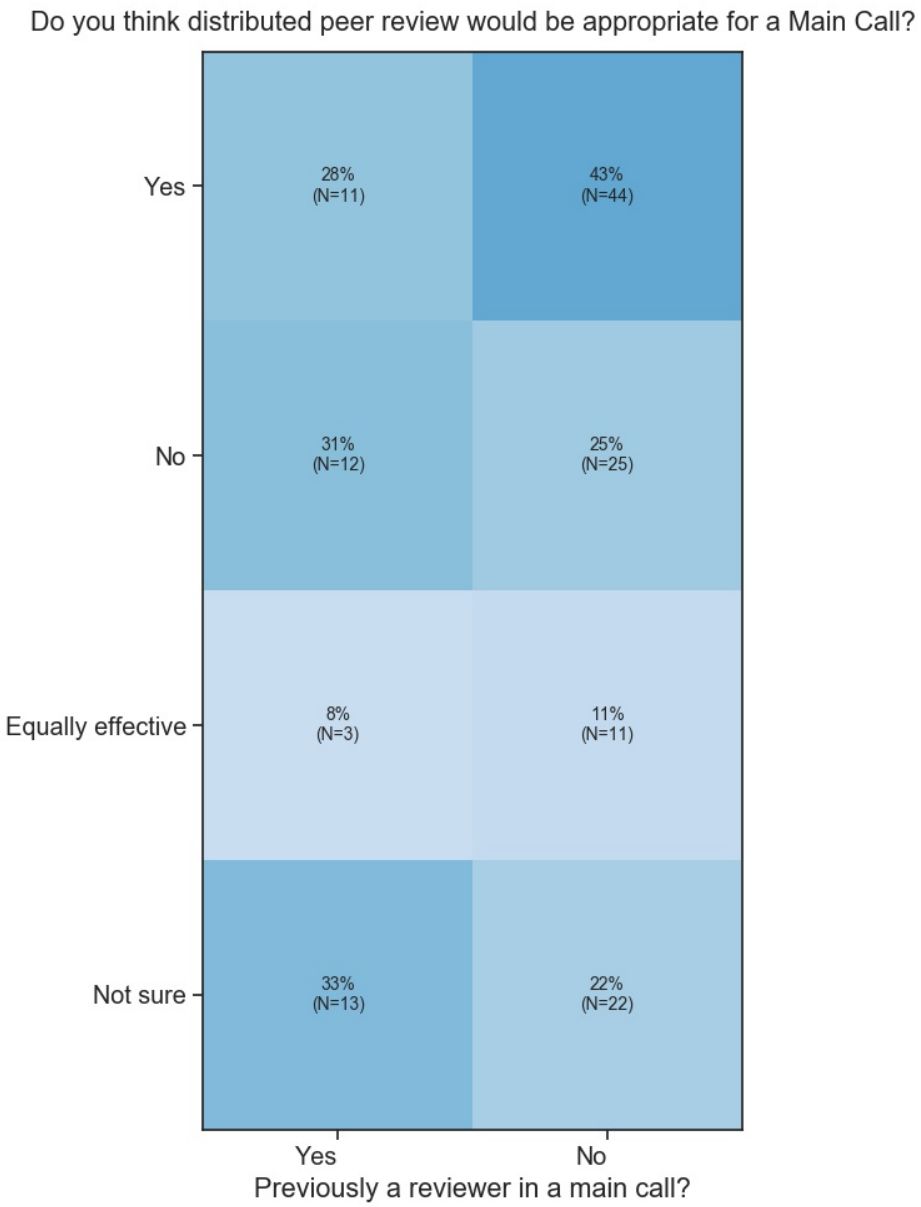

Figure 22: Distribution of reviewer opinions from the supplemental call on the viability of DPR for regular proposals in the main call. The responses are grouped by previous participation in the main call review. Responses from those who have not previously served on a main call panel are shown if they received their PhD 4 years ago or more. The results are normalized separately for each column.

- Most PIs (69%) indicated the comments are fully or mostly accurate scientifically (Figure 24). Only two PIs that responded indicated that the comments are not accurate overall. Younger PIs are more likely to indicate that the comments are accurate than experienced PIs.

- Most PIs (54%) indicated the comments will fully or mostly help them improve future proposals (Figure 25). Only 5% of the PIs indicated that the comments will not be helpful. Younger PIs are more likely to indicate that the comments will help them improve future proposals than experienced PIs.

- The vast majority of PIs ($> 82\%$) indicated that the reviewer comments are professional overall (Figure 26). The trend is present across all career levels. Only $\sim$1% of the PIs indicated the comments are unprofessional overall.

PIs in the Cycle 7 main call answered similar questions on the quality of their consensus reports. The main difference in the syntax of the questions/answers between the two surveys is that the main call referred to the singular consensus report and the supplemental call referred to the overall responses of the individual reviewers. Figures 27–30 compare the responses of the two calls. PIs from the main call are about twice as likely to find the consensus report "fully" clear, scientifically accurate, or helpful compared to supplemental call style of comments. However, PIs from the main call are also more likely to find the consensus reports unclear, inaccurate, or unhelpful. The lower percentage of fully helpful or unhelpful responses in the supplemental call is likely because given that there are 10 individual comments, PIs will find some clear and others unclear, leading to more intermediate responses. Considering the sum of responses for "fully" and "mostly", a similar percentage of the PIs in the main and supplemental calls found the reviewer comments clear, accurate, and helpful.

PIs in the supplemental call assessed the quality of the reviewer comments in the supplemental call versus the consensus reports in the main call. Figure 31 shows the results by the career level of the PI. A slightly higher fraction prefer the consensus reports than the style of comments in the supplemental call (34% vs. 31%). Given that 25% of the PIs indicate the quality of the comments are similar, there is no clear preference for one style over the other. No significant trend is present with career level of the PI.

5.2 PI assessment of individual comments

To evaluate the helpfulness of the reviews across all career levels, PIs in the supplemental call assessed the helpfulness ("very helpful", "somewhat helpful", "inaccurate or not helpful", or "inappropriate/unprofessional") of each reviewer comment. The replies have been correlated with the career level of the PI, the career level of the reviewer, and the self-assessment of the reviewer expertise to determine the fraction of comments that are helpful and the source of the helpful comments.

As indicated in Figure 32, PIs rated 75% of the individual comments as very or somewhat helpful. Younger PIs tend to rate a higher percentage of their assessments as "very

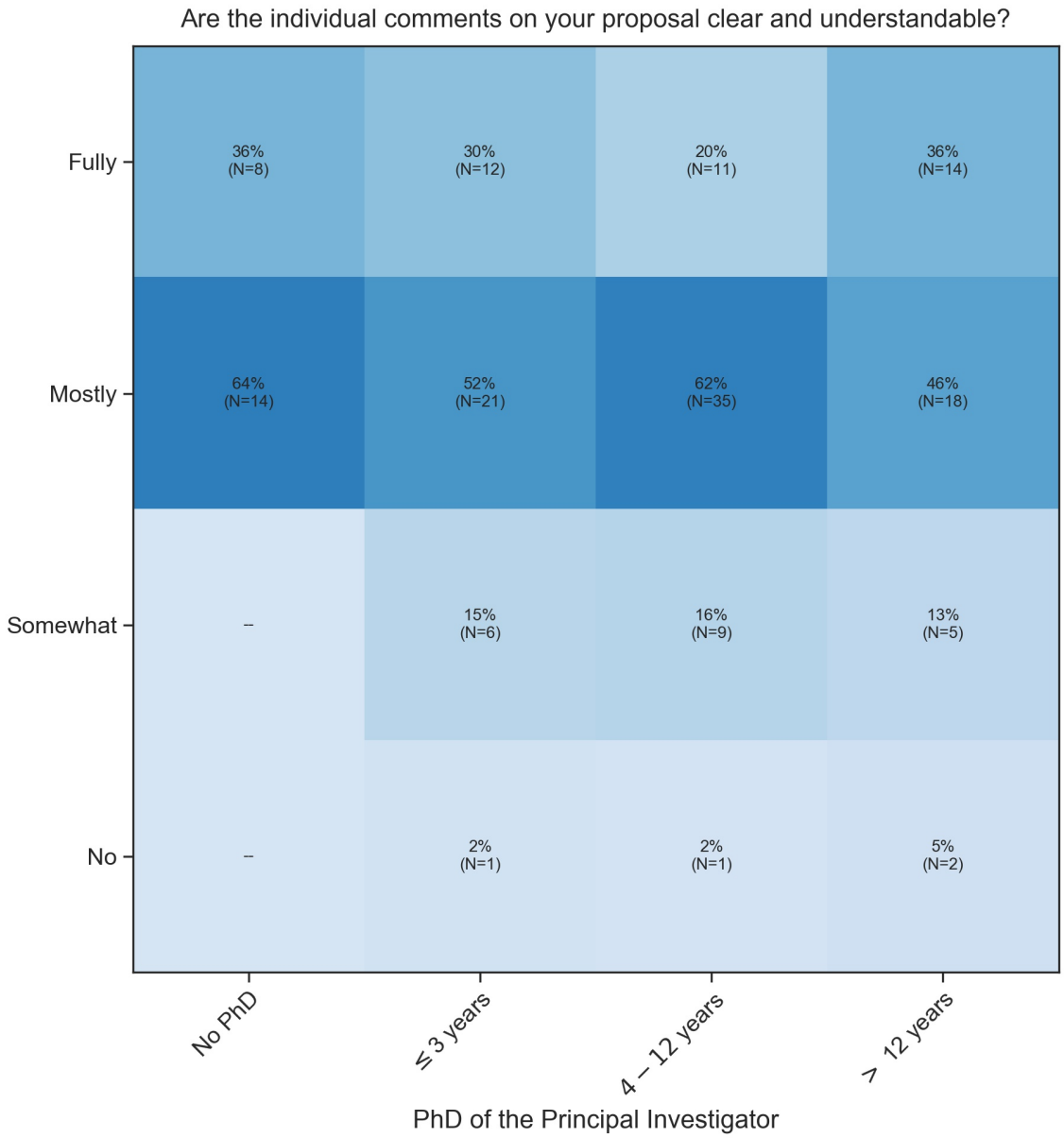

Figure 23: Percentage of PIs that found the reviewer comments clear in general, even if they disagreed with the comments scientifically. The results are normalized for each career level of the PI.

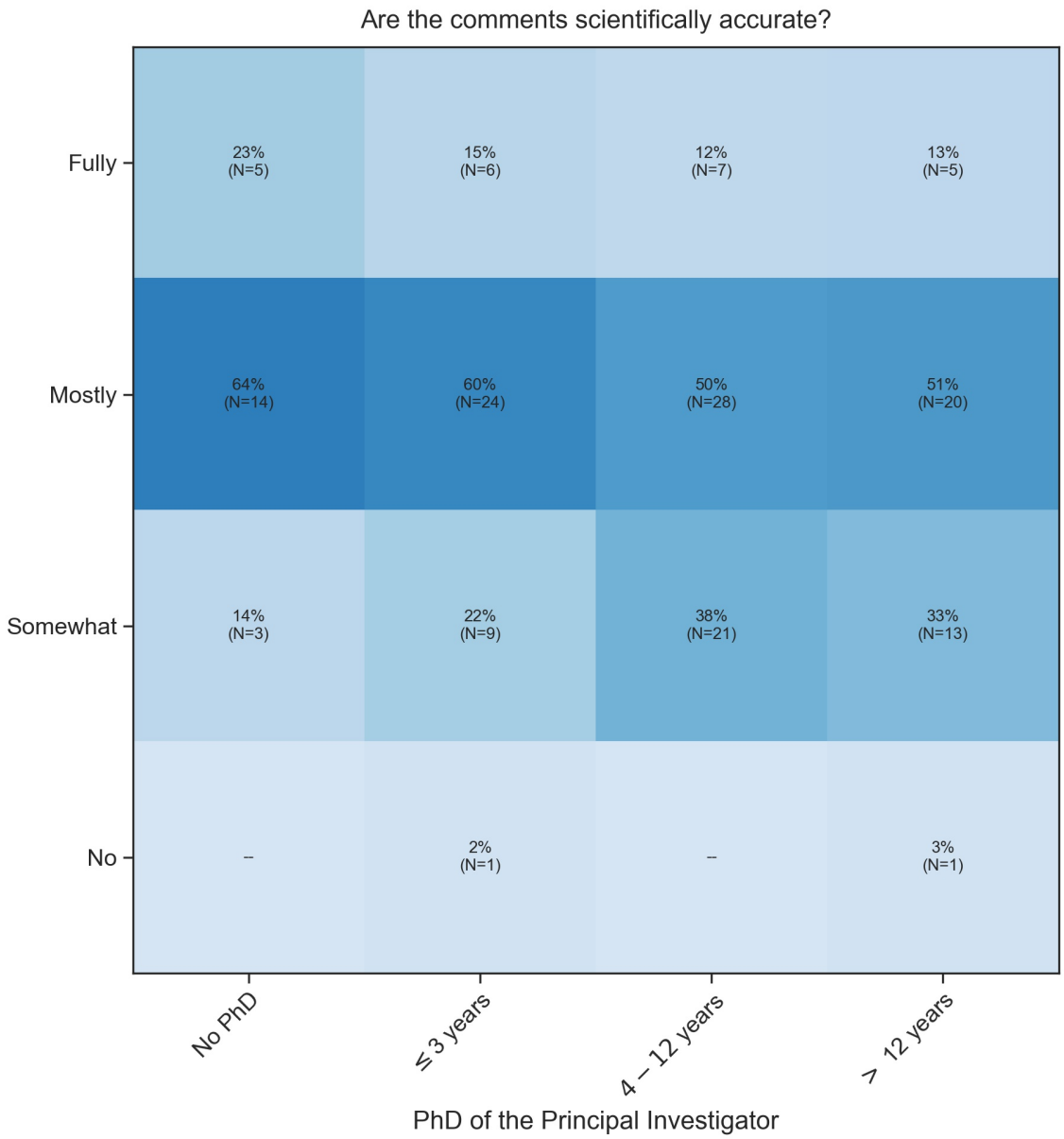

Figure 24: Percentage of PIs that found the reviewer comments accurate in general. The results are normalized for each career level of the PI.

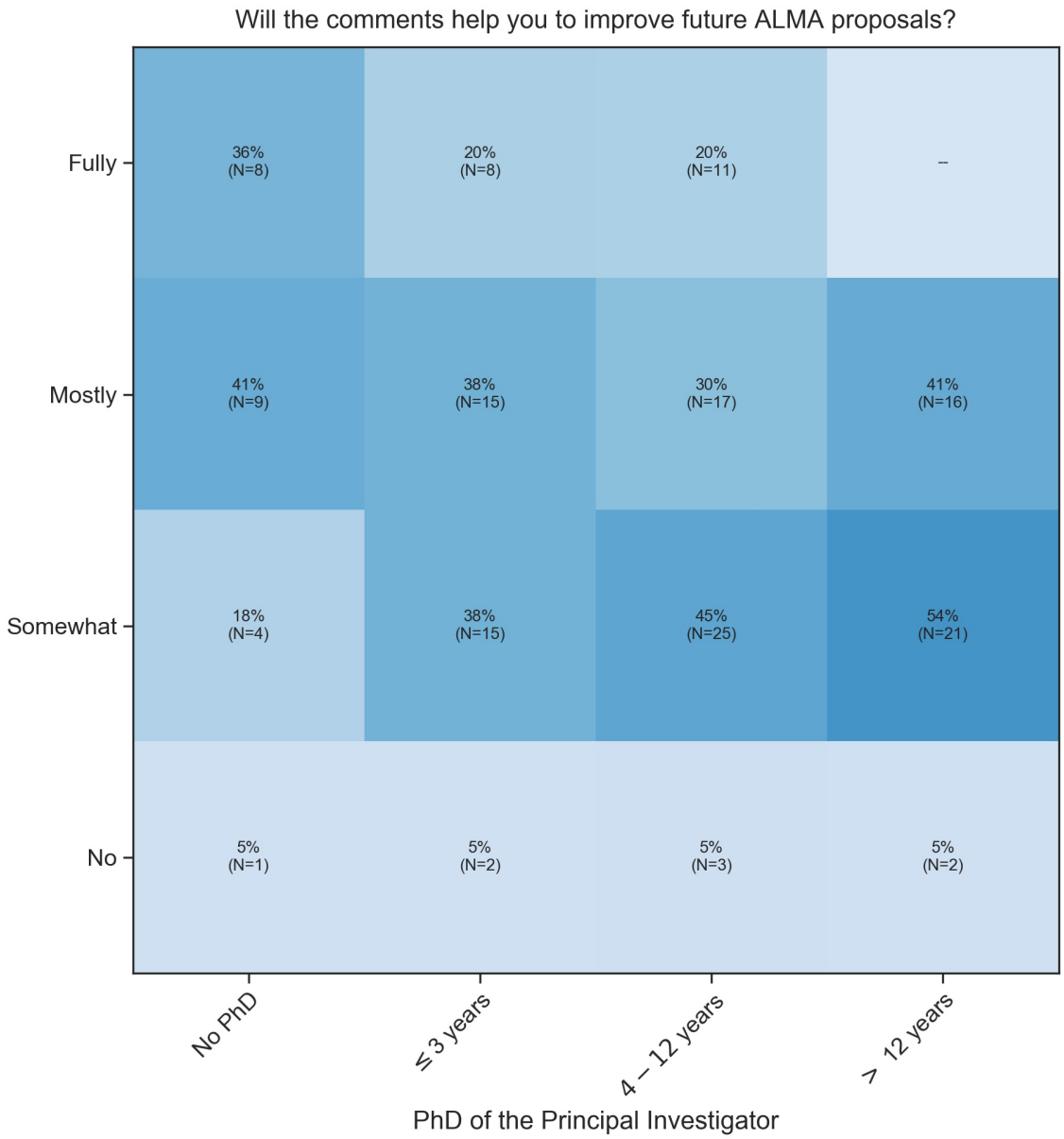

Figure 25: Percentage of PIs that found the reviewer comments can help improve future proposals. The results are normalized for each career level of the PI.

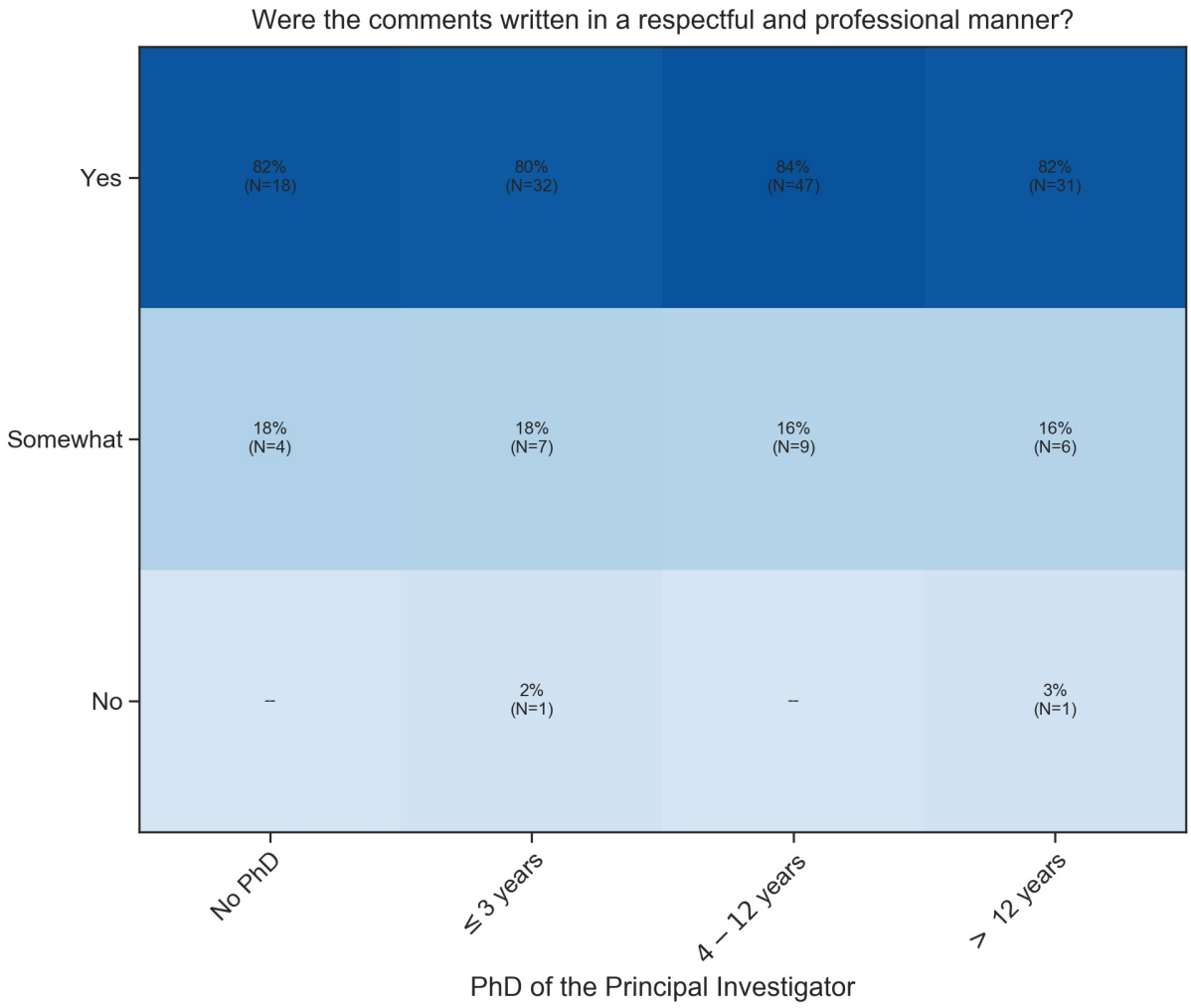

Figure 26: Percentage of PIs that found the comments professional overall. The results are normalized for each career level of the PI.

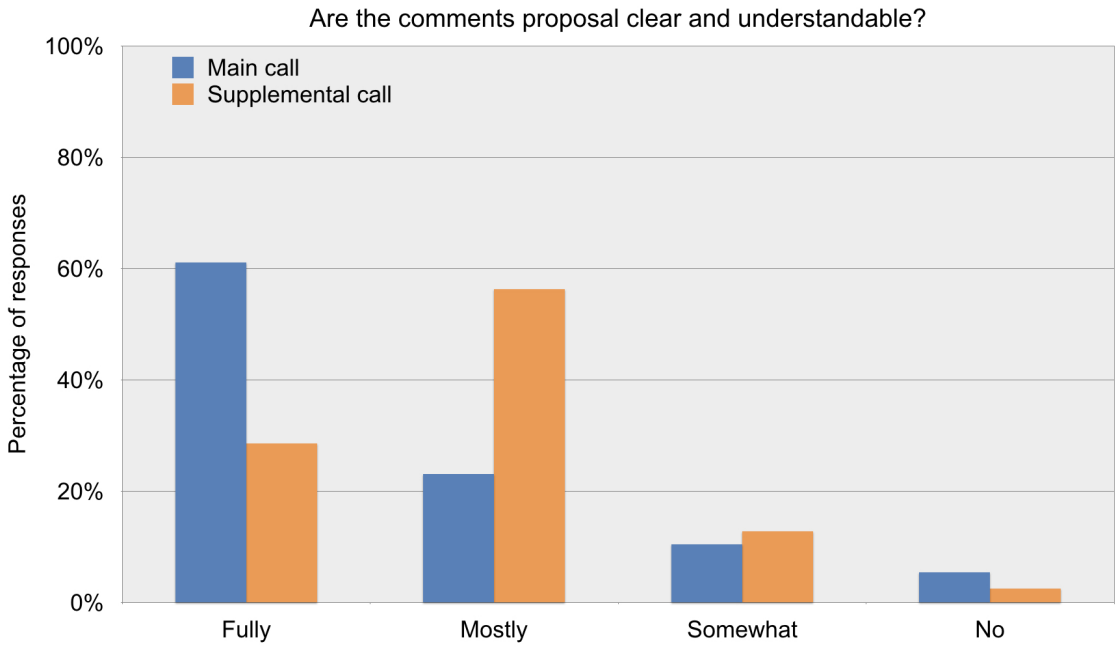

Figure 27: Comparison between the Cycle 7 main and supplemental calls on the clarity of the reviewer comments, regardless of the scientific accuracy of the comments.

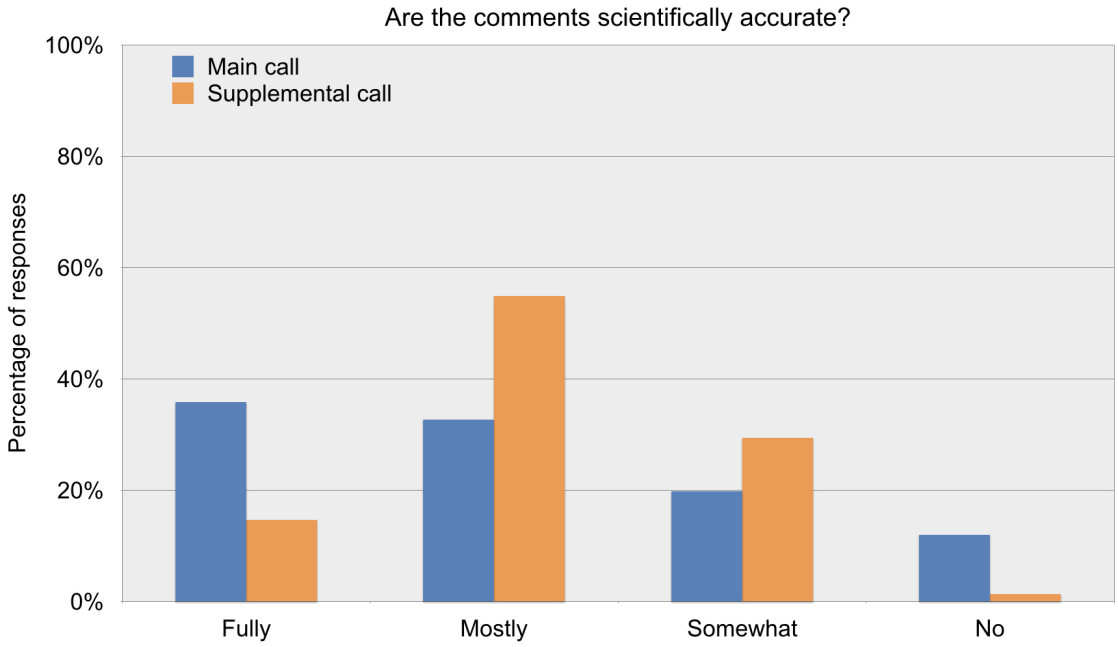

Figure 28: Comparison between the Cycle 7 main and supplemental calls on the scientific accuracy of the reviewer comments.

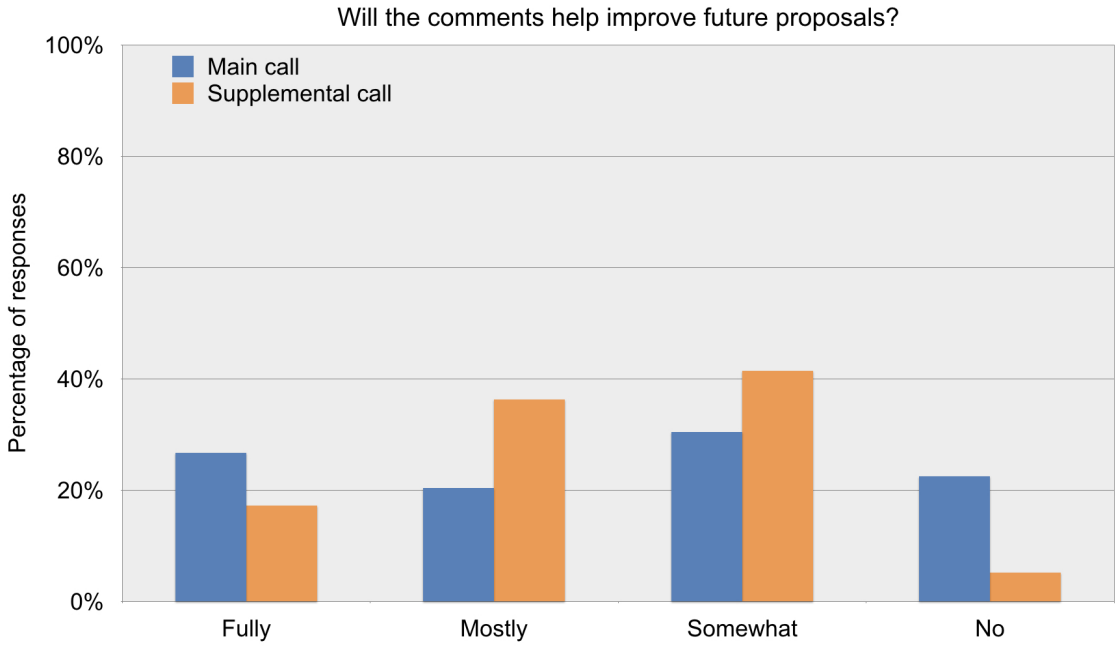

Figure 29: Comparison between the Cycle 7 main and supplemental calls on the helpfulness of the reviewer comments.

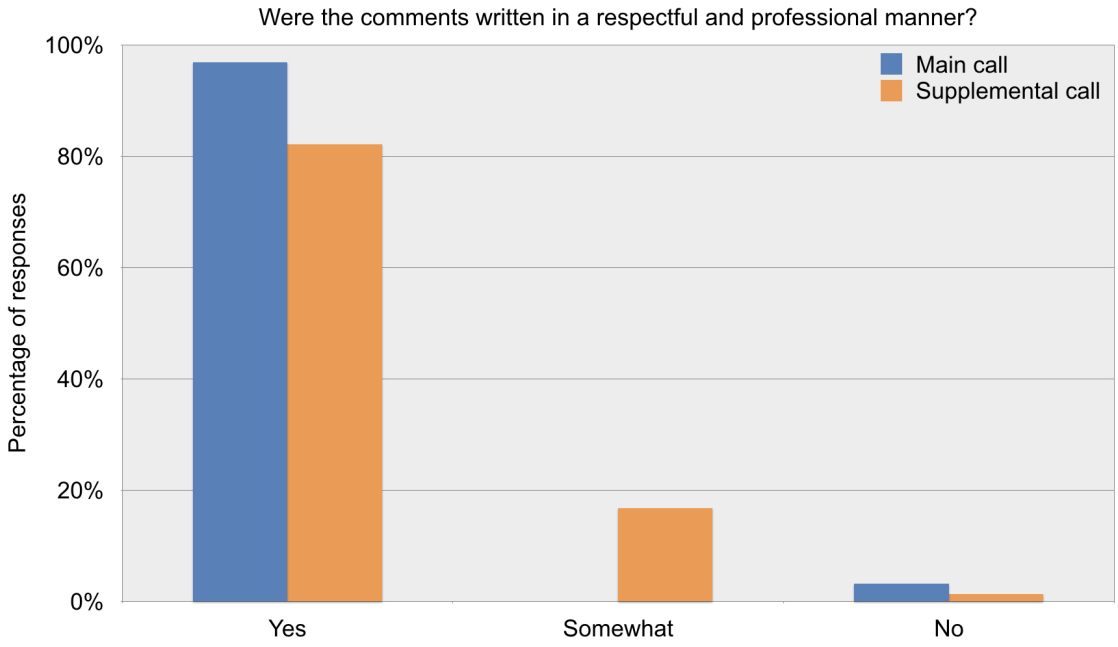

Figure 30: Comparison between the Cycle 7 main and supplemental calls on the professionalism of the reviewer comments. For the main call, the only available options were "yes" and "no".

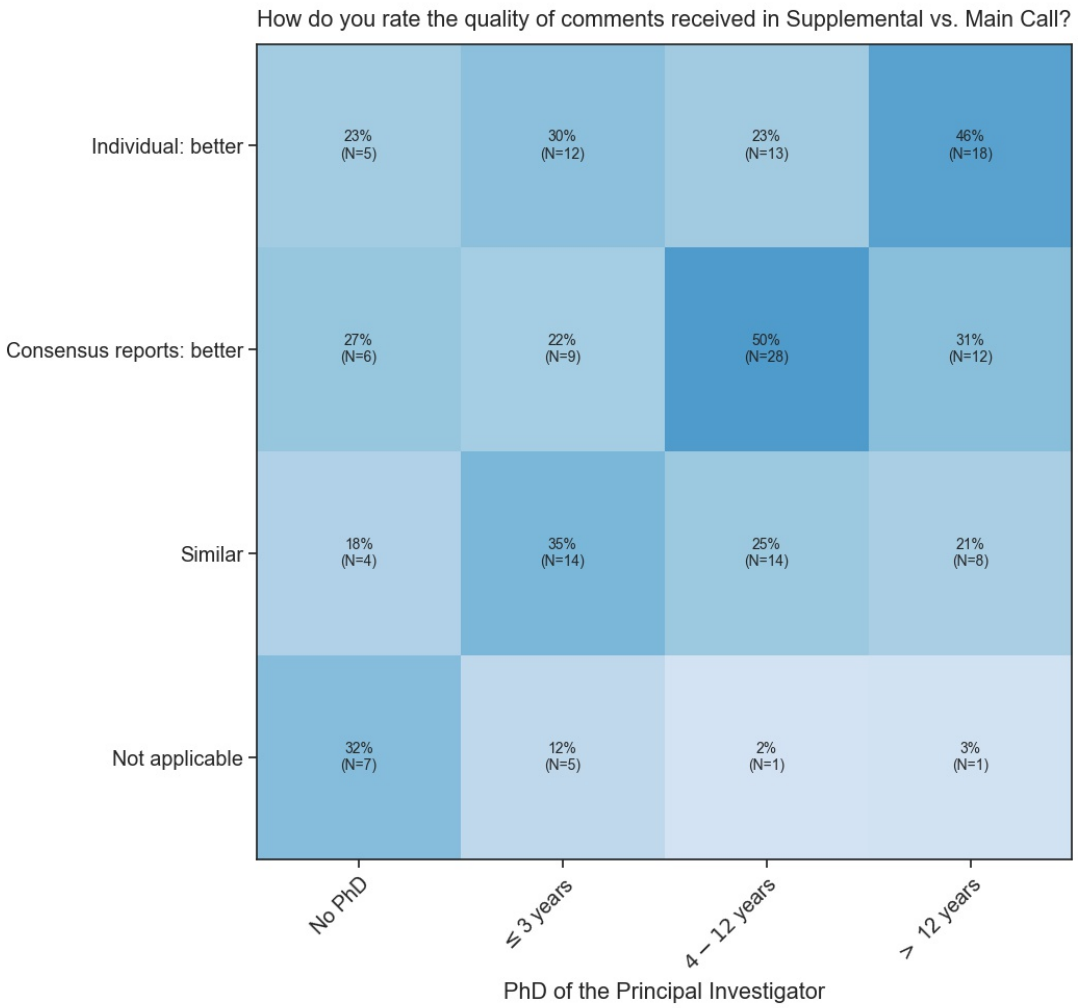

Figure 31: PI assessment of the quality of the reviewer comments relative to the consensus reports from the main call. The results are normalized for each career level of the PI.

helpful" compared to the most senior PIs. In terms of who wrote the comments, non-PhD reviewers had the highest percentage of very helpful reviews and the lowest percentage of inappropriate/unprofessional reviews, but the helpfulness of the comments show a similar distribution across all reviewer career levels (Figure 33). A χ^2 analysis of the matrix indicates that the results are statistically consistent across all career levels ($p = 0.44$). Since most non-PhD reviewers consulted with the mentors on the science evaluation and writing the comments (see Appendix A.10 and A.11, respectively), this suggests the mentor role either added value, or the students were able to provide quality feedback on their own. The helpfulness of reviews from young postdocs (i.e., less than 3 years with a PhD) are on par with experienced researchers with no assistance from a mentor.

Figure 34 correlates the helpfulness of the comments with the self-declared expertise of the reviewers. Again the results are consistent across of levels of expertise. Reviewers with little or no expertise in the proposal topic had slightly less helpful reviews, but the results are all consistent with the same parent distribution ($p = 0.64$). Figure 35 shows the the percentage of reviewer comments rated in various degrees of helpfulness by the PIs. The typical case is that a reviewer contains a mix of feedback on the quality of the comments, with most of the comments somewhat helpful.

5.3 Suitability of DPR in future calls

After receiving the outcome of the review process, PIs assessed the suitability of DPR in future proposal calls. The survey included questions to gauge their level of concern about biases and confidentiality in DPR, and the type of proposals (if any) that would be appropriate for DPR.

Figure 36 shows the correlation between the career level of the PI and their concerns about biases in DPR relative to panel reviews. More PIs indicate that panels are more robust against biases (29%) than DPR (22%), but 50% feel that the biases are similar in the two review processes or have no strong opinion. The results are similar across career level of the PI, with younger PIs tending to think panels are more robust against biases than DPR. However, any differences are not statistically significant.

Figure 37 shows the correlation between the career level of the PI and their concerns about confidentiality of the review process. More PIs (27%) are concerned about confidentiality in DPR than in the main call (7%). The trends are consistent across all career levels, and based on Poisson statistics, the higher level of concern abut DPR is statistically significant. However, the majority of PIs (64%) feel confidentiality concerns are the same between DPR and panels or have no strong opinion. The youngest researchers tend to have no strong opinion, which may not be surprising given their lack of experience in review processes.

Similar to the reviewer survey, PIs indicated which type of proposals would be suitable for DPR. The options were "small" ($<$ 25 h on the 12-m array), "medium" ($25 - 50$ h), "large" ($>$ 50 h), ACA standalone in a supplemental call, or none. Figure 38 shows the correlation between proposal type and the career level of the PI. A majority of PIs support using DPR to evaluate small (61%) and ACA standalone proposals in a supplemental call

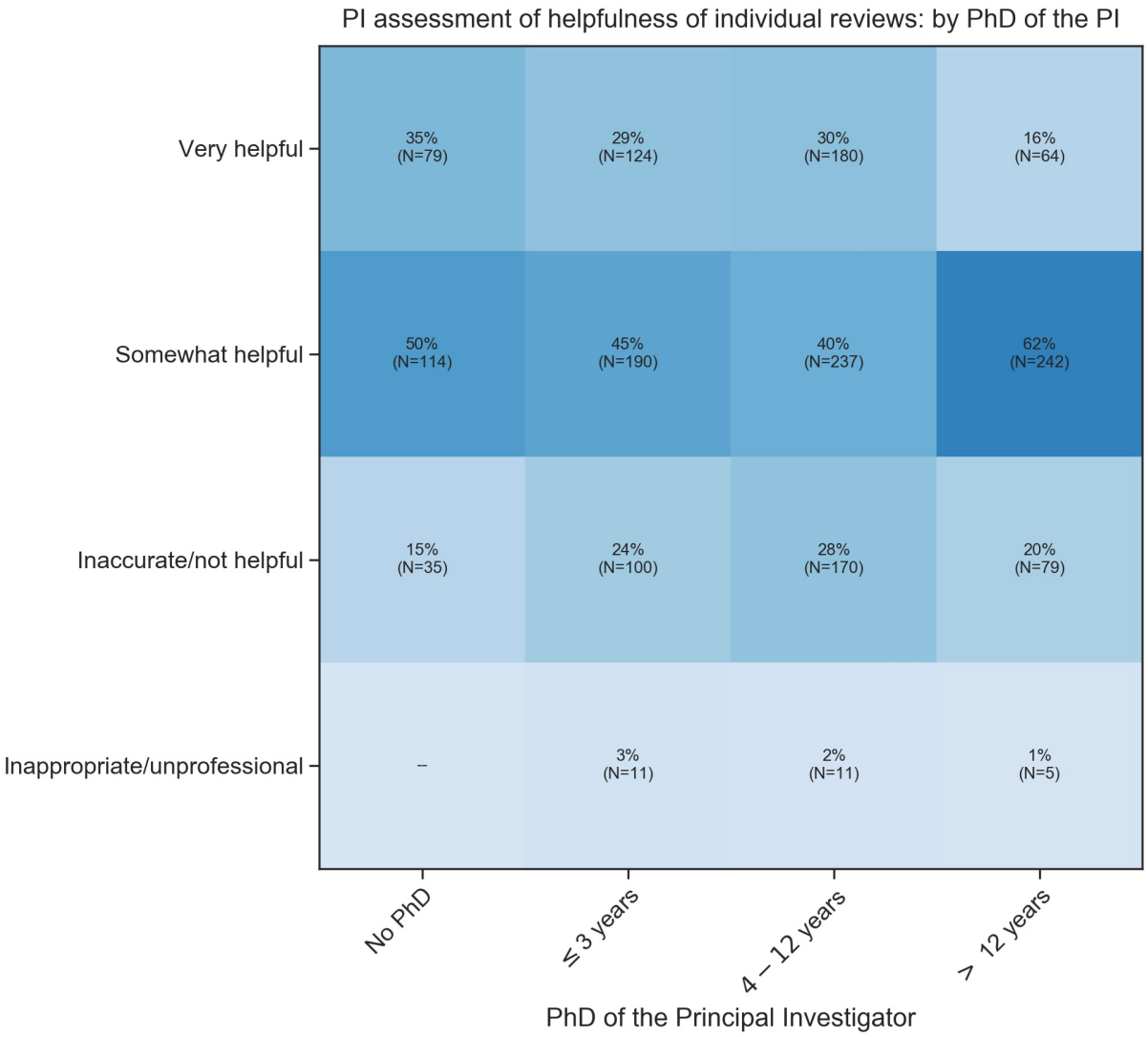

Figure 32: PI assessment of the helpfulness of individual comments by the experience level of the PI. The results are normalized for each career level of the PI.

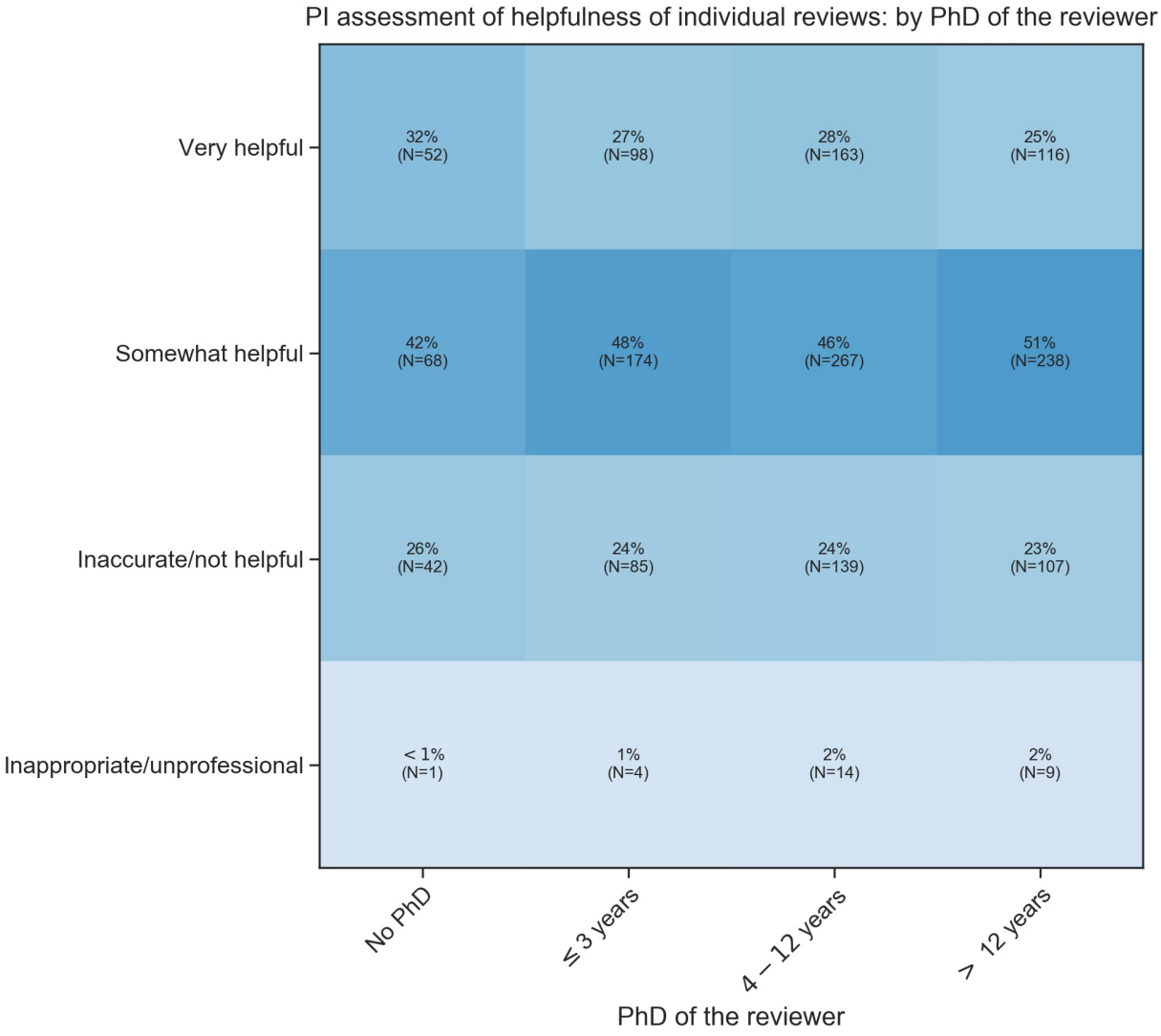

Figure 33: PI assessment of the helpfulness of individual comments by the experience level of the reviewer. The results are normalized for each career level of the reviewer.

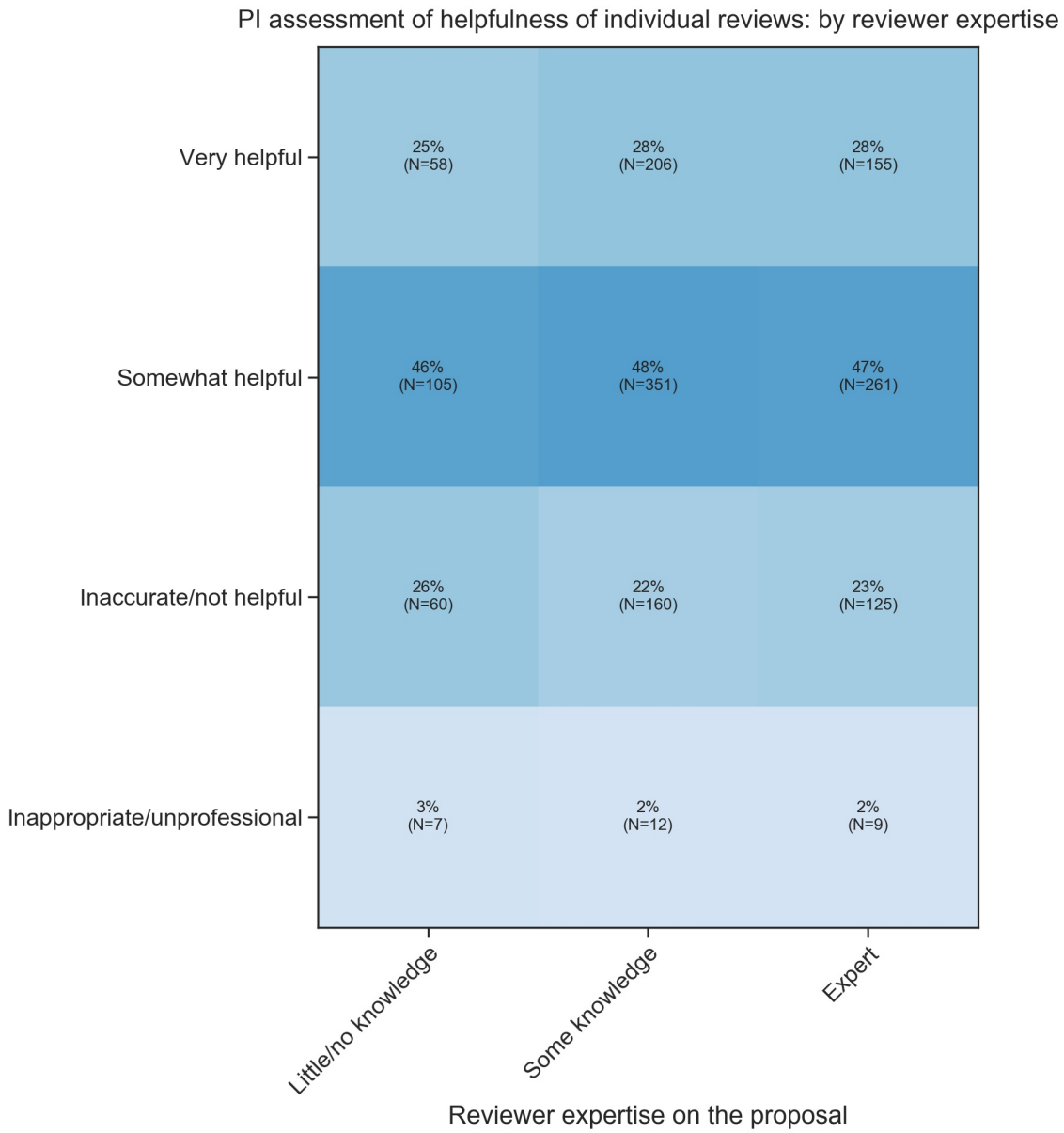

Figure 34: PI assessment of the helpfulness of individual comments by the expertise level of the reviewer for the proposal. The results are normalized for level of reviewer expertise.

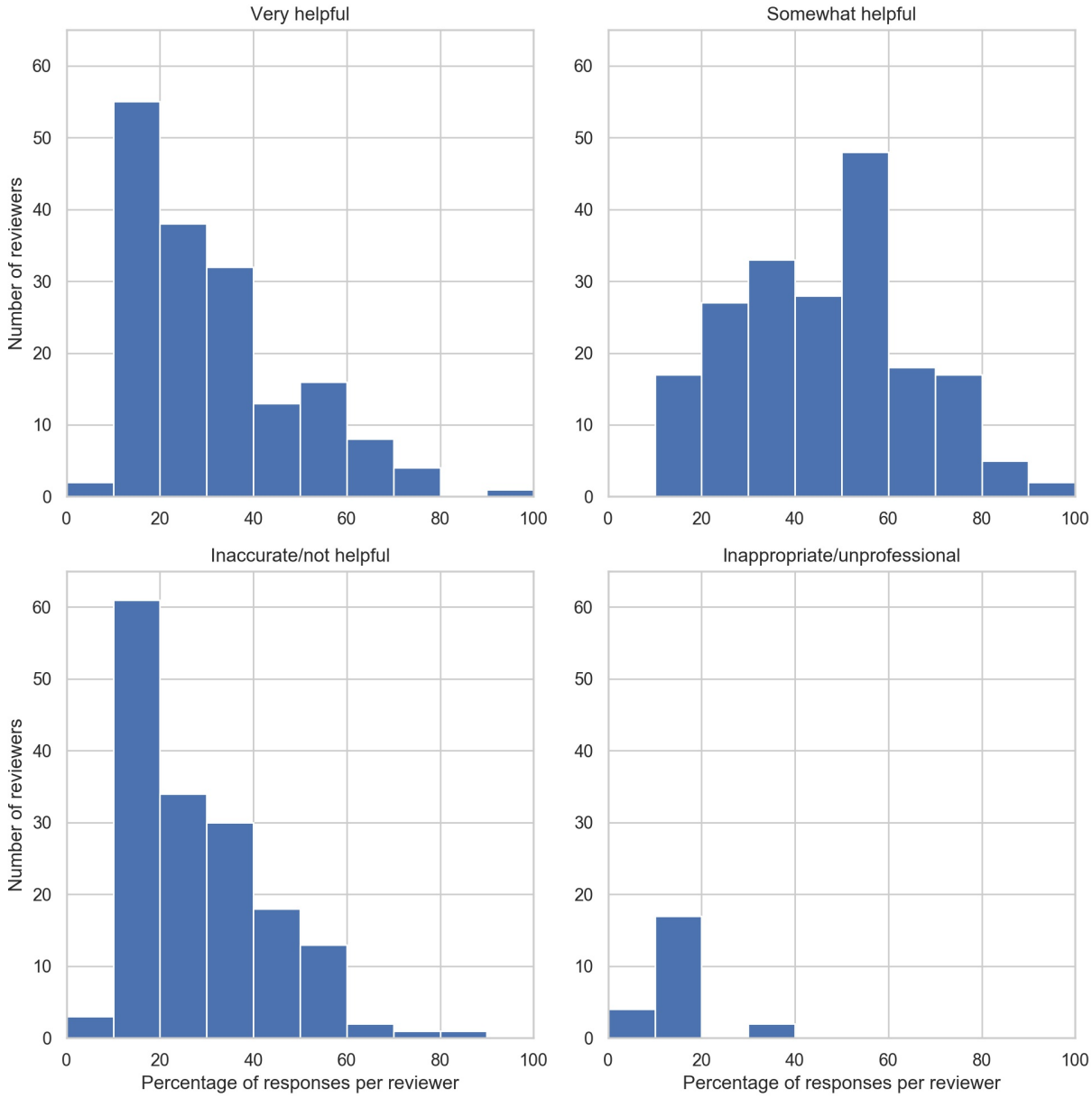

Figure 35: Percentage of comments per reviewer rated very helpful, somewhat helpful, inaccurate/not helpful, or inappropriate/unprofessional by PIs. Results are shown only for reviewers which had 6 or more of their comments rated by PIs.

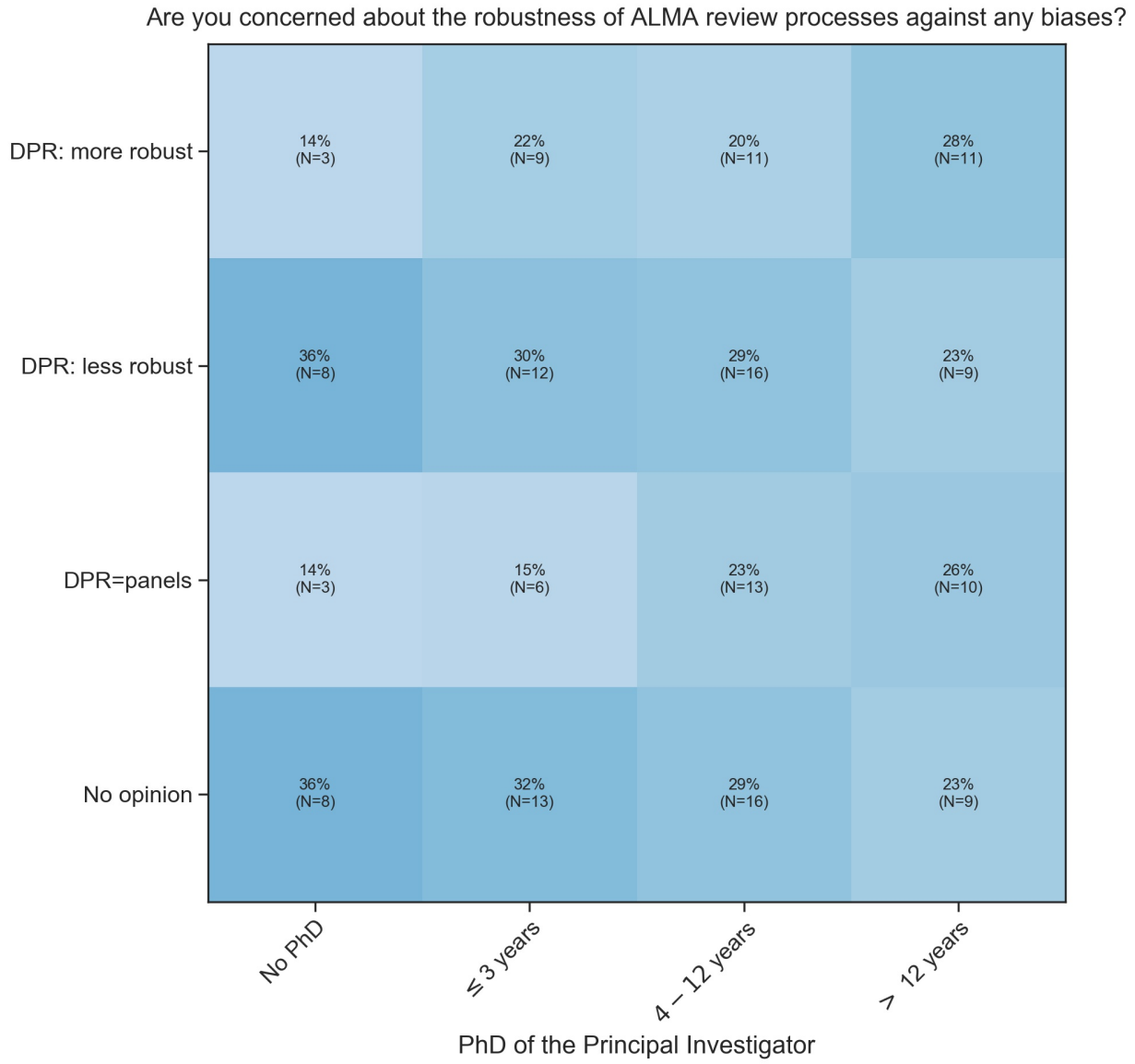

Figure 36: PI concerns about biases in the review process, relative to the main call. The results are normalized for each career level of the PI.

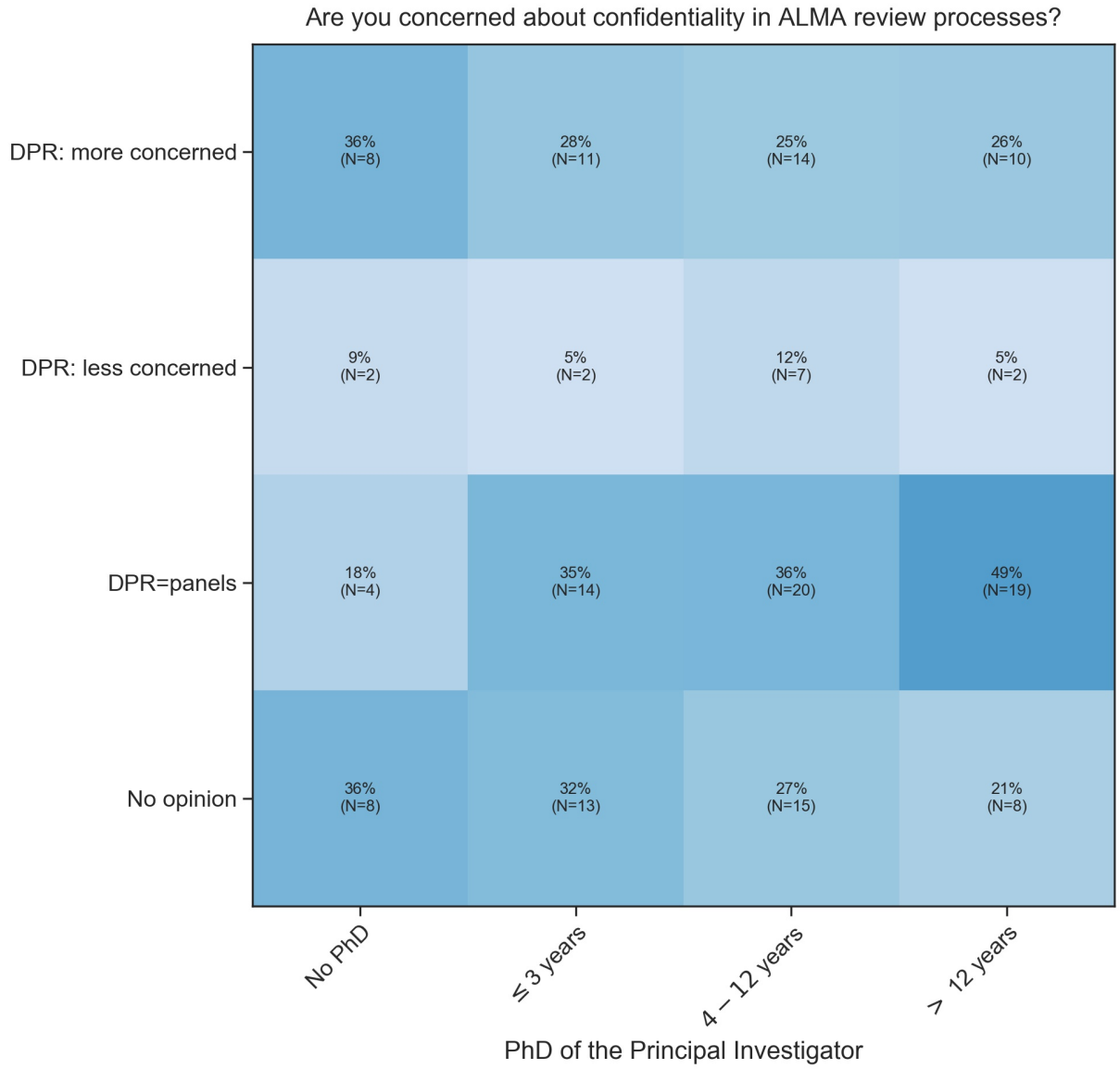

Figure 37: PI concerns about confidentiality in the review process, relative to the main call. The results are normalized for each career level of the PI.

(71%). Less than half of the PIs supported using DPR for medium (41%) and large (13%) proposals, while 20% of PIs did not support using DPR for any proposals regardless of the request. The percentage of PIs that do not support DPR is similar to that in the reviewer survey (27%). The trends are broadly consistent across the career levels of the PIs, although experienced PIs are more supportive of DPR than junior colleagues. The differences with PhD demographics are not statistically significant though.

6 Lessons learned from the Cycle 7 Supplemental Call

This section provides some perspectives on the lessons learned from the supplemental call based the analysis presented in this document, experience from the PHT with the process, and comments from the community. For completeness, we list many of the suggestions that have been proposed but withhold judgement concerning what changes should be implemented pending further discussion.

6.1 Dispersion in the individual ranks

PIs often expressed surprise at the large range of ranks given to their proposal. As shown in Section 3.5, the scatter can be attributed primarily to the different opinions among the individual reviewers, as opposed to the limitation in reviewing 10 proposals. Figure 39 shows that the same issue is present in the main call as well. While the dispersion in the scores decreases after the panel discussions, a significant spread in the scores still exists. Since the range of scores is not given to the PIs in the main call (only the quartile rank, the priority grade, and the consensus reports), the range of opinions is not exposed to the PI. A number of steps could be taken to mitigate the frustration of the PIs in the large dispersion of their ranks.

1. Analogous to the main call, only the quartile rank could be provided to the PIs instead of the individual ranks. However, this would reduce the transparency of the process.

2. Proposals with a large dispersion could be passed to a review panel, which can discuss the discrepant reviews and reach a consensus. As shown in Figure 9, there is no natural break in the dispersions and a somewhat arbitrary cut will need to be made as to which proposals are passed along to such a panel. Given the large dispersions, the final resolution of such a proposal will depend critically on who is selected for the panel.

3. Clearer review criteria may help reduce the dispersion in some cases. For instance, some reviewers weighted heavily if the proposed observations were uniquely suited to the ACA, while for others this did not appear to be a major factor. Even if clear guidelines are written and followed, many proposals will continue to have large dispersions.

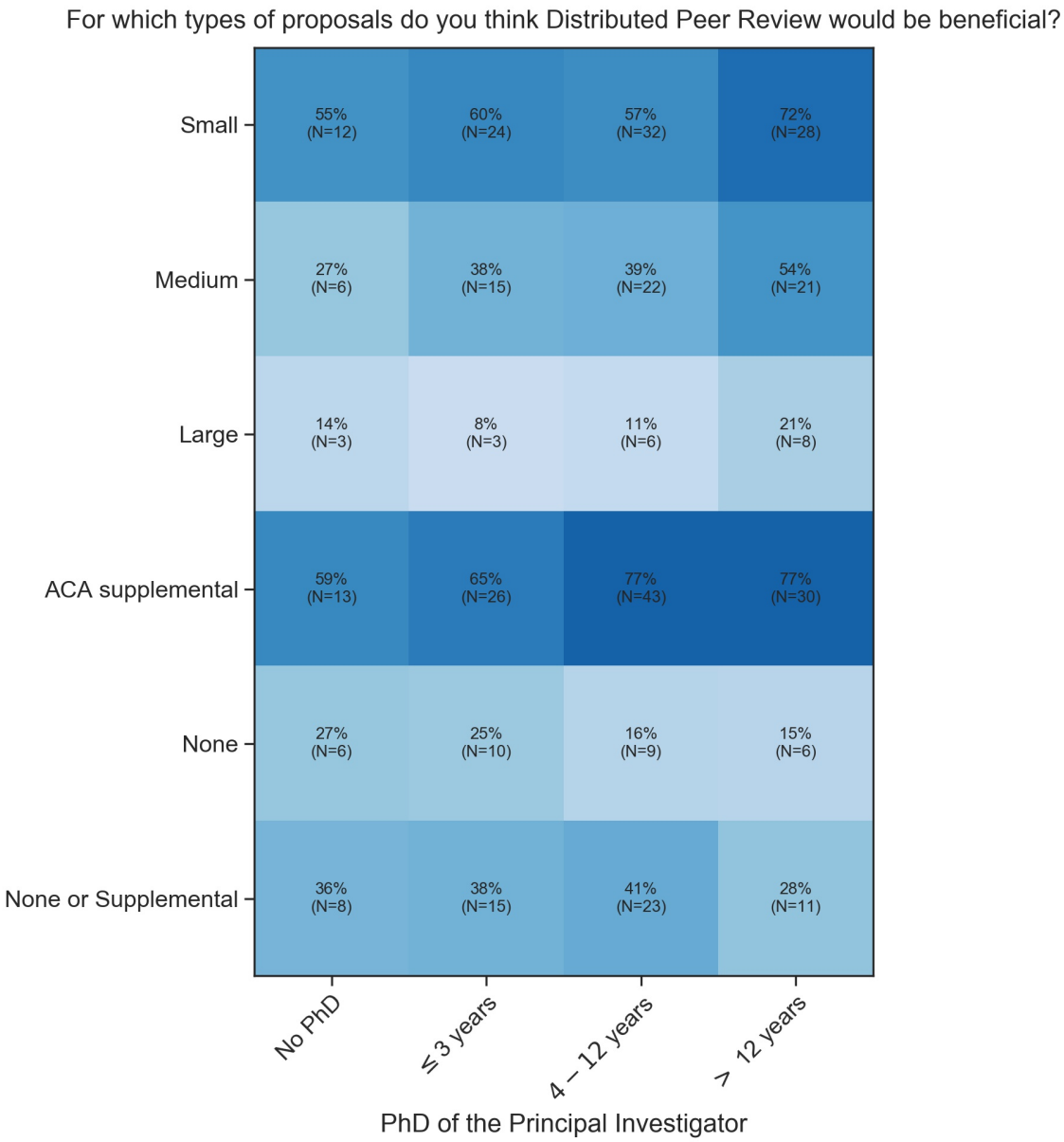

Figure 38: Percentage of PIs that found DPR suitable for different proposal types. The results are normalized by career status of the PI. PIs were able to select multiple options.

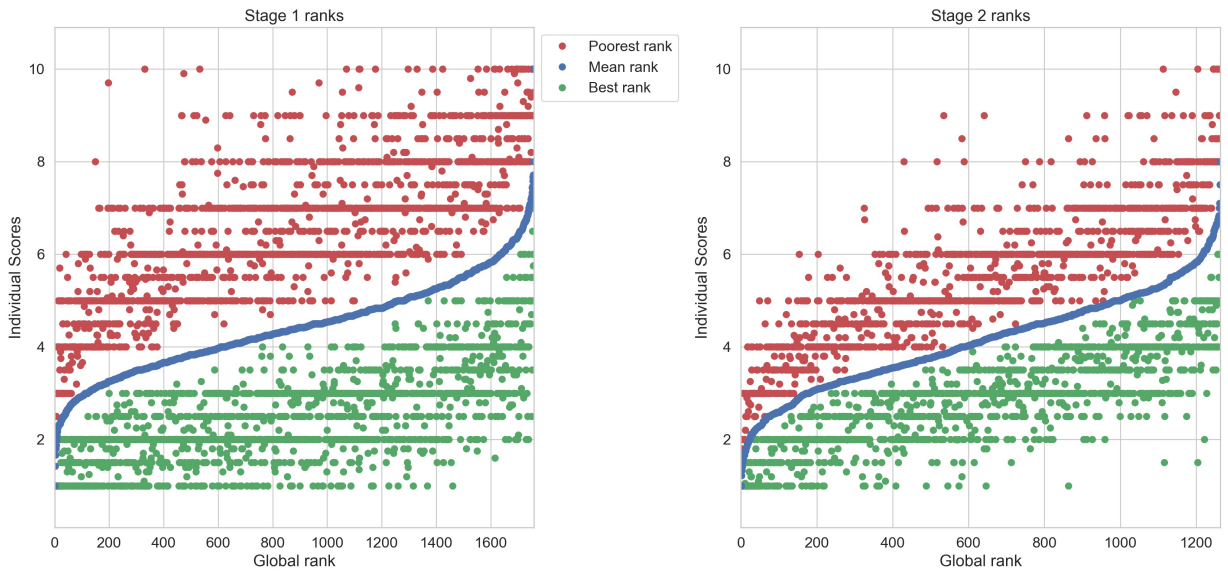

Figure 39: Distribution of mean scores (blue), best score (green), and poorest score (red) for each of the proposals submitted to the Cycle 7 main call. The left panel shows the results for the Stage 1 scores and the right panel for the Stage 2 scores.

6.2 Relative ranks versus absolute scores

A frequent suggestion from reviewers is to use absolute scores that rate the absolute scientific merit of the proposal instead of relative ranks. The main argument is that ranks do not properly differentiate the quality of proposals. Also, reviewers with multiple proposal sets reported having difficulty ranking the proposals from different proposal sets separately, or found that an assignment in one proposal set may have been ranked at the top of the list while only being a middling ranked proposal in another proposal set.

Absolute scores were considered when implementing DPR for the supplemental call. One difficulty is that reviewers will adopt different scoring scales: some reviewers will use the full range of possible scores while others will use a limited range. This can be mitigated by either normalizing the scores to the same mean and dispersion (which partially defeats the purpose of using absolute scores) or just allowing the relatively large number of reviews per proposal to average over such differences.

6.3 Creating the global ranked list

The global ranked list was created through a straight average of the individual ranks. Outlier rejection was explored, both using the median and by removing the high and low scores before averaging. The latter had a minor effect on the overall rankings since the scatter in ranks is nearly uniform across the entire list of proposals. Given the impact was low and there was no a priori reason to reject a particular review, it was decided to use a straight average. Using the median for the global rankings would have a more significant

impact since it generated more ties among rankings; i.e., there is a small number of possible median values since the ranks can only be integer values of 1 through 10. This combined with the tie-breaking criteria can change the order of the rankings appreciably.

If ranks are used in future implementations of DPR, alternative approaches to producing a global ranked list of proposals from the individual ranks should be explored. Reviewers are making a preference choice in the relative rankings, and the Plackett-Luce model (Plackett, 1975; Luce, 1959) presents a formal mathematical approach to convert a set of ranked lists into a global ranking even when each individual ranked list is only for a subset of the full list.

6.4 Improvement in the tools

Based on the reviewer survey, a few improvements in the tool are needed in any future use of DPR, including:

1. A common feature request is for a batch PDF download of the proposals for a given proposal set. The feasibility of such an option will be explored.

2. The auto saving feature needs to be more robust.

3. Erroneous ranks were submitted in a few cases, perhaps related to the autosave feature mentioned above.

4. Upon review of the text submitted by all reviewers, it was apparent that a handful of reviewers likely reversed the ranking scale. This was clearly frustrating to a few PIs when the comments did not match the rankings, and it reduces the robustness of the final rankings. While the vast majority of reviewers understood the ranking system, the directions at the top of the page in a contrasting color were not sufficient to ensure that all reviewers read them. This situation can be further mitigated by, for example, using a drop-down menu of selectable ranks in the reviewer tool and state in the menu that #1 is the best proposal and #10 is the worst proposal in that set.

5. Some reviewers suggested that the ranking system could be more automatic when tweaking the individual ranks; e.g., when a rank of a proposal is improved, the other ranks could be decreased automatically.

6. Some reviewers requested to have the investigator teams listed directly in the tools to make it more convenient to identify conflicts of interest. With the dual anonymous system coming in Cycle 8, this will not be necessary.

7. The display format of the reviewer comments in SnooPI needs to be improved to make it easier to read long comments that do not have any line breaks.

6.5 Evaluating the appropriateness of the comments

In the supplemental call, the PHT read all the reviewer comments to identify inappropriate remarks and, in the case of one reviewer, some edits were made. The problems encountered were generally with the tone of the reviews, not words that could be found with an automated search of the text. If DPR would be implemented in a main call, there will be ∼18,000 individual comments to review. It would be feasible to review this number of comments if JAO staff within DSO assisted the PHT.

While the PHT reviewed the comments in the supplemental call, 29 comments were still flagged as unprofessional by the PIs. For context, the 29 comments are shown in their entirety in Appendix C. Only three of these comments were flagged by the PHT in their review, but the PHT did not feel it was warranted to edit the comments. This clearly indicates there is considerable subjectivity in determining which comments are inappropriate or unprofessional. Clear guidelines would need to be established on what type of comments should be flagged and modified.

Another issue encountered was that some reviewers referenced the proposal teams and made few comments on the proposal itself. This will be mitigated in Cycle 8 with the implementation of dual-anonymous review.

6.6 Improving the proposal assignment process

Proposals were assigned to reviewers under the assumption that a reviewer is an expert in the category, and especially the keyword(s), specified on the submitted proposal. This does not appear to be universally true, especially given the broad topics that are present within a category (see Section 4.3). While in practice, there is little correlation between the PI rating of the helpfulness of the comment and the reviewer expertise (see Figure 34), some reviewers were clearly frustrated reviewing proposals from other categories. There is also some evidence of bias in the ranks for proposals where the reviewer is a non-expert (see Section 3.6). Possibilities to improve the proposal assignment include:

1. Ask reviewers to specify the keywords corresponding to their expertise when the proposal is submitted. Especially in cases where reviewers specify multiple keywords and categories, it can provide considerably more flexibility in the proposal assignments.

2. Use a matching algorithm that looks at the reviewer's publication record in the literature and matches words in it to the proposal text up for review, which has been pioneered at ESO (Patat et al., 2019).

3. Use machine learning to identify similar proposals based on a reviewer's submitted proposal text (e.g., the PACMan tool at HST; Strolger et al., 2017).

A general problem for DPR and the panel reviews is that some communities are small and generate a few proposals; e.g., the Sun, the solar system. Selecting reviewers within the same category is not sufficient to identify experts for DPR (especially given the large

number of conflicts that come with a small community), and there are too few proposals to warrant a separate panel in the traditional review process. In these cases, it is difficult to assign only expert reviewers.

6.7 Confidentiality and conflicts of interest

One complication of the proposal assignment process in Cycle 7 was handling conflicts of interest. In one case, a PI and their student went through four rounds of assignments before settling on acceptable proposal sets without conflicts, due to the fact that the PI is on several large science teams. This issue should be mitigated in Cycle 8 by the dual anonymous review process. The PHT has also outlined an algorithm to identify more objectively close collaborators based on PIs and co-Is who are frequently investigators on the sme ALMA proposals.

More generally, PIs are more concerned about confidentiality in DPR than in review panels. The probability that a given proposal will be used unethically depends on the number of reviewers and the number of proposals that they review. In this regard, the level of confidentiality concerns between DPR and panels reviews should be comparable unless it is thought a great fraction of DPR reviewers behave unethically. The specific concerns from PIs were not revealed in the surveys, but some comments alluded to competitors having access to their proposals. A possible mitigation strategy is to allow PIs the option to specify in the Observing Tool a limited number of people who are not allowed to review the proposal because they are viewed as competitors.

6.8 Improving the quality of the reviewer feedback

Another frequent concern expressed by PIs was the quality of the feedback, which is a persistent issue in the main call as well. The overall ratings of the quality of the feedback in DPR is similar to the review panels from the main call (see Section 5.2), indicating that both processes face similar challenges in improving the feedback. One suggestion from the reviewers is to provide more specific examples of both good and inappropriate feedback in the documentation.

6.9 Providing feedback to the reviewers

One element missing from the current implementation of DPR is a feedback process that would inform reviewers how well they completed their reviews and if they provided useful feedback to the PIs. Possible options to provide feedback to the reviewers include:

1. Implement a Stage 2 process
   Reviewers would access the (anonymous) comments of each of the other reviewers assigned to the same proposals after all comments are submitted. Reviewers would then review the feedback and optionally be able to change their ranks and reviews. In

this way, inexperienced reviewers and experienced reviewers alike can compare their thoughts and determine if they missed something substantial. The downside is that reviewers would incur a non-trivial amount of additional work to evaluate 90 other proposal comments and then revise their own reviews and ranks.

2. Reviewer education
   Reviewers could receive copies of the comments made by other reviewers on the proposals that they were assigned, but will not be allowed to change their own comments or ranks. The comments would be listed anonymously (as they are for PIs), and they would be sent at the same time as the PIs are sent their comments. In this manner, inexperienced reviewers and experienced reviewers alike can learn what others thought about each proposal. This could be an opt-in feature of the process for the reviewers.

3. PI review evaluation
   It could be a standard feature for PIs to evaluate the helpfulness of each comment and have that information send to reviewers. Similarly, if reviewers are given access to the comments in a Stage 2 process as indicated above, reviewers could also be asked to rate the helpfulness of comments and even identify inappropriate reviews. However, while the response rate of PIs to the supplemental call was strong, it would seem unlikely such a high level of voluntary participation would be maintained year after year.

6.10 High-risk / high-reward proposals

A frequently raised concern is that reviewers are inherently conservative and tend to rate proposals with a guaranteed result over high-risk / high-reward proposals. Such proposals may include observations that are challenging observationally (e.g., high frequency), require large amounts of observing time, or particularly innovative ideas with high scientific payoff if successful. While DPR in the supplemental call did not seem to hinder approval of proposals in the first two instances (see Section 3.7), DPR does not allow for discussion and debate of proposals that promote innovative ideas. One option is to allow reviewers to flag proposals that they view as high-risk / high-reward. These proposals could receive special consideration by the ALMA Director if they do not receive time in the proposal call.

6.11 Risks in using DPR in the main call

There are two main operational risks in extending the use of DPR in the main call. First, the main call will have about 5× the number of concurrent users. As shown in in Section 3.4, most of the submissions of the reviews and ranks will occur close to the review deadline. Automated stress tests will need to be run to ensure the system can handle the increased load, similar to the tests run for the OT submission server. The second risk is that DPR will likely be run in parallel with a panel review process for the larger programs. This will challenge the PHT staff to operate both review processes simultaneously. Further improving both the documentation and the robustness of the tools will help reduce the volume of questions asked of the PHT.

References

Carpenter, J. M. 2019, arXiv e-prints, arXiv:1908.09639. https://arxiv.org/abs/1908.09639

Luce, R. D. 1959, Individual Choice Behavior: A Theoretical Analysus (New York: Wiley)

Patat, F., Kerzendorf, W., Bordelon, D., Van de Ven, G., & Pritchard, T. 2019, The Messenger, 177, 3, doi: 10.18727/0722-6691/5147

Plackett, R. L. 1975, Journal of the Royal Statistical Society, 24, 193, doi: 10.2307/2346567

Strolger, L.-G., Porter, S., Lagerstrom, J., et al. 2017, AJ, 153, 181, doi: 10.3847/1538-3881/aa6112

A Reviewer survey results

This appendix lists the questions and responses for the Cycle 7 supplemental call reviewer survey. Many of the survey questions were adopted from a similar ESO survey in their pilot review using DPR (Patat et al., 2019).

A.1 How helpful were the guidelines on writing comments to the PI?

- The guidelines were clear and appropriate and helped me write the comments. (166 responses)
- The guidelines were mostly clear, but I did not fully understand what I should include in the comments. (33 responses)
- It was unclear what should be included in the comments to the PIs. (5 responses)
- I did not read the guidelines. (9 responses)

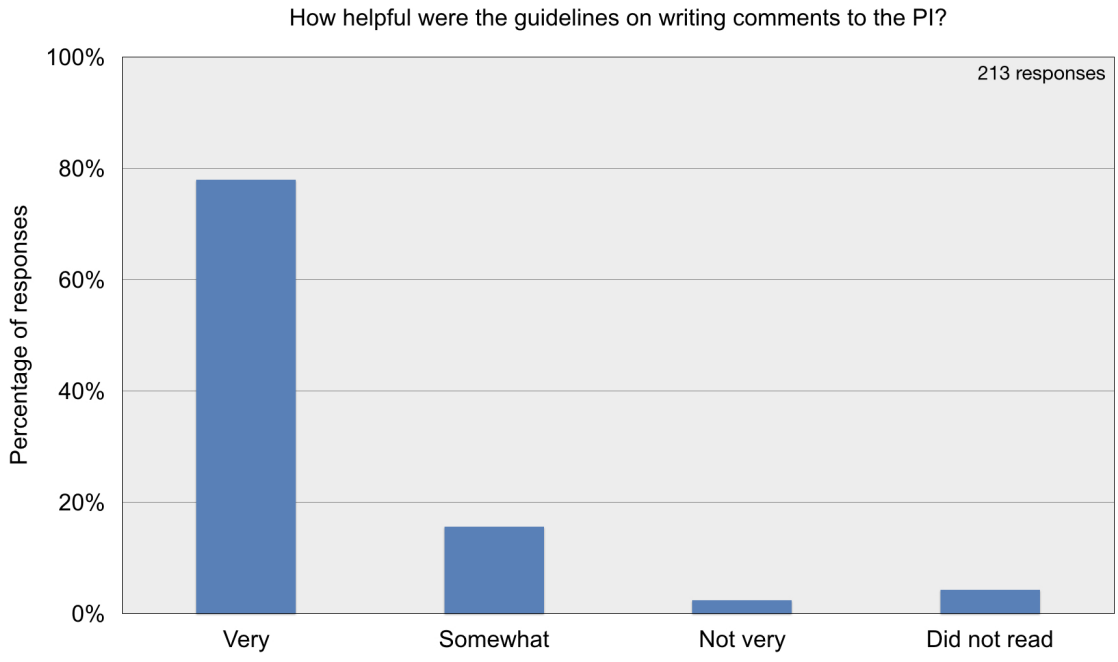

Figure 40

A.2 How relevant were the review criteria to evaluate the proposals?

- Fully relevant; the criteria were clear and easy to apply. (94 responses)
- Mostly relevant; the criteria were clear but not easy to apply. (94 responses) item Somewhat relevant; the criteria can be clarified but were applicable. (21 responses)
- Not relevant; the criteria should be changed. (1 responses)
- Did not reply. (3 responses)

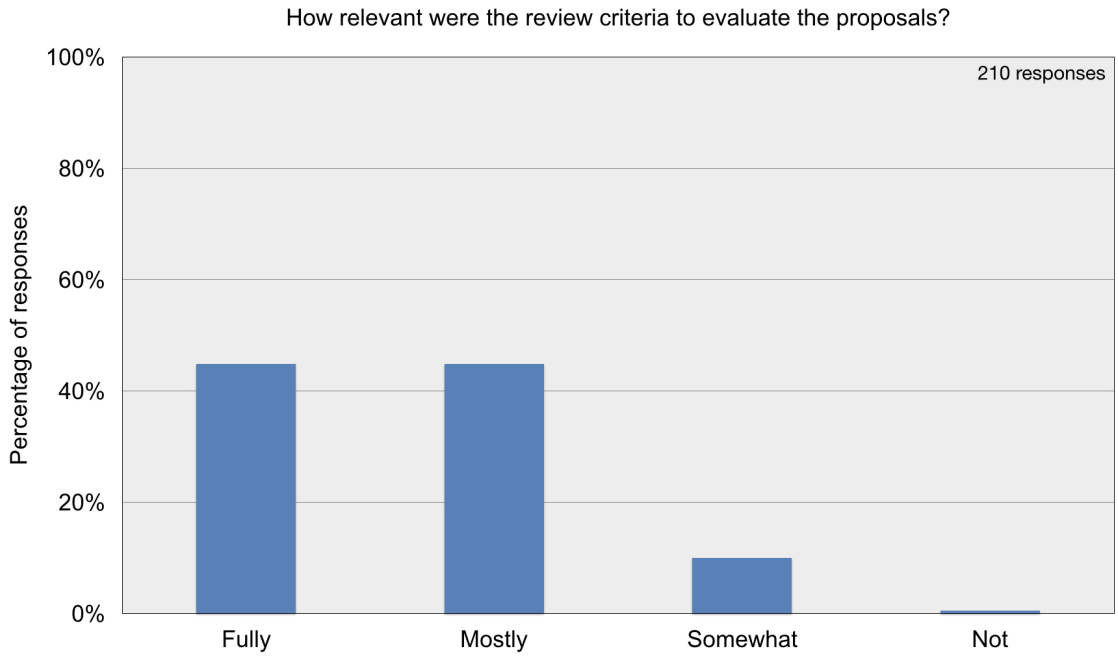

Figure 41

A.3 How much time did you spend, on average, reviewing each proposal (including writing comments)?

- Less than 15 minutes. (5 responses)
- Between 15 and 30 minutes. (48 responses)
- Between 30 and 45 minutes. (85 responses)
- Over 45 minutes. (75 responses)

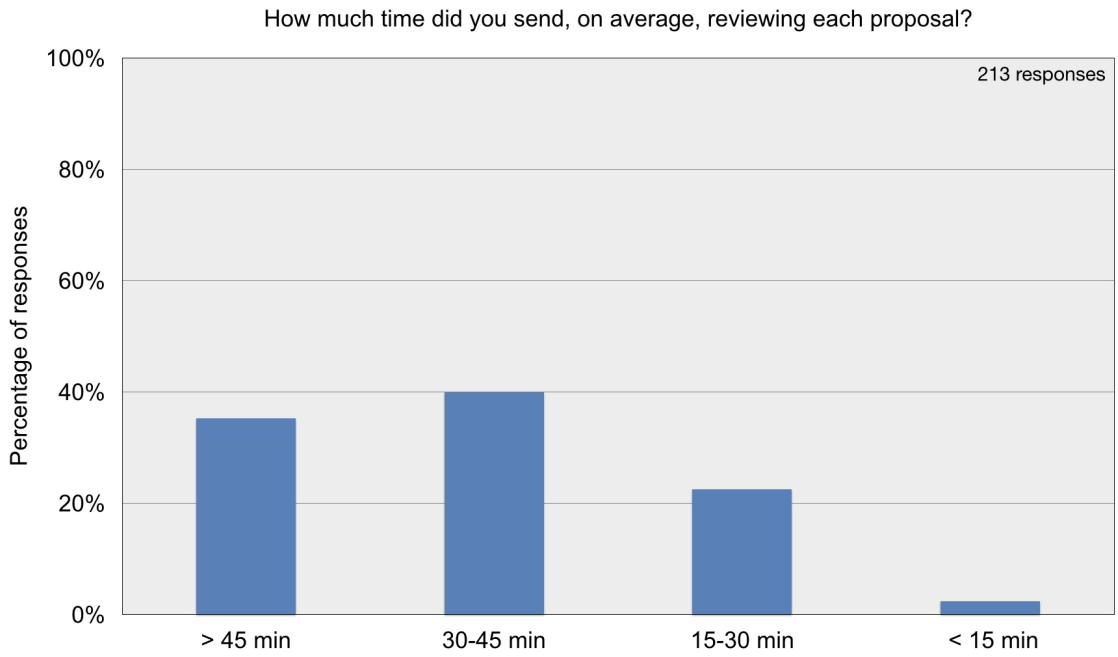

Figure 42

A.4 How satisfactorily were you able to evaluate the proposals for which you were a non-expert?

- Fully; I could evaluate well and fairly as a non-expert. (17 responses)
- Mostly; I did not appreciate some details but I was still able to evaluate fairly. (109 responses)
- Somewhat; I struggled and might not always have been able to evaluate fairly. (66 responses)
- Not satisfactorily; I might have unintentionally provided an unfair evaluation. (18 responses)
- Not applicable; I consider myself to be an expert on all proposals which were assigned to me. (3 responses)

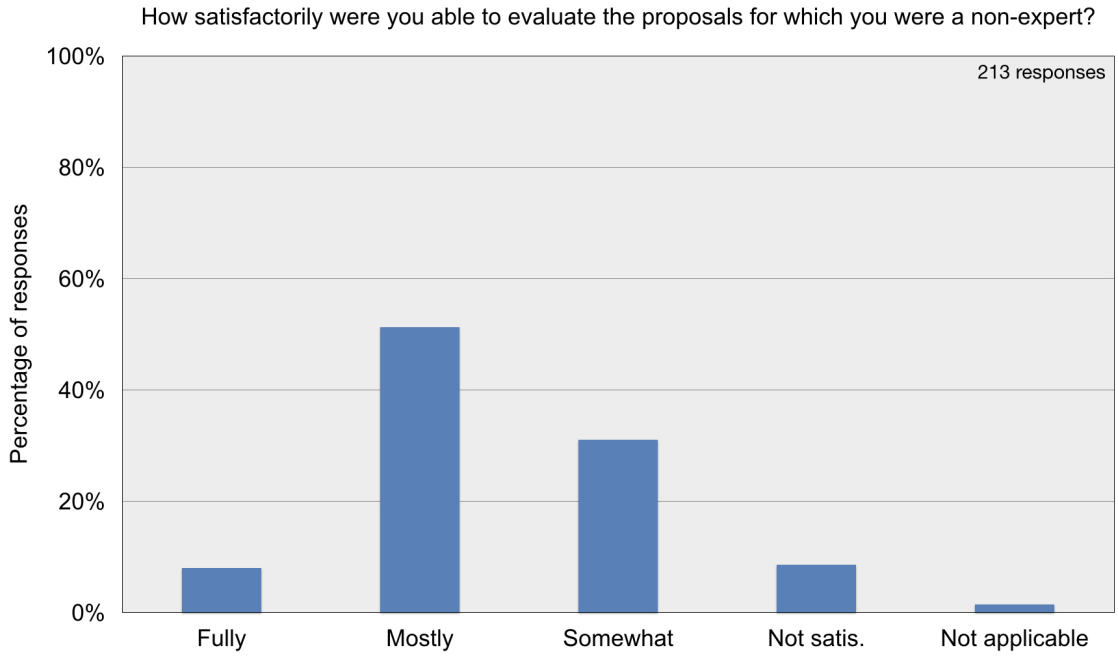

Figure 43

A.5 How easy was it to navigate the interface to review the proposals?

- Easy; no problem in usage. (172 responses)
- Mostly easy; some improvements possible. (37 responses)
- Somewhat difficult; some improvements needed. (4 responses)
- Difficult; several problems in usage. (0 responses)

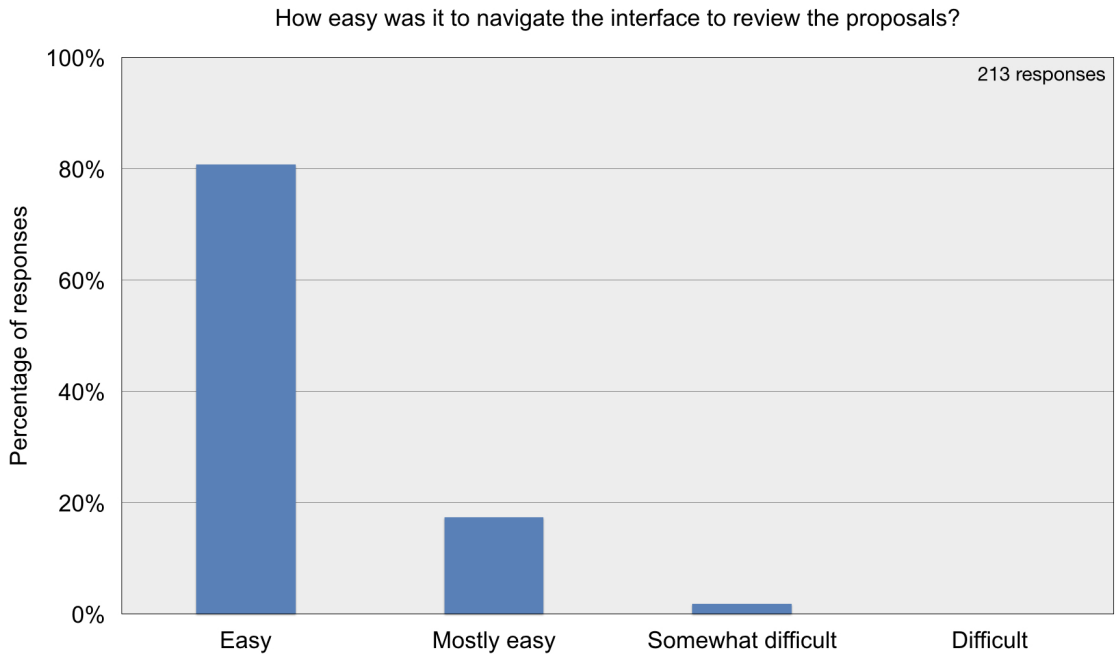

Figure 44

A.6 Would you submit ALMA proposals in future cycles if you were required to review 10 proposals for every proposal submitted?

- Yes; I would submit just as many proposals as I do now. (126 responses)
- Yes; I will submit proposals, but perhaps not as many as I do now. (76 responses)
- No; I would not submit any proposals. (2 responses)
- I don't know. (9 responses)

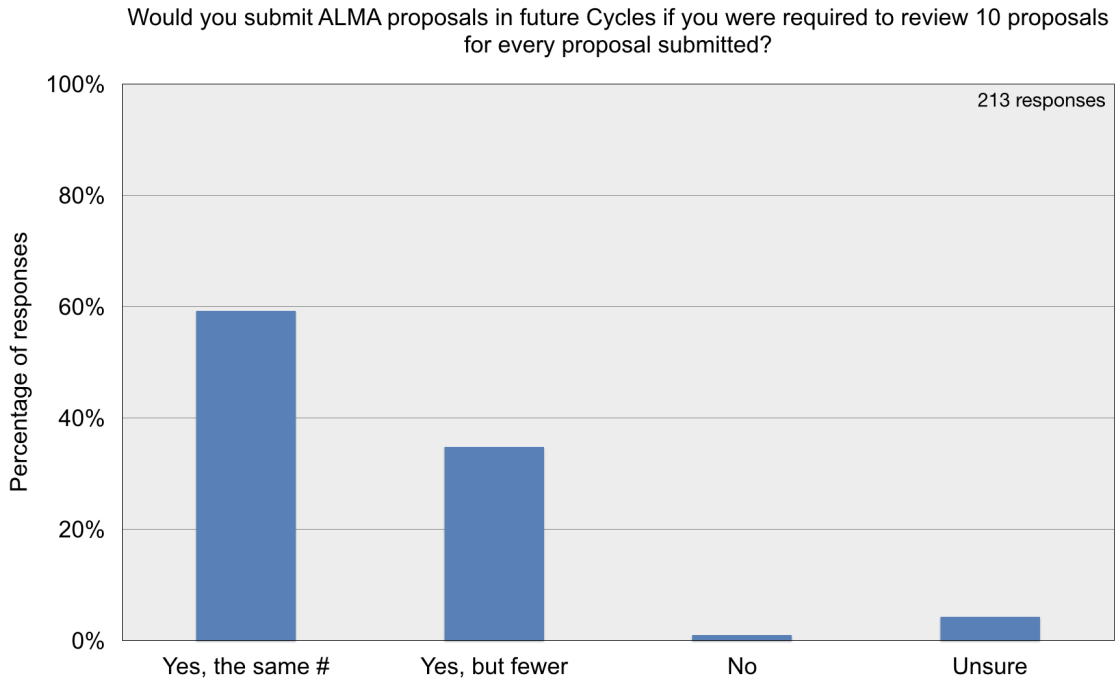

Figure 45

A.7 How many years has it been since you obtained your PhD?

- I do not have a PhD yet. (23 responses)
- Three years or less. (50 responses)
- Between 4 and 12 years. (75 responses)
- More than 12 years. (65 responses)

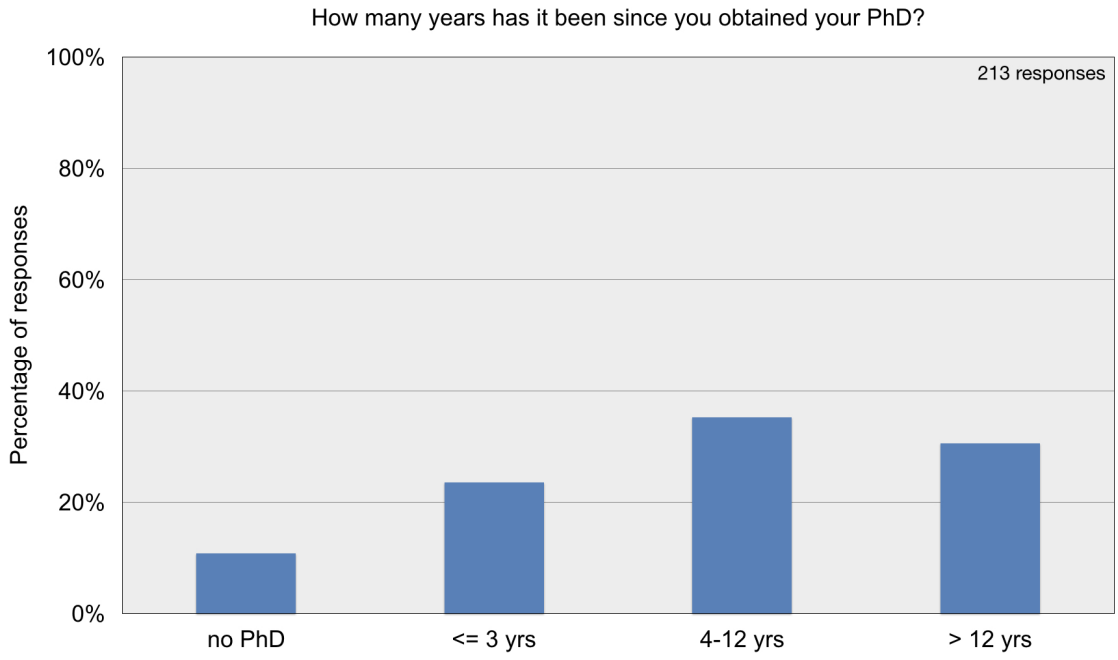

Figure 46

A.8 Have you ever participated in the ALMA review panels as a Science Assessor or Chair for the main proposal call?

- Yes, I have been an ALMA Science Assessor or Chair at a panel review meeting. (39 responses)
- No, I have never been an ALMA Science Assessor or Chair at a panel review meeting.)174 responses)

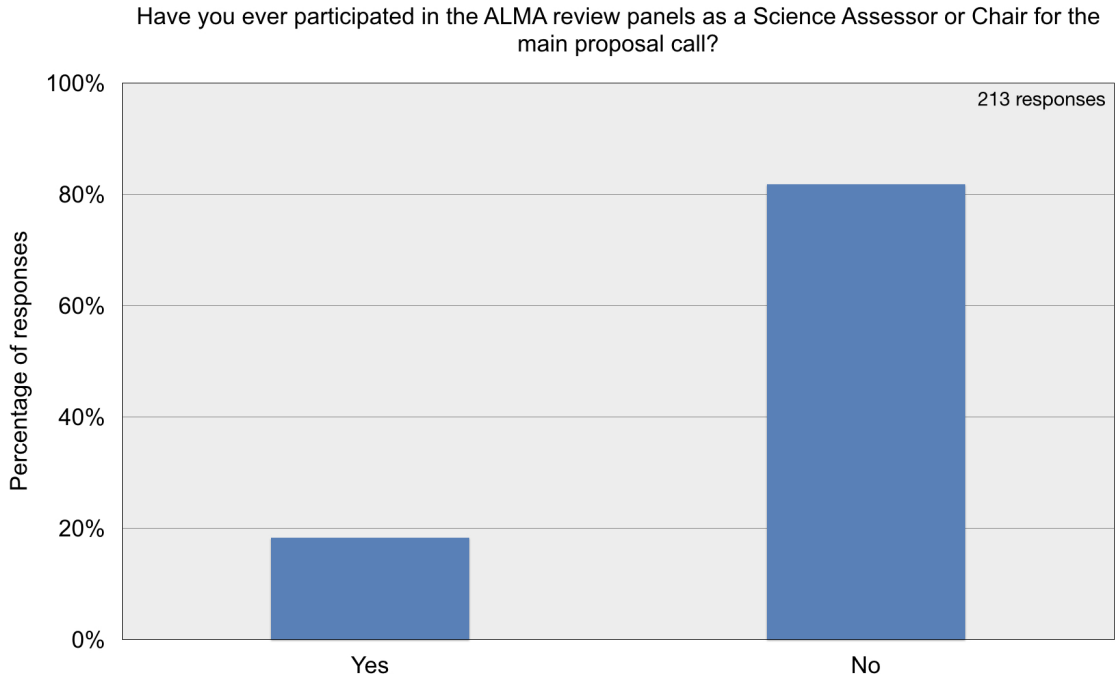

Figure 47

A.9 Proposals submitted to the Cycle 7 Main Call for Proposals were reviewed by one of 25 review panels, while the Cycle 7 Supplemental Call used distributed peer review, requiring each proposal team to designate a person to review 10 proposals. ALMA is considering using distributed peer review in an upcoming Main Call for regular proposals, while continuing to have a face-to-face review panel for Large Programs. Do you think distributed peer review would be appropriate for a Main Call?

- Yes, I think that distributed peer review is a viable review process for the Main Call. (77 responses)
- No, I think ALMA should continue to use review panels for the Main Call. (57 responses)
- I think both review processes are equally effective. (20 responses)
- I am unsure. (59 responses)

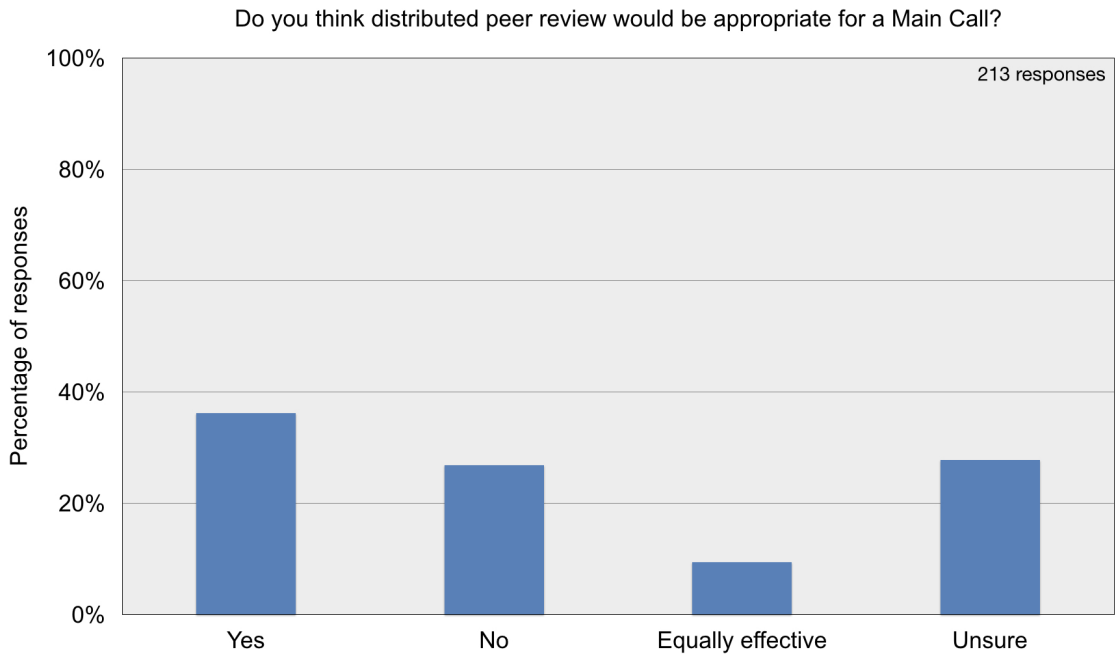

Figure 48

A.10 How extensively did you consult with the mentor on the science evaluation? [If you do not have a PhD.]

- I consulted with the mentor on at least half of the proposals. (15 responses)
- I consulted with the mentor on less than half of the proposals. (5 responses)
- I did not consult with the mentor on any proposals. (3 responses)

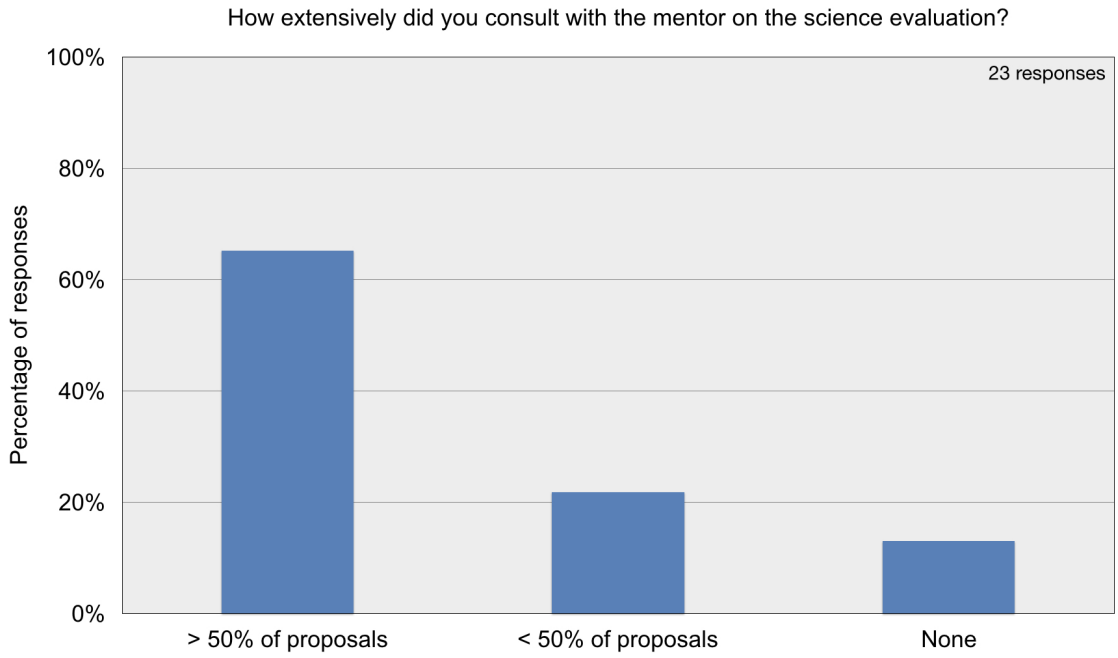

Figure 49

A.11 How extensively did you consult with the mentor on writing the comments to the PI? [If you do not have a PhD.]

- The mentor helped me with the comments on at least half of the proposals. (12 responses)
- The mentor helped me with the comments on less than half of the proposals. (7 responses)
- The mentor did not help me with any comments. (4 responses)

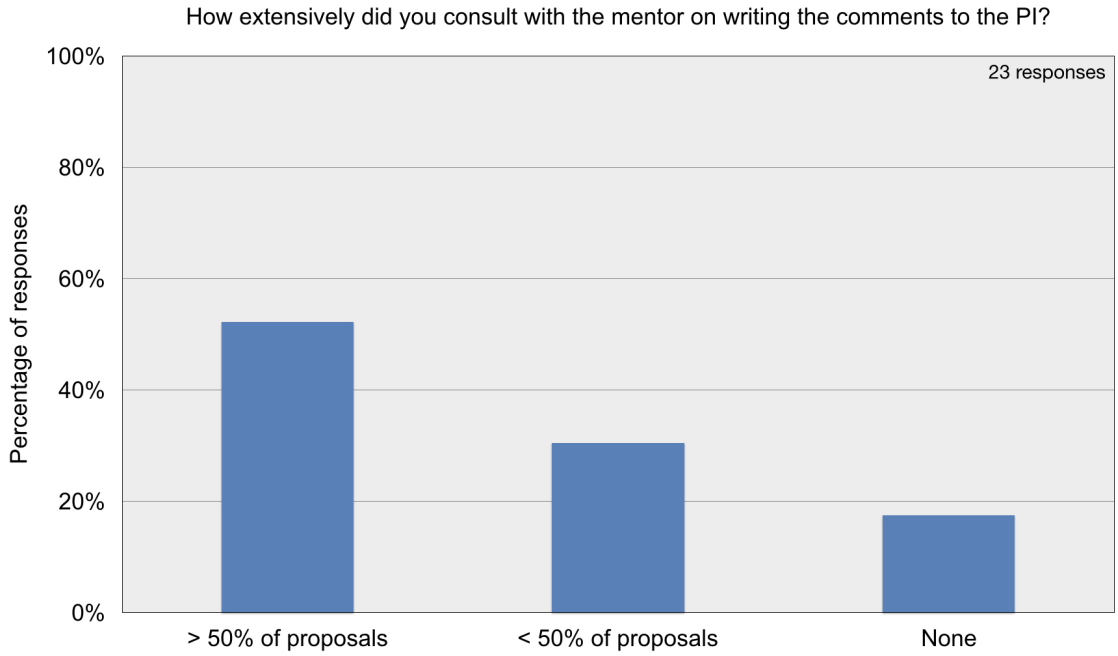

Figure 50

A.12 Please rate your level of expertise on each of your review assignments, using the text of your reviews for reference.

- This is my field of expertise. (813 responses)
- I have some general knowledge of this field. (1099 responses)
- I have little or no knowledge of this field. (362 responses)

Figure 51

B PI survey results

This appendix lists the questions and responses for the Cycle 7 supplemental call PI survey. Many of the survey questions were adopted from a similar ESO survey in their pilot review using DPR (Patat et al., 2019).

B.1 Are the individual comments on your proposal clear and understandable?

- Fully: The comments are clear, whether I agree with the comments scientifically or not. (45 responses)
- Mostly: Most of the comments are clear. (89)
- Somewhat: Some of the comments are clear. (20 responses)
- No: Few, if any, of the comments are clear. (4 responses)

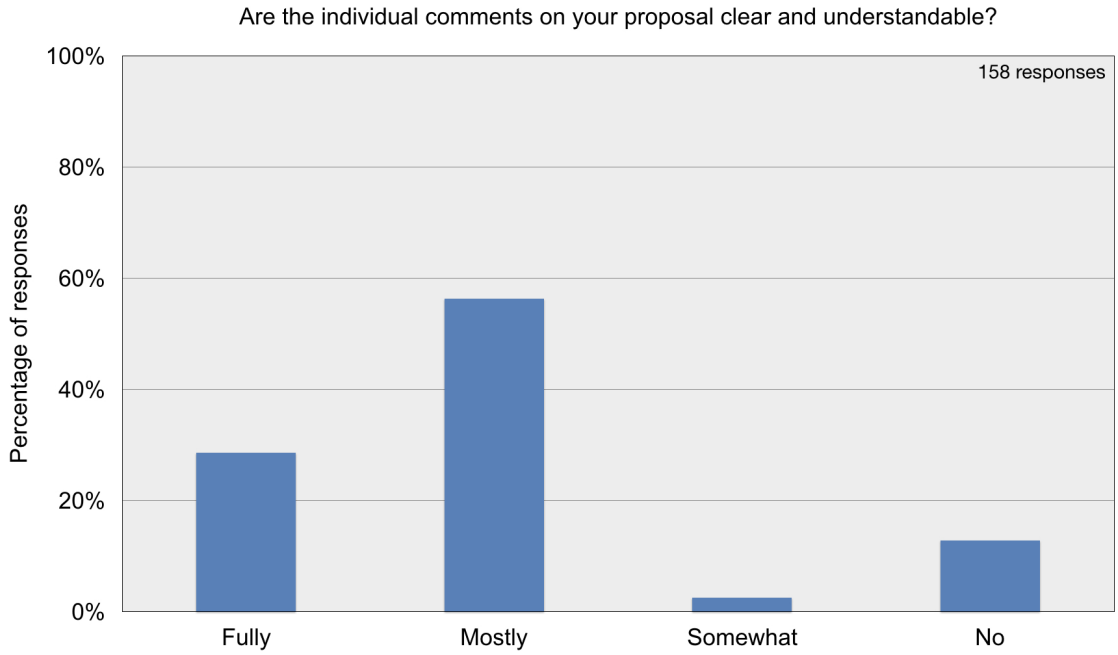

Figure 52

B.2 Are the comments scientifically accurate?

- Fully: The comments are scientifically accurate. (23 responses)
- Mostly: Most of the comments are scientifically accurate. (86 responses)
- Somewhat: Some of the comments are scientifically accurate. (46 responses)
- No: Few, if any, of the comments are scientifically accurate. (2 responses)

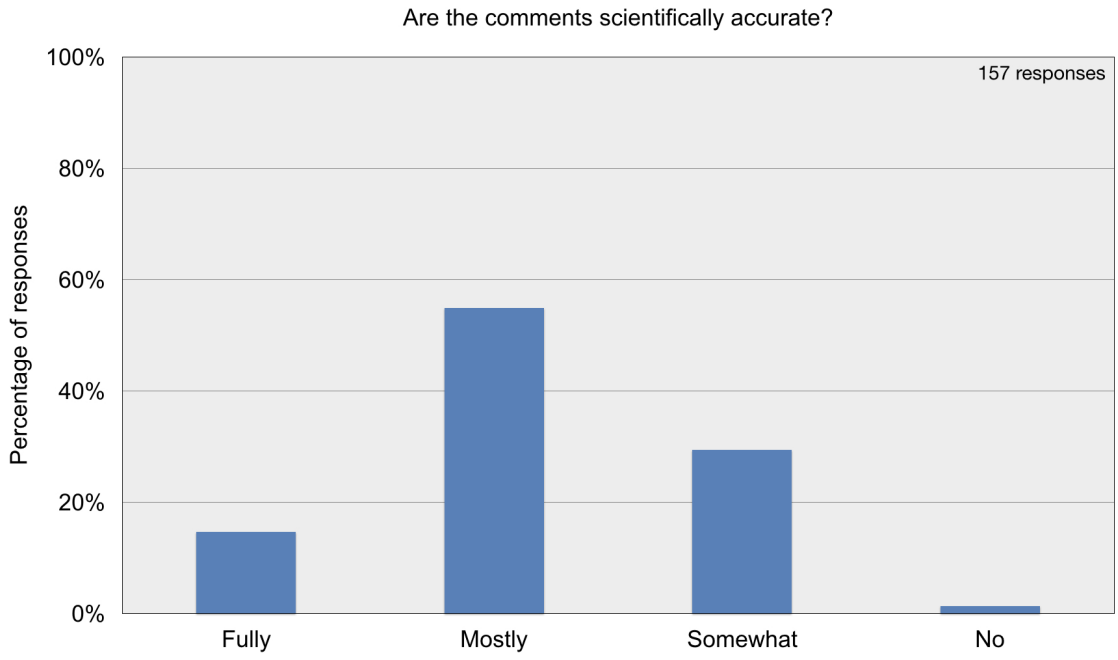

Figure 53

B.3 Will the comments help you to improve future ALMA proposals?

- Fully: The comments will allow me to improve future proposals. (27 responses)
- Mostly: Most of the comments will allow me to improve future proposals. (57 responses)
- Somewhat: Some of the comments might help me strengthen future proposals. (65 responses)
- No: The comments will not help me improve future proposals. (8 responses)

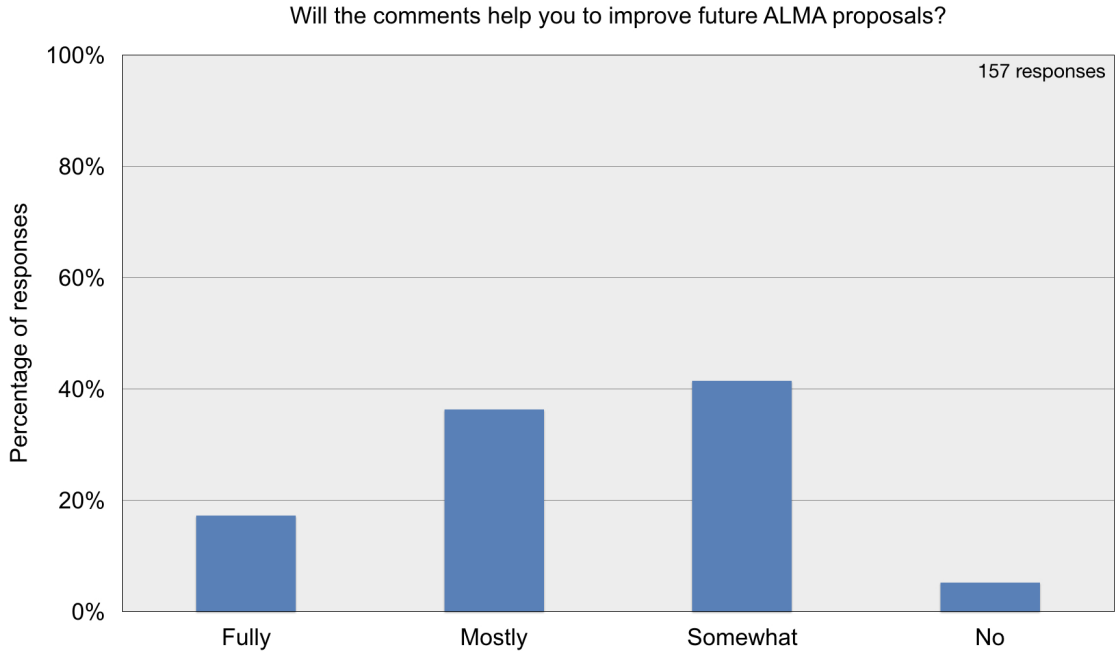

Figure 54

B.4 Were the comments written in a respectful and professional manner?

- Yes: All of the comments were respectful even if I do not agree with the comments scientifically. (128 responses)
- Somewhat: Most of the comments were respectful. (26 responses)
- No: Many of the comments contain unprofessional remarks. (2 responses)

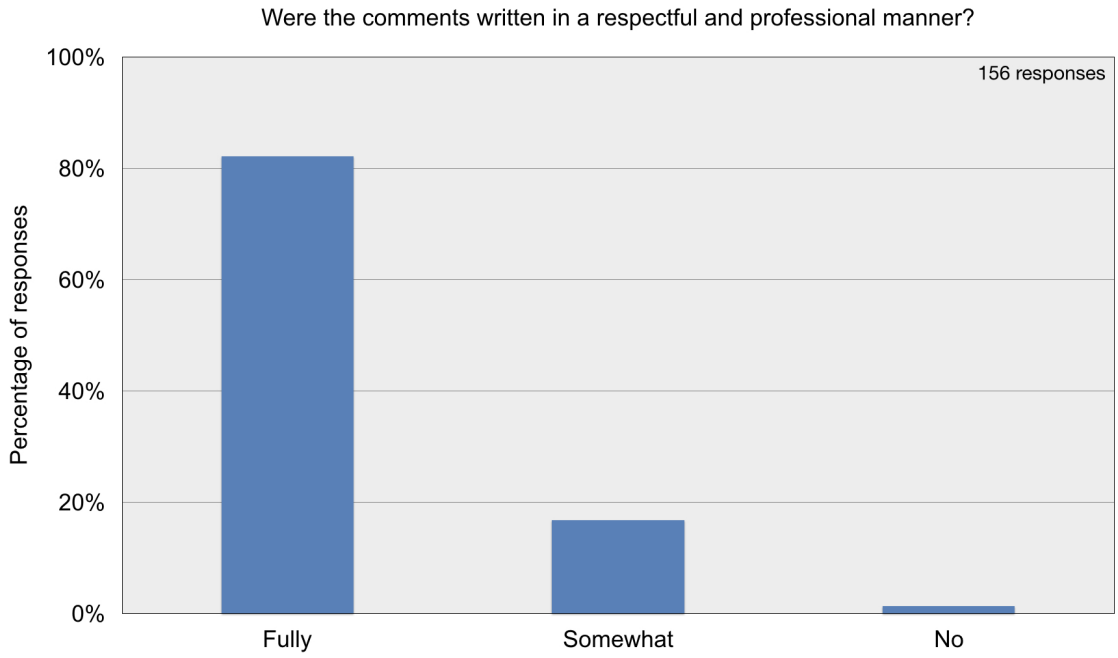

Figure 55

B.5 If you have ever submitted a proposal to an ALMA Main Call, how do you rate the general quality of the comments you have received in the Supplemental Call versus consensus reports you have received in a Main Call?

- The individual comments in the Supplemental Call were better overall; the consensus reports in the Main Call are not as helpful. (48 responses)
- The consensus reports in the Main Call are of better quality than the individual comments in the Supplemental Call. (55 responses)
- The comments from the Main Call and Supplemental Call are of similar quality. (40 responses)
- Not applicable: I have never submitted a proposal to the Main Call as a PI. (14 responses)

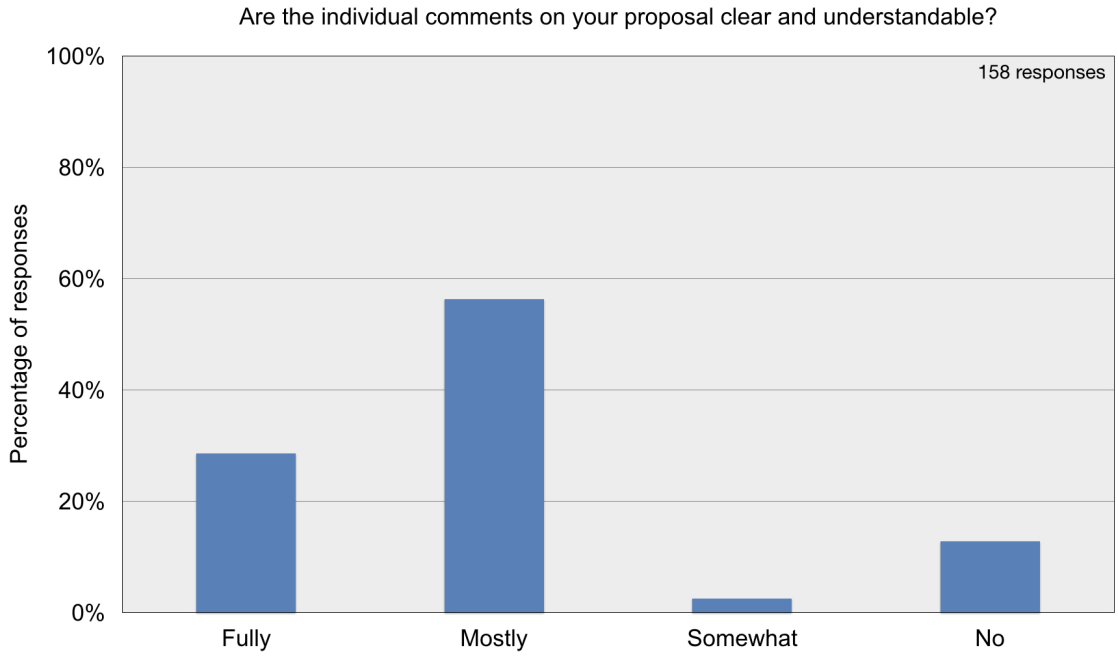

Figure 56

B.6 For which types of proposals do you think Distributed Peer Review would be beneficial? (check all that apply)

- Small proposals, with requested 12-m array time less than 25 hours. (94 responses)
- Medium proposals, with requested 12-m array time between 25 and 50 h. (63 responses)
- Large proposals, with requested 12-m array time greater than 50 h. (20 responses)
- ACA standalone Supplemental Call. (111 responses)
- None of the above. (31 responses)

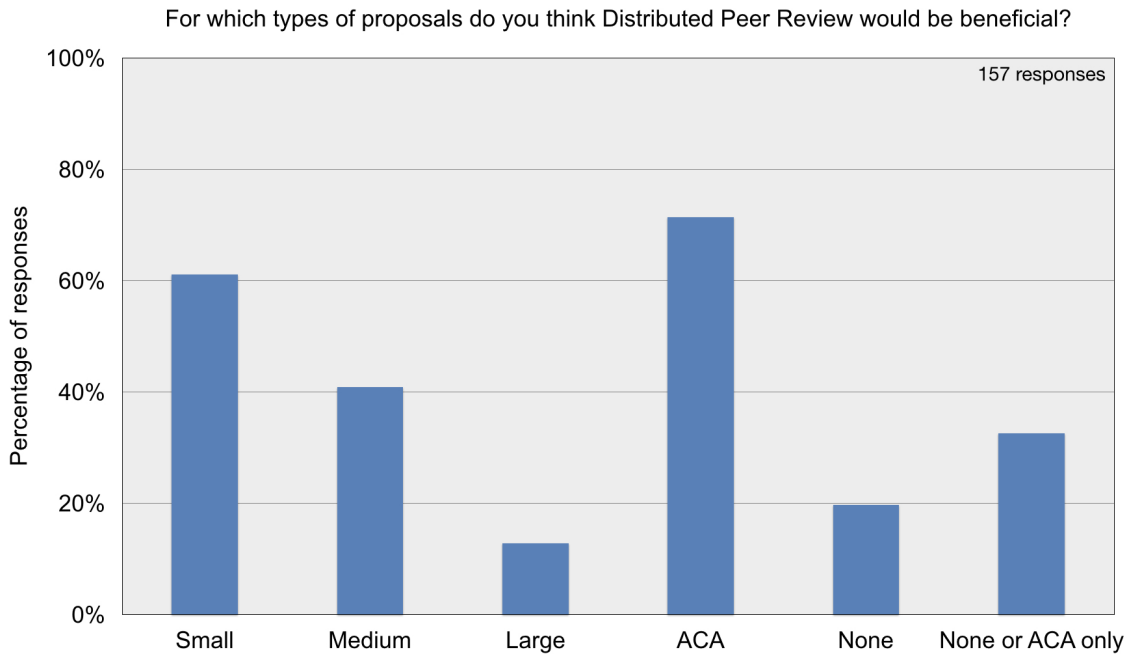

Figure 57

B.7 Are you concerned about confidentiality in ALMA review processes?

- I am neither more nor less concerned about confidentiality issues in the Distributed Peer Review process compared to the Panel Review process. (57 responses)
- I am more concerned about confidentiality in the Distributed Peer Review process. (43 responses)
- I am less concerned about confidentiality in the Distributed Peer Review process. (13 responses)
- I have no strong opinion on this point. (44 responses)

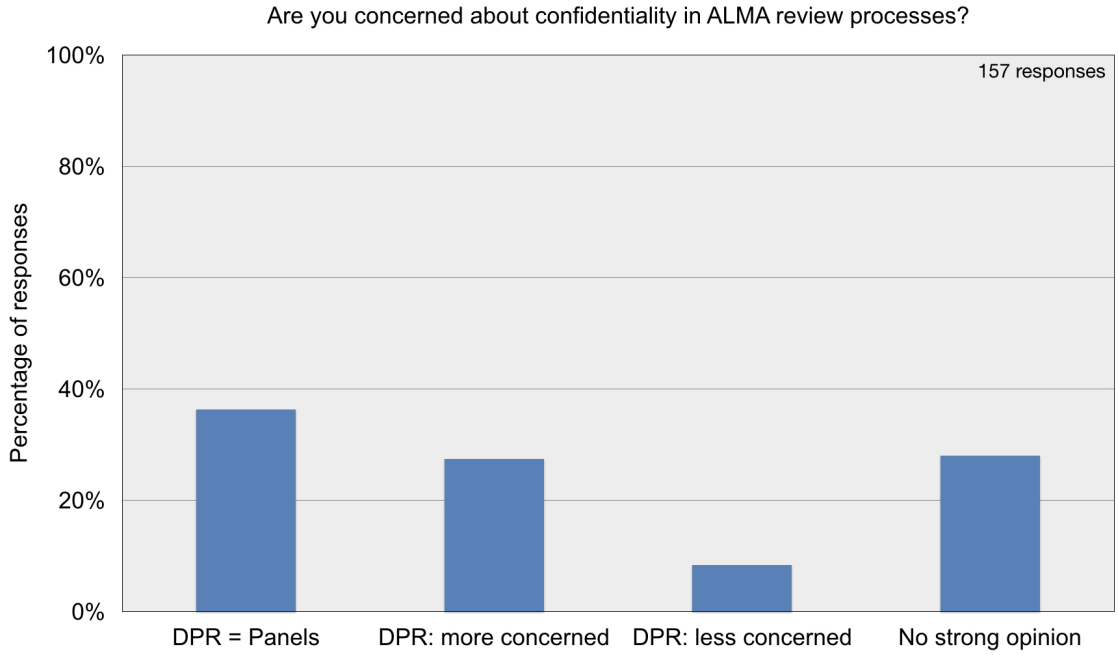

Figure 58

B.8 Are you concerned about the robustness of ALMA review processes against any biases?

- I think that the Distributed Peer Review process is as robust against biases as the Panel Review process. (32 responses)
- I think that the Distributed Peer Review process is more robust against biases compared to the Panel Review process. (34 responses)
- I think that the Distributed Peer Review process is less robust against biases compared to the Panel Review process. (45 responses)
- I have no strong opinion on this point. (46 responses)

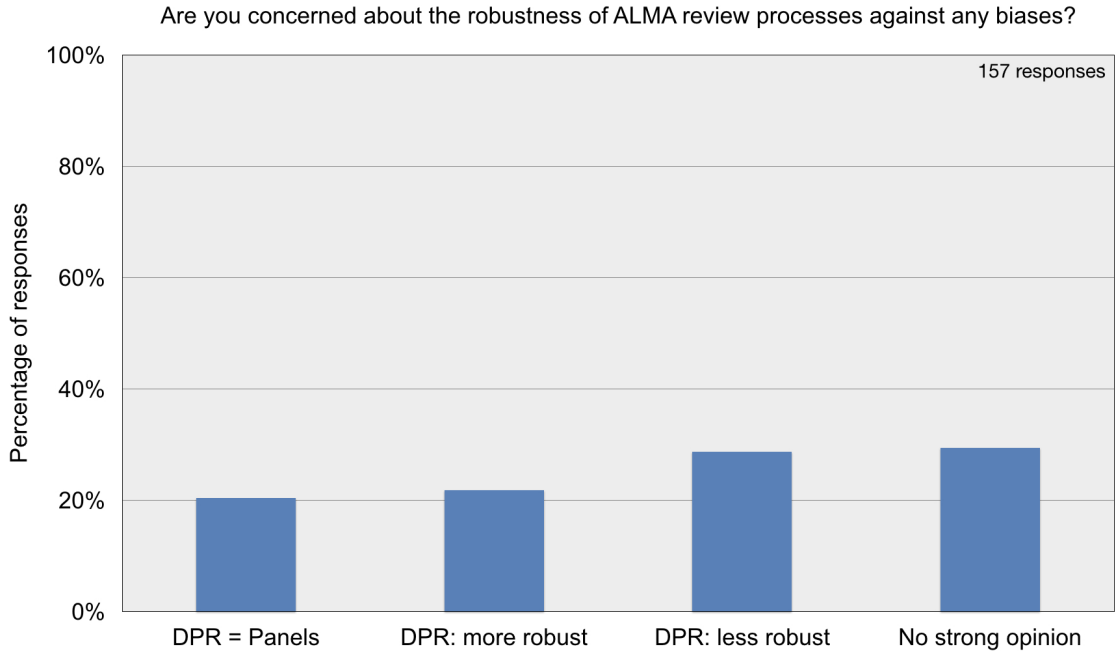

Figure 59

B.9 Would you submit ALMA proposals in future cycles if you were required to review 10 proposals for every proposal submitted?

- Yes; I would submit just as many proposals as I do now. (93 responses)
- Yes; I will submit proposals, but perhaps not as many as I do now. (53 responses)
- No; I would not submit any proposals. (4 responses)
- I don't know. (7 responses)

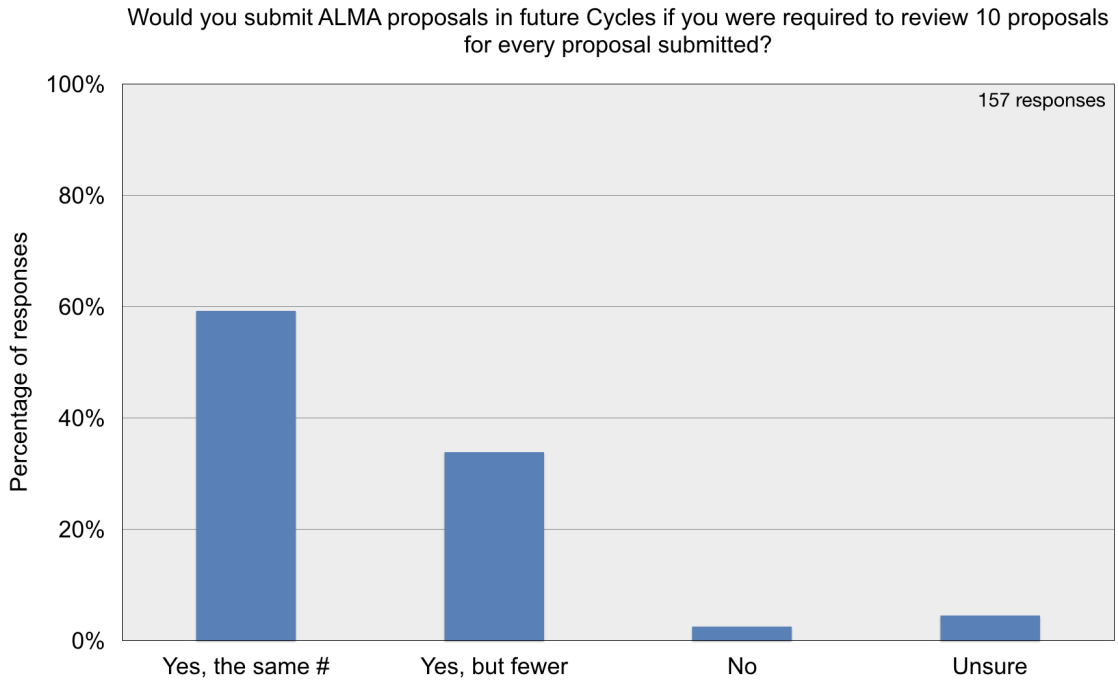

Figure 60

B.10 How many years has it been since you obtained your PhD?

- I do not have a PhD yet. (22 responses)
- 3 years or fewer. (40 responses)
- Between 4 and 12 years. (56 responses)
- More than 12 years. (39 responses)

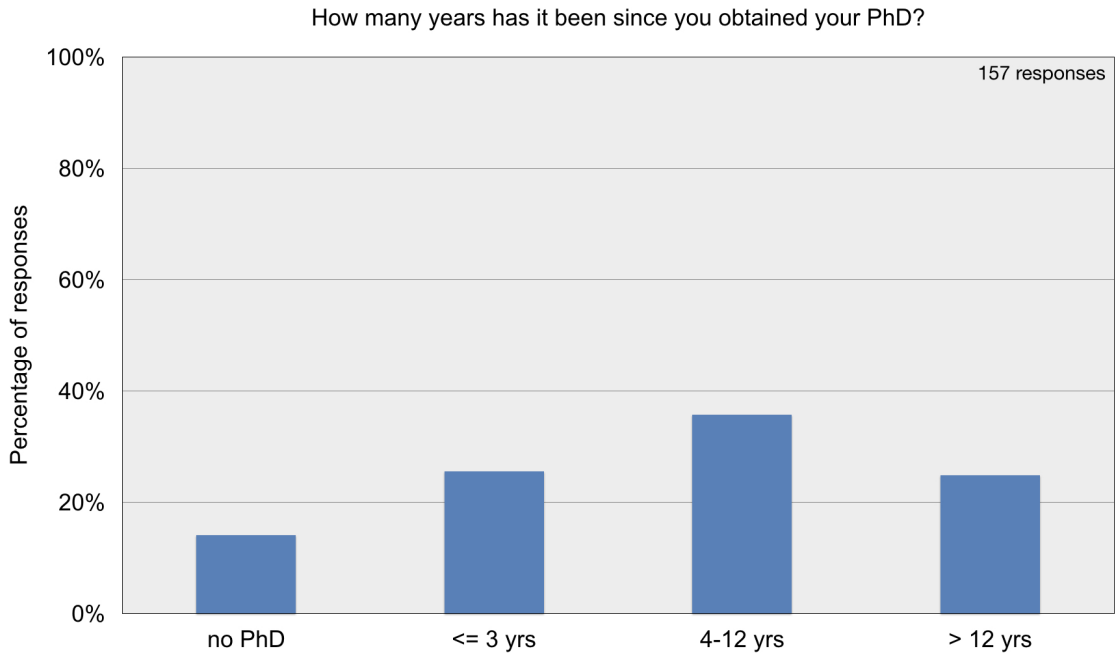

Figure 61

B.11 Please take a few minutes to rate the helpfulness of each review that you received, indicating the extent to which this comment will help to improve your proposal in the future. Positive comments like "best proposal I ever read" can be ranked as not helpful as it does not improve the proposal further.

- This review is very helpful. (450 responses)
- This review is somewhat helpful. (792 responses)
- This review is inaccurate or otherwise not helpful. (390 responses)
- This review is inappropriate or unprofessional. (29 responses)

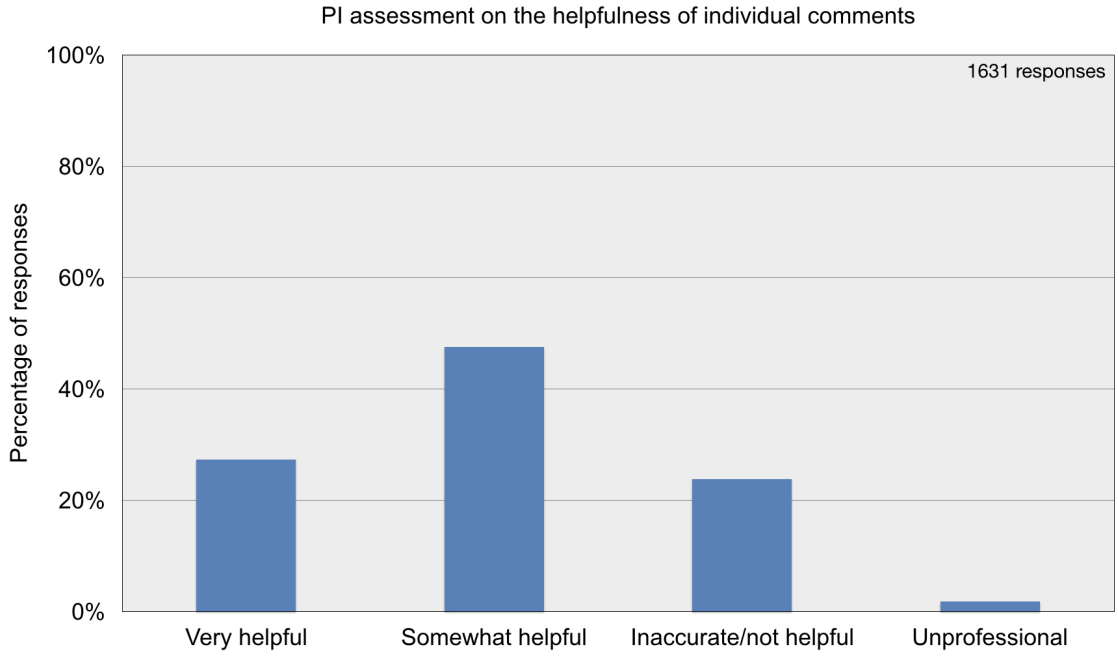

Figure 62

C Reviewer comments marked by the PI as inappropriate or unprofessional

This section lists the 29 reviewer comments that were classified as "inappropriate or unprofessional" by PIs. The comments are grouped by the experience level of the reviewer that wrote the comment. Only three of these comments was flagged as potentially offensive by the JAO in their review of the comments, but no edits were made.

The comments have been withheld because of confidentiality.