ALMA BOARD

ALMA EDM Document	
Distribution	Ordinary Session

Subject: Alternative Proposal Review Models for ALMA

Authors: J. Carpenter

Purpose of Document: To provide the ALMA Board with alternative proposal review models

Status: To be presented to the ALMA Board for the April 2018 face-to-face meeting

Change Record

Version	Date	Affected Section(s)	Reason/Initiation/Remarks
0.0	2018-02-03	All	New document

Executive Summary

The large number (~1600) of proposals that are submitted to the annual ALMA proposal call has stretched the resources of ALMA and the community, who volunteer an appreciable portion of their time to review the proposals. Both the reviewers and the ALMA Science Advisory Committee (ASAC) are concerned that the heavy workload may negatively impact the quality of the scientific assessments, and could make it difficult to sustain this review model. This memo explores alternative approaches to the classical panel-based, peer-review system that aims to maintain the current quality of the scientific reviews while reducing the burden on individual community members.

One class of the proposal review models retains a face-to-face meeting, but instead of a single venue, the panel meetings are decentralized to multiple venues world-wide. A second class of models foregoes the face-to-face meetings and relies on remote reviews. The reviewers in this scenario would be either (i) volunteers from the community following current practices, or (ii) PIs of the submitted proposals, who would be required to review a set of proposals in a distributed peer review model.

Distributed peer review has a strong advantage is that is easily scalable regardless of the number of proposals received. This model may improve the quality of scientific assessments by significantly reducing the workload on individual reviewers, and may improve the overall quality assessment by providing more reviews for each proposal. Two of the main concerns that have been raised about distributed peer review is that a significant fraction of PIs may be too inexperienced to review proposals, and PIs may attempt to game the system by downgrading good proposals in order to boost their own proposals. Possible mitigation strategies are discussed. Given distributed peer review has not been implemented on the scale required by ALMA, the biggest challenge will be to gain the confidence of the astronomical community in the process. We also describe other steps that could be taken to improve the ALMA proposal review process.

Table of Contents

1. Introduction

The current ALMA proposal review process follows the classical panel-based model that has been adopted at many observatories. Indeed, the ALMA review process has its roots in the ESO proposal review system. As with other leading observatories, the increasing number of observing proposals submitted by the community is stressing the ALMA review process. The problem is particularly acute at ALMA, where ~1600 proposals are submitted each cycle, which far exceeds that of other telescopes (although JWST may face similar challenges in the near future).

The growing number of submitted proposals has required a steady increase in the number of reviewers, which has stretched the resources of the community, where 146 people devote approximately 1-3 weeks to review ~90 proposals each and spend an additional week in a face-to-face meeting. Thus, any individual reviewer devotes up to 5-10% of their work time per year to review ALMA proposals. Moreover, reviewers are requested to provide their services for three years. While the number of submitted proposals appears to be stabilizing at around 1600 per cycle, the outstanding concern is that the review process is exhausting the ALMA community and that the large workload negatively impacts the quality of the scientific assessments. Indeed, ALMA Proposal Review Committee (APRC) and the ALMA Science Advisory Committee (ASAC) have repeatedly expressed concern about the high workload on reviewers.

In response to the community concerns, the ALMA Director charged the ALMA Observatory Scientist to evaluate alternative approaches for the ALMA proposal review process. The memo begins by reviewing the current proposal review process and its perceived strengths and weaknesses that motivated the charge from the Director. We then discuss the principles and goals of any new process, since it is essential that the integrity of the proposal review process be maintained and have the confidence of the astronomical community. We then outline four alternative approaches to the review. The Appendix discusses other steps that could be taken to further improve the proposal review process.

2. Current proposal review process

Overview

The current ALMA proposal review process follows the classical panel-based, peer review process adopted at many observatories. Each reviewer is assigned to a panel in one of six proposal categories, with up to 4 panels per category. In Cycle 5, the ALMA review committee consisted of 146 reviewers distributed over 18 panels to review 1661 proposals, which implies a typical workload of ~90 proposals per panel. Each panel contains 8 or 9 reviewers.

The review proceeds in a two-stage process. In Stage 1, the reviewers read all proposals[1] assigned to their panel for which they do not have a conflict, assign preliminary scores, and

[1] In earlier cycles, reviewers read/scored/commented on ~2/3 of the proposals in the Stage 1 process, and then read any of the remaining proposals that survived triage. This process was modified in favor of the current procedure to make triage more robust (since there will be more reviews per proposal) and to have more robust discussions at the face-to-face review (since all reviewers will have read all proposals).

provide comments on the proposals. An individual reviewer is assigned as primary reviewer on ~12 proposals. The Stage 1 scores are normalized and averaged, and the ~25% proposals with the poorest scores are triaged and not discussed further; in practice, the triage level varies with each region. The non-triaged proposals are advanced to the Stage 2 process where they are discussed in a face-to-face meeting held over 3.5 days, and then re-scored to produce a final ranked list. The primary reviewer summarizes the collective opinions of the panel in a written consensus report, which is forwarded to the Principal Investigator (PI) by the Joint ALMA Observatory (JAO).

Figure 1 compares the normalized ranks from Stage 1 and Stage 2 for the non-triaged proposals in Cycle 5, where Stage 1 ranks are based on the initial scores from the reviewers, and Stage 2 ranks are based on the face-to-face discussions. This figure shows that 28% of the top quartile proposals (corresponding to the 4:1 oversubscription rate) in the Stage 1 process were replaced as a result of the face-to-face discussions.

Figure 2 compares the mean scores and standard deviation of the scores as a function of the proposal rank. The standard deviation of the scores are nearly constant as a function of proposal rank in both Stage 1 and Stage 2, with a higher median dispersion in Stage 1 (σ=1.2) than in Stage 1 (σ=1.0). Thus, the discussion improves the consensus among the panels. However, while the panels may agree on which are the best and worst proposals, the dispersion in the scores is nearly the same whether the proposal is among the best, is average, or is among the worst.

The review panels also recommend which subset of Large Programs are forwarded to the ALMA Proposal Review Committee (APRC), which consists of the panel chairs. The APRC reviews the Large Programs in a 2-day face-to-face meeting that is held immediately after the regular review. The APRC recommends which Large Programs should be scheduled to the ALMA Director.

Strengths

1. With 146 reviewers, the ALMA proposal review process engages a significant fraction of the ALMA community and facilitates communication between the different regions.
2. About 28% of proposals in the top quartile change as a result of the face-to-face review compared to the preliminary (Stage 1) scores. This is arguably a significant change, and presumably reflects better science is scheduled on the telescope. (Although it should be noted that this is an assumption that has not been tested by having the same proposals reviewed in separate panels.)
3. The review model has general acceptance by the community, although with well-acknowledged shortcomings.

Weaknesses

1. The workload on the reviewers is high in that they need to review ~90 proposals. Overall a reviewer will spend 2-4 weeks of their time participating in the proposal review, which represents a significant demand on their time. The immediate risk is that a high workload may prevent a thorough review of the proposals, and thus the best science may not always be selected. A long-term risk is that it will be become increasingly difficult to identify qualified reviewers willing to serve on the panels.

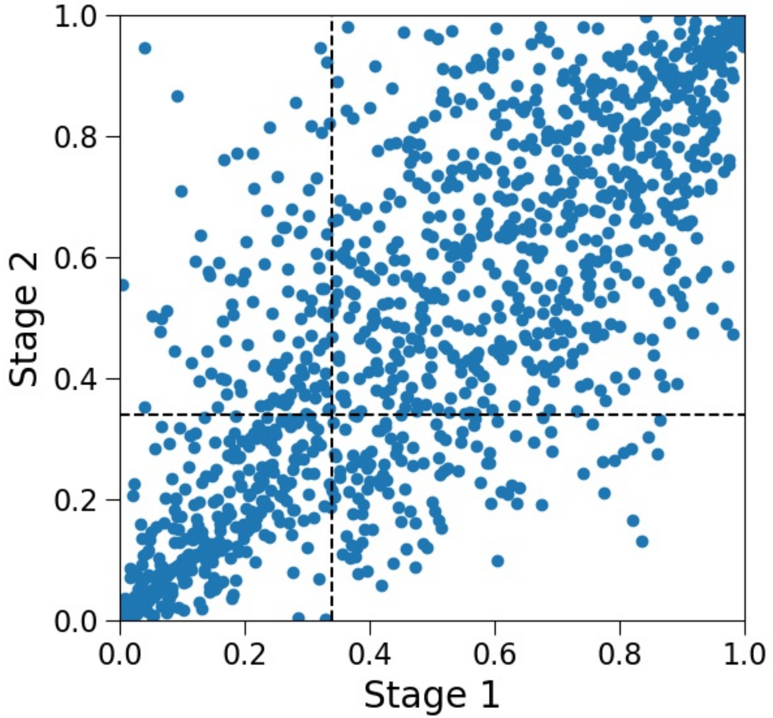

Figure 1: Comparison of the normalized Stage 1 and Stage 2 ranks for the Cycle 5 non-triaged proposals, where rank=0 indicates the top proposal and rank=1 indicates the worst proposal. The dashed box indicates the cutoff for the top 400 proposals, which corresponds to the 4:1 oversubscription limit. A total of 113 proposals (28%) in the top 409 proposals in the Stage 1 reviews were replaced after the face-to-face discussions.

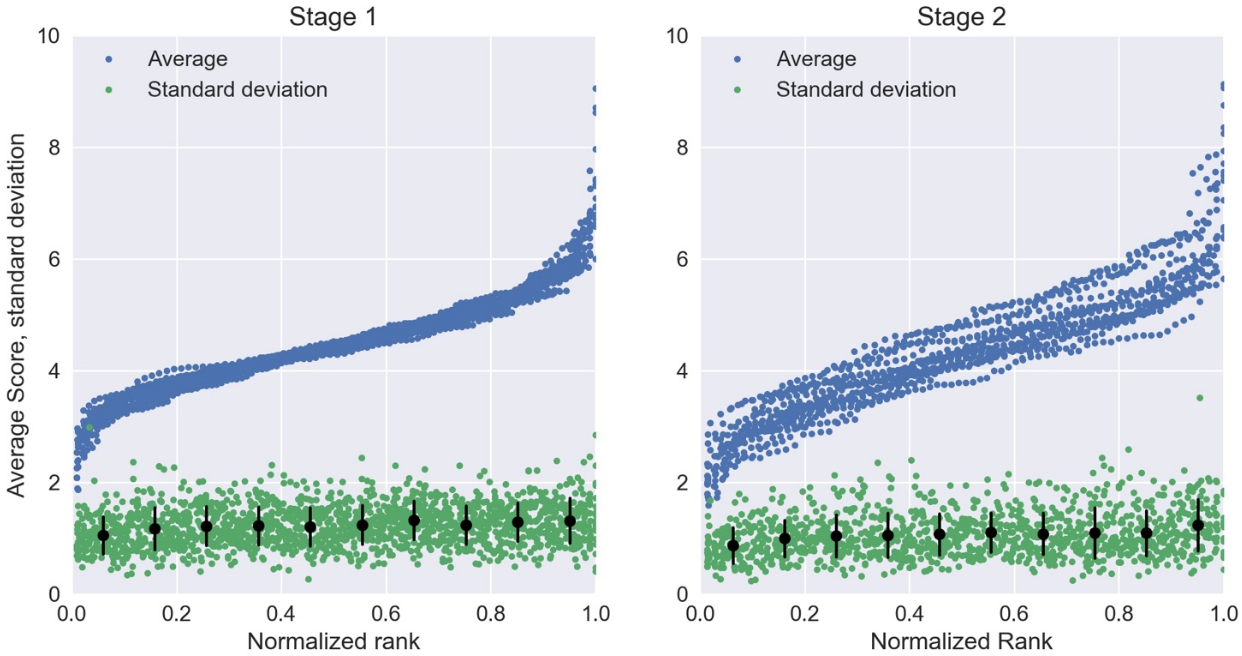

Figure 2: The average score (blue) and standard deviation of the scores (green) versus the normalized proposal rank for Stage 1 (left) and Stage 2 (right) process in Cycle 5. The black circles and error bars indicate the mean values and RMS of the standard deviation. The scores do not strictly increase systematically with rank because there are 18 individual panels.

2. It requires significant effort for the JAO to find 146 reviewers each year and organize the logistics for a large meeting.
3. If the number of proposals continues to increase, the review model cannot be sustained by continuing to increase the number of reviewers, if that point has not been reached already.

3. Principles

Any new proposal model must have the confidence of the astronomical community that the best science projects are selected. We therefore establish the principles that any new process.

1. The quality of the overall scientific assessment must be maintained or improved compared to current standards.
2. The review process must be fair and transparent to the community.
3. The review process must be scalable to change in the number of submitted proposals.
4. The review process must be sustainable, both in terms of involvement by the community and with the current level support provided by ALMA.

In considering alternative models, various assumptions were made.

1. No review process is perfect.

   Whether considering the traditional panel-based review system or alternative models, any model will contain weaknesses.
2. ALMA should retain peer review.

   Any review model will be based on the principle that peer review produces the most compelling scientific program. Thus any system which relies on a lottery to select proposals were not considered.
3. Most reviewers behave in an ethical manner.

   Any review model hinges upon the basic assumption that most reviewers behave in ethical manner. If most reviewers do not make honest scientific assessments, then no peer review process will be successful.
4. It is not practical to add more volunteer reviewers.

   ALMA already recruits 146 people each year to serve on the review panels. To reduce the workload by a factor of two while retaining the number of reviewers per panel would require doubling the number of reviewers to nearly 300 people. Based on experience with the review process to date, this is not considered feasible.
5. Large Programs will still be reviewed in a face-to-face meeting.

   Given the considerable investment of resources that ALMA devotes to individual Large Programs, these proposals will still be reviewed in a face-to-face meeting
6. A teleconference, held world-wide for 18 panels over 3 days, is not a viable alternative to a face-to-face meeting.

   It has been suggested to hold a telephone or video conference in place of the face-to-face review, which would reduce the travel and venue costs substantially. Many telecons are successful, but they are generally for 1-2 hours. Extending them for 2-3 days for 8h at a time, for people who are all across the world in different time zones, will be challenging. With 146 reviewers distributed over 18 panels, the risk of a poor audio or video connection is rather high. Also, personal experience with telecons suggests it is difficult for reviewers to fully disengage from responsibilities at the home institution and therefore capture their full attention over 3 days. The other possibility is to have a short telecon (< 1 day per panel) and restrict the discussion to only the high dispersion proposals or proposals near the boundary of being scheduled. Given that the scheduling is a complex function of proposal pressure as a function of configuration, frequency band, right ascension, region and proposal category, identifying proposals in the middle-ground is non-trivial.

4. Alternative models with face-to-face review

If retaining a face-to-face review is required or desirable, the process could be modified to the convenience of the reviewers. This section describes two variants of this approach, where instead of a single location, the review panels meetings are held at multiple locations worldwide.

4.1. Regional review

Overview
Each region conducts a review of the proposals in their region and forwards the results to the JAO, which will merge the results and create the observing queue using current procedures. The details of the review will be determined regionally, but there would be a common framework that will enable a global ranked list to be created from the regional ranks. There would still be a common proposal deadline and a single proposal submission system developed and maintained by ALMA. A single face-to-face meeting to review the Large Programs will still be held and organized by the JAO.

As an example, in Cycle 5, there were 695 proposals from Europe, which was the most from any region. This would imply an average of 155 proposals per category, and that 3 panels per category would be needed to reduce the number of proposals per reviewer to ~45.

Strengths

1. Any regional biases in the review process would be reduced.
2. Travel will be more convenient for the reviewers.

Weaknesses

1. Holding regional reviews is contrary to the notion of one observatory. In particular, the scientific merit of proposals in different regions would no longer be compared and ranked.

The ALMA Integrated Science Team (IST) rejected the regional model on the basis that there should be a common, international review process for ALMA.

4.2. Decentralized review

Overview

XMM-Newton has a peer review system where panel members meet to discuss proposals, as in the current ALMA system, but the system is "decentralized". This and other differences compared to the current ALMA model could potentially reduce the workload for reviewers.

The decentralized review will require the following changes.

1. Reduce the number of reviewers per panel from 8 to 5.
    a. If the total number of reviewers is kept the same, the number of proposals per panel would decrease to ~60 from ~90, which would be a significant reduction in the workload.
    b. If the total number of panels is kept the same, each panel would still review ~90 proposals, but the process would be more sustainable since fewer reviewers would be needed.

    There are of course intermediate solutions between these two extremes.
2. Decentralize the ARP meetings: instead of making the ARPs meet at the same place, each panel organizes itself.
    a. In the XMM-Newton review system, each panel chair writes an e-mail to the other panel members well in advance of the meetings and proposes a place to meet, which is usually the institute/university where it is easiest to travel to for all members. This could be the home institute/university of the panel chair or alternatively the easiest place to reach by most panel members.
    b. Because the members meet in the most convenient place for all members of the panel, the travel burdens will be reduced. The most convenient option would be for all panel members to be from the same region, but would be evolving into a regional review model, which the IST rejected. Therefore, in practice, each ALMA panel would include a member from each region (Chile, East Asia, Europe, and North America).
    c. Technical and administrative support during the meetings would come via phone and email primarily from the local ARC. It would still be an option to send a technical secretary to each review panel from the regional ARC.
3. Rather than have the primary assessor and *all* secondary assessors (i.e. the entire panel) write comments on every proposal, the following procedure could be used following the Chandra scheme[2]:
    a. One primary assessor per proposal
    b. One secondary assessor per proposal
    c. All members of the panel read every proposal but reports are only written by the primary assessor and checked by the secondary assessor, although the consensus report is of course agreed by all members of the panel and includes comments from all assessors.

[2] For reference, in Chandra's 17th observing cycle, 578 proposals were submitted and reviewed by 13 panels containing 96 reviewers total. The average panel therefore consisted of 7 reviewers to review 44 proposals, which is half the workload of ALMA panels.

The Large Program proposals will still be reviewed by the panel chairs in a separate face-to-face meeting held over 2-3 days.

Strengths

1. The face-to-face meetings are kept, which is assumed to select the best proposals by reaching a consensus after a discussion.
2. It preserves a review model that has broad acceptance by the astronomical community.
3. It retains the ability for PIs to have proposal feedback in the form of consensus reports[3].
4. If there are 29 panels, the ARP meeting could be completed in two days to further reduce the burden on the reviewers.
5. If we have 29 panels, then the work should is reduced to ~ 60 proposals/reviewer, which may result in better reviews. It may also be able easier to recruit panel members if the workload is reduced, which makes the process more sustainable.
6. It may contribute to confidentiality since in principle one ARP member knows only his/her ARP members and thus no pressure from the community should be received on individual members.
7. Little if any software development should be needed from ALMA.

Weaknesses

1. If the number of panels remains the same, each panel will still review ~90 proposals. Reviewers will formally enter comments for ~36 proposals, but not for the remaining ~54 proposals. It seems likely that reviewers will still need to record comments for the remaining ~54 proposals in some manner so that they can remember their comments for the face-to-face meeting. Therefore, the workload will likely not be significantly reduced if the number of panels stays the same.
2. If the total number of reviewers remains the same but the number of panels is increased, it may be difficult to identify ~29 people with the relevant experience who are willing to serve as panels chairs and lead the organization at their home institution.
3. Most of the proposal categories cover a broad range of topics. With 5 reviewers per panel, it will be difficult to cover the range of topical expertise, and there is a greater chance a panel will have a single (or no) expert on a subtopic.
4. The mean scores with 5 reviewers will have a larger uncertainty than with 8 reviewers, and thus there will be less robustness in selecting the best proposals. This may be compensated by better individual reviews if there is a reduced workload with fewer proposals per panel.
5. It would be difficult to provide centralized support from the JAO since the panel meetings will be held throughout the world around the clock. The immediate support will need to come from the ARCs, while there may be delays in answering questions that require JAO staff.

[3] It is unclear what fraction of the community finds consensus reports useful overall, and some people have advocated against any consensus reports. However, those opinions may change if the consensus reports were of higher quality.

5. Alternative models without face-to-face review

This section describes two approaches to the review model that do not require a face-to-face review. The first approach follows the current practice, where members of the community are invited to participate on review panels, and thus participation is voluntary. The second approach relies on the PIs of the submitted proposals to review the proposals: PI participation is mandatory, or else their proposal will not be considered for scheduling.

5.1. Current review model ex. Stage 2

Overview

This model follows the classical approach where the JAO identifies 146 reviewers to serve on as panels, and each panel is assigned ~90 proposals. The current Stage 1 review is kept as-is. Instead of a face-to-face review, Stage 2 would consist of the reviewers reading the comments from the other reviewers. The Stage 2 process could potentially be a message board discussion, although that may be cumbersome with 1600 proposals and 90 proposals per reviewer. The reviewers would then enter their final scores, and the average score among the reviewers would be used to determine the final ranks. A face-to-face review for Large Programs and potentially medium-sized (~25-50 h) proposals will still be held.

Strengths

1. The workload could be reduced by not requiring travel to a face-to-face meeting.
2. It may be easier to identify reviewers willing to serve on the panels if they did not have to travel to a face-to-face meeting.
3. A triage process will no longer be needed. Poor proposals can be rejected more robustly (e.g., proposals consistently ranked in the bottom 10%), while all others are eligible to be scheduled. This will avoid the issue now where proposals in low-pressured regions are triaged and cannot be scheduled.

Weaknesses

1. Unless the number of reviewers is increased, each panel would still need to review ~90 proposals, which would remain a significant burden.
2. To reduce the workload by a factor of 2 would require a total of 300 volunteers to serve on the panels. It would likely be rather difficult to identify such a large number of volunteers even if travel was not required.
3. It is unknown if the review of the Stage 1 comments and/or a message board would stimulate critical discussion of a proposal. Overall such forums would be less efficient than a face-to-face discussion. Also, the quality of the current Stage 1 comments is quite heterogeneous, perhaps because of the current workload.
4. There will be less accountability for reviewers to justify their scores and comments without a face-to-face review.

5.2. Distributed peer review

Overview

The distributed peer review model follows the procedure described by Merrifield & Saari (2009, Astronomy and Geophysics, v60, p4.16). They proposed this model after experiencing a heavy workload when reviewing proposals for ESO, which is the same circumstances that pertain to the ALMA review. In distributed peer review, every PI is obligated to review and rank *N* proposals for each proposal they submit. If a PI does not submit their ranks by a set deadline, their proposal is rejected.

The number of proposals to review per submitted proposal is a compromise between (i) having a reasonable workload that would allow for careful reviews, (ii) having a reasonable number of proposals for the PI to remember and produce a relative ranking, and (iii) producing a sufficient number of reviews per proposal to obtain a good statistical sampling of ranks. Considering these factors, I suggest *N* ~8-16 (see accompanied document for further justification).

PIs can delegate responsibility for the proposal review to another member of their team at the time of proposal submission. The main purpose would be to allow PIs who do not have the time to conduct the review to pass the responsibility to a team member who does.

Distributed peer review would proceed in either one or two stages. In the simplest implementation, the reviewers would read their proposals, write brief comments on the strengths and weaknesses of the proposal, and rank the proposals in their set from 1 (best) to *N* (worst). In the two-stage process, the reviewers would only provide comments in Stage 1, and in Stage 2, they would review all comments on the proposals in their set, and then rank the proposals. Merrifield and Saari (2009) do not have a Stage 2 process, or any comments at all, in their model.

To provide an incentive for PIs to conduct careful reviews, Merrifield and Saari (2009) proposed that reviewers who produce a ranked list that closely match the consensus rankings receive a boost in the rankings of their own proposal. This will discourage PIs from downgrading otherwise good proposals in an attempt to game the system. In blog discussions of distributed peer review, this was one of the more controversial aspects of the process.

For Large (> 50 h) and potentially medium sized proposals (~25-50 h), there will still be a face-to-face review held over 2-3 days. For at least the medium-sized proposals, the rankings and comments from the PI reviews will be available for the face-to-review meeting.

Distributed peer review has been adopted in a few forums. Gemini Observatory has used distributed peer review for their fast-turnaround proposals since 2015, for which they receive ~15 proposals/month. The NSF had a pilot program in the Sensors and Sensing Systems proposal call, and reported positive results in that the number of applications (131 submitted proposals) increased over the previous year and the length of the reviews increased by 40%. The NSF pilot also sparked more detailed theoretical analysis of this model (Naghizadeh & Liu 2013, arxiv 1307.6528) and variants thereof (Kurokawa et al. 2015, Proceedings of the International Joint Conference on Artificial Intelligence, 582). The National Institute of Food and Agriculture has also started a pilot study of distributed peer review for 3 programs. However, none these programs will have implemented distributed peer review on the scale of the ALMA proposals.

Strengths

1. This model significantly reduces the workload on reviewers. In Cycle 5, 67% of the PIs submitted one proposal and 22% submitted two proposals. Therefore 89% of the PIs would have needed to review 8-32 proposals (assuming N~8-16) if this model was in place for Cycle 5.
2. Reviewers will be self-selected and eliminate the task of identifying reviewers each year.
3. Each proposal would have more reviews than the current practice if $N > 8$, which will better measure the mean rank and potentially allow outlier rankings to be rejected.
4. A triage process will no longer be needed. Poor proposals can be rejected more robustly (e.g., proposals consistently ranked in the bottom 10%), while all others are eligible to be scheduled. This will avoid the issue now where proposals in low-pressured regions are triaged and cannot be scheduled.
5. It should be possible to reduce potential conflicts of interest in the review assignments by imposing more stringent criteria, for example, by not allowing PIs to review proposals from any of their coIs, from the same institution, or proposals observing the same object.
6. It may encourage larger proposals with coherent ideas rather than breaking a large proposal into multiple submissions.
7. Since all PIs participate in the review process, the overall transparency will increase.
8. Young students and postdoctoral students gain a valuable educational experience in reviewing the proposals.

Weaknesses and mitigation

1. The review model has not been implemented on a large scale, and therefore may not have immediate acceptance by the community.

   Mitigation: Implement the review model on a small scale (e.g., a call for ACA standalone proposals) would be one way to introduce the model to the community.
2. Some PIs may try to game the system by demoting good proposals and promoting bad proposals, so that their own proposal rises in the ranks.

   Mitigation: Reject outlier ranks before averaging the results, especially if $N > 8$.
3. Some PIs may not take the review process seriously and randomly rank the proposals just to complete the task.

   Mitigation: Allow PIs to designate another reviewer who would be responsible for the review.
4. Some PIs may have inexperience with reviewing proposals or are inexperienced in astronomy.

   Mitigation: Student PIs can be allowed designate a mentor on the proposal who will assist in the review process.

   Analysis: ALMA does not track the experience level of the PIs (as measured by year since PhD), but it can measure how often PIs are submitting ALMA proposals.

While this does not measure experience per se, it does reflect experience in writing ALMA proposals and general familiarity with ALMA's capabilities. In Cycles 5, 71% of the PIs have submitted proposals in at least half of the six cycles, and 14% of the PIs submitted a proposal for the first time. By comparison, of the 146 reviewers in Cycle 5, 76% have submitted an ALMA proposal in at least half the cycles, and 14% of never been PI on an ALMA proposal. Thus, the PI pool and reviewers have comparable experience in submitting ALMA proposals. The high fraction of repeat PIs also indicates that experience in reviewing proposals will increase rapidly with each passing cycle.

5. One reviewer may have a "killer" argument missed by the other reviewers that may promote and undermine a proposal.

   Mitigation: This would be mitigated if there was a two stage review process, but not the single stage model.

   Analysis: It should be noted that in Cycle 5, no proposal that was in the bottom quartile in the Stage 1 review were promoted to the top quartile after the face-to-face discussion, and only 0.7% of the all proposals advanced from third to top quartile. Similarly, 0.9% of proposals went from top to third quartile as a result of the face-to-to face review. This suggests that it is relatively rare for a proposal to have a major change in their ranks as a result of the face-to-face review.
6. It will not be practical to review the comments to ensure the remarks are not offensive, since there will be ~1600 x *N* comments. Therefore, PIs will not receive consensus reports. Either they need to receive the unedited comments, or not receive comments at all. Regardless, PIs could be given the distribution of ranks (e.g., top, bottom, mean, dispersion, or a histogram of the ranks) so that they can understand how the proposal was perceived.
7. The review process may disfavor high risk-high reward proposals, because the reviewers will tend to be conservative.

   Mitigation: This is true of the current review process as well, and could be mitigated by explicitly calling for high-risk/high-reward proposals, and encouraging such proposals for DDT.
8. There may be considerable overhead in dealing with ~1100 reviewers and answering their questions with a set deadline. The software would need to be robust and well tested beforehand. It would be beneficial to have ARC support to help answer questions. However, this is all true for the proposal submission deadline and that largely runs smoothly, and the instructions for the distributed peer review will be considerably more straight forward than proposal submission.
9. There will be significant (onetime) software development to implement the system.

6. Discussion

This section compares the review models discussed in Sections 3 and 4. To facilitate comparison, the following goals were set that were consistent with the principles outlined in Section 3.

1. Reviewer workload

   Reviewer workload should be ≤ 45, which is half the current workload.

2. Number of reviews per proposal

   The goal for the number of reviews per proposal was set to a minimum of 8, consistent with the current practice, on the assumption that more reviews and thereby more opinions leads to a firmer consensus on the proposal quality. It is debatable what is the minimum number of reviews to provide an adequate consensus. This metric then only implies there is an increase in the number of reviews.

3. Review discussion

   For the review discussion metric, a face-to-face meeting is assumed to be a better forum to reach a consensus than a blog-type discussion or by simply averaging the review scores, all else being equal.

4. Scalability

   The review process can readily scalable if the number of proposals continues to increase without overburdening the community or the JAO. In determining if the process is scalable, it was assumed that it would be difficult to added more volunteer reviewers beyond the current 146 if the requirement is to read more than 90 proposals.

Table 1 summarizes the analysis of the different review models, including the current review process as a benchmark. Two versions of the decentralized model are shown: (i) the same number of panels (18) as currently but fewer reviewers in total, and (ii) the same number of reviewers currently but distributed over more panels (29). The regional review is not shown since it is rejected in principle by the IST.

The color coding in Table 1 has the following meanings:
Green : Fully compliant with goals or improvement relative to current review process
Yellow : Partially compliant with goals (within a factor of 2)
Red : Non-compliant with goals (> 2) or less effective than current review process
White : Undetermined

Table 1: Assessment of proposal review models against goals

Goal	Current Process	Decentralized: 18 panels	Decentralized: 29 panels	Current ex. Stage 2	Distributed peer review
Reviewer workload	🟥	🟥	🟨	🟥	🟩
Reviews per proposal	🟩	🟥	🟥	🟩	🟩
Review discussion	🟩	🟩	🟩	🟥	🟥
Scalability	🟥	🟨	🟨	🟥	🟩

The assessment quality is a subjective measure that reflects if the resulting ranked list of proposals would better identify the top ranked proposals. The assessment quality is judged relative to the current practice of 8 reviewers per proposal and a face-to-face discussion. Given this is a subjective judgement, an explanation is provided for each model.

1. Decentralized with 18 panels

This model preserves the face-to-face discussion, but the workload on the reviewer is essentially the same as the current model. The scientific rankings would be negatively impacted by having fewer reviews per proposal with no reduction in overload workload. There are no clear reasons to believe that the overall scientific assessment would be markedly improved with this model.

2. Decentralized with 29 panels

   The workload will be reduced by ~38%, which is a significant reduction and should improve the quality of individual reviews. However, there will also be fewer reviews per proposal, which would negatively impact the robustness of the scientific rankings. The net impact is unclear.

3. Current ex. Stage 2

   Excluding the Stage 2 face-to-face review should detract from the assessment and thus reduce the overall quality of the scientific assessment.

4. Distributed peer review

   While there is no face-to-face discussion to reach a consensus, there could be up to twice as many reviews per proposal to better sample the range of opinions. The workload of the reviewers is markedly lower, which should improve the quality of individual assessments. Some reviewers will not be as experienced. The net impact on the scientific assessment is unclear.

The distributed peer review model is the only model considered here that truly resolves the current heavy workload on the reviewers. This model is attractive in several other respects as an alternative to the current review process. There are potentially more reviewers per proposal, which would better measure the mean rank. Each reviewer in the distributed peer model has far fewer proposals to read, and thus presumably the quality of individual assessments will improve. While ALMA reviewers have been chosen for their overall scientific expertise, they have not necessarily previously submitted an ALMA proposal or are familiar with ALMA observations; on the other hand, all reviewers in the distributed peer review model have at least submitted ALMA proposals by design and should be familiar with some aspect of ALMA's capabilities. How the quality of the scientific assessments in the distributed model compares with the current panel-based discussed with a face-to-face review is more subjective.

One possibility to judge the merit of distributed peer review would be to run it in parallel with the current system for one or more cycles. However, the conclusion may not clear. If the two approaches give similar ranked lists, that would provide confidence in the distributed peer review model. If the ranked lists are significantly different, it would not imply the distributed peer review model is inferior, but only that one or both models have their shortcomings. A more informative assessment would likely require not only running both reviews in parallel, but also have a subset of proposals reviewed by multiple panels in the classical model. Thus, there will be two independent panel-based rankings of the proposals, as well as the distributed-based rankings, which may help assess which model produces more consistent results. A metric would be needed to establish beforehand on how one would judge if the distributed peer review gave "better" or "worse" results. Of course, this only further increases the workload on the community and the JAO, and such experiments would need to run over multiple cycles to reveal the trends.

Finally, while Gemini has a positive experience with distributed peer review with their fast-turnaround proposals (Markus Kissler-Patig, private communication), such a system has not been implemented on the scale of ALMA and has not been used as the primary means to select proposals. As discussed in Section 5.2, distributed peer review as potential weaknesses, although mitigation strategies can be implemented for many of them. Given distributed peer review has not been implemented on a large scale, perhaps the biggest challenge would be to convince the community that the collective opinions of their peers, without a face-to-face discussion of selected experts, produces a credible science program.

Appendix

General approaches to reduce workload

Even if the basic proposal model is kept the same, other steps can be taken to reduce the workload on reviewers.

1. Reduce the number of pages per proposal to a maximum of 3/5 (for regular/LPs, respectively). This would reduce the number of pages that reviewers need to read by 25% reduction. This is a trivial change to implement.

   Some members of the Cycle 5 APRC recommended to reduce the number of pages, which was supported by the ASAC, iSOpT, and IST. While some Board Science Committee members expressed concern that 3 pages was too short, the committee left the decision to the JAO. Ultimately the Director decide to maintain the current page limit for Cycle 6 out of concerns that a reduced page limit may negatively impact medium-sized proposals, and because more comprehensive approaches to reduce the workload will be explored for Cycle 7. Nonetheless, this can be reconsidered for future cycles along with other steps to reduce the workload.

   It should be noted that ESO has a shorter proposal length than ALMA, but their advisory committees report very similar complaints about the workload and potential impact on the review quality. Thus, reducing the page length alone will not solve the current concerns about the workload on the reviewers.

2. Remove comments. Writing comments is a significant part of the workload, and there are general complaints that the comments are not useful (although it is not clear that is a universal opinion). However, PIs presumably will make notes anyway on each proposal, so it is not clear how much this actually reduces the workload in practice.

Other potential improvements

1. Remove the proposers name from cover sheet may also help to remove unconscious biases. Alternatively, one could provide just an alphabetical list of investigators, as is the current practice at HST.
2. It is often discussed that gender biases encountered during review processes of other telescopes may be explained simply by seniority. In fact, seniority could be an important bias too (panels could tend to award more time to senior PIs regardless of proposal quality). Both biases would be minimized if proposals were anonymized.

3. Regarding language biases towards non-native English speakers, this can only be achieved by giving training/explicit instructions to assessors and making sure to include a variety of nationalities in the panels (which is the current practice). In general, any assessor would benefit from training to overcome conscious/unconscious bias during their ratings/discussions.
4. Additional modifications that could contribute to a better review for Large Programs, by reducing conflicts of interest:

   > Chandra model: “Prior to the review, LPs are distributed to a group of “pundits”. Pundits are experienced scientists with broad research interests who focus exclusively on large projects. Pundits are asked to read all LPs and to provide written reports on specific proposals assigned to them. The pundit reports are made available to the topical panels and incorporated into the panel discussion. LPs are discussed by the topical panels and ranked along with the GO, archive and theory proposals. The recommendations from topical panels are recorded and passed to the Big Project Panel (BPP), which includes all topical panel chairs and the pundits.”