ALMA BOARD

ALMA EDM Document	
Distribution	**Ordinary Session**

Subject: Distributed Peer Review

Authors: J. Carpenter

Purpose of Document: To describe possible implementation of Distributed Peer Review Model

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

Summary

Merrifield & Saari (2009, Astronomy and Geophysics, 50, 4.16) proposed a distributed peer review to address the growing number of observing proposals submitted to leading observatories. The basic premise is that the Principal Investigators (PIs) review the proposals and assess which projects should be scheduled on the telescope. This approach may improve the overall scientific assessment by reducing the workload on reviewers an enabling more reviews for each proposal. This memo outlines how distributed peer review could be implemented for ALMA.

1. Overview

Merrifield & Saari (2009, Astronomy and Geophysics, 50, 4.16) described a distributed peer review model to address the deficiencies in the widely practiced panel-based, face-to-face peer review model. In particular, the proposed distributed peer review addresses the increasing number of proposals submitted to leading observatories such as ALMA, which leads to a large burden on members of the review panel.

The basic steps of a distributed peer review model are as follows:

i. Principal Investigators (PIs) submit their proposals by the normal process.
ii. After the proposal deadline, each PI is sent a set of *N* proposals for each proposal they submit. The proposals will generally be in the same scientific category as their proposal. It is expected that *N* ~ 8-16.
iii. PIs identify any conflict of interests they have in their set of proposals; any such proposals will be replaced by an alternative proposal.
iv. PIs review their proposals and write brief comments that summarize the strengths and weaknesses of the proposal. They rank the proposals 1 from (best) to *N* (worst).
    a. As a variant to the model, the review could proceed in two stages. In Stage 1, PIs submit their comments only. In Stage 2, they review all comments for their proposals, and then rank the proposals.
v. If the PI do not submit their assessments by the deadline, their proposal is rejected.
vi. The rankings from all PIs are combined to produce a globally ranked list of proposals that is used to generate the queue.
vii. The comments and individual proposal ranks are sent to the PIs in lieu of consensus reports. If the comments are not sent, then PIs will be sent statistics on the ranks; e.g., min, max, mean, standard deviation, or a histogram showing the distribution of ranks.

For Large Programs (> 50 h) and potentially medium sized proposals (~25-50 h), there will still be a face-to-face review as we have now, held over 2-3 days.

There are two main difference between the proposed implementation for ALMA compared to Merrifield and Saari (2009). First, PIs enter comments for each proposal, which are then sent to the PI. This serves two purposes: it allows some feedback to the PIs and allows quality control afterwards to verify that low proposal ranks are scientifically justified. Second, Merrifield and Saari (2009) proposed that PIs should be rewarded for a particularly good job in their reviews by boosting their proposal in the rankings; this step was omitted in the proposed implementation for reasons discussed in Section 2.

In the following section, a possible implementation of distributed peer review for ALMA is discussed in more detail.

2. Distributed peer review

The goal of distributed peer review is for the community to be produce a scientifically-ranked list of proposals, just as in the current review process. Thus, subsystems downstream from the APRC rankings should not be affected by changes in the review model itself. This section describes the processes between the OT and the end of the APRC that may need to be modified for distributed peer review.

2.1. Proposal submission

Merrifield and Saari (2009) suggested that a proposal team have the collective responsibility for reviewing their assigned set of proposals. In this manner, the PI could review the proposals alone, delegate the responsibility to a team member, or collectively as a team decide on the proposal ranks. However, such broad distribution of the proposals would make it more difficult to control conflict of interests and proposal confidentiality.

Nonetheless, there are circumstances where it is desirable to have someone other than PI participate in the review process. The most obvious case is if the PI knows in advance that they will not have the time to commit for a thorough review, and thus would prefer to delegate the task rather than conduct a poor review. Another case would be if the PI is inexperienced overall with ALMA or astronomy in general (e.g., a student). In this case, either delegating the review task, or designating a co-reviewer, should improve the overall review. Moreover, it may encourage more accountability from the PI to submit careful reviews for their assigned proposals.

The proposal is that a PI is the given the review assignments by default unless designated otherwise at the time of proposal submission in the OT. A student PI may designate one person on the proposal team who will be allowed to discuss the proposals with the student and provide advice. The designated reviewers cannot be changed after the proposal deadline since it impacts the conflicts of interests and thus the review assignments.

2.2. Proposal handling by the JAO

This section describes the role that the JAO plays between the proposal deadline and submission of the final ranks by the PIs. Most of the work is frontloaded to cull the proposals for erroneous submissions and assigning the set of N proposals to each PI. The critical role at the end of this period is to remind the PIs that the deadlines are upcoming.

i. After the proposal deadline, the JAO will take one week to identify and remove duplicate proposals and other erroneous submissions, as is the current practice.
ii. After removal of erroneous proposal submissions, a set (S_i) of N proposals is assigned to each submitted proposal (P_i). The assignment algorithm is described in Section 2.3.
    a. If proposal P_i is later identified to be an erroneous submission, the proposal set S_i is dropped and does not need to be reviewed.
    b. However, the invalid proposal P_i will still be in other proposal sets. It may be difficult to replace the proposal with a new proposal since the reviewer may have already started, and potentially finished, their reviews. One option is to

simply remove the proposal so that S_i contains N-1 proposals. That will slightly bias the rankings, but is perhaps inevitable.

iii. After all review assignments have been made, the JAO will send an email to each PI and co-reviewer (if applicable), with a cc to the relevant ARC manager, indicating that their proposal set is ready for review.
iv. A confirmation email is sent to the PI (cc to the relevant ARC manager) when the Stage 1 comments and ranks are submitted.
v. The PI shall receive a periodic reminder near the Stage 1 deadline if their ranks have not been submitted. The frequency of emails remains to be determined, and as whether the emails are sent automatically from the JAO or sent by the ARCs.
    a. How do we deal with emails that are sent to spam or not delivered?
    b. There is a step where the PI needs to confirm that they do not have conflicts of interest on a proposal. There could also be a reminder if this step has not been completed after two weeks.

2.3. Assigning proposals

The JAO will assign *N* proposals for each proposal set to minimize potential conflicts of interest. This procedure will never be a perfect, and if needed, PIs will also be able to identify conflicts with their assigned proposals. Because we only need to identify *N* proposals per proposal set, and there is a large pool of potential proposals, the criteria for a conflict can be stricter than the current practice. Primary and secondary conflicts are defined as follows:

Primary conflicts

i. The PI or designated reviewer is a PI, coPI, or coI on the proposal.
ii. The PI or designated reviewer is at the same institution as any one of the PIs, coPIs, or coIs on the proposal.
iii. The PI or designated reviewer has proposed to observed the same object as the proposal with a tolerance of 10 arcminutes or with the same source name (case and space insensitive).
    a. Ideally overlap of mosaics should be checked as well, but it is more time consuming to determine if mosaics overlap and by how much. A search radius of 10 minutes should capture most overlapping mosaics.
    b. Solar proposals should not be checked.
    c. A PI who submits a Target of opportunity (ToO) proposal with unspecified coordinates (ra=0, dec=0) should not review other ToO proposals. This will flag proposals that are not real conflicts, but it is otherwise difficult to identify ToO proposals that are in conflict and are triggering on the same class of objects.

Secondary conflicts

i. A proposal which has a common coI as any proposal submitted by the PI or designated reviewer.

The proposals are assigned to a set as follows:

i. Proposals for Large Programs are not assigned. They will be reviewed in a separate face-to-face review.

a. In addition to the *N* proposals, we could have non-conflicted reviewers review one Large Program each, and rate it as Excellent, Very Good, Good, Fair, or Poor. This would be provided as input to the APRC.

ii. Any proposal with a primary conflict cannot be assigned to the PI.

iii. Proposals are preferentially selected that are in the same science topic.
   a. Ideally, proposals are selected from the same category and keyword as the PI proposal.
   b. For each category/keyword, an acceptable list of categories/keywords that could be added to the proposal set shall be defined, with priority order. In general, any proposal in the same category is suitable, but some keywords in other categories may be suitable as well.
   c. Alternatively, machine learning techniques can be used to identify similar proposals based on the proposal abstract and text. HST, for example, has been experimenting with such an approach (Strolger et al. 2017, arxiv 1702.03324).

iv. A PI may not be assigned a proposal more than once, even if they submit more than one proposal.

v. A proposal with a secondary conflict should not be assigned if at all possible.

vi. All review assignments are confidential.

vii. There is no explicit requirement that the reviewers for proposal P_i are balanced according to executive share, as they are for the current review process. This cannot be achieved since the regional distribution of PIs are not necessarily in executive balance.

2.4. Stage 1

Once the PIs receive their proposal set, they begin the Stage 1 process. The Stage 1 process consists of (i) reviewing their assignments for potential conflicts, (ii) entering in comments for each proposal, and (iii) and ranking the proposals.

Identify conflicts of interests

i. The reviewer clicks a button acknowledging that all review materials are confidential and that they will be behave in an ethical manner.

ii. The reviewer reads the abstract and authorship list of the proposals to identify any potential conflicts not caught in the automated assignments. If the PI declares a conflict of interest on a proposal, they must declare the reason for the conflict.
   a. A new proposal could be assigned automatically and immediately after the conflict if declared. The concern is that reviewers will declare conflicts merely to get another proposal; e.g., they do not feel they are qualified to review the assigned proposal. The other option is to submit the conflict so that the justification can be reviewed by the JAO, and then a new proposal is assigned.
   b. Declaring a conflict does not extend the deadline to submit the comments. It is the PI's responsibility to identify the conflicts early.

iii. After a proposal set has been declared free of conflicts for both the PI and the co-reviewer, they can view the PDF files. It is possible that a PI may declare a conflict at this stage as well.

Proposal comments

iv. The PI reads each proposal and enters brief comments on the strengths and weaknesses into the proposal tool.
v. A PI can modify and resubmit their comments at any time up until the Stage 1 deadline.
vi. PIs may request a technical assessment for any proposal. The assessments will be posted on the web page for each proposal so that all reviewers for that proposal can read the feedback.
vii. The Stage 1 deadline will be 4 weeks after the JAO releases the proposal assignments.
viii. The reviews must be completed by the Stage 1 deadline, or else the PI's proposal will be rejected.

Proposal ranks

ix. For each assigned proposal set, the PI will rank the proposals from 1 to *N*, where 1 is best proposal in the set and *N* is the worst. If a PI submits more than one proposal, they will assign ranks to each set separately. Only integer values can be assigned, and no rankings may be duplicated. A ranking, and not a score, is used to ensure all reviewers use the same scale and avoids the need to renormalize the scores.
    a. The Gemini scoring system is that each proposal is ranked 0 (poor proposal) to 4 (excellent proposal). There does not appear to be any requirement that all scores are used or balanced.
x. A PI can modify and resubmit their ranks at any time up until the Stage 1 deadline.

2.5. Stage 2 (optional)

This is a variant of the basic model where the PIs enter comments only in the Stage 1 process and do not rank the proposals. Stage 2 then proceeds as follows.

i. The PI reviews the comments from the other reviewers for the proposals in their set.
    a. One could imagine an interactive message board, but I envisioned here it is a static list of comments.
    b. The comments are display anonymously, but the PIs own comments should be clearly indicated.
ii. After reviewing the comments, the proposals rank the proposals as described in Section 2.4.
iii. The deadline to submit the Stage 2 ranks is two weeks after the comments are opened for viewing.
iv. The web page should have a "Save" and "Submit". "Save" saves the rankings as currently entered by the PI without any validation. When the PI clicks "Submit", the rankings are validated and saved.
v. A reviewer can modify and resubmit their ranks at any time up until the deadline.

The two-stage process allows PIs to determine if other reviewers had particularly compelling arguments in favor or against the proposal that may alter their perception of the proposal. However, it adds a fair more work to the review process since the PI will need to read N^2 comments, and it lengthens the review process overall since there are now two deadlines.

2.6. Global ranked list

The average proposal rank shall be used to produce a global ranked list of proposals. Merrifield and Saari (2009) suggested using the modified Borda count. I believe the procedure below is equivalent (but that needs to be confirmed), and takes into account that not all proposals will necessarily have the same number of reviews.

i. Rankings are only used for proposal sets that have been submitted and validated.
ii. The rankings for set S_i are normalized by the number of proposals to range from 0 (best) to 1 (worst).
iii. The preliminary average rank and standard deviation are computed for each proposal.
iv. A parameter Q_i (see Merrifield and Saari) is computed for each proposal set that measures how close the reviewer rankings match the consensus rankings.
    i. $Q_i = 1 - 1/\text{int}(0.5\, N^2) \sum_{j=1}^{N} | j - rank\ of\ j\ in\ global\ list |$
    b. Q_i=0 implies there is no correlation between the reviewers list and the consensus rankings and Q_i=1 implies a perfect correlation.
v. The average rank is re-computed for proposals after rejecting outliers. This will likely require experimentation, but possible algorithms to exclude outliers are:
    a. Exclude reviewers with the lowest 10% (TBD) of Q_i.
    b. Combined with (a), exclude the lowest and highest rank before averaging.
vi. A list of proposals sorted by the average rank is generated.
    a. Proposals are removed for PIs that have not submitted their ranks.
    b. Ties are broken by the following order:
        i. Proposal with the better median score
        ii. Proposal with the best individual rank
        iii. In order of proposal number

    Simulations (see Section 3.2) suggest that if N=16, then 13% of the proposals will have duplicate ranks after apply (i) and (ii).
vii. Merrifield and Saari (2009) proposed that Q_i be used to identify PIs that did a good job in the review process, which is measured by how well their ranked list matches the consensus ranked list from all reviewers. The reward is that their own proposal will be bumped up in the rankings by a modest amount. The purpose of the award system is to discourage PIs from trying to game the system by assigning poor ranks for good proposals in an attempt to boost their own proposal. A criticism of this reward system is that it penalizes reviewers who may have justifiably downgraded a proposal for a reason other PIs may have missed. I propose that a reward system is not implemented until more experience is gained.
viii. No specific balancing by category is attempted by this procedure. The alternative is to generate a ranked list of each category, and then merge the lists as we do now, but this would seem to be an unnecessary complication.

3. How many proposals should a PI review?

A critical decision point is how many proposals (*N*) should a PI review in each proposal set. This section considers both analytic consideration and simulations to suggest that *N* ~ 8-16, with a preference for the upper range.

3.1. Considerations

The value of *N* is a compromise between several factors:

i. *N* should be large enough to produce an accurate measure of the mean rank;
ii. *N* should not be too large or it would discourage the PI from conducting a careful review;
iii. *N* should not be too large or it would hinder the PIs from recalling the proposals and producing a carefully ranked list.

Since reviewers are required to rank proposals from 1 to *N*, every reviewer will be required to specify a top ranked proposal regardless of the intrinsic merit. I posit that *N* should be large enough that each set likely contains at least one meritorious proposal, which is defined as the top quartile since the oversubscription rate is approximately 4. This is partially for perception: if a set contains only poor proposals, it may undermine their confidence in ALMA science. For example, if a set contains *N*=4 proposals, only 68% of the sets will have an intrinsically top quartile proposal. Similarly, for *N*=8 proposals, 90% of the sets will contain a top quartile proposal, 97% for *N*=12, 99% for *N*=16, and 100% for *N*=32.

As an additional consideration, from the reviewer perspective, one quarter of the proposals in a set will be ranked high enough to be scheduled given the oversubscription rate is ~ 4. I propose that *N* be large enough such that it is likely that *N*/4 proposals are in the top half of intrinsically meritorious proposals so that a reviewer is not forced to promote a poor proposal in the bottom half of the rankings. For *N*=4, 93.8% of sets will have ≥ 25% of the proposals in the top half of the intrinsic rankings. The corresponding values for *N*=8, 12, 16, and 32 are 96.5%, 98.1%, 98.9%, and 99.9% respectively.

If we arbitrarily require that > 95% of the sets contain (i) an intrinsically top quartile proposal and (ii) ≥ 25% of the proposals are in the top half of the intrinsic rankings, the minimum number of proposals in a set should be N~12. The maximum number of proposals that would be needed in a set is N~32, since essentially all proposal sets would satisfy these criteria.

A critical question is at what is how many proposals can a PI reasonably manage in a set. We want a minimum of 8 reviews such that the statistical averaging at least matches the current practice, but prefer a large number of allow more rigorous outlier rejection and compensate for a lack of a face-to-face review. Based on these grounds, I advocate for *N* ≥ 12.

3.2. Simulations

To simulate distributed peer review (see also Merrifield and Saari 2009), I created a list of N_p=1600 proposals that are assigned with an intrinsic ranking of 0 (best) to 1 (worst). I assumed that the probability that any reviewer can recover the intrinsic rank is given by a Gaussian with dispersion σ. I assumed there are 3 types of reviewers: i) experienced reviewers with σ=0.15N_p; ii) inexperienced reviewers with σ=0.3N_p; iii) biased reviewers with σ=0.15N_p, who systematically place their best proposals at the bottom of the rankings and the worst proposals at the top in an attempt to manipulate the outcome. I assumed 45% of the reviewers are experienced, 45% are inexperienced, and 10% are biased. (Personally, I believe the percentage of biased reviewers will be far lower, but I wanted to present a pessimistic simulation.) In addition, to reflect that some proposals are misunderstood by the reviewer, I assume that 25% of all proposals are misjudged and are randomly assigned a rank between 1 and *N*. This allows for a wider tail to the distribution of ranks than a pure Gaussian.

Figure 2 compares the intrinsic ("true") ranks vs. the PI ranks for various values of *N* (top row) for the simulation. The second row shows the histogram of *Q* for the reviewers, where the peak in the distributions around *Q*≈0.7 reflects the experienced and inexperienced reviewers, and the peak near *Q*≈0.1 represents the biased reviewers. The third row compares the true and PI ranks after removing reviewers with outlier proposal rankings by removing the lowest 10% of *Q* values. The bottom row shows the cumulative distribution of true proposal ranks for the high ranked proposals from the review panels.

These simulations show that for *N*=8, about 29% of the top ranked proposals will be replaced because of the imprecise nature of the review process. This is similar to that replaced in the Cycle 5 proposal review process assuming the Stage 2 ranks represent the "true" rankings. However, it should not be interpreted that the model presented here accurately reflects the Cycle 5 reviewers. In particular, the correlation between the Stage 1 and Stage 2 rankings for high ranked proposals is stronger in the Cycle 5 review process than present in the simulation.

Figure 2 shows that the filtering outlier ranks by the *Q* parameter reduces the number of replacement proposals to 25% for *N*=8. The replacement level improves with large *N*, even without implementing outlier rejection. Also, with *N*=12, only one proposal with a true ranking in the bottom half has promoted to the top quartile, while with *N*=16, all of the proposals rated in the top quartile are also in the top half of true ranks.

Based on the above analysis, I propose that *N*=12 or 16 for the proposal review process. This will ensure that nearly all reviewers have a proposal set that samples the full range of proposal ranks. *N*=32 would provide for more robust rankings, but I feel that this is too many proposals for a given reviewer.

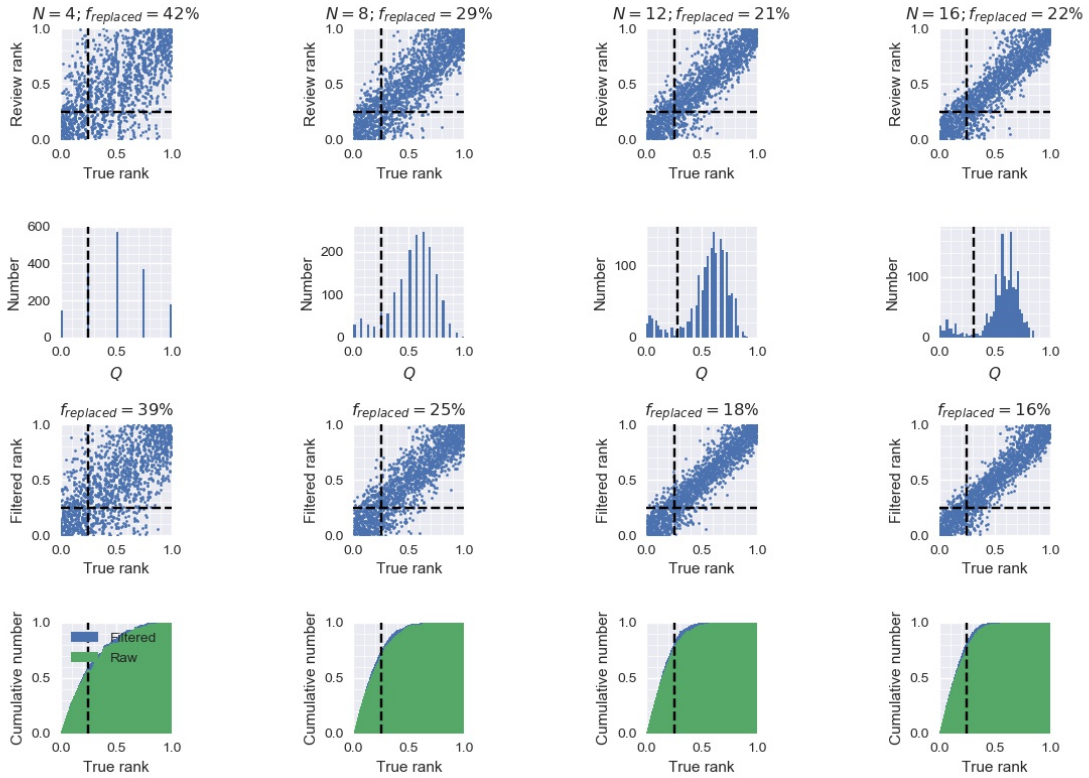

Figure 1: Simulated proposals ranks with the distributed review model for N_p=1600 proposals. Each column indicates the results of the simulation for a review of N proposals per reviewer (N=4, 8, 12, 16). *Top row*: Comparison of the true rank versus the rank from the review process; see text for details. A proposal rank of 0 is the top proposal. The dashed line indicates the 4:1 oversubscription rate typical for ALMA. $f_{replaced}$ indicates the fraction of truly high ranked proposals that were replaced by lower ranked proposals as a result of imperfect review ranks. *Second row*: Histogram of the Q parameter that measures how the ranks of individual reviewers matches the consensus ranks (Q=1 implies perfect agreement). The vertical dashed line marks the lowest 10% of the reviewers that were removed before computing revised ranks. *Third row*: Same as the top row, but after removing reviewers with the lowest 10% Q values. *Fourth row*: Cumulative distribution of the true proposal ranks for proposals that would be deemed high rank by the review process. The green histogram are the ranks for no outlier rejection, and the blue histogram is with outlier rejection.

4. Review of Large Programs

Given the considerable time investment, Large Programs will undergo a face-to-face review so that the strategic value of the proposals can be carefully considered. The review shall closely follow current practices.

Membership

i. There will an APRC chair to moderate the Large Program discussion. The APRC chair will be a non-voting member.
ii. The review committee will consist of 4 experts each in cosmology, galaxies, star formation, and disks. In Cycle 5, there were no Large Programs for stars, so it is a bit unclear whether reviewers should be included for this category.

iii. Each committee member will serve two years, with approximately 50% replacement from cycle to cycle.
iv. The reviewers should be in approximate balance according to executive share.

Review process

i. The Large Program review committee will meet two weeks after the Stage 1 deadline to review the Large Programs.
    a. This is the same as the Stage 2 deadline, if there is two stage process. While some of the committee members will have submitted regular proposals, they will be able to delegate their review responsibilities.
ii. At the Stage 1 deadline for the regular proposals, Large Program reviewers will enter comments for the proposals in their category.
iii. The comments on all Large Programs are made available to the APRC at the Stage 1 deadline.
iv. The reviewers will meet for 2-3 days to review the Large Programs.

Comments

One concern is that the Large Program reviewers will not have any context with the regular proposals, especially the medium-sized proposals. There is also the possibility that an approved Large Program will duplicate in spirit the medium-sized proposals. One possibility is to allow the Large Program review panel to view the medium-sized proposals and their rankings. The critical aspect is what will they do with this information? Will they able to reject a medium size proposal because it overlaps scientifically a Large Program? Can they reject a medium size program regardless of its rank?

5. Concluding remarks

5.1. Timeline summary

The timeline from the proposal deadline to the PI notifications is as follows:

i. 1 week for the JAO to remove erroneous proposal submissions.
ii. 1 day to create proposal sets for each PI
iii. Stage 1 review: 4 weeks
iv. Large Program review held one week after the Stage 1 deadline.
v. 3 weeks for the JAO to make the queue after the Large Program review
vi. 1 week for the Director's Council (DC)/Chile to approve the results
vii. 1 day for the JAO to release the results after DC/Chile approval.

Thus, the results are released 10 weeks after the proposal announcement and 5 weeks after the reviewers submit their rankings. This compares to ~12 weeks for the current review process, and implies the proposal deadline should be pushed back by two weeks and otherwise maintain current schedule for PIs to complete Phase 2 preparation. Options to reduce the time further is to have the Large Program review earlier or to further streamline the queue building and the DC/Chile approval process.

5.2. Critical questions

A number of critical questions need to be answered to fully define the process for distributed peer review. Here is a summary that were raised in the memo.

i. How many proposals (*N*) shall a PI review for every submitted proposal?
ii. Should this be a two-stage process where comments are entered and then ranks are determined, or should comments and ranks be entered in a single step with no review of comments?
iii. To encourage good reviews, should a PI be rewarded with a boost in the rankings if the rankings in their proposal set match the consensus?
iv. Should non-conflicted PIs review and rate (Excellent / Very Good / Good / Fair / Poor) one Large Program in addition to ranking *N* regular proposals?
v. Should medium-sized proposals receive any review by the APRC?
vi. Should the reviewer comments be sent as-is to the PIs? (It will not be practical for the JAO to moderate all 1600**N* comments.)