Peer Review and the HST TAC

Neill Reid

SMO

Allocating Observing Time

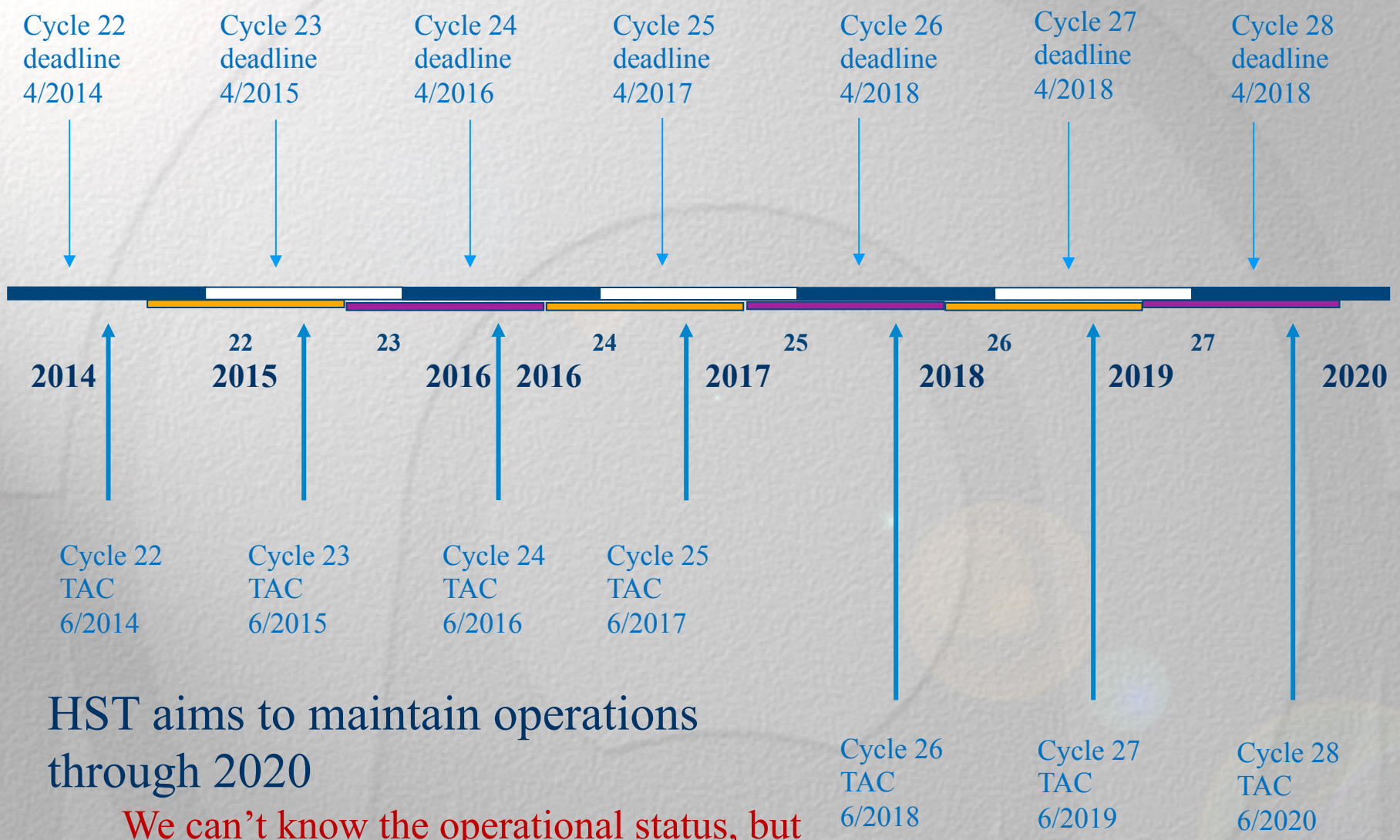
The prime task facing every observatory is devising an equitable and efficient process for allocating observing time to it's user community

- The time allocation process aims to optimise the scientific return from the facility.
- The process is defined by the institute running the observatory, but is only successful if supported by the community

Most observatories employ a form of peer review to select proposals

- HST convenes a committee comprising ~140 scientists drawn from the US & international community
  - ~1,000 – 1,100 proposals per cycle
  - ~60-90 proposals per panelist
- ESO, Chandra, ALMA are facing similar workload
- Looking forward, the HST community will face an additional challenge

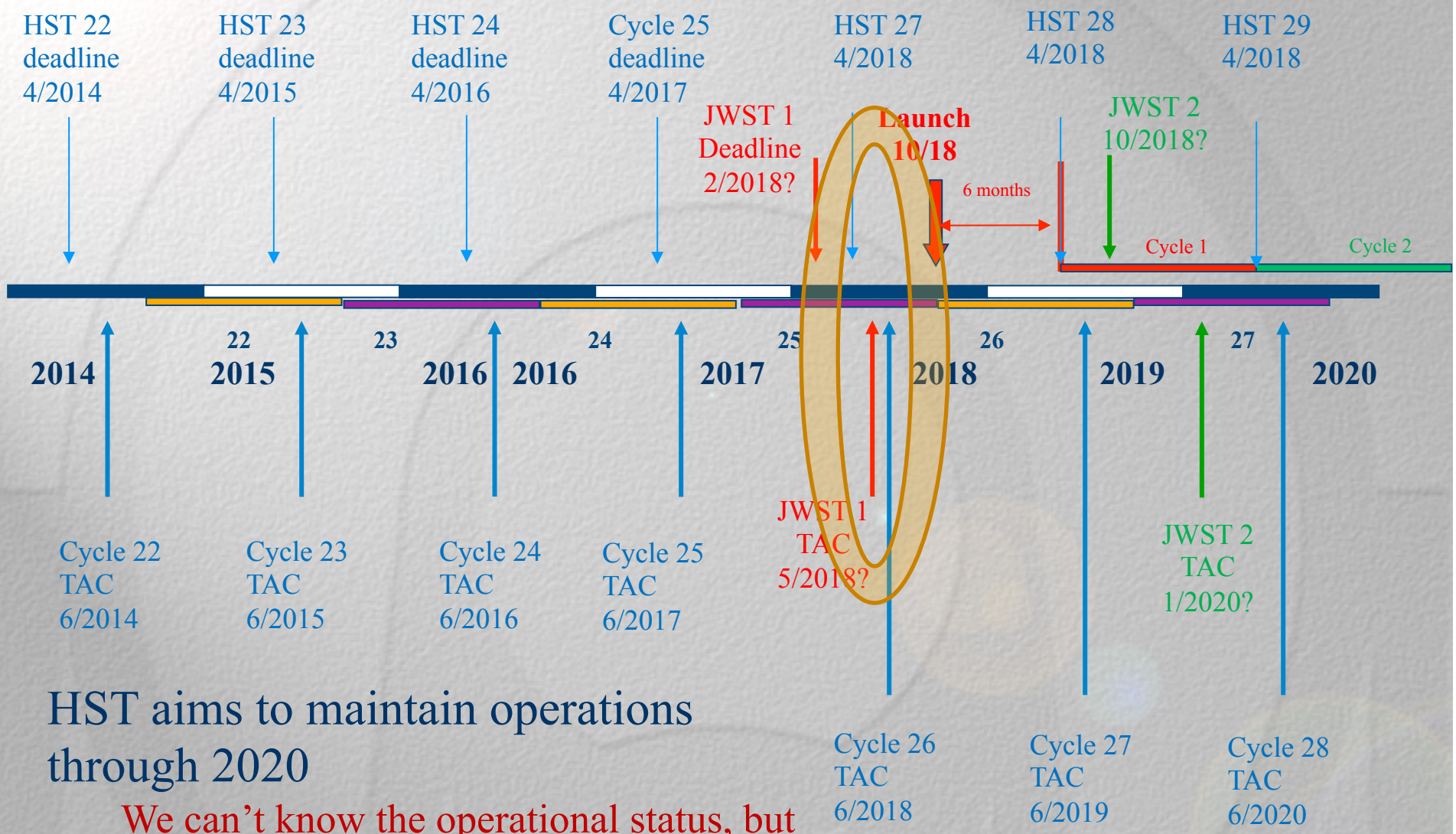
Taking the long view



HST aims to maintain operations through 2020

We can't know the operational status, but need to plan for a best case scenario

Now add JWST



HST aims to maintain operations through 2020

We can't know the operational status, but need to plan for a best case scenario

Allocating Observing Time

The prime task facing every observatory is devising an equitable and efficient process for allocating observing time to it's user community

- The time allocation process aims to optimise the scientific return from the facility.
- The process is defined by the institute running the observatory, but is only successful if supported by the community

Most observatories employ a form of peer review to select proposals

- HST convenes a committee comprising ~135-145 scientists drawn from the US & international community
  - ~1,000 – 1,100 proposals per cycle
  - ~60-90 proposals per panelist
- ESO, Chandra, ALMA are facing similar workload
- Looking forward, the HST community will face an additional challenge:
  - JWST will clearly attract significant attention

We are currently developing the JWST proposal process

- There are no plans at present for revisions to the HST process, but we may need to consider options in the future

The main purpose of this presentation is twofold,

- Inform the STUC what we have learned from analysis of results from the HST process
- Lay a foundation for future discussions of possible procedural changes

The Peer Review process

Peer review is frequently used as a selection process by scientists:

- Grant funding
- Telescope time
- Refereed publications
- High-level postdoctoral fellowships

Peer review can be implemented in three ways:

- **Individual reviews**, usually written, submitted to a central source (editor, review coordinator)
  - Quasi-independent assessments (editors may provide 2nd referees with the 1st review)
  - Direct feedback to proposers
- **Remote grading** by multiple reviewers/referees
  - Independent assessments
  - Decisions are based on averaged grades/ranks, i.e. an electoral system
  - Limited options for feedback
- **Panel reviews**, conducted (at least partly) in an interactive environment
  - Decisions are based on a consensus ranking
  - Consensus feedback to proposers

In all cases, it is important to bear in mind the human element
Peer Review is a SUBJECTIVE process

The HST Process

The current process incorporates remote grading and panel discussions:

- Proposals are assigned to panels based on science topic and distributed to all panel members
- Panelists submit preliminary grades for all proposals (except in cases of conflicts) ~1 week before in-person meeting
- Preliminary grades are used to construct a rank-ordered list and the lowest 40% of proposals marked for potential triage
- Panelists meet to discuss and independently re-grade proposals, including those revived from triage
- Once the final ranked list is available, panels can re-balance to allow for science balance and duplications with mirror panels
- Final list is accepted as the consensus view of the panel and passed to the Director for approval

The Distributed Review Process

Merrifield & Saari (2009) have proposed a variant of the remote grading approach to reduce the workload for reviewers:

- Reviewers are drawn from the Principal Investigators
- Each PI agrees to review ~10 proposals for each proposal submitted
  - i.e. Each proposal receives 10 independent grades
- Failure to submit reviews results in disqualification of the PI proposal
- PIs are given an incentive to set aside individual bias by rewarding those whose proposal ranking closely matches the final averaged rank order
- Modified Borda statistic

Equation 1

$$Q_i = 1 - \frac{1}{\text{int}(0.5m^2)} \sum_{\substack{\text{applications} \\ \text{in } i\text{'s list} \\ j=1}}^{m} |\text{rank of } j \text{ in } i\text{'s sub-list} - \text{rank of } j \text{ among these } m \text{ in global list}|$$

High Q indicates a strong correlation between the individual ranked list and the averaged rank order list.

The DRP incentive plan

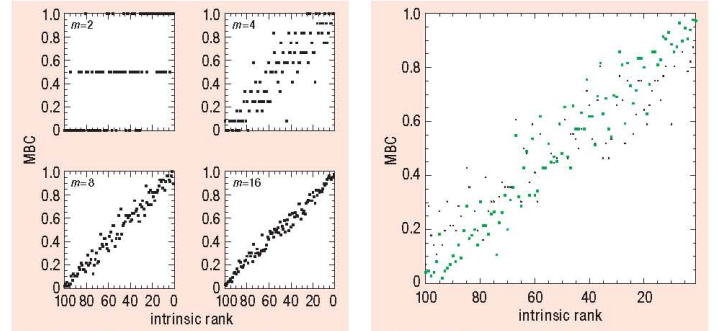
Merrifield & Saari state "there is no objective right answer in this kind of peer review process" – i.e. there is no absolute ranking *de jure.*

However, they assume that, given 100 proposals, "reviewers are fundamentally unable to distinguish between applications within Δn=10 places of each other..but can otherwise rank each application fairly against its competitors" – i.e. they assume a *de facto* consensus ranking with individual uncertainties of ~10%.

Under those assumptions, competent referees score $Q > 0.6$

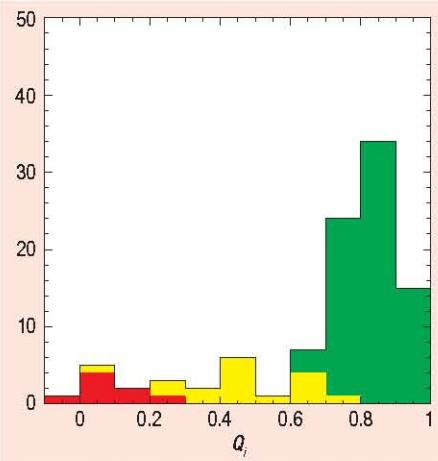
Evil referees score $Q < 0.3$, incompetent referees score $0.3 \leq Q \leq 0.6$

Average scores quickly converge



2: Modified Borda count returned when a sample of 100 applications of known rank are sent to differing numbers *m* of the applicants for assessment.

3: Modified Borda count returned in a more realistic mix of 85% competent, 10% incompetent, and 5% deliberately mis-ranking referees. Green points show the improvement in fit obtained after the identified "evil" referees are clipped from the scoring.



5: Distribution of measured refereeing quality as defined by the *Q* parameter. Competent referees are coloured green, incompetent ones are yellow, and evil ones are red.

Results from the HST Process

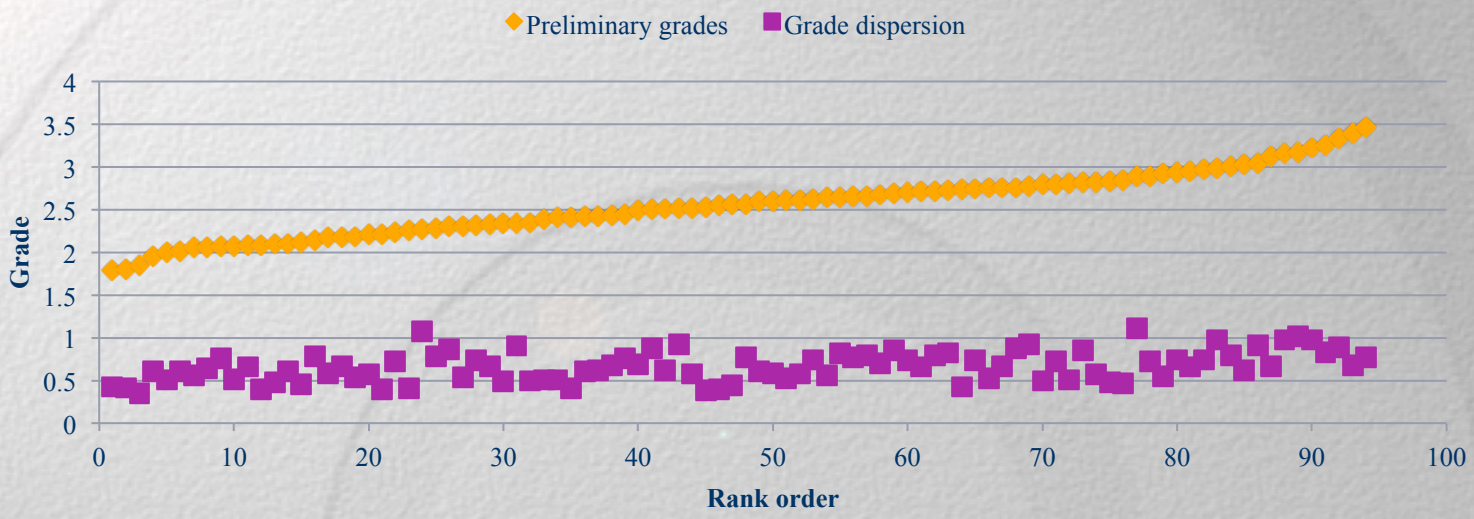
Details regarding discussion and specific grades of individual proposals submitted to the HST TAC remain proprietary

- However, provided we maintain anonymity, we can use the distribution of results to probe statistical quantities, such as the dispersion in grades with proposal rank and the range of ranks assigned by individual reviewers; we can track how those parameters change through the review process, and compare results from the preliminary grading against the final rank-ordered list.
- We cannot directly test the Distributed Review Process; *we can, however, use the present results to set a baseline.*

We have analysed data from the Cycle 21 TAC, paying particular attention to the results from one panel

- Analysis of other panels and data from previous TACs shows that these results are representative

Dispersion in average grades

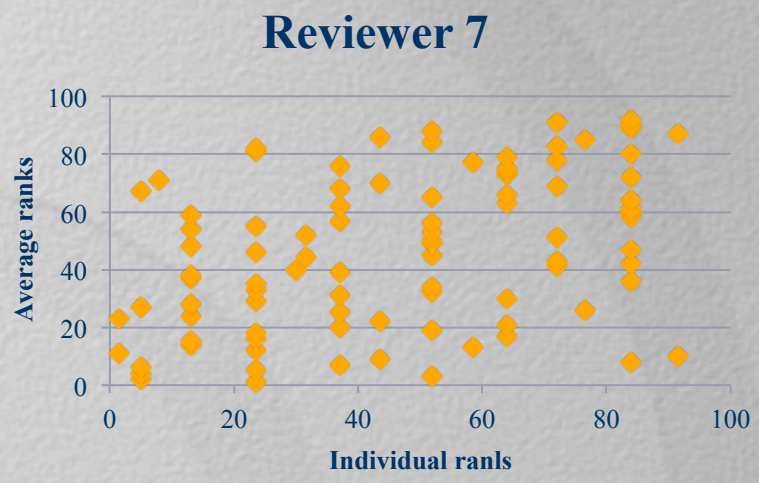


Panelists are asked to provide preliminary ranks from 1-5 for proposals, where 1=good, 5=poor.

We do not impose a particular system, but ask that panelists use the full range available.

The dispersion in grades tends to be lower for highly-ranked proposals, and increases (slowly) towards lower rankings; there is significant dispersion

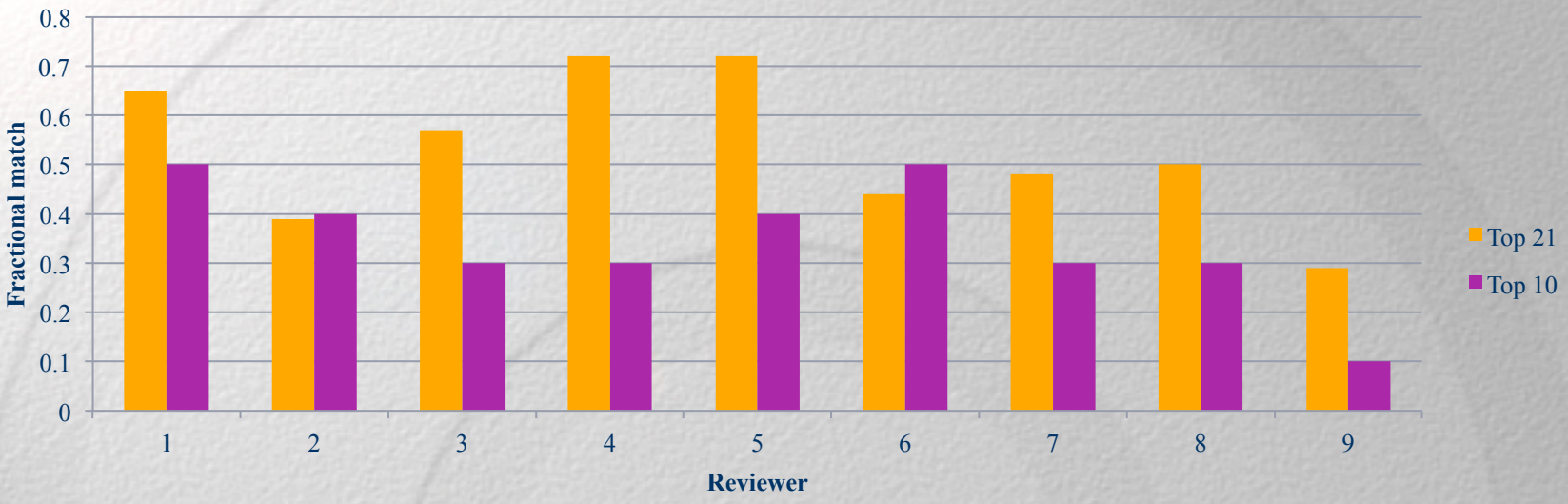
How well do panelists agree on the preliminary ranking?



We can match the rank-ordered lists from individual reviewers against the list based on averaging the preliminary grades

➔ There are substantial differences in how each reviewer ranked the proposals

Do panelists agree on the top proposals?



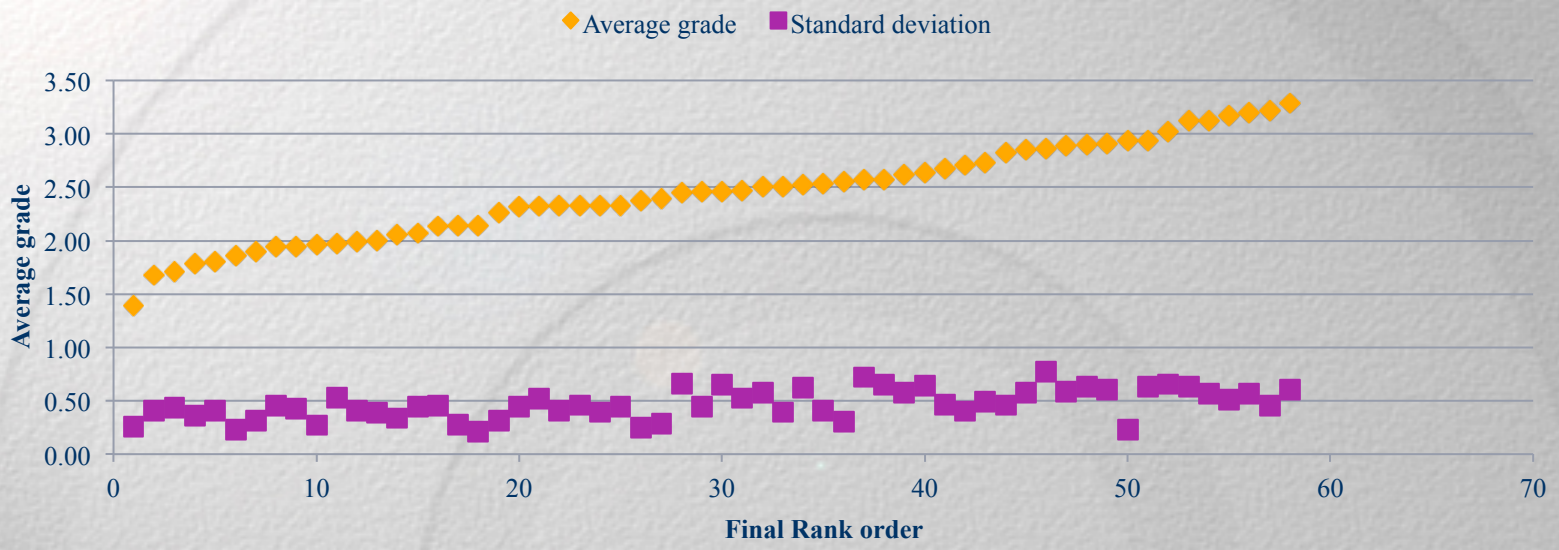
Tradition holds that the best proposals are easiest to identify

- We identified the Top 10 and Top 21 proposals based on the averaged preliminary grades from Panel X
- The figure shows the fractional overlap between the averaged list and the Top 10 and Top 21 lists for individual panelists

None of the panelists had more than 5 Top 10 proposals in common with the averaged ranking

- i.e. panelists come to the TAC meeting with significantly different views on which are the strongest proposals.

Dispersion in final grades



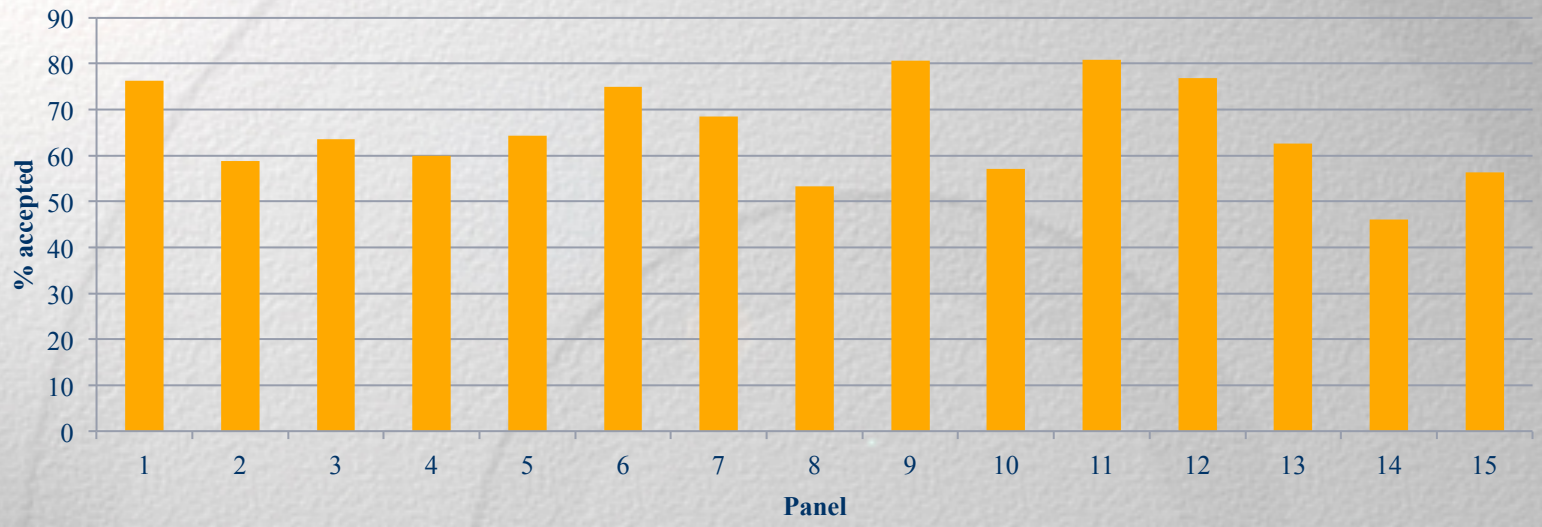
Dispersions for the proposals ranked by Panel X

Note that ~35 proposals were triaged

Overall, the dispersions decrease showing greater agreement among the panelists, with a milder trend to increased dispersion at lower ranks.

However, only a handful of proposals have $\sigma < 0.3$

How well do the preliminary and final ranked lists agree?



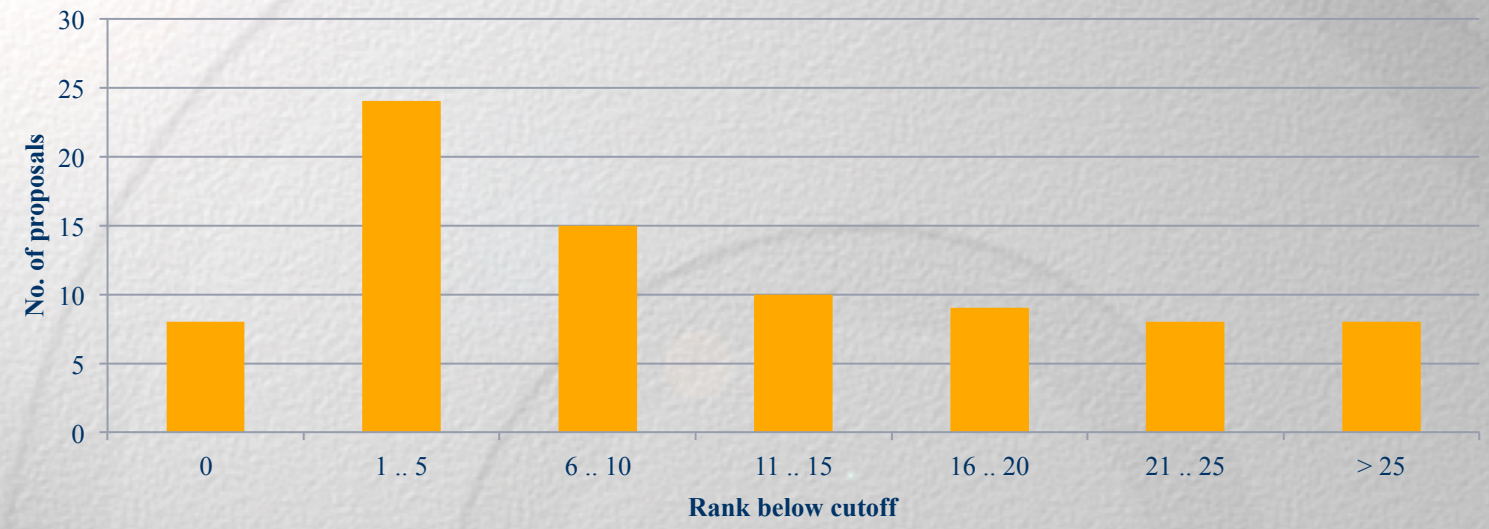
Each panel allocates time to N proposals

What fractions of those proposals would have been awarded time had we used the preliminary grades to select accepted proposals?

Overall, 252 proposals were accepted in Cycle 21; 170 (67%) would have been accepted based on the preliminary ranking

The overlap ranges between ~45% and ~80% for the individual panels

What happens to the "missing" proposals?



What happened to the 82 proposals that didn't make the final cut?

The figure shows their ranking relative to the cutoff in each panel:

- 8 proposals were highly ranked, but eliminated as science duplicates
- 24 proposals slid just below the cutoff
- 35 proposals were ranked at least 10 below the cutoff, reflecting a significant re-assessment based on the panel discussion

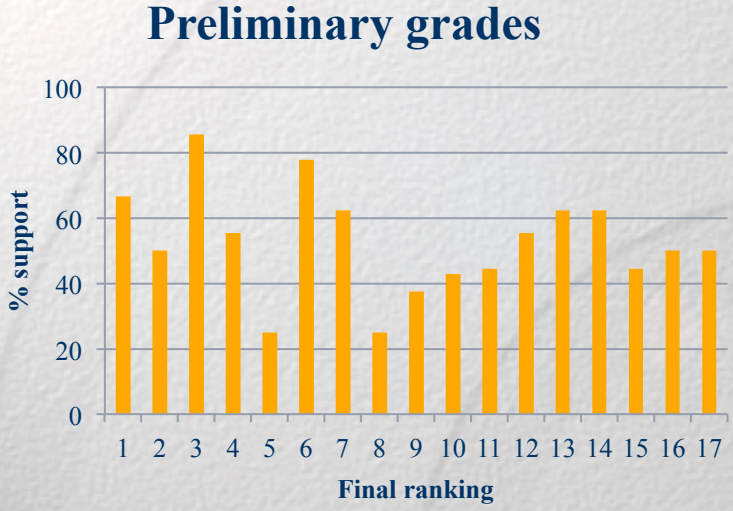
Thus 14% of the proposals selected based on preliminary grades were ranked close to the third quartile or lower in the final grades.

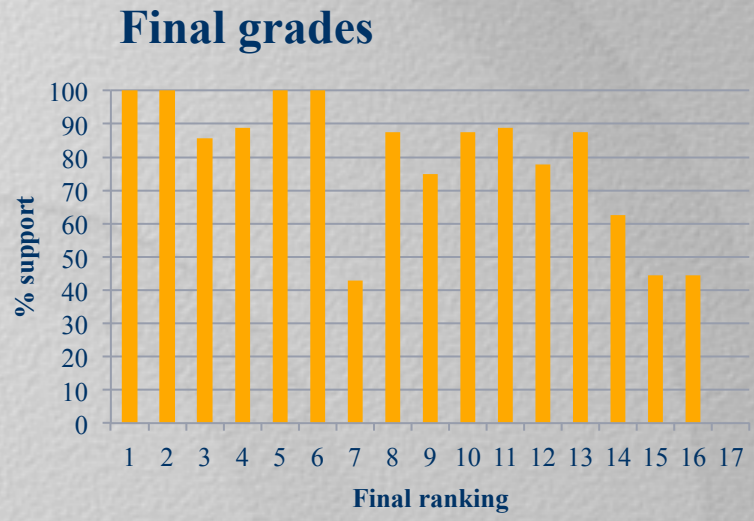
How well do individual panelists agree after the discussion?

As with the preliminary grades, we can compare the final ranked list against the results from individual reviewers

Overall, the agreement is closer, but significant differences remain in the rankings by individual reviewers.

Panelists build consensus on the top proposals



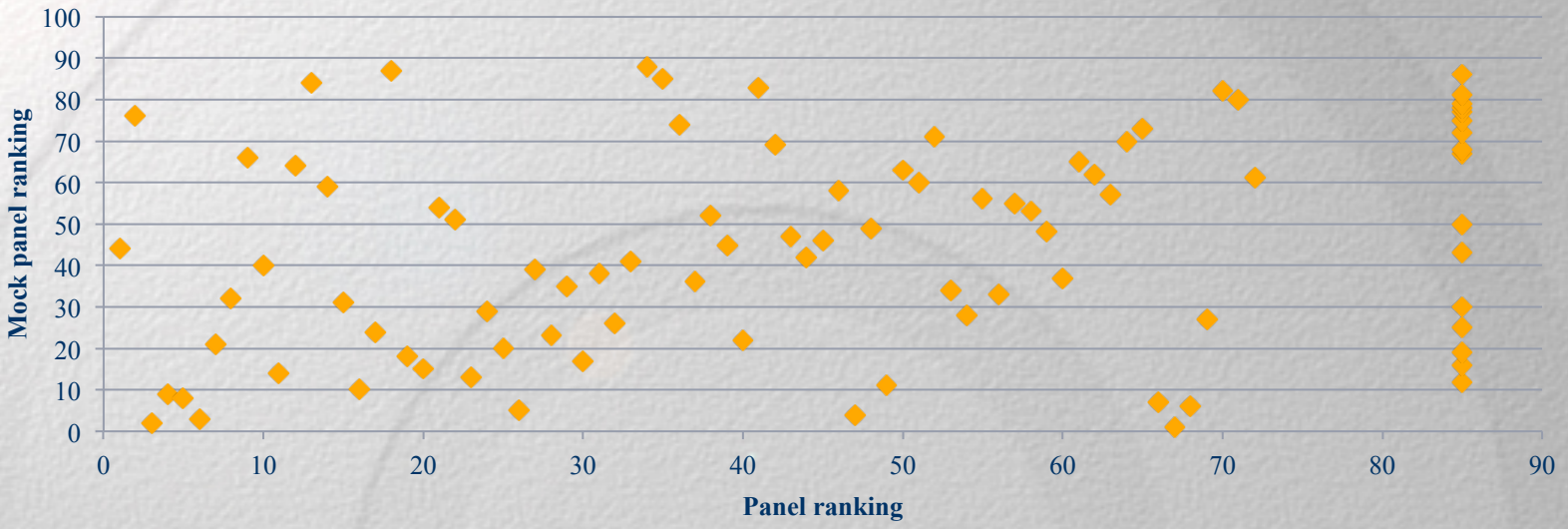


Panel X recommended 17 proposals for acceptance

- The figures show the panelists support, proposal by proposal, based on the preliminary grades and the final grades: i.e. we show the fraction of panelists who put each proposal in their individual top 17.
- Support for the top proposals clearly grows following the discussion.
- Every panelist included at least 60% of the top 17 in their personal top 17

Viewed as a package, each panelist "won" more than they "lost" in the final results from the panel.

What happens if different panels consider the same proposals?



One test case:

- Mock panel comprised of postdocs & research staff at STScI in the late 1990s
- Reviewed same proposals as an HST TAC panel
- Individual rankings show significant differences – BUT
- 16 of the top 31 proposals (accepted or sent to TAC) are in common
- ➔ Combinatorics indicates a probability of ~10^{-8} that this happens by chance

Greater agreement in the package than in the individual rankings

HST and the Distributed Review Process

We can analyse the preliminary and final grades using the modified Borda statistic – the table also shows the top 21 and final acceptance fractions for each panelist

Panelist	1	2	3	4	5	6	7	8	9
Q_{prelim}	0.55	0.56	0.58	0.66	0.61	0.60	0.52	0.51	0.51
N_{T21}	65%	39%	57%	72%	72%	44%	48%	50%	29%
Q_{final}	0.75	0.79	0.70	0.70	0.82	0.73	0.66	0.70	0.73
F_{acc}	87%	91%	76%	82%	88%	77%	62.5%	65%	76%

borderline competent based on the preliminary grades

•Even after the discussion, only one panelist achieved Q > 0.8

The results indicate that the assumption of 10% agreement in the individual ranking underestimates the dispersion in ranking among HST panelists

The HST process does not include a personal incentive for matching the average, but these results clearly indicate the modification in behaviour that is required for the DRP incentive to be a useful addition.

The Distributed review process is being utilised by the NSF's Sensors & Sensing Systems Program in late 2014, and is being considered as part of Gemini Observatory's rapid turn-around program, starting in 2015. We look forward to seeing statistical analysis of the results from those programs.

Lessons Learned

- Discussion changes the results
  - One-third of the highest-ranked proposals from the preliminary grades are not recommended for acceptance
  - ~15% are demoted significantly in ranking
- Individual reviewers differ significantly in their indepednent assessment of the relative rankings of proposals
  - Those differences persist at a lower level after discussion
- Discussion leads to support coalescing around a set of proposals
  - No-one supports every proposals, but everyone supports ~2/3rds of the proposals recommended for acceptance

Overall,

Peer Review remains a SUBJECTIVE process

Panels develop support for a package

Implications for HST

There are no plans to change the HST TAC system at present.

Any future changes should aim to:

- Recognise the dispersion in views of individual reviewers
- Retain the "package" aspect of the review process

One possibility is a two-phase approach

- Phase I: remote grading by pre-selected reviewers
  - Each reviewer receives ~30-35 proposals
    - Provides a perspective on the proposal quality
  - Proposals are graded against specific criteria, e.g.
    - Timeliness of the proposed science
    - Overall science impact
    - Suitability of the proposal team
- Phase II: panel discussion of a subset of the proposals, e.g.
  - Top ~20% of standard proposals? Top 30% of Large/Treasury programs?
  - ~5% of borderline proposals with significant dispersion?
  - Allow for overall science balance and duplications

A topic for discussion at a future STUC meeting