ESO TIME ALLOCATION WORKING GROUP

REPORT

Presented to the ESO Director for Science

December 2016

1 Composition of the Working Group

The **T**ime **A**llocation **W**orking **G**roup (hereafter TAWG) was composed as follows:

Almudena Alonso Herrero – Centro de Astrobiología (CAB), Spain (STC delegate)

Antonio Chrysostomou – SKA, Science Operations Planning, UK

Stefan Janssen – Paul Scherrer Institute, User Office, Switzerland

Martin Kessler – ESA, Science Operations, Spain

Arvind Parmar – ESA, Science Operations, Spain[1]

Rachel Mason, Gemini Observatory, USA[2]

Ferdinando Patat – ESO, Observing Programmes Office, Germany (chair)

Neill Reid – Space Telescope Science Institute, Science Mission Office, USA

Damien Sègransan, Université de Genève (UC delegate)[3]

Stephen Smartt, Queen's University Belfast (UC delegate)[4]

[1] Replacing M. Kessler as of September 2016.
[2] Member until January 2016.
[3] Member until July 2016.
[4] Replacing D. Sègransan as of August 2016.

2 Introduction

The TAWG reviewed the whole time allocation process currently implemented at ESO, examining the procedures and analyzing a number of statistical indicators. The recommendations that follow were formulated in the attempt of achieving two goals:

1) Maximize the scientific return of ESO by selecting proposals that have the promise of leading to significant advancements;
2) Improve the level of feedback provided to the community by the OPC.

Given the current load of the panel members, the TAWG reached the conclusions that the most urgent measure ESO needs to take in order to achieve these goals is a drastic reduction in the number of proposals they have to review. Although there is no clear indication in the literature about the optimal number of proposals per referee, the current number (70 and more) is certainly above what is reasonable for performing a thorough and consistent review.
Most of the recommendations that follow, and in particular the first five, address this matter, and aim at encouraging: *a)* the reduction in the total number of proposals, *b)* the increase of the average telescope time requested per proposal, and *c)* the quality of the review process and the feedback provided.

The TAWG discussed at length possible radical changes to the way proposals are evaluated, also considering substantial departures from the peer-review schema (e.g. random selection in the mid-range of the rankings). Notwithstanding the known limitations of this paradigm, the TAWG concluded that peer-review still remains the most satisfactory way for selecting time applications. The choice of any time allocation committee (TAC) is unavoidably subjective, and a different set of referees would most likely provide a different list of top ranked proposals, with only a fraction in common. However, in a schema in which the panels are nominated by the community, that selection still represents the expression of the underlying population. Based on

these considerations, the TAWG does not recommend changes in the detailed process, although it is advocating some substantial modifications.
It is also the opinion of the TAWG that distributed approaches that "penalize" the reviewers (like the Merrifield-Saari), should be avoided by ESO. Nevertheless, the TAWG is in favor of having a non-penalizing distributed review, which is a valid possibility to sustain the fast-track channel proposed here.

For each recommendation the advantages and disadvantages seen by the TAWG are listed. Suggestions for a possible implementation plan for the proposed recommendations are presented in Section 4.
When applicable, reference to relevant statistical analysis performed during the TAWG activities is provided in Sections 5 to 11. Section 12 provides the essential bibliography, while Appendix A contains the TAWG terms of reference.

For convenience, a summary of the previous OPC Working Group recommendations (Brinks et al. 2011) is given in Appendix B.

3 Recommendations

Recommendation 1 - Decrease the frequency of the call for proposals to one per year.

This recommendation was included in the OPC WG report (Brinks et al. 2011), and was rejected by ESO based on operational reasons (mostly related to the increase of the load on the Phase 2 review)[5]. In sight of the artificial pressure that a semester call puts on the community, the typical time scale of publication return, and the large number of resubmissions, there is no compelling scientific reason supporting a semester-based call (Sects. 8, 9, 10). The decreased frequency of

[5] The TAWG notes that of the seven recommendations made by the OPC WG (Brinks et al. 2011; see also Appendix B here), only one was introduced, i.e. the creation of Monitoring Programmes.

standard calls can be compensated with the creation of a new, fast submission and review channel (see next recommendation).

Advantages:

- Decrease the total number of proposal (<2x900. For instance, resubmissions caused by the artificial RA boundary set by the ESO periods would disappear).
- Presumably increase the number of larger time requests per submission, given that there is more time available at a given call.
- Decrease the annual load on the TAC and produce better reviews and feedback quality.
- Give the possibility of stretching the Phase 2 (now 4 weeks) to alleviate the higher peak created by the yearly cycle.
- Improve completion rate because of more observing flexibility during the allocation periods (by removing the artificial pressure at the semester boundaries).
- Potentially significantly reduce the cases (mostly B-rank in the current schema) that are scheduled in one semester, rejected (or not scheduled) in the next semesters, and re-submitted a third, fourth, … time to complete.
- Relax the timelines of proposal selection and long-term scheduling processes, so that they can be conducted in a more thorough way.
- Give the possibility of having a larger number of distinct referees per meeting (at the same cost), and so decrease the number of proposals per reviewer. This also enables the possibility of covering a larger set of scientific areas, hence allowing for a better proposal-referee matching (see also Section 4).
- Increase the reviewers' acceptance rate (committing for a yearly meeting is easier than for two meetings per year).
- Align ESO to what is already done at other comparatively large facilities (ALMA, HST, XMM, Chandra and Spitzer).

Disadvantages:

- Introduce a loss of some flexibility and response time (but see next recommendation).
- Increase the difficulty of planning ahead technical activities. This could be mitigated by a devoted group in charge of dynamical scheduling (as is already the case at other large facilities like HST and ALMA. See also Section 4).

Recommendation 2 - Create a fast-track channel (FTC) for the VLT, for fulfilling the need of a shorter duty-cycle.

This new peer-reviewed channel is in addition to the existing DDT channel and is not meant to replace it. This should provide a quick duty-cycle (timescales of months), allowing the users to obtain data for amounts of time below a given threshold (e.g. 20 hours). Like in the cases of already existing faster channels (e.g. Gemini, see Section 11), specific criteria should be applied to the eligible programmes as to why they could not be submitted through the regular channel. To ease the review and scheduling processes, the deadline should be periodic (i.e. once a month) and not continuous.
ESO should fix a maximum amount of time to be allocated to these programmes (see also Recommendation 3).

Advantages

- Compensates for more extended response time caused by the introduction of the yearly cycle.
- Potentially decreases the number of proposals submitted through the yearly call. The possibility of submitting short programmes (which are the majority in the current implementation, see Section 6) at the next fast-track deadline will naturally lower the pressure, as users will not feel compelled to submit. The case of the fast-turnaround

channel created at Gemini (Section 11.1) shows that approximately 20 proposals/months should be received via this channel.

Disadvantages

- The review of the applications has to take place within a relatively short amount of time (depending on the final periodic deadline cadence).
- Requires, at least to some extent, a schema for a dynamic scheduling in order to accommodate the newly approved fast-track programmes.

This recommendation was part of Recommendation 1 in the OPC WG report (Brinks et al. 2011).

Recommendation 3 – Radically change the proposal review procedure for the yearly cycle to decrease the load on the panel members.

In the current implementation, all proposals assigned to a panel are reviewed by all non-conflicted members of that panel. This turns into 85% of the proposals being reviewed by 5 or 6 referees. These referees are required to read and grade the proposals. This translates into a load of 70 or more proposals per panel member, reaching up to about 100 for OPC-proper members. It is widely acknowledged that this number is too high to allow for a thorough review and a satisfactory feed-back.

This recommendation goes along the lines already in place at other facilities (also outside Astronomy), in which only a fraction of the panel reads and grades the proposals in the pre-meeting phase. Here follows an outline of the proposed schema:

- Panel members review a subset of proposals assigned to their panel at the pre-meeting phase (properly matched to their expertise) and provide pre-meeting grades. At least 3 members of the panel should grade a proposal.
- Based on the pre-meeting grades, a ranking list is drawn up by telescope and a substantial triage (≈30%) is applied to the bottom, similarly to what is currently done.
- The ranked lists per telescope are circulated before the meeting. Panel members can then read all non-triaged proposals assigned to their panel. The discussion of triaged proposals at the meeting is strongly discouraged but not forbidden.
- The proposals ranked in the first quartile (or another suitable and configurable quantile) are identified from the pre-meeting grades, and are not discussed at the panel meeting. Their relative ranking should just be confirmed collegially by the panel. These programmes will compose the equivalent of the current A-rank class.
- The panels will then discuss and rank the proposals in the middle quartiles, which will mostly compose the equivalent of the current B-rank class.

This proposition is based on the observed fact that the grade dispersion for higher graded runs is lower than in the middle and bottom ranges (see Section 7).

Advantages

- Reduce the number of proposals to be read by each panel member. In the most extreme assumption that the total number of proposals per yearly cycle doubles the current value per semester, and that the number of referees is also doubled, this produces a factor 2 reduction in the pre-meeting work load. As the number of proposals is not expected to double, the decrease will be larger, hence reducing the number of proposals/reviewer below 35.
- Increase the time dedicated to the discussion of the most promising proposals in the central part of the distribution, and to their final ranking (which decides their priority at observing time).

Disadvantages

- The first, pre-meeting screening is based on a smaller number of reviews than in the current schema. It may therefore be more prone to strong opinions.

Recommendation 4 - Change the policy of Large Programmes on the VLT, allowing them more massive allocations and shorter completion time.

The current VLT Science policy includes a 30% upper limit for Large Programmes (LP) at the VLT. However, the time allocated for LPs is on average around 17% (a remarkable exception being the two VIMOS spectroscopic surveys implemented at UT3, which will be completed by P100).
LPs at the VLT can span up to four semesters. However, in a schema in which LPs can only be requested on a yearly basis (see Recommendation 1), there is no compelling reason as to why all the time should not be requested and allocated in one single cycle (barring time monitoring needs).
As the statistics show, LPs are the most productive programmes (even taking into account the large amount of time allocated to them; see Section 9), and should therefore be promoted and completed in the shortest amount of time compatible with their scientific requirements. Therefore, the TAWG recommends that:

- LPs should normally span only one year. Larger time spans have to be scientifically justified.
- ESO considers increasing/removing the 30% limit if the demand goes in that direction as a consequence of the changes proposed here.

Advantages

- Faster scientific return.
- Encourage larger applications.

Disadvantages

- Favoring larger/stronger collaborations, and forcing small groups to join forces (not necessarily a disadvantage, but may be perceived as a disadvantage).
- Less chances for ESO (via the OPC) to stop non-productive programmes.

With reference to the last point, it must be noticed that this situation is already present in the current schema for LPs requesting time on two semesters only, and this was never considered as an issue. In addition, and to the best of the TAWG's knowledge, the OPC has never terminated a LP based on poor productivity.

Recommendation 5 – Redistribute the time balance of time allocated to favor larger programmes and limit the fraction to be allocated to programmes shorter than a given threshold.

The rationale behind this recommendation is that the change in balance will encourage the submission of larger and more comprehensive requests. Conversely, this recommendation will limit the number of submissions of short proposals at the yearly cycle by increasing their rejection rate.

One can argue that the community will react just artificially increasing the requested times in order to get above the threshold. However, if this is set for example to 3 nights, about 25% of the current proposal would have to at least triple their times to "escape" the restriction introduced by this recommendation. Although not forbidden in principle, this will require a well substantiated scientific justification (as opposed to a purely artificial inflation of exposure times). In the current schema 50% of the programmes at the VLT request less than 2 nights, 95% less than 5 nights. There is a gap in the region 5-12 nights, with a very small request, and only about 17% of the time at the VLT is allocated to Large Programmes (Section 6). With this recommendation the TAWG is proposing ESO to consider a paradigm change. The TAWG is aware

this will not be a popular recommendation. However, after more than 15 years of operations and with about 50% of the NOR programmes that do not publish a refereed paper (Section 9), the TAWG argues it is time to reconsider the way time is distributed. Although it is certainly true that ESO serves a wide community (about 3000 distinct scientists every semester), the time allocation is very fragmented and the projects are all comparatively small, mostly producing incremental science. In the current situation, ~80% of the programmes request less than 3 nights: setting a maximum fraction of accepted programmes with t≤3 nights (for instance) will encourage the community to propose larger and more ambitious projects. The observed growth in the number of collaborators clustering around an almost constant number of PIs (Section 6) goes already in this direction, which needs to be encouraged by ESO.

In the same spirit, ESO should consider allowing the community to request Target of Opportunity (ToO) time extending over several cycles. Very often the targeted events are rare and lead to multiple, identical resubmissions just to reach the sample size required by the science case. Multi-cycle ToO allocations would reduce the number of proposals to be reviewed and ensure continuity to programmes related to comparatively rare and transient phenomena.

Advantages:

- The PIs would be incentivized to complete their programme at once (complete samples with no RA limitations, which becomes possible thanks to the yearly cycle).
- Possibly produce more productive programmes, increasing the VLT scientific return.
- Populate the 5-12 nights desert.
- Decrease the total number of proposals.

Disadvantages:

- Possibly discourages small teams and favors larger collaborations. Partial compensation is given by the fast-track channel.

The TAWG is fully aware that this is not a decision that can be taken by ESO alone, and it will require thorough discussions with the community, via the governing and advisory bodies. However, ESO is definitely in the position of initiating the process.

Recommendation 6 - Introduce a new programme type for fillers, with a light review process.

In the current implementation there is no filler channel. Fillers (aka C-rank) are selected by ESO purely based on their relaxed conditions and simple/standard observing strategy. Often, and especially for some telescope, the Observatory ends up having no or very few fillers.
Proposals submitted through this channel will have a very light review (in order to encourage submission), with a binary accepted/rejected outcome, with no ranking and no feed-back to the PI. Only proposals asking for loose constraints (to be defined by the observatory and instrument dependent) will be accepted (proposals violating the constraint on conditions can be automatically rejected). The filler proposals enter the queues with no priority. The programmes selected in this way would compose the equivalent of today's C-rank class.

Very good science cases that require loose conditions would be submitted through the regular channel. The choice of the channel is left to the PIs, who will choose the regular channel only if they believe they have a strong case.

The assessment of filler programmes must be kept simple:

- Technical feasibility and filler status confirmation (by the Observatory)
- Scientific assessment run by a single referee (accepted/rejected).

This is in order to ensure a minimum scientific level for fillers. Current statistics show that when C-rank programmes are completed to a reasonable fraction, they are as productive as B-rank

programmes (see Section 9). Therefore, in addition to producing a reasonable scientific return, they make good use of conditions that would go otherwise unused (i.e. leaving telescopes idle).

Advantages

- Encourage the submission of poor conditions proposals.
- Further decrease the load on the panels.

Disadvantages

- None.

Recommendation 7 - Create a joint ESO-ALMA channel to guarantee an effective synergy between the two facilities.

The TAWG perceives that the link between VLT and ALMA is probably stronger than with any other existing astronomical facility. In the current implementation, PIs applying for projects that require data from both sets of telescopes need to submit two proposals to two different TACS, out of sync and with no cross talk. This often causes discordant outcomes, hence penalizing projects that were judged as very promising by one of the two TACs, probably depending on their experience and focus. For this reason, the TAWG recommends that a joint channel is created, similar to the one currently in place for the VLT-XMM proposals. The WG acknowledges that the scale of joint VLT-ALMA proposals is going to be significantly larger than that of VLT-XMM, and it therefore requires a more structured implementation.

Advantages

- Avoids double submissions of similar science cases.
- Avoids cases of partially approved projects that need to be delayed for obtaining (if any) the missing part.

- Enables stronger, multi-wavelength science cases.

Disadvantages

- It requires coordination between two very large TACs, which are currently out of sync.

Recommendation 8 – Introduction of a high-risk channel for the submission of high-risk/high-gain projects requiring significant amounts of time.

ESO should consider introducing a high-risk channel (HRC), for the submission of very un-orthodox projects requiring significant amounts of time that no TAC would ever approve. The HRC should be allocated by a special, external committee (not necessarily composed by astronomers only) reporting directly to the DG, but should be kept separated from the DDT Committee. Only requests of large amounts of time for very risky, potentially highly rewarding proposals should be considered for this channel.

Advantages

- Allow for major break-through
- Overcome the conservativeness typical of TACs dealing with a large number of requests.

Disadvantages

- Risk of investing significant amounts of time with uncertain returns.

Recommendation 9 – Gender issues should be considered and actions taken to mitigate them.

The study conducted in the framework of the TAWG activities (Patat 2016) has given a statistical overview on gender systematics at ESO. The analysis has shown that there is a strong, possibly

dominating component related to the professional seniority of the applicants (the career level of male PIs is, on average, higher that those of their female counterparts). However, one cannot exclude the presence of conscious or unconscious effects in the review process itself (which is suggested by the different behavior seen for male and female referees). Although the TAWG is aware that implementing limitations on the level of information regarding the proposing team accessible to the reviewers is a very delicate matter (see for instance Reid, 2014), it recommends ESO addresses this topic, in coordination with other large scientific facilities world-wide. Possible actions are described in Section 4.

Advantages

- Maintain the trust in the community about fair treatment in the time allocation procedures at ESO
- Provide regular statistics to the community at large, where these are not always available for homogeneous and statistically significant samples.

Disadvantages

- None

Recommendation 10 – ESO should develop a strategic plan for the usage of the various facilities during the E-ELT era. When this is available, a new working group should be reconvened.

The terms of reference explicitly tasked the TAWG with examining *"the foreseeable evolution of the proposal selection and time allocation processes into the E-ELT era"* (see Section 13).

However, the TAWG feels that the times are not mature for providing recommendations on this matter without knowledge on the foreseen strategy of the VLT in the ELT era. Also, it reckons it is not within its mandate to make recommendations on the way time will be allocated at this unique facility. It is reasonable to imagine that operations will start with a relatively small number

of large programmes, while retaining a fraction for smaller projects. However, given the scale of the facility and the tremendous pressure it will experience, the working group deems it is inappropriate to express opinions on a matter that still needs to be shaped by the main stakeholders.

The TAWG would like to see the impact of the proposed changes on the VLT before making any recommendation for the E-ELT. It therefore recommends the Observatory considers the development of a strategic plan for the usage of the various ESO facilities at the start of E-ELT operations. When this is ready, a new working group should be constituted and tasked with the review of the recommendations presented in this document and implemented by ESO, their effects, the lessons learnt and their possible applicability to the E-ELT case.
This applies especially to the reduction in the number of small programmes and the allocation to larger, more coherent projects that are being proposed here.

Advantages

- Set the stage for the next working group.

Disadvantages

- None

The TAWG is aware that the vast majority of the recommendations presented by the OPC Working Group (Elias et al. 2011) were not implemented by ESO, mostly because of operational reasons. The situation has significantly changed in these last five years, especially because of the approval of the new Phase 1 proposal submission system. This creates the ideal conditions for the implementation of the recommendations presented in this report, both because it will represent a significant change in the work-flow, and because the project is in a phase in which it is still possible to include the necessary requirements.

4 Suggestions for an implementation plan

This section discusses a possible implementation plan for the proposed recommendations.

Recommendation 1: *Move to a yearly cycle*

Recommendations 1 and 2 need to be deployed simultaneously, in order to immediately address the loss in the time response for those programmes that do require it.
Both recommendations have a significant impact on the current operations and workflow. Because of this, they will be best deployed together with the new Phase 1 proposal submission system. The expectation is that the number of proposals will not double, while the time available for the various phases will. The TAWG reckons this should be sufficient to compensate for the increased workload peaks at Phase 1 and Phase 2, and should address the concerns raised in 2011, following the recommendations of the OPC WG, which led to the final negative outcome.
In the following some implementation indications are presented.

Phase 1

- ESO could in principle recruit 156=78x2 referees within the current budget. Nevertheless, because of the current level of period-to-period overlapping, this would require to manage about 40% more distinct referees in one cycle. Depending on the actual number of proposals, and considering that recommendation 3 proposes a factor 2 reduction in the number of proposals per referee, it will probably be possible to avoid doubling the number of referees, and limiting to 130-140 scientists (which is approximately the current size of HST and ALMA TACs).
- In the current implementation, the changes proposed here were very hard to implement, because of its rigidity and fragility. The new Phase 1 system is built with much more flexibility, and will therefore allow easy adjustments in the numbers of panels, members

per panels and so on. For this reason, it is fundamental that the recommendation is incorporated as soon as possible in the software requirements (the same applies to recommendation 2).

- A yearly long-term scheduling will be more prone to uncertainties on the technical activities, which may be postponed/anticipated or even inserted/cancelled with short notice. The same applies to unexpected, extended instrument/telescope failures. The TAWG suggests that this is taken care by separating the proposal selection process from the long and short-term scheduling, as it is the case at ALMA and HST. A dedicated group should take care of the changes, guaranteeing a faster reaction time, in consultation with the Observing Programmes Office (OPO) for guaranteeing scientific consistency with the OPC review. In an intermediate approach, the long-term schedule could be kept within OPO, while the short-term scheduling should be transferred to an operational unit.

Phase 2

- Like for Phase 1, the yearly cycle will in principle produce higher activity peaks for Phase 2. However, this is strictly true only if the Phase 2 reaction time is kept as is (four weeks for preparation plus four weeks for checks and certification before final delivery to the Observatory). The TAWG does not see a reason why this should be the case. With the yearly cycle the duration of this phase can be extended in time, or split into two (depending on target RAs, for instance). What is important is that all the Phase 2 material pertaining to a given programme is submitted within the same time, to avoid having the support scientists returning multiple times on the same cases (unless this is strictly necessary).
- One of the possible issues with the yearly cycle is related to the probable increase in the programme change requests (and, more specifically, the target change requests). One may expect that having a longer time baseline, PIs will want to request changes more frequently than in the current semester-based cycles, with repercussions on the workload

on the parties involved in processing the requests. This can be mitigated by introducing stricter rules and more rigidity in the way the requests are handled. The deployment of the fast-track channel (recommendation 2) will also help alleviating this problem as, by construction, no changes are expected for those programmes.

Recommendation 2: *Introduce a fast-track channel*

The fast-track channel (FTC) is meant as an addition and not as a replacement of the DDT, which serves a completely different scope.

- The exact frequency of the deadlines needs to be decided by ESO, but the TAWG tentatively suggests these should be spaced by 2 to 4 months, and out of sync with the yearly cycle (i.e. there should never be a FTC deadline close in time to the yearly deadline).
- Fast-track (FT) proposals need to be justified in terms of response time for getting the data to the PI, that is to say that FT proposals should not be just a shorter versions of proposals submitted at the yearly deadline.
- FT proposals should be allowed to request only a maximum amount of time (smaller or equal to the threshold fixed for the yearly cycle, see recommendation 4).
- Unless properly justified (e.g. explicit OPC requests of conversion from normal to FT), proposals rejected at the yearly cycle should not be allowed to resubmit through the FT (as is the case at the HST for the mid-term proposals, see Section 11.2). A machine-learning mechanism for tracking resubmissions should be put in place (see also Section 10). This places additional requirements on the new Phase 1 proposal submission system (see also recommendation 3).

The FT proposals need to fulfill the following conditions:

- They must include a strong justification as to why they could not be submitted through the yearly cycle.
- In case of projects related to PhD theses, a support letter from the relevant institution (and not just by the supervisor) is compulsory.
- Only standard modes and no visitor instruments can be requested.
- Pilot studies can be accepted, provided they do not qualify for the DDT channel.

Not meeting these criteria should lead to immediate rejection.

The TAWG has considered the following options for FT proposals review:

- If the total number of proposals decreases, and given that the number of distinct serving members per cycle will increase, the most natural candidate is the OPC itself. This would ensure a coherent review process, that guarantees the consistency between the yearly cycle and the FTC. It would also preserve history and maintain the continuity of the science programme. Based on the Gemini statistics (see Section 11.1) extrapolated to the VLT case, it is estimated that ESO will receive 20 FT proposals/month, yielding 240 proposals per year. If the TAC will be composed by, say, 120 scientists and each proposal is assigned to 3 referees (see also recommendation 3), this implies an additional load of 6 proposals per referee/year. Considering that this will be diluted across the year, this appears to be a very reasonable number.
- If the load turns out to be significantly higher, ESO may consider resorting to a standing panel, completely external to the OPC and nominated by ESO.
- Alternatively, ESO can devolve the review process to the PIs from the yearly cycle, without implementing any referee-penalization mechanism, i.e. similar to what is being done at Gemini. In this schema, yearly cycle PIs will implicitly accept to serve as proposal reviewers for the year that follows the submission. Given the very large number of PIs (currently more than 700 every semester), the FT proposals can be sent to a larger

number of reviewers (within specific categories and even sub-categories) to increase the statistical baseline.

Once accepted, the FT programmes will obey to the following rules:

- They will be scheduled as A-rank fast-track, at the same level of normal A-rank from the yearly cycle.
- They will not be carried-over. Their life-cycle duration will be limited to the time that separates two FT calls.
- They will benefit from a shorter proprietary period (3-4 months).

ESO should fix a maximum total amount of time to be allocated to these programmes. As in the case of Gemini, the TAWG suggests starting with a comparatively small fraction (e.g. 10%). The limit can then be adjusted based on the demand. The presence of a maximum limit also allows the control of the oversubscription of this channel, in order to avoid a possible inflation of short programmes submitted through the FTC, should the success rate of this channel be significantly higher than that of the yearly cycle. To facilitate the deployment of this new channel, ESO should start allocating FT proposals in pre-fixed slots. This will make the scheduling process easier, although it will introduce some rigidity in the request (depending on moon conditions during the chosen time slots, for instance). Once the channel is consolidated and depending on the demand, ESO may consider moving to a schema in which the FT runs are inserted in the ordinary service mode queues.

These programmes require a more dynamical scheduling, which deviates from both the static semester-based paradigm and the 5% floating provision currently in place for the DDT. This is best done by a dedicated operational group.

Recommendation 3: *Radically change the proposal review procedures*

The implementation of the recommendation strictly requires the new Phase 1 proposal submission system, because the change is very significant and cannot be deployed within the current system.
The proposed schema sets the following requirements on the system:

- Better proposal-expert matching. At the moment proposals are distributed randomly after the panel composition is decided. This is sub-optimal (irrespective of whether or not any of these recommendations will be implemented). As a minimum requirement the system should consider multiple keywords (e.g. 5 keywords for the proposals, 10 for the referees are used at HST), but ESO should explore machine-learning matching based on textual analysis (abstract of papers for the reviewers, abstracts and scientific rationale for the proposals).
- The Phase 1 system must be specified in such a way that the text can be easily parsed (at least the abstract and the scientific rationale).
- The system must be flexible in terms of configurability of number of panels, members per panel, number of reviewers per proposal and so on.

The TAWG recommends the following changes are considered for the review process of the yearly cycle:

- Panel members should rank the runs, not grade them. This removes the need for the grade normalization, which leads to controversies and confusions, in addition to being possibly incorrect (depending on how much the grade distributions deviate from Gaussian. See Section 7). Although the grade scale is normally considered as a linear scale, it is not: the same grade difference corresponds to different inter-percentile intervals for different grades. This is not the case for the ranks.

- If the grading is kept, a transformation to ranking must be devised and applied before merging the rankings.
- More structured figures of merit for ranking the proposals should be introduced. For instance, each proposal can be given separated evaluations judging the following aspects:
    - Scientific merit
    - Impact in the field
    - Broader impact
    - Need for the specific facility

  These are then combined to a single figure of merit, which is eventually converted to a rank.
- The publication record related to programmes previously allocated to the PI should be automatically logged in the proposal form. The record has to be complete.
- An automatic, machine-learning way of identifying re-submissions should be put in place. A genuine resubmission to complete an incomplete programme (A-rank, or more likely visitor mode) should not be seen again by the referees, just by the Observatory.
- The feedback to the PIs is provided in a more structured way (e.g. Strengths, Weaknesses, Relevance to the field, …).
- For the B-rank class only, and below a given (and configurable) rank per telescope, the Observatory is given freedom in terms of taking the final decision as to what is scheduled. While in the current implementation scheduling follows thoroughly the merged ranking (per telescope), it is here proposed that this should be rather done by the Observatory, matching the constraints of the selected runs to the statistical properties of the site and the RA-DEC accessibilities.

The last point places a different requirement on the scheduling tool, that should behave in a different way above and below a given (and configurable) rank for the given telescope. This will require a dedicated algorithm. The TAWG is aware that this is not possible within the current scheduling tool, which uses the OPC ranking as the driving parameters for scheduling runs. It is

recommended that ESO considers this when preparing the requirements of the next version of the scheduling tool.

Recommendation 4: *Promote the submission of larger and coherent large programmes*

This recommendation does not have a direct impact on operations. It may require a change of policy (regarding the maximum limit of time to be allocated to LPs), but its is mostly a matter of encouraging the community to submit larger time requests concentrated in shorter amounts of time. This may be done through Messenger articles and Science Newsletters, and explicitly promoted through the Users Committee.

Recommendation 5: *Redistribute the time balance of large and short programmes*

This recommendation only affects the scheduling process, while it does not have any direct influence on proposal submission, Phase 2 and observations. The scheduling tool needs to be modified, in order to control the amount of time allocated to programmes shorter than the given threshold (by telescope or globally).
Therefore, the implementation of this recommendation does not have any significant impact on operations. It certainly does have an impact on the community, though.

The threshold value should be set by ESO. The TAWG suggests that this is around 3 nights.

Recommendation 6: *Introduce a filler channel*

This is a very simple recommendation. However, because it implies a different review process (in terms of proposal distribution and review), it will naturally fit in the context of the new Phase 1 system.

Recommendation 7: *Create a joint ALMA-VLT channel*

From the implementation point of view this is probably the most complex recommendation, as it involves two independent facilities, TACs, and schedules. ESO will have to necessarily coordinate this with ALMA and what follows only provides basic indications.
The implementation of this recommendation is in principle completely independent from the others, but since it requires a new concept for the review process, it will require the new Phase 1 proposal submission system to be deployed first.

- Joint ALMA-VLT proposals can only be requested at the yearly deadline.
- The two facilities need to agree beforehand on the mutual fractions of time that will be allocated through this channel.
- Pre-selection of joint proposals is first done by one of the two TACs.
- A second pass is run by the second TAC, only for the proposals promoted by the first. The move to a yearly cycle should in principle allow to complete the decision process before the long term schedules of the two facilities are completed and the outcome communicated to the community.
- The outcome of the joint selection may be delayed with respect to the release of the schedules at the two facilities (which may, and probably will, be asynchronous).

This schema is complex, as it requires a coordinated effort from both facilities, which have systems that do not talk to each other (in terms of exchanging proposal information). This will have to be investigated by ESO. The TAWG is aware that this may constitute a show stopper.
An alternative, more easily implemented possibility is to follow, though on a larger scale, what is currently done within the VLT-XMM joint channel, with some ad hoc variation:

- Proposals are submitted to the TAC of the facility for which the request is largest.
- Each TAC allocates time to its facility and makes a recommendation for the other facility, for the latter only up to a maximum, pre-agreed amount of time.

- Only top-ranked, A-rank equivalent programmes will be allocated time.
- The results are mutually communicated by/to the two facilities, so that the telescope schedules can be prepared accordingly. This will require coordinating the two yearly deadlines, at least to some extent.
- A policy should be established as to what happens if data are successfully obtained at one facility but not at the other.

Recommendation 8: *Introduction of a high-risk channel*

In terms of proposal submission and distribution this is very similar to the DDT, the only operational deviation being proposal distribution. Given that the channel is designed for large requests, one can reasonably expect that observations will not be very urgent. This mitigates the problem of inserting a large allocation into an already fully scheduled and running cycle. In turns, it removes the need for making a very cumbersome provision for an event that may happen only very rarely.

For this reasons, the TAWG suggests that a HRC proposal approved in a given cycle, is scheduled for observations only in the subsequent cycle (under the assumption that the move to the yearly deadline is implemented). The necessary time can be reserved ahead of time, before the schedule of the starting cycle is prepared (similar to the current case of public surveys).

Recommendation 9: *Gender issues should be considered and actions taken to mitigate them.*

The results obtained for ESO (Patat 2016) are similar to those reported for HST (Reid 2014). It is therefore very useful to briefly describe here the actions that were taken (or are under study) at that facility, so that they may be considered in the context of the implementation of Recommendation 9.

Working in consultation with the Space Telescope Users Committee (STUC), a number of steps have been taken at HST in an attempt to mitigate the observed gender systematics: the potential for implicit bias has been highlighted during the TAC orientation; reviewers are asked to focus on well-defined criteria, prominently displayed during the proposal discussion; and the proposal format has been changed to reduce the prominence generally accorded to the PI. In the last case, investigators are now listed alphabetically with initials rather than first names, and the PI is not identified explicitly. These modifications have been generally well received by the TAC panelists; as yet, however, the outcomes have not changed, suggesting the possibility that the results are driven by deeper influences within the given research field.

HST also considered the possibility of making the proposals fully anonymous to the TAC. This was discussed with the STUC, but it was felt it would be too much of a stretch for the community to take that step now. In addition, there is the valid point that the team composition can be important with regard to expertise.

While no other modification to team information will be implemented in the course of 2017, HST is seeking the advice of outside experts, who might be able to help us better interpret what has been observed so far.

Based on this, the TAWG suggests to follow three lines of work:

- Panel orientation: proactively raise awareness with a dedicated introductory session at the meeting. This should include the discussion of other possible discrimination (countries, ethnicities, scientific fields, …). A summary of the results of the analysis conducted by the TAWG should be made available to all panel members as part of the regular set of documents they receive when they accept to serve. All panel members should be reminded of the results during the pre-meeting phase, and at the meeting itself. This is an easy addition to what ESO is currently doing, and can be implemented any time.
- Statistical monitoring: ESO is in the position of collecting uniform data for a sample that is highly representative of the astronomical population world-wide. For this reason, ESO should introduce in the User Portal indicators that will allow more sophisticated analyses

in the future. These include gender and seniority level, which should be made mandatory for all applicants (both PIs and coIs). This will be best done with the deployment of the new Phase 1 proposal submission system, which marks a natural time for transitioning into the new schema.

- Further measures: although the TAWG fully appreciates that limiting the amount of information about the proposing team is a delicate matter, it recommends that the actions taken at HST (and possibly at other large scientific facilities) are considered and discussed between ESO and the community. For instance, different levels of "team obfuscation" could be used at different stages of the review. This poses some extra requirements on the new proposal submission system, which should include the concept of configurable *proposal views*. These would enable the possibility of giving access to a different level of proposal content depending on the user's role (PI, co-I, ESO staff, reviewers, …) and proposal handling stage (e.g. submission, pre-meeting distribution, technical feasibility, panel meeting, …). For instance, the proposals could be made completely anonymous to the referees during the pre-meeting phase, while an alphabetically ordered list of co-Is (with no indication of the PI) could be offered to the Panels for the discussions.

Recommendation 10: *Development of a strategic plan for the usage of the VLT in the E-ELT era*

Given the timescale foreseen for the E-ELT, ESO should address this matter and reconvene a new time allocation working group in about five years.

5 Supporting material

The tasks assigned to the TAWG were divided into two groups: proposal submission and proposal review (see Sect. 13). The first is concerned with the way proposals are collected (frequency of calls, offered programmes, etc.), while the second deals with the way proposals are reviewed (panel composition, grading scheme, load on the reviewers, etc.).

Although they are linked to distinct parts of the process, they are deeply interrelated, with aspects that cannot be considered as completely independent. A clear example is related to the number of proposals, which carries the imprint of a culture in the community (i.e. submit many, relatively small time requests) and at the same time generates the load on the reviewers. In turn, the way in which the proposals are selected may be favoring small requests, hence starting a negative feedback process that pushes the community to fragment even more.

In the following two sub-sections, we will briefly summarize the problems we identified within the two main parts. The issues are then discussed and quantified in more detail in the remaining sections of the report, which provide the statistical grounds for the recommendations presented in Sect. 3.

5.1 Proposal submission

5.1.1 Frequency of the calls

In the current implementation, ESO proposals are called for twice a year. The only exception is the public surveys, which are called for much less frequently (typically every five years).

Most astronomical facilities of similar scales (in terms of number of proposals, users, oversubscription factors), like HST, ALMA, CHANDRA and XMM-Newton have annual calls (one remarkable exception is Gemini, which issues two calls per year).

More frequent calls create two kinds of pressures:

a) On the users: when the deadline is approaching, scientists feel compelled to apply, as this somehow signal their “presence” to the rest of the community, in many cases without

real need for new data (also, about 50% of the data are still unpublished after more than 3 years (see Sections 8 and 9).

b) On the panel members: having to serve twice in a year is demanding and more difficult to accept than a yearly call.

Part of the resubmission fraction (see Sect. 10) is generated by the artificial period boundaries, so that even successful proposals are resubmitted to complete the sample, adding to the workload without a real need. As discussed in the OPC WG report (Brinks et al. 2011), a call per year would most likely decrease the total number of proposals to be reviewed, although it would increase the number of proposals per cycle. Therefore, although desirable, moving to one cycle per year would not completely solve the workload problem. Additional measures need to be taken.

5.1.2 Types of programmes

ESO offers a large number of programme types: Normal (NOR), Large (LAR), Target of Opportunity (TOO), Monitoring (MON), Calibration (CAL), Guaranteed Time Observations (GTO), Director Discretionary Time (DDT). Nevertheless, the large majority of the request (≥80%, both in number of proposals and in requested time; ~90% when considering the VLT alone) comes in the form of NOR.

While the spectrum of offered programmes appears to be sufficient to cover all the needs of the community, there seems to be a possible issue with the large number of NOR programmes and their relatively low performance in terms of scientific return.

5.1.3 Time request

The time request distribution is strongly skewed towards small amounts of time (see Sect. 6 and Table 1). The Observatory has never taken an active role in steering the request, and the community was left free to choose. It is possible that the selection process itself (if biased against larger requests) has created this situation, or that this is what is perceived by the community.

Either way the result is that the request is extremely fragmented and this may be one of the reasons for the observed return rate in terms of programmes that do not publish (see Section 9, and in particular Section 9.1).

This is an issue ESO should look into, in concert with the community, through both the governing and advisory bodies.

5.2 Proposal review

5.2.1 Review load

As it was pointed out in the ESO OPC Working Group report (Brinks et al. 2011), the main issue with time allocation at ESO (and at other highly oversubscribed astronomical facilities) is the exceedingly large number of proposals that the Observing Programmes Committee (OPC) has to review. This is typically about 100 and 70, for OPC-proper and panel members, respectively.

This is not just a problem of the reviewers: the overload has effects on the quality of the judgment, on the feedback that is given back to the community and, in the end, on the final scientific product.

Although there is no quantitative study indicating what is the ideal number of proposals per referee, it is obvious that for a reviewer with the current load it is very hard, if not impossible, to produce a meaningful ranked list, simply because keeping in mind the relative merits of such a large amount of cases is very cumbersome. In addition, it is very difficult (and demanding) to provide the applicants with proper and constructive feedback.

Although we could not find literature about this very specific topic (while there is something about the number of reviewers per application, see for instance Snell 2015), there is a general consensus that this number should not exceed 30 applications. Here we note that, for instance, at the Paul Scherrer Institute, the usual number of proposal assigned per reviewer is between 25 and 40. In each panel (composed by 10 scientists), the proposals are reviewed by 2-3 referees, and they are then discussed in the panels only if the grades dispersion is larger than a given threshold (otherwise the average is taken as final grade).

In the current ESO schema each proposal is read by all non conflicted members of the panel (the average number of reviewers per proposal is 5.6). With this setup, lowering the load to 30 proposals/member would require about 180 members. Although ALMA and HST TACs include around 140 scientists, a further increase in the number does not seem a logistically/practically viable solution, certainly not within a semester-based schema.

5.2.2 Alternative review systems

Within the peer-review paradigm, variations can be considered. The OPC WG (Brinks et al. 2011) discussed various possibilities (e.g. decreasing the number of referees per proposal, eliminating the meeting, etc.). None of these alternatives was judged to alleviate the problem without introducing additional problems (e.g. lack of a robust statistical basis, lack of commitment, etc.). The TAWG is in substantial agreement with the conclusions reached by its predecessor.

One add-on to the current system would be the introduction of external consultants (EC), who would be asked to give an expert opinion on a limited number of proposals. Their grade (or rank) would be then properly weighted and merged with that of the panel. The ECs would have to be recruited in addition to the panels, and would not participate to the physical meeting.

One draw-back is that if too much weight is given to the EC, the final fate of the proposal may be influenced by one single opinion; if a small weight is given to the EC, this would probably have no measurable effect.

An interesting concept is that of distributed review, in which by submitting a proposal, the PIs accepts to referee a number of other proposals submitted during the same cycle (Merrifield & Saari 2011; see also Kurokawa et al. 2014 for further developments). In addition to removing the need for a standing committee, this approach limits the number of submissions. The algorithm includes a self regulating mechanism that penalizes PI-reviewers with outlier behaviors (e.g. trying to promote themselves by downgrading everybody else).

The Merrifield & Saari method was implemented experimentally by NFS (see Mervis 2014), and by Gemini, which deployed it for its new Fast Turnaround channel (see Sect. 11.1). While the current scale of the Gemini experiment is small (nominally 10% of Gemini time is allocated in this

way; the number of proposals is comparable to what ESO manages via the DDT channel), the NFS experiment (its CMMI Division) has a scale that starts to be interesting for our purposes.

This approach clearly reduces the costs for the Observatory (no travel/lodging/per-diem expenses)[6], it simplifies its life (the recruitment process is cumbersome and time consuming), and it definitely lowers the load on the reviewers. However, there are no quantitative assessments yet as to whether this system produces better science.

One of the most frequent criticisms to this system is that it tends to promote "average science", i.e. proposals that play safe. The lack of a proper meeting is also seen as a weakness, because the discussions in the panel requires preparation and contribute to form a consensus.

In general, the TAWG is against any system that would penalize the reviewers. The Merrifield-Saari concept is based on the assumption that most referees act in a malevolent way, why we are convinced of the opposite.

5.2.3 Recent criticism to peer-review

Although peer-review review certainly has its drawbacks, it is widely used and it still encounters the support of the majority of scientists. While there is a certain level of awareness about the fact that the global process has a significant stochastic component[7] (see Sect. 7 for a quantitative analysis), peer-review can be seen as a democratic process, as theorized by Kitcher (2011).

In this approach, which Kitcher calls "*well ordered science*", the science that society should choose to carry out and support is the science that is most favored by the scientific community. The panelists serve as representatives of that community, and they are being asked to use the knowledge and skills that they have acquired over their careers to assess what they judge to be the most productive observing proposals. There is no existential "right" answer out there that they are trying to match against – by definition, their answer is the "right" answer; and a different

[6] These currently add up to about 150,000 EUR per year at ESO.

[7] It is important to remark that the stochastic part does not necessarily relate only to the "imperfection" of the referees seen as measurement instruments, but it also reflects the intrinsic fuzziness in the quality of a proposal, which depends on a number of aspects, which include the context, the time and other proposal properties.

set of panelists would come up with a different set of "right" answers, that will likely overlap by probably around 50%. The system is inevitably subjective to some extent – what one needs to focus on is how that subjectivity is applied (e.g. as a subjective assessment of the characteristics of the PI of a proposal, rather than the science proposed).

That said, the lack of full reproducibility and the possible/probable introduction of selection biases is often used against the peer-review as a concept. For instance, Gillies (2014) argues that a random selection would be better, because it avoids what he calls the *systemic bias*, which tends to "*favor mainstream research programmes*", leading to "*the stifling of new ideas and of innovation*". Other problems of the peer-review are related to the progressively degrading quality of the reviews in connection to the increasing load (see for instance Fox & Petchey 2010; Bohannon 2013).

Most of the criticism comes from the stochastic component that is implicit to the process. Although applicants base their evaluation of the peer-review paradigm on their own experience, and normally do not have means of estimating how much the selection is really aleatory, this perception is more and more diffused and creates a significant level of frustration. This is normally coupled to the feedback that unsuccessful applicants get, judged as unsatisfactory and, at times, as provided by "incompetent" referees.

For obvious reasons there are not very many quantitative studies that try to assess the repeatability of the selections operated by a given panel. The NIPS experiment is one of those (Price 2014). In this experiment the organizers of an important conference series in theoretical computer science assigned 10% of the applications (166) to two different panels, without the panel members knowing which ones were in common[8]. The two committees were tasked with a 22.5% acceptance rate. The final result was that 57% of the papers accepted by the first panel were rejected by the second (and vice versa). This level of disagreement can be compared to what is expected for a purely random selection, in the case of the NIPS experiment 77.5%.

[8] A similar experiment was proposed by the OPC WG (Brinks et al. 2011), but was not run by ESO.

This confirms in a quantitative way that some level of fuzziness is to be expected, and that submitting the same proposals to two different panels would lead to a ≈50% agreement. As the statistics show (see Section 7.3), the dispersion in the grades increases at poorer grades, which confirms the common belief that in the mid-range the selection becomes fuzzier. Based on this result, the TAWG reached the conclusion that for the proposals ranked below a certain threshold (which will depend on the given telescope), the TAC could give some more freedom to the Observatory. In the current implementation the scheduling is done thoroughly following the OPC ranking, even though the uncertainty in the middle range is large, and deprives relative ranking in that area of the distribution of a statistically significant meaning. Indeed, in that region, the fact that a proposal gets scheduled or not may depend on details that have no relation to the scientific merit.

Things being so, in this ranking range, the Observatory could select the runs that better match the available conditions, in statistical terms.

By declaring up-front that proposals ending in this range will be selected following constraint-matching criteria, will also address the user's frustration caused by the perception that the outcome was largely aleatory, while being told that it was based on scientific arguments.

Of course, this will require some "education": the community needs to be made well aware of the workings of the process, and of the limitations of the peer-review paradigm in a quantitative way.

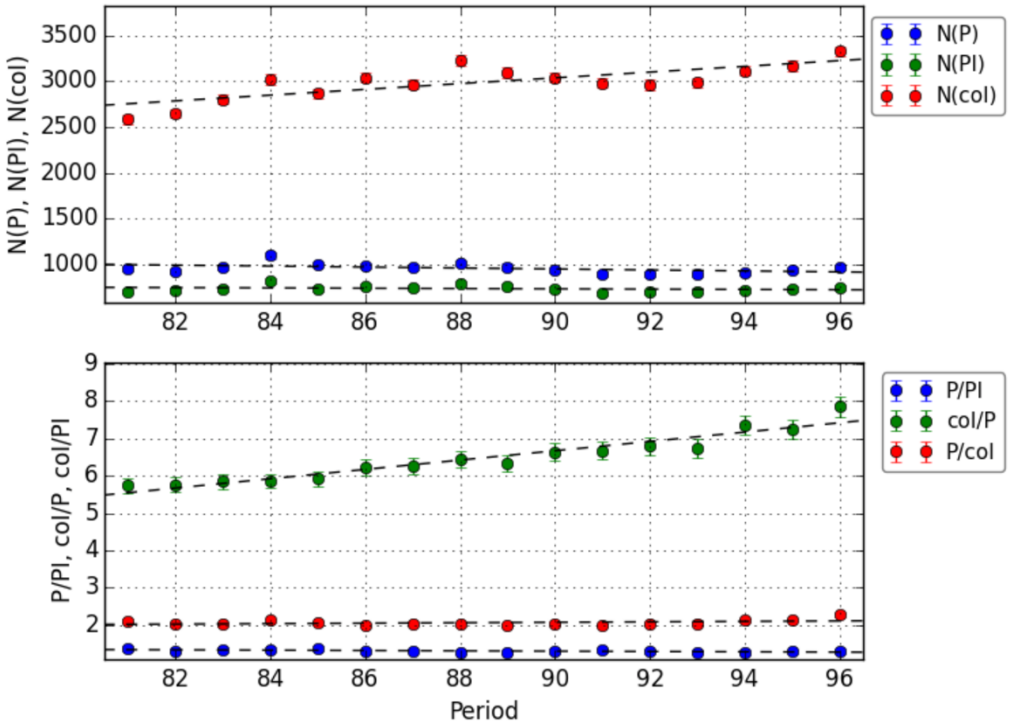

Figure 1 - Proposal submission trends between P82 and P96. Upper panel: number of proposals, distinct PIs and distinct cols. Lower panel: proposals per PI, non distinct cols per proposal and proposals per col. The dashed lines trace linear squares fits to the data

6 Telescope Time Request distribution

The TAWG has studied in great detail the distribution of the telescope time request in the period range 81 to 96. The results were provided in an internal report (*Telescope time request statistics for ESO periods 81 to 96*, August 2015). The main findings can be summarized as follows:

1. About 72% of the ESO proposals are submitted for the VLT, ~11% for LS. The VLT request shows a small but steady increase (~10% in the period range covered by this report).

2. While the total number of proposals is roughly constant (and so are the number of proposals per PI, and proposals per col), the number of cols (and cols per proposal) is still steadily increasing (Figure 1). This implies a systematic growth in the team sizes over time (~40% in the reporting period range).

3. At the VLT, the request is largely dominated by NOR programmes (89% in proposals, 77% in time). LAR programmes contribute to 1.7% of the proposals (16.5% in time). The time success rates of the two programme types is 0.30±0.01 and 0.17±0.03, respectively.

4. For LS telescopes 41% of the time is requested via LAR programmes, with a success rate of 0.30±0.06. The amount of time requested through LAR programmes shows an increase during the second half of the reporting period range.

5. At the VLT, about 90% of the PIs (88% of the coIs) request time through NOR programmes. The fraction for LAR programmes is ~2% (7.6% of the coIs). While the PI fraction is similar for LS telescopes, the corresponding coI fraction is definitely larger (16.5%) implying that, for this telescope group, applicants tend to cluster more around larger time requests.

6. The median NOR time request for the overall sample is 2.0 nights (when considering proposals) and 1.8 nights (when considering time).

Type	Proposals									
	Requested					Assigned				
	05	25	50	75	95	05	25	50	75	95
NOR	0.3	0.9	1.6	2.7	4.9	0.3	0.9	1.5	2.5	4.9
LAR	10.4	15.0	20.0	24.4	43.1	11.0	14.4	19.6	24.6	45.8
TOO	0.2	0.6	1.1	2.0	4.0	0.2	0.5	0.9	1.6	3.6
GTO	0.4	0.8	1.5	2.6	6.7	0.4	1.0	2.2	4.0	7.5
MON	0.5	1.2	2.4	5.1	8.5	0.6	1.4	2.5	4.9	9.6
CAL	0.1	0.5	0.9	1.2	3.1	0.1	0.5	1.0	1.5	3.1

Type	Time									
	Requested					Assigned				
	05	25	50	75	95	05	25	50	75	95
NOR	0.2	0.8	1.4	2.3	4.1	0.2	0.7	1.2	2.1	4.0
LAR	2.0	9.6	13.6	20.3	35.6	2.4	9.2	12.2	18.1	35.4
TOO	0.2	0.6	1.1	2.0	3.6	0.2	0.5	1.0	1.8	3.6
GTO	0.4	0.8	1.5	2.5	6.1	0.4	1.0	2.3	4.0	7.0
MON	0.5	0.7	1.7	4.6	5.8	0.5	0.5	0.8	1.7	5.0
CAL	0.0	0.2	0.9	1.0	2.1	0.1	0.2	1.0	1.0	2.1

Table 1 – Time distribution (requested/assigned) for proposals (upper table) and time (lower table) for the VLT by programme type. Times are expressed in nights. The columns indicate the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} percentiles of the distributions.

7. At the VLT the median NOR time request is 1.6 nights (proposals) and 1.4 nights (time). The 95-th percentile of the distribution is 4.9 nights (4.1 in time). The requested and assigned time distributions are very similar. The median LAR time request is 20 nights (13.6 in time).

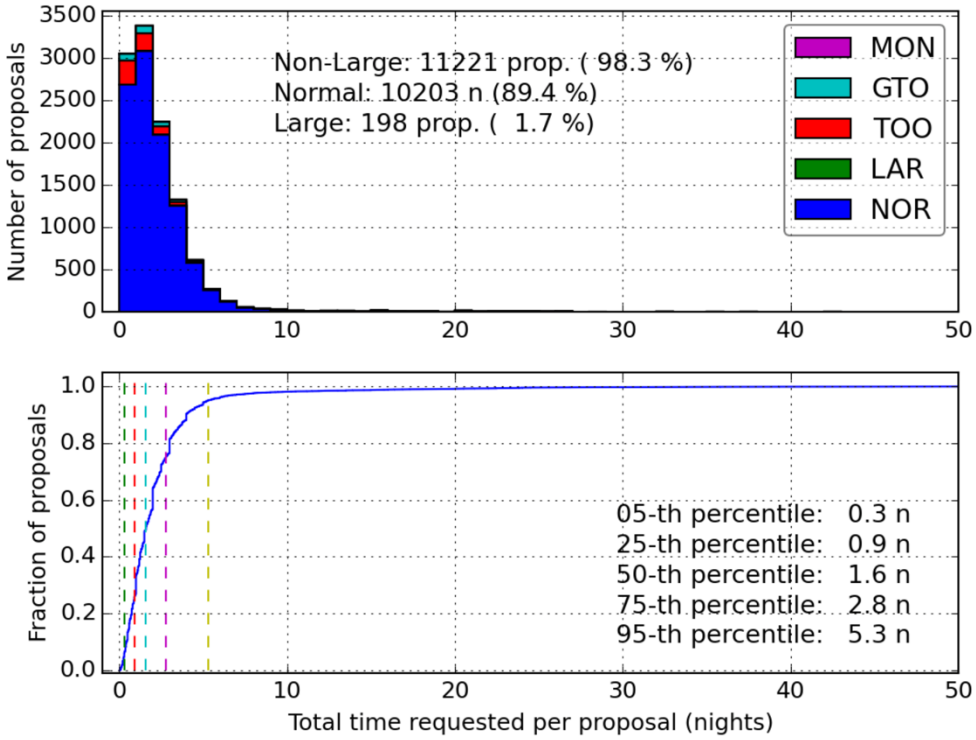

Figure 2 - Time request distribution for the VLT and by programme type (upper) and cumulated distribution function (lower). The vertical dashed lines indicate the 5th, 25th, 50th, 75th and 95th percentiles.

8. At LS telescopes the median NOR time request is 4.0 nights (both in proposals and time). LAR programmes are nominally larger than at the VLT, with a median request of 40.2 nights (36 nights in time). Requested and assigned time distributions are more similar to each other than in the VLT case, signaling a milder selection effect during the proposal review process.

9. At the VLT, the NOR median time request shows a steady decrease with time, with a peak-to-peak amplitude of ~0.5 nights. The corresponding values for LAR programmes show marked fluctuations around the median value (≈20 nights) and no trend is detected.

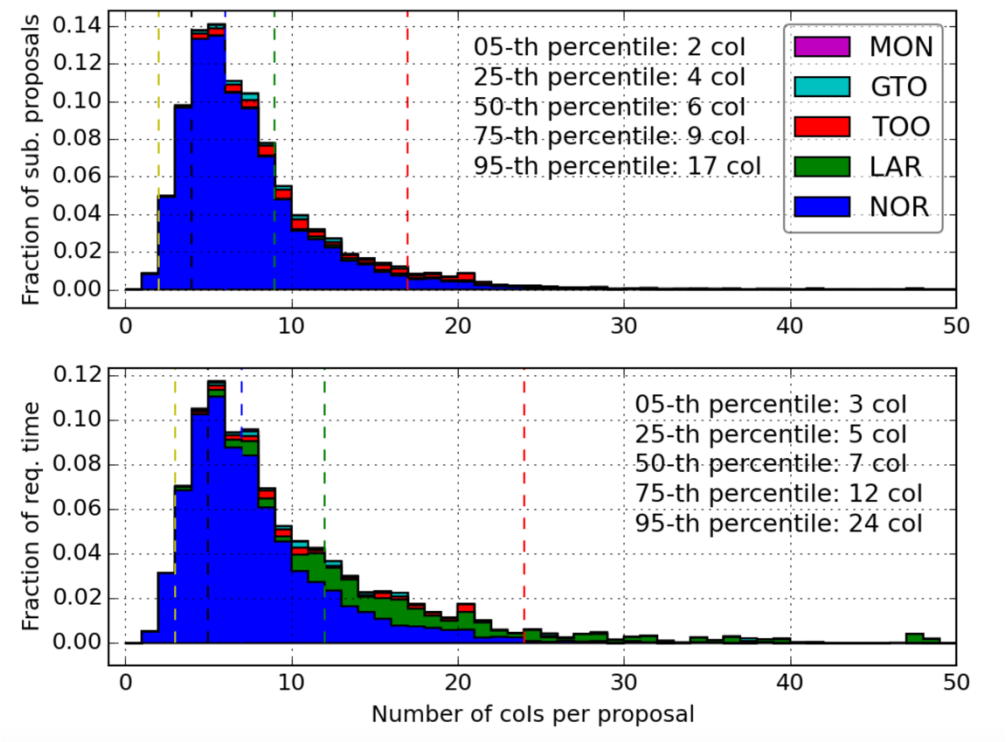

Figure 3 - Distribution of number of cols per proposal by fraction of proposals (upper) and requested time (lower).

10. At LS telescopes the NOR time distribution is stable during the reporting period range. LAR programmes display a possible evolution towards larger requests.

11. The success rate of VLT-NOR programmes decreases with increasing time request; this behavior is particularly pronounced for the time success rate. At face value, a programme requesting less than one night has 30% more chances of being scheduled than one requesting 4 to 6 nights.

12. The acceptance rates for LS telescopes are higher than those of the VLT, in all time request bins. No clear trend is seen in the 3-10 nights range and the rates remain roughly constant up to 20 nights.

13. At the VLT, NOR programmes involve 88% of the total users in a given semester, while the corresponding fraction for LAR is 8%. For LS telescopes the two numbers are 84% and

17%, respectively. The median proposing team includes 6 coIs affiliated to 3 different countries.

14. While LAR programmes are submitted by larger teams (typically a factor of two larger), there is no significant correlation between the amount of requested time and the size of the team.

The bottom line is that a very significant amount of VLT time is allocated to programmes that request less than 2 nights and are submitted by small groups of researchers (see also Figure 2 and Figure 3). The question arises as to whether, after more than 15 years of VLT operations, it still makes sense to maintain the status quo or the community should rather move to a different approach. While it is clear that time fragmentation allows ESO to keep the community at large happy by fostering very numerous and small collaborations, it is not so obvious that this paradigm can still provide transformational science.

Another open question regards the possible presence of a negative feedback mechanism in the submission/selection process. As there are signs that larger requests are disfavored[9], it is not at all clear whether the observed time distribution is dictated by the scientific needs (i.e. the distribution naturally evolved towards an optimal configuration that guarantees a maximum scientific return) or it is rather the result of a convenient adaptation of the community to the "environment". In other words, there is the possibility that the time request distribution is what it is simply because it is easier to get small amounts of time.

It is a very diffused opinion that time allocation committees judging proposals serving very large communities tend to be conservative and less committal, when it comes to approving larger requests. Whether this is a fact or a pure perception (or something in between) remains to be quantified.

[9] With the available information it is not possible to decide whether this is related to a conservative approach of the OPC or rather to a lower intrinsic quality (and/or poorer justification) of proposals with larger requests.

The fact remains that so far and in terms of time request, ESO has been following the community needs (the OPC is an expression of the same community that submits the proposals), without trying to influence them. It may now be time to ask the community to revise the status quo, for properly transitioning the VLT science into the ELT era.

7 Proposal review process

In the context of the TAWG activities, a thorough statistical analysis was carried out, using the grades collected between P82 and P97, for more than 14,000 proposals reviewed by 527 distinct referees on more than 1200 referee sessions, for a total of ~140,000 pre-meeting grades. The results are described in detail in a separate document (*Report on proposal grading at ESO*, May 2016). The report goes into great detail in the statistical analysis. The following summary collects only the aspects that are relevant for this discussion.

7.1 Pre-OPC grade distributions

1. The grade distributions of the single referee sessions show a great variability, which can only be characterized in terms of global statistical properties (quartiles, Skewness, Kurtosis, ...).

2. In general, the distributions do deviate from Normal, with a tendency for negative Kurtosis and positive Skewness (i.e. excess at poor grades).

3. Interesting correlations between the various indicators are detected. For instance, the dispersion is well correlated with the average, and so is the Skewness. Skewness and Kurtosis show a parabolic correlation, which is typical of complex, non-Gaussian systems.

4. The overall distribution of pre-OPC grades is non-Gaussian, with a positive Kurtosis (0.50) and a positive Skewness (0.71). The first, second and third quartile grades are 1.8, 2.2 and 2.8, respectively.

5. Given the great variety of observed distributions and their typically non-Gaussian nature, the justification and the effectiveness of the normalization procedure currently in place (which assumes Normal distributions) is questionable.

7.2 Post-OPC grade distributions

1. In the OPC meeting phase, the identity of single referees is not stored in the database, and so the analysis in only possible at the panel level, i.e. grouping together all the grades assigned within a panel, which is treated as a super-referee.

2. All statistical estimators show less dispersed distributions. The median Kurtosis becomes slightly positive, while the median Skewness increases.

3. When considering the overall distributions (i.e. not by referee or by panel), the pre- and post-OPC data do not differ substantially.

7.3 Run grades distribution

1. The comparison between the pre- and post-OPC data shows that the run grade distribution becomes more dispersed and more skewed towards poorer grades, while the median grade remains unchanged.

2. The panel discussions tend to populate the regions of good and bad grades, with an excess at the poor grades end.

3. Pre- and post-OPC run grades are well correlated (the correlation coefficient is 0.82). The correlation is tighter in the interval 1.5-3.0.

4. The overall slope is 1.2, indicating a tendency to have poorer grades after the Panel discussion. The data reveal a S-shaped dependency. At the two ends of the distribution (<1.7 and >2.8) the slope is lower than in the central region.

5. The general effect of the panel discussions is the promotion of well pre-OPC graded runs, and the demotion of poorly pre-OPC graded runs.

6. There is no statistically significant overall tendency to degrade or promote runs. Promotions are balanced by demotions.

7. The pre- vs. post-OPC correlation is tighter at the hi-end than at the low-end.

8. Pre-OPC data show that the dispersion of grades given by the referees to the same run shows a mild correlation with the average grade. In general, the dispersion increases with the grade value. Although the dispersion around the best fit line is smaller in the post-OPC data, the slope is a factor ≈2 larger in the pre-OPC data.

9. This finding partially supports the diffused belief that it is easy to identify the top proposals. It does not support the same belief about the other extreme of the distribution, where the data clearly show a systematically larger dispersion.

10. The consensus reached at the high end is partially artificial because, by definition, when a run has an average grade close to 1.0 the single referee grades must be all close to that value, even in the case of a completely random process. In other words, it is practically impossible to have a top grade run if the dispersion is large.

11. Although mild, the correlation found here is in partial contrast with the conclusions reached by Brinks et al (2011), who report complete uncorrelation between grade (or rank) and dispersion at both ends of the distribution. This is probably due to the much larger sample used in the present analysis.

12. The pre- and post-OPC dispersions for the same run show some degree of correlation (0.60), but the slope of the best fit relation is about 0.5. On average, the post-OPC dispersions are a factor ≈2 smaller than their pre-OPC equivalent.

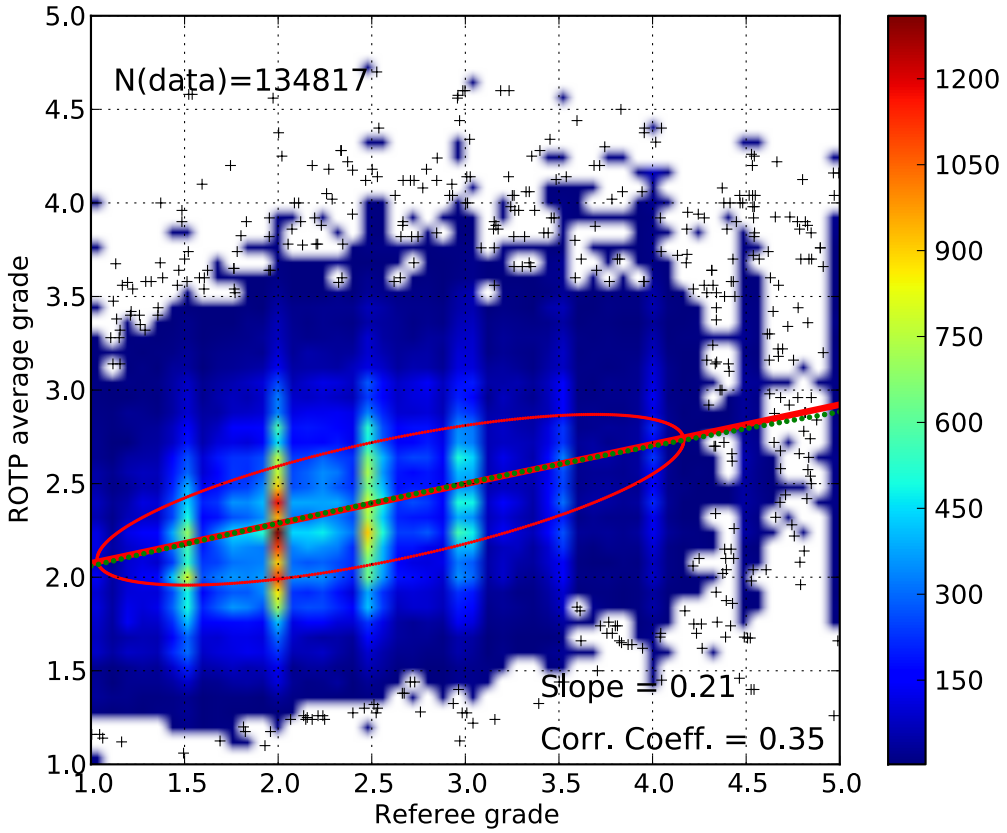

Figure 4 - The R-ROTP correlation for the pre-OPC grades. Only runs with 5 or 6 referees were included. The red and the green lines mark 1-st and 2-nd order polynomial fits to the data. The ellipse encircles 50% of the data. The vertical density enhancements corresponding to 1.0, 1.5, ..., are produced by the referees giving grades only within 0.5 bins.

7.4 Referee-Rest of the Panel correlation

1. The implementation of the review process does not allow the comparison of the grades given by different panels to the same proposal, as there is no intersection between the proposal lists of different panels.

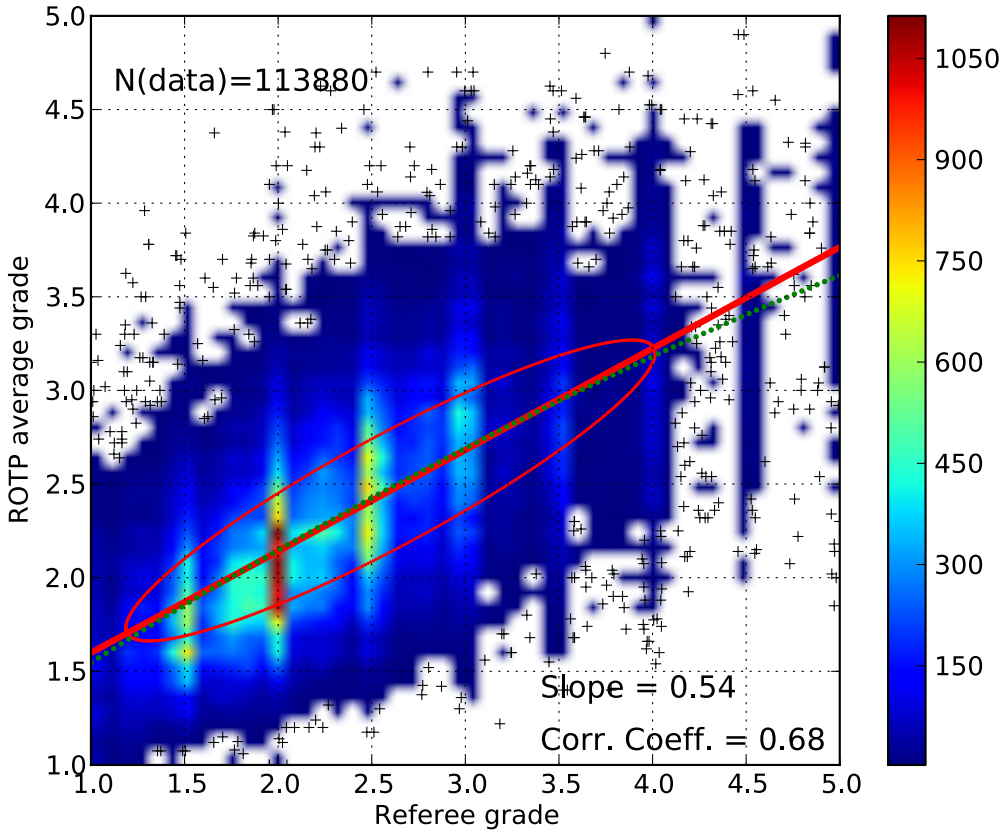

Figure 5 - Same as Figure 4 for the post-OPC grades.

1. All proposals assigned to a panel are read and graded by all non-conflicted panel members. As the majority of the proposals is reviewed by 5 or 6 referees, this allows a referee-to-referee correlation study.

2. To this aim, we analyzed the correlation between the grade of single referees and the average grade computed for the **rest of the panel** (ROTP). The pre-OPC data shows a very poor correlation (0.35), with a great dispersion around the best fit line (*Figure 4*).

3. The referee-ROTP correlation becomes stronger in the post-OPC data (0.68) and so does the slope of the best fit line. This indicates a significant increase in the statistical consensus.

4. This last finding undermines to a large extent all peer-reviews without meetings (including for instance the distributed approach by Merrifield & Saari 2009). The data clearly show that before the meeting the dispersion is high, and it is only with the meeting that a consensus is reached. Statistical consensus (i.e. small dispersion) does not necessarily mean that the panel attaches the proper scientific merit to a given proposal, but at least fulfills the mandate implicitly given by the community to the panel to vote and reach a "democratic" consensus. There are also obvious reasons: having to meet in a panel, the referees get better prepared (they have to discuss in person and be confronted with other peers). In a blind grading, no-meeting system, it is far easier to give a grade and be done with it without too much commitment.

5. Given that the data were obtained with 5-6 referees, even moving to a distributed review with, say, 10 reviewers would not change significantly the result in terms of increasing the statistical consensus.

7.5 The scientific return

1. There is a very weak correlation between proposal grade and paper impact (citations per year) within any of the rank classes. Top A-ranked proposals (i.e. having grades better than the median) have a median publication rate of 4.5±0.8 papers/year, to be compared with that of bottom A-ranked proposals, 2.5±0.7. The two values differ at the 2-sigma level (the overall A-rank citation rate is 3.5±0.6 papers/year).

2. Even considering extreme cases (e.g. >10 citations/year), no statistically significant differences are seen in the grade distributions within the same rank class.

3. For B-rank proposals, the citation rates are 3.3±0.6 and 2.4±0.4 citations/year, for grades above and below the median grade. The difference is within the 2-sigma level.

4. For A-rank proposals (TOO excluded) the non-publishing (NP) fraction is 56.3±3.0%. The

grade distributions of publishing (P) and NP proposals in this rank class are statistically indistinguishable, implying that the reviewing process has no predictive power (at least considering this simple P/NP binary metric).

5. The strongest correlation between OPC grade and publication rate is seen for TOO proposals (scheduled in A-rank by construction). The grade distributions of proposals that published is significantly different from that of proposals that did not publish. This suggests that the proposal review has some predictive power in terms of assigning higher merit to TOO proposals that will publish results (which includes the capability of finding suitable targets).

6. The case of VM proposals is similar to that of A-rank proposals: no statistically significant difference is seen between the post-OPC grades of P and NP proposals. The NP rate is slightly smaller than for A-rank (49.5±2.3%).

7. For B-rank proposals the NP fraction is 64.3±2.7%, which is likely affected by the lower completion rate. Within the class, better graded proposals (both in pre- and post-OPC data) are correlated to smaller NP fractions. Proposals with post-OPC grades below and above the median have NP fractions 59.6±3.6% and 68.9±4.0%, which differ at the ≅2.5-sigma level. Since B-rank execution is blind to the OPC grade (which was not propagated down the observing chain during the period range considered here), this suggests a mild predictive power in the review process.

7.6 Demographics

This section gives a short summary on the statistical properties of the ESO PI/coI community, as derived from the proposals submitted between P82 and P97. The statistics include all programme types (surveys excluded), after cleaning some spurious cases.

The sample includes 16,375 proposals, submitted by 3,300 distinct PIs (~5 proposals/PI) and including 11,007 distinct CoIs, for a total of 11,572 distinct users. Of the distinct CoIs, 8,272 (75.2%) were pure CoIs, meaning that never submitted proposals as PIs. Of the distinct PIs, 565

(17.1%) were pure PIs (i.e. never submitted a proposal as CoIs), 2,735 (82.9%) were also CoIs. The number of successful PIs (i.e. the PIs that had at least one proposal approved) is 2,065 (62.6%), implying that 37.4% of the PIs **never got any time**. Similarly, the number of successful CoIs (in the above definition) is 7,889 (71.7%), implying that 28.3% of the "community" never got time. The median number of periods in which PIs submit a proposal is 2, with 33% of them submitting only in one single semester.

What is still missing is an investigation to establish whether those scientists that did not get the time are those that submitted only one time (kind of "just tried"). The two fractions seem to match (37% of the PIs never got time, 33% of the PIs submitted only one proposal). The main question to be answered is: "Is there a fraction of the community that systematically does not get time?". This requires some more sophisticated analysis, which goes through the identification of "teams", while the above crude analysis was run on an individual PI basis.

8 Publication Delay

One of the main arguments that are often made against moving to one cycle per year (OCPY) is the loss of flexibility by the community, in terms of a diminished reaction time between the development of an idea and the publication of final results. In the current semester-based implementation, the worst case scenario is one in which the PI conceives the idea the day after the last proposal submission deadline, waits ~6 months before being able to submit a proposal for observations starting 6 months after, which will possibly take the whole period to be completed. This produces a pessimistic duty cycle of ~1.5 years. The opposite case is one in which the PI comes up with the idea a few days before the deadline and the observations are completed at the beginning of the semester, reducing the duty cycle to ~0.5 years.

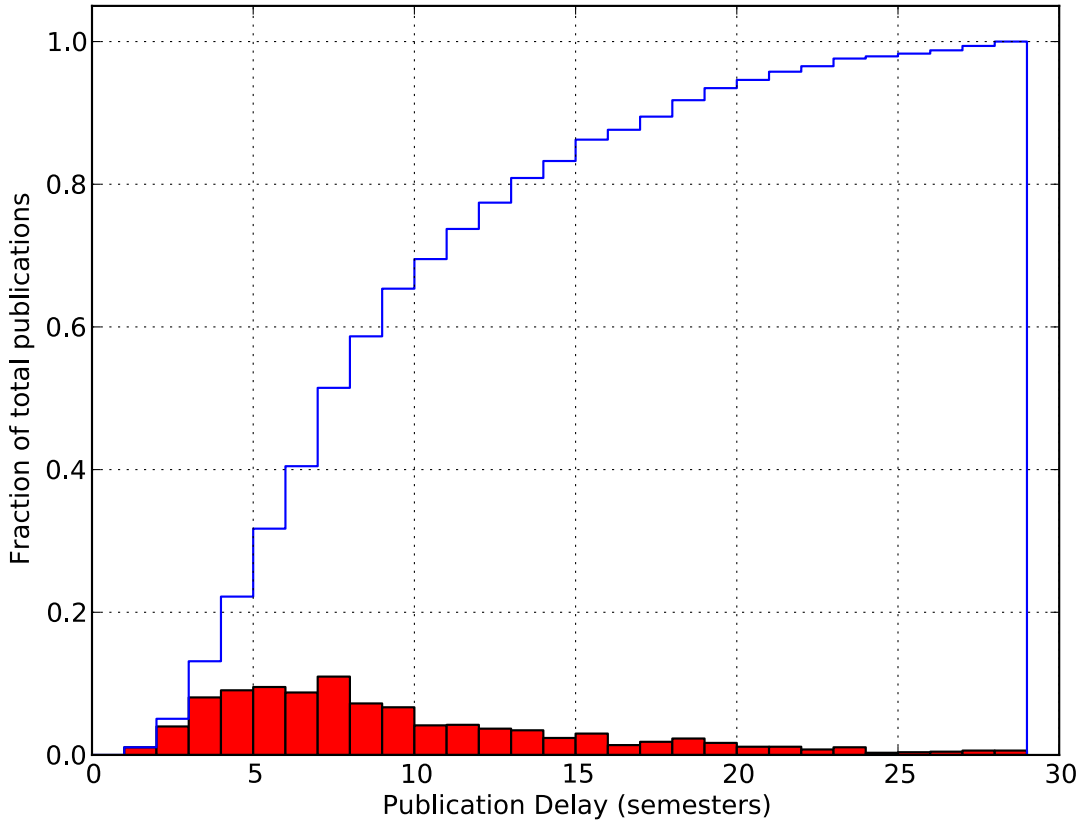

Figure 6. Publication delay time distribution. The blue line traces the cumulative distribution function.

So, in the current schema, the waiting time ranges between 6 and 18 months. Moving to OCPY would imply a waiting time from 1 to 2 years. In this process, the only quantity that can be objectively measured is the time delay between the date on which the last science frame was taken and the publication date[10].

For an exploratory analysis, we run a simple study using the ESO Telbib interface[11]. For each year in the interval 2008-2015 we extracted the publications produced by Paranal telescopes resulting

[10] Rigorously speaking, one would need to use the paper submission date, because the publication date includes the refereeing process and the delays introduced by it, which are not strictly related to the operational procedures adopted by the observatory. However, the submission date is not easily extracted.

[11] This is publicly available at: http://telbib.eso.org/.

from data produced by single programme IDs. Large Programmes and Director Discretionary Time proposals were excluded. A total of 1,303 refereed papers were selected.

Each publication is characterized by a publication year (t_p) and a scheduling period (P). For any given publication year one can derive the most recent period that produced a publication (P_0). For the data in the selected sample we found that:

$$P_0 = 2\,(t_P - 2008) + 81$$

which means that publications are produced at most by programmes scheduled in the central semester of the given year, which runs from Apr 1st to Sep 30th. The only exception is 2008, for which P_0 is 80 (rather than 81). The publication delay (in semesters) is then computed as:

$$\Delta P = P_0 - P$$

The data show that only 1.1% of the papers are published with a null delay (in the above definition), while the fraction grows to 11% for a delay of 6 semesters. It then steadily decreases for larger delays. This is illustrated in Figure 6 which also includes the cumulative distribution function. At face value, it takes 7 semesters to reach 50% of the publications, while the 95% level is reached in 20 semesters. A similar result was reported by Sterzik et al. (2016).

One may argue that since this analysis is based only on published papers, it does not give any information on those programmes that did not publish because the duty cycle was too long. However, the ESO-SNPP survey (see Sect. 9.1) revealed that only 2.3% of the users indicated obsolescence of the science case as the reason for non publishing. This fraction is negligible.

In this context, it is interesting to look at the Director General Discretionary Time (DDT) programmes. The DDT proposals can be submitted any time, and are reviewed internally by a standing committee that reports directly to the DG. The maximum processing time is 10 days and the observations are executed with top priority, typically within one month (or less in case of transient events). ESO receives on average two DDT proposals per week (about 100 per year).

The 50% DDT publication level is reached in 1.5 years from proposal submission, while the 95% level is reached in 5 years (Patat et al. 2014). Therefore, although the DDT has a very fast reaction time (both in terms of proposal review and observations execution), the publication return time is only a factor ~2 shorter than for regular proposals, while the proposal processing time is at least a factor of 10 faster.

In the light of the above facts, it is difficult to argue that a change to the OCPY would dramatically impact the productivity of the observatory. For the small number of proposals that do publish with a delay of 6-18 months (~15%), things can be made more efficient than in the present system by introducing a fast-track channel, which would in principle enable a return similar to that of the current DDT channel.

9 Scientific Impact

ESO has recently published a comprehensive study on the scientific return of the VLT (Sterzik et al. 2016). These are the main conclusions relevant to the present discussion:

1. Large Programs are by far the most productive, both in terms of number of publications and citations. Even after normalizing by the allocated time, their impact is still higher than that of normal programmes.
2. Normal programmes that turn into publications produce, on average, 1.5 refereed publications after 15 years.
3. DDT programmes are the most productive when considering their allocation (typically <5 hours).
4. GTO programs have on average a higher impact than normal programs.
5. TOO programs have on average a higher impact than normal programs.
6. VM programs exhibit high productivities and impact, in particular for normal programs with telescope time allocations of a few nights.
7. A-rank programs produce, on average, 2 refereed publications per program. C-rank

programs produce typically 1.6 publications per program, while B-rank programs reach 1.1 publications per program, which is 45% lower than for A-ranks.

8. B-rank programmes suffer from lower completion rates when compared to A-rank programmes (by construction). However, when considering the sub-set of fully completed B-rank programmes, their publication rate is 37% lower, indicating a correlation between productivity and OPC ranking.
9. C-rank programs perform well in terms of normalised productivity and impact, especially considering that they make use of poor conditions.
10. The publication ratio for all VLT programs (i.e. the number of programs that publish at least one refereed paper, divided by the number of all VLT programs scheduled) reaches 58% after 15 years. Depending on the mode and type of a program, about 33%-50% of all scheduled programs *do not* contribute to refereed publications.

As part of the TAWG activities, a dedicated survey was run in order to understand the reasons for the observed publication rate. The preliminary results are presented in the next section.

9.1 The SNPP survey

The Survey of Non Publishing Programmes (SNPP) targeted 1,278 programmes, scheduled between P78 and P90, that did not publish a refereed paper [yet]. The survey was launched in May 2016, and returned 967 valid responses. While the analysis of the data will be published elsewhere, here we provide a brief overview, and highlight the points that are relevant to this discussion.

9.1.1 Sample selection

The sample included NOR, GTO and TOO programmes scheduled in A-rank or Visitor Mode (VM) between period 78 and 90, to cover a significant part of the publication delay distribution (see Sect. 8). In order to minimize ambiguity in the results, only programmes with all runs scheduled as A-rank and/or VM were included. In addition, to remove obvious spurious cases, only

programmes that had some data in the archive where included; the threshold was set at one science frame per allocated hour. In the selected period range 2,716 proposals obeying to the above criteria were scheduled (90.7% of A+VM). Of these, 1,278 (47%) did not produce a refereed paper according to the ESO telbib system (Grothkopf et al. 2015). In terms of PIs, 1,143 could have published a paper (in the above definition), while 755 (66%) did not, implying that there is some systematic behavior. This can be better quantified looking at the fraction of PIs that published all proposals, some proposals or no proposal at all: 34%, 29% and 37%, respectively. About 45% of the PIs had only one proposal that could have produced a paper; of these, about 55% did not publish. On average, 1.1 proposals per PI did not produce a refereed paper [yet].

9.1.2 The questionnaire

The questionnaire was discussed within the TAWG and with the various ESO groups involved in operations. The PIs were asked to answer the following question: "*Why were you not able to publish the results of your observations in a refereed paper?*". Ten options were offered to the users:

1. I did publish a refereed paper (provide a hyperlink in the comments)
2. Insufficient data quality (observations out of required specifications)
3. Insufficient data quantity (partially completed programme)
4. Inadequate ESO data reduction tools
5. Null or inconclusive results
6. Lack of resources on the PI side
7. Science case no longer interesting
8. I am still working on the data (provide time estimate in the comments)
9. I published a non-refereed paper (provide a hyperlink in the comments)
10. Other

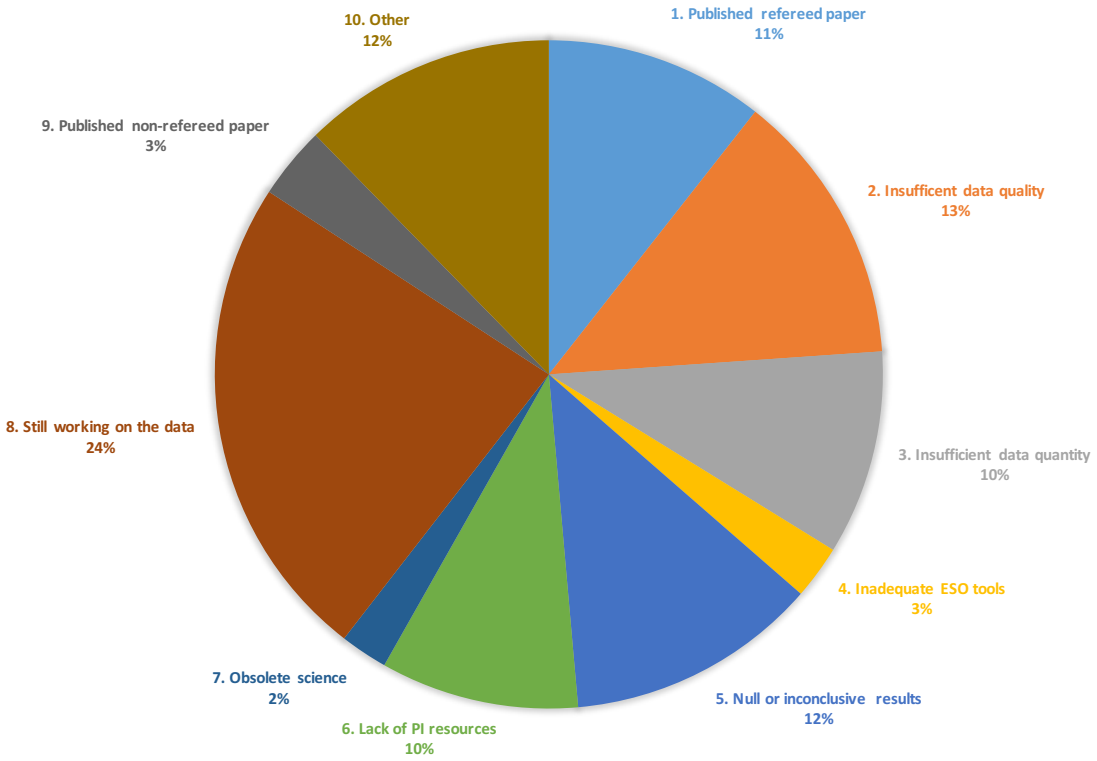

Figure 7 - Results of the SNPP survey.

The web form included a free text comment field. The responses were tagged with the programme ID, to allow the study of correlations between the selected options and the properties of the programme (e.g. time, constraints, instruments, scientific category, team composition, …). The analysis is ongoing at the time this report is being written. Therefore, here only basic results are presented and discussed.

Of the 1,278 targeted programmes we received replies for 967. Accounting for the fact that about 70 notification messages bounced back (PIs left astronomy or changed institute without updating their user portal profile), the response return is 80%, which is much higher than what is normally measured in web-surveys (~10%, see Fan & Yan 2010. Surveys that reach 40% are considered very successful). This indicates that the community is well motivated, despite of the somewhat inquisitory content of the survey and the fact that it was not anonymous.

The users were allowed to select more than one option. The distribution by number of selected options was: 55.5% (1), 31.1% (2). 9.8% (3), 3.6% (3<n≤5). The most popular single-option

response was n. 8 (24%; still working on the data), followed by n. 1 (9%; published). The most popular two-option response was n.6+n.8 (5%; lack of resources + still working on the data), followed by n.2+n.3 (3%; insufficient data quality + insufficient data quantity).

The final fractions were computed giving equal weight to the various options within the same response. The result is presented in Figure 7. The most popular answer is n. 8 (24%; still working on the data), followed by n.10 (12%; other), n.5 (12%; null or inconclusive results), n.1 (11%; published). The remaining options are all below 10% each.

A first interesting aspect emerges from the most frequent response: for about 24% of the non-publishing proposals the PIs declared they are still working on the data, although the selected programmes were all completed more than six semesters before the survey was run. The SNPP result is fully compatible with the publication delay time distribution obtained here (see Sect. Figure 6). The expected fraction of programmes that will eventually produce a publication but for which work is still in progress can be estimated as the weighted cumulated publication fraction over the look-back semesters in the SNPP (7 to 19), where the weights are the number of proposals in each semester. This yields 24.5%, which is indeed very similar to the measured value. This indicates that SNPP fraction of option 8 gives a realistic representation of the situation and is not just an “easy answer” from PIs trying to “defend” the lack of publications.

Another remarkable finding is that, although statistically significant, the fraction of null or inconclusive results is comparatively small (12.2%). At face value, this does not constitute a dominant reason for not publishing. One can imagine a certain reluctance on the PI side to “admit” a null or inconclusive result, which may bias the responses, although null/inconclusive results are definitely part of the scientific process and are not a negative fact per se.

One of the first explanations that was brought forward for explaining the ~50% publication rate was that, on average, only one out of two scientific ideas are successful and would therefore lead to a publication.

The SNPP gives a different indication. When considered together, options 6+8 (lack of PI resources + still working on the data) account for 33% of the cases, which points to a significant difficulty to sustain the data production rate. In other words, a significant fraction of the community has access to more data than it can deal with.

The other dominant aspect is that related to the data delivered by the observatory. When counted together, options 2 and 3 account for 23% of the cases. What the users have declared will have to be verified against the observatory records before drawing solid conclusions.

Most interestingly, only 2.3% of the cases were indicated as obsolete science. These occurrences can be identified with cases in which the data delivery duty-cycle was too long compared to the evolution in the given field. In any case, the fraction is very small.

A similar value is observed for the cases that did not publish a refereed paper, but did publish a non-refereed article. The implication is that, with very few exceptions, when a project does not produce a refereed publication, it does not produce any publication at all.

Finally, about 11% of the programmes declared that, contrarily to what is reported by telbib, a refereed publication was produced. This is under investigation, but a coarse inspection of the comments accompanying option 1 shows that in most of the cases the paper was not associated to the programme ID because this was either wrong or missing in the article acknowledgments. If this fraction is confirmed, it implies that the non-publication rate is ~5% smaller than the original estimate (42% vs. 47% in the SNPP sample). It also demonstrates that telbib is accurate to about 95%.

10 Proposal Resubmission

During the TAWG activities, an exploratory study on proposal resubmission was conducted. The results were presented in an internal report (*Report on proposal resubmission*, January 2016). The study provided the first quantitative estimate of proposal resubmission at ESO between periods 69 and 97, with a look-back time window of six semesters. It is based on abstract similarity analysis and it provides a lower limit to the true resubmission rate. The main conclusions can be summarized as follows:

1. During the period range covered by this analysis, the global resubmission rate steadily increased from ~25% to ~35%.
2. The growth is a probable consequence of the increased overall proposal rejection rate

that, in recent periods, on average exceeds 60% (see Figure 8).

3. The ratio between rejection and resubmission rates remained constant at about 60%, implying that on average ~40% of the rejected proposals are either not resubmitted or their abstracts are radically modified.
4. The largest fraction of resubmissions (>80%) originates from NOR programmes. GTO and TOO programmes contribute to ~6% and ~8%, respectively.
5. On average, about 50% of the LAR programmes are resubmitted.
6. Although there was some evolution with time, in the last ~12 periods about 50% of the resubmitted proposals originate from rejections.
7. The fraction of resubmissions stemming from previously accepted proposals is about 35%.

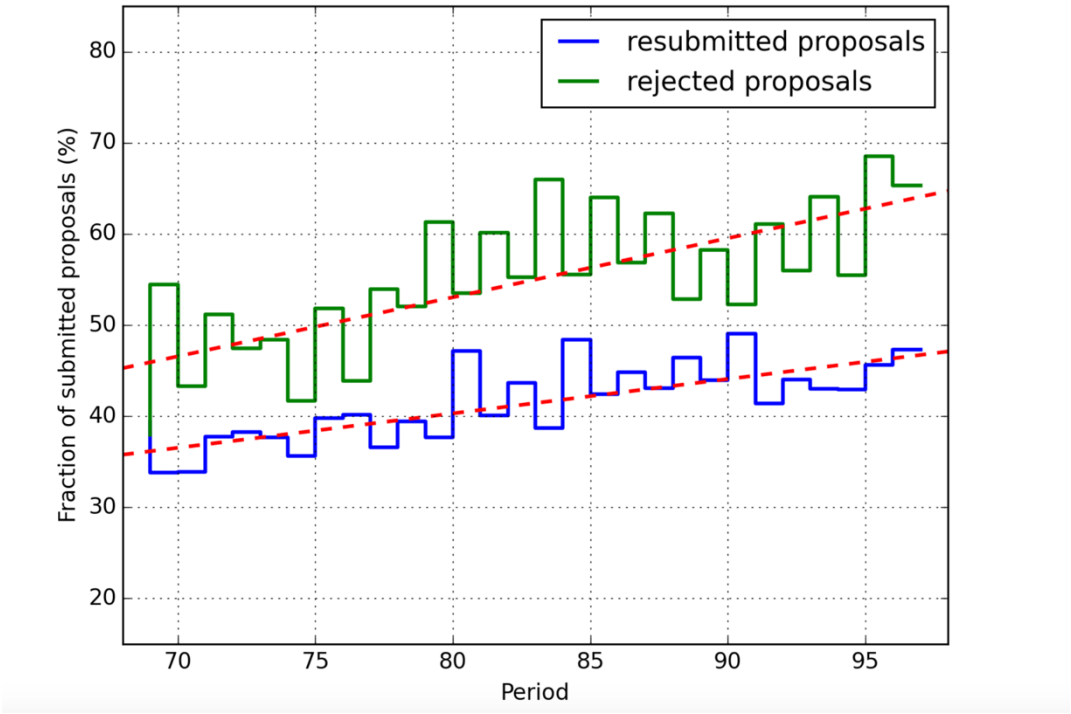

Figure 8 - Resubmission history over the period range 69 to 96. For comparison, the proposals rejection rate in the given semester is also plotted (green). The dashed lines trace least squares fits to the data.

8. Resubmissions of mixed-outcome proposals account for ~15% of the total resubmission.
9. After an initial evolution, the fractions of resubmitted, previously accepted proposals per rank class stabilized on roughly constant values.
10. The fractions of resubmissions from A-rank, B-rank and VM programmes are comparable (25%-30%).
11. The fraction of B-rank resubmissions is not larger than that of A-rank resubmissions.
12. The resubmissions of C-rank (filler) programmes only accounts for less than 10% of previously approved, resubmitted proposals.
13. In ~85% of the resubmissions, the PI of the proposal remains unchanged.

11 Existing inter-cycle proposal submission channels

11.1 Gemini Fast Turnaround

In the context of fast-reaction channels, it is important to briefly discuss the fast turnaround (FT) channel deployed at Gemini in 2015[12]. This channel features a monthly deadline for short, self-contained proposals, with criteria that overlap those of the ESO DDT Channel (pilot/feasibility studies, follow-up of unusual/unexpected astronomical events, speculative/high risk - high reward observations of fairly short duration). Proposal review is carried out by the PIs of other proposals submitted during the same round. When the number of proposals submitted in a given round is smaller than 7, an external standing committee runs the proposal review.

At variance with the ESO DDT time (for which 5% of the science time is reserved in a statistical way), the FT time is reserved in pre-allocated slots in the schedule.

The number of FT proposals per round in 2015 (Jan to Dec) was 17, 12, 7, 8, 3, 4, 7,4, 16, 16, 12, 5 (9.5 proposals/month), while in 2016 (Jan, Feb, May, Jun) was 15, 15, 16, 17 (15.7 proposals/round, 10.5 proposals/month)[13].

[12] http://www.gemini.edu/sciops/observing-gemini/observing-modes/fast-turnaround
[13] http://staff.gemini.edu/ft/

To give an idea of the scale, the average FT numbers are slightly above the typical value for the ESO DDT channel (8 proposals/month). Taking into account that ESO has 4 VLT units, and assuming that the above numbers scale linearly and that the communities behave in a similar way, one expects about 20 proposals/month.

11.2 HST Mid-Cycle

NASA's goal is to operate Hubble through 2020 and beyond, providing at least one year of overlapping science observations with the James Webb Space Telescope. This corresponds to reaching Cycle 28, providing the community with five further opportunities to propose for Hubble time. Those opportunities occur with an annual cadence. To enable the community to maximize Hubble's scientific potential in these final cycles, HST introduced a new type of GO proposals that allows a faster response to new discoveries.

Since Cycle 1, Hubble has provided the community with the opportunity to apply for in-cycle Director's Discretionary time observations. However, those programs are generally limited to observations of transient phenomena. Proposals to observe astrophysically-interesting but non-transient objects, such as newly discovered exoplanets or high redshift galaxies, are generally recommended for submission at the next annual proposal deadline.

Mid-Cycle GO proposals are designed to address phenomena that fall into the latter category. Up to 200 orbits will be made available for small-scale, high-impact science programs to be executed during Cycle 23. The proposal category is modeled to some extent on the Fast Turnaround observing mode offered by Gemini Observatory.

11.2.1 Proposal Criteria and Characteristics

Mid-Cycle GO proposals must meet the following prime criteria:

1. Proposers must provide an explanation of why the proposal could not have been submitted in response to the standard annual Call for Proposals: for example, the target source may have been identified subsequent to the most recent proposal deadline.
2. Proposers must provide a clear description of the scientific urgency of these observations and why they should be executed in the present cycle.

Mid-Cycle GO proposals will also have the following characteristics:

- Proposals are limited to requesting no more than 5 orbits;
- Observations should have minimal constraints to maximize scheduling flexibility;
- Observations taken for accepted programs will have a proprietary period of no more than 3 months;
- Proposals may request only HST time - joint proposals are not permitted;
- Proposers may apply for all available instruments. Proposals must be compliant with the technical restrictions described in the Cycle 23 Call for Proposals.

11.2.2 Review Process and Schedule

Members of the STScI Science Policies Group undertake an initial review of GO MC proposals to determine whether they meet prime criterion #1. Proposals that do not meet that criterion are not distributed for further review; the Principal Investigator is informed of that decision, and is free to submit the proposal at the next standard cycle deadline. Re submissions of rejected Cycle 23 proposals is rejected automatically.

Mid-Cycle GO proposals receive scientific review by members of the community who have contributed to recent HST TAC reviews. Each proposal is distributed to four reviewers.

Reviewers are required to address the following issues:

1. The scientific merit of the proposed investigation and its potential contribution to the advancement of scientific knowledge

2. The program's importance to astronomy in general
3. The scientific urgency of the observations and their suitability for a Mid-Cycle proposal
4. The program science goals can only be achieved by observations with Hubble Space Telescope – yes/no.

Reviewers are also required to provide brief text feedback on the strengths and weaknesses of the proposal, and a rationale for their grades. The STScI Science Policy group will collect the reviews and rank order the proposals. The Director will make the final selection of proposals to be awarded time.

12 REFERENCES

Bohannon, J., 2013, *Who's afraid of peer review?* Science, 342, 6154, p. 60-65, DOI: 10.1126/science.342.6154.60

Brinks, E., et al., 2011, *ESO OPC Working Group Report*, presented to ESO Director General, September 12, 2011

Fan, W. & Yan, Z., 2010, *Factors affecting response rates of the web survey: A systematic review*. *Computers in Human Behavior*, *26*(2), 132-139.

Fox, J. & Petchey, O. L., 2010, *Pubcreds: fixing the peer review process by "Privatizing" the reviewer commons*, Bulletin of the Ecological Society of America, DOI: 10.1890/0012-9623-91.3.325

Gillies, D., 2014, *Selecting applications for funding: why random choice is better than peer review*, Roars Transactions, A Journal on Research Policy and Evaluation, v. 2, n. 1, May 2014

Grothkopf, U., Meakins, S. & Bordelon, D., 2015, *If we build it, will they come? Curation and use of the ESO telescope bibliography*, Science Operations 2015: Science Data Management - An ESO/ESA Workshop, held 24-27 November, 2015 at ESO Garching.
http://adsabs.harvard.edu/abs/2015scop.confE..26G

Kitcher, P., 2011, *Science in Democratic Society*, Prometheus Books

Kurokawa, D., Lev, O., Morgenstern, J. & Procaccia, A.D., 2014, *Impartial Peer Review*, *EC'14*, June 8–12, 2014, Stanford, CA, http://www.seas.upenn.edu/~jamiemor/papers/peer-review.pdf

Merrifield, M. R. & Saari, D. G., 2009, *Telescope time without tears: a distributed approach to peer review*, Astronomy & Geophysics, 50, 4.16

Mervis, J., 2014, *Want a grant? First review someone else's proposal*, http://www.sciencemag.org/news/2014/07/want-grant-first-review-someone-elses-proposal

Patat, F., Hoppe, E., Cristiani, S. & Hussain, G., 2014, *Director General Discretionary Time – A Statistical Report over the period April 1, 1999 – March 31, 2014*, ESO internal report presented to the Director General, November 2014

Patat, F., 2016, *Gender Systematics in Telescope Time Allocation at ESO*, The Messenger, vol. 165, p. 2-9

Price, E., 2014, *The NIPS Experiment*, Moody Rd, http://blog.mrtz.org/2014/12/15/the-nips-experiment.html, downloaded on Jul 28, 2016

Reid, N., 2014, *Gender correlated systematics in HST proposal selection,* PASP, 126, 923

Snell, R.S., 2015, *Ménage a Quoi? Optimal Number of Peer Reviewers*, PLoS ONE 10 (4): e0120838. doi:10.1371/journal.pone.0120838

Sterzik, M., et al., 2016, *The impact of science operations on science return of the Very Large Telescope,* SPIE, 9910-02, http://dx.doi.org/10.1117/12.2232521

13 APPENDIX A – TERMS OF REFERENCE

The tasks assigned the TAWG are divided into two main categories:

Proposal submission

a) Critically review the proposal submission channels currently implemented at ESO and identify possible areas of improvement and/or new channels in relation to the different scientific cases they are supposed to serve.

b) Investigate the need for decreasing the response time between proposal submission and the communication of the proposal selection, possibly differentiating among the various programme types.

c) Provide guidelines for the implementation of incentives as to foster larger time requests from a smaller number of teams.

d) Review the frequency of calls for proposals and evaluate the possible differentiation among different program types. The concept of rolling-deadline should be considered, at least for some proposal types.

e) Review the concept of joint proposals to improve the synergies with other large, present and future astronomical facilities (ALMA, JWST, PLATO, EUCLID, ...).

f) Examine the foreseeable evolution of the proposal selection and time allocation processes into the E-ELT era.

g) Consider whether the status quo, in terms of proposal submission or proposal review, is preferable to any proposed alternative models.

Proposal review

a) Critically analyze the implementation of peer-review in the ESO system (including time allocation committee structure and selection) and evaluate alternative approaches (distributed review, external consultancy, ...) in the framework of decreasing the administrative load on the referees without impacting the quality of their assessment.

b) Review and study alternative approaches (e.g. random selection) in the light of the recent criticisms to the peer-review paradigm and their applicability to telescope time allocation.

c) Consider and evaluate mixed proposal review approaches to be applied to different proposal types.

The TAWG will produce a final report presenting the analysis and the recommendations for implementation. The report should also include an implementation plan that takes into account the change management necessary to allow a smooth community transition into the new features of the system.

14 APPENDIX B – Recommendation of the OPC Working Group

1) ESO should use the current system to perform a (partially) blind test of Peer Review, e.g., by having a fraction of proposals ranked by multiple panels where those panels in principle see the same cross section of proposals, i.e. within each of the Categories A, B, C, & D.
2) ESO is to consider introducing a category of monitoring proposals (MPs) for projects that are long-term (several years) but at low impact per semester.
3)
    a) ESO should consider soliciting proposals only once per year. The expectation is that although there will be an increase in proposals, the numbers will not double. A mechanism akin to DDT will need to be provided to allow telescope access at short notice when warranted.
    b) ESO is to consider inviting proposals for La Silla telescopes and for LPs/MPs on a yearly basis, in alternate semesters.
4) ESO might consider, either on all telescopes/instruments or a subset, to offer observing time exclusively in large allocations, akin to the process used for large surveys. This will have the effect of outsourcing the process of selecting and consolidating projects to the user community, resulting in fewer, but larger, bids, by consortia rather than individuals or small groups.
5) ESO is encouraged to perform a larger scale experiment, more closely linked to the observing proposal process, to explore distributed Peer Review as an alternative method.
6) The OPC-WG suggests that those proposals where the Phase 2 time request exceeds by some set margin that requested originally in Phase 1, be removed from the scheduling process, irrespective of their post-OPC ranking.
7) ESO should convene a similar working group in three years time, to assess the impact of changes made in response to this report, and to analyze the data produced by the internal calibration tests (Recommendation 1), and larger-scale experiments (Recommendation 5).