DOI: 10.18727/0722-6691/5147

The Distributed Peer Review Experiment

Ferdinando Patat[1]
Wolfgang Kerzendorf[2,3,4]
Dominic Bordelon[1]
Glen Van de Ven[5]
Tyler Pritchard[2]

[1] ESO
[2] Center for Cosmology and Particle Physics, New York University, USA
[3] Department of Physics and Astronomy, Michigan State University, USA
[4] Department of Computational Mathematics, Science and Engineering, Michigan State University, USA
[5] Department of Astrophysics, University of Vienna, Austria

All large, ground- and space-based astronomical facilities serving wide communities face a similar problem: in many cases the number of applications they receive in response to each call exceeds 1000. This poses a serious challenge to running an effective selection process under the classic peer-review paradigm, in which the proposals are assigned to pre-allocated panels with fixed compositions. Although, in principle, one could increase the size of the time allocation committee, this creates logistic and financial problems which place a practical limit on its maximum size, making this solution unviable beyond a certain volume of applications. For this reason, alternative solutions must be sought. One of these is the so-called Distributed Peer Review (DPR) in which, by submitting a proposal, the Principal Investigators (PIs) agree both to act as reviewers and to have their proposal reviewed by their peers. In this article we report the results of a DPR experiment run by ESO in Period 103, in parallel with the regular review by the Observing Programmes Committee (OPC).

Introduction

Following the start of VLT operations in 1998, the number of applications to use ESO telescopes has been steadily growing, exceeding 1100 proposals in Period 84. After this peak, the number of submissions per semester stabilised at around 900 (Patat et al., 2017). Despite the significant growth of the user community, which has made ESO one of the largest astronomical facilities in the world, the way telescope time applications are reviewed has remained substantially the same since 1993. Barring the necessary increase in the number of reviewers, the procedure has changed in the details, but not in its substance. Following steady growth in the numbers of submissions, the current review load is about 70 proposals per panel member and up to 100 for OPC-proper members (the latter serve on a second panel which reviews the recommendations across all science categories). These numbers have reached critical levels, requiring a re-evaluation of the procedures and an examination of the effectiveness of peer review.

The pressure on the peer review process has been the subject of a study by the ESO OPC Working Group (Brinks et al., 2012) and the Time Allocation Working Group (TAWG; Patat, 2018a). Both studies identified the excessive number of proposals per referee as the most urgent problem that ESO needs to tackle. Not only does the workload severely affect the referees (also increasing the rejection rate during the recruitment phase), but it can also have an impact on the quality of the reviews and the feedback provided to the applicants, with potentially serious consequences. The feedback has been repeatedly and consistently identified as a major problem by the OPC and the Users Committee, and via direct communications from numerous individual users. Problems with the peer review could ultimately affect the scientific productivity and impact of the Organisation itself. A number of recommendations have been proposed by the working groups, some of which are interdependent.

As a first step, since Period 102 ESO has decreased the number of referees (from six to three) who review a proposal ahead of the OPC meeting. Triage is then applied using the three pre-OPC meeting grades, with about the lowest 30% of proposals being rejected. At the meeting all non-conflicted panel members are then asked to discuss and grade only the surviving proposals. While this measure has successfully reduced the workload of the panel members, it has become cumbersome to manage in practice. For example, late dropouts during the review process can reduce the number of pre-meeting reviews per proposal, making the triage procedure less robust. While this change was relatively easy to implement, experience gained during Periods 102 and 103 suggests that the negative consequences outweigh the benefits. It is clear that further and more drastic and structured actions need to be taken; these include a move to an annual cycle and the deployment of a fast track channel (FTC; see Patat, 2018a).

By construction, the FTC requires a short duty cycle during which referees are continuously on duty. The most suitable mechanism for reviewing the proposals is a Distributed Peer Review (DPR), one of the most innovative schemes through which the load on referees can be alleviated (Merrifield & Saari, 2009). This concept has been successfully applied to the Fast Turnaround channel deployed at the Gemini Telescope, which has processed over 1000 proposals in this way since 2015. The Gemini Observatory has published a report (Andersen et al., 2019) and updates are continuously provided on its webpages[1].

Depending on the fraction of total telescope time that is allocated via the FTC, this channel may also serve to decrease the load on the OPC, which would then focus only on proposals with larger time requests. ESO has conducted a systematic study aimed at better evaluating the application of DPR to its programmes. In Period 103, in parallel with the regular OPC cycle, a DPR experiment was run involving a subset of submitted proposals. This article presents a brief description of the experiment setup and summarises an analysis of several statistical indicators. More details can be found in Kerzendorf et al. (2019).

Distributed Peer Review and the DPR Experiment

Different measures to alleviate the load on the reviewers have been and are being considered by various facilities. These include drastic solutions, like the one deployed by the National Science Foundation (NSF, USA) to limit the number of applications (Mervis, 2014a). The

Distributed Peer Review (DPR) concept is simple; in submitting a proposal the PI agrees to review *n* proposals submitted by peers, and to have her/his proposal/s reviewed by *n* peers. Also, if s/he submits *m* proposals, s/he accepts to review $n \times m$ proposals, hence essentially limiting the number of submissions through a self-regulating mechanism. Following this idea, the Gemini Observatory deployed the DPR for its Fast Turnaround channel (Andersen et al., 2019), which is capped to 10% of the total time. The NSF also explored this possibility with a pilot study in 2013, in which each PI was asked to review seven proposals submitted by peers (Ardabili & Liu, 2013; Mervis, 2014b). The NSF pilot was based on 131 applications submitted by volunteers within the Civil, Mechanical and Manufacturing Innovation Division, but the outcome is unknown as no report on the study was published. Interestingly, a similar pilot experiment was carried out in 2016 by the National Institute of Food and Agriculture[2]; in this case too the results were not published. Despite the general acceptance that followed the deployment of this channel at the Gemini Observatory, to the best of our knowledge the Fast Turnaround channel is the only example of DPR being employed by a large-scale astronomical facility.

In the specific case of ESO, the TAWG tasked to address these issues has produced a set of recommendations. The core aim is to reduce the number of applications per reviewer, which has been identified as an urgent action that ESO needs to take (Patat, 2018a). The deployment of DPR falls within the recommendations. As a first step, and after consulting the advisory bodies, ESO decided to run a test during the ESO Period 103 in parallel to the regular OPC review. The experiment was designed in line with the implementation at Gemini, enhancing the process by means of Natural Language Processing (NLP) and Machine Learning (a different method of using NLP for proposal reviews can be found in Strolger et al., 2017).

The DPR experiment was announced in the Call for Proposals for Period 103, released on 30 August 2018. A total of 172 PIs — representing 23% of all distinct PIs in that semester — volunteered to participate in the experiment. This implied that each would review eight proposals submitted by peers and have their proposal refereed by the same number of peers. The participants were given two weeks to complete their reviews and were informed that the outcome of the DPR would have no effect on the fate of their proposals. By the deadline (22 October 2018) 167 (97.1%) had completed their task. In a real implementation the five PIs who did not meet the deadline would have had their proposals automatically rejected. In this experiment however, their proposals were kept in the sample, but the PIs did not receive the final feedback. Additionally, the participating PIs were asked to fill in a web-based questionnaire covering various aspects of the experiment. A total of 140 (83.8% of the DPR sample, 19% of the total PI sample of P103) returned the completed form.

The proposal distribution was performed using two channels, which we will call OPC Emulate (OE) and DeepThought (DT). In both cases the reviewers were assigned eight proposals each. For the OE channel, 60 volunteers were selected at random and assigned, on the basis of the category of the proposal each submitted, to the four scientific categories: A (Cosmology), B (Galaxy Structure and Evolution), C (Planets, Star Formation and Interstellar Medium) and D (Stellar Evolution). The underlying (and reasonable) assumption is that a scientist submitting a proposal for a given category is an expert in that same area. This emulates the case of the real OPC, in which a person only receives proposals within her/his area of expertise.

For the remaining 112 volunteers selected for the DT channel, the process was as follows. For each scientist, a knowledge vector was built based on their publications, which were downloaded from the public SAO/NASA Astrophysics Data System database (ADS) and processed by a machine learning algorithm (Kerzendorf, 2017). The same approach was used for the proposals and applied to their scientific rationale. The match between the referee expertise and the area covered by the proposal was then quantified through the "cosine distance", which is directly related to the angle formed by the two hyper-vectors; a null cosine signals a complete mismatch (orthogonal knowledge vectors), while a unit cosine indicates a case of perfect match (parallel knowledge vectors). For the purposes of the statistical analysis, each DT referee received four proposals with the largest similarity, two proposals with median similarity, and two proposals with the lowest similarity.

The participants were not aware of the distribution mechanism just described. They were just provided with a simple web-based interface giving them access to the eight assigned proposals and allowing them to review, grade and comment on the applications. Before accessing the proposals, the referees were asked to sign a non-disclosure agreement, very similar to that signed by the OPC and Panel members.

During the review phase, the participants were also asked to declare any scientific/personal conflicts, while institutional conflicts were automatically taken into account by the distribution software, based on the affiliations recorded in the User Portal database. For each proposal, the referees had to fill in a comment (with a minimum length of 80 characters), and also provide a self-evaluation of their expertise level (high/medium/low) for each proposal assigned to them.

Once the review process was completed, the grades of the various referees were combined using a simple average (similar to the regular OPC process), and a final ranking list was compiled. The PIs were then provided with the quartile rank and the individual, unedited anonymous comments. Finally, they were asked to provide feedback on the experiment via a web-based form; this included a request to express the usefulness of each comment they received on their proposal.

General statistics and demographics

Although, in principle, each proposal should have been reviewed by eight scientists and each scientist should have reviewed eight proposals, because of the scientific/personal conflicts declared during the refereeing process (and to a much smaller extent because five participants did not complete the process),

both these numbers were on average smaller than eight. The number of reviewers N_r ranged from 4 to 8, with an average of 7.3; in 95% of the cases the number was $N_r \geq 6$. The number of proposals N_p varied from 5 to 8, with an average of 7.6, and $N_r \geq 6$ in 98% of cases. The DPR produced a total of 4055 distinct grade pairs, to be compared with the maximum number of pairs $172 \times 8 \times 7/2 = 4816$ (see below for more details) one would obtain in the case of no conflicts and no dropouts.

The F/M gender distribution of the DPR participants (32/68) and the scientific seniority distribution derived from the DPR questionnaire (see Figure 1) reflect the underlying PI population of ESO users (Patat, 2016). Since participation in the experiment was on a completely voluntary basis, we cannot exclude the presence of self-selection biases. For instance, one could argue that researchers who already had a positive opinion of the DPR concept would be more willing to participate than opponents, hence introducing systematics into the final analysis. On the other hand, if the community were strongly against the paradigm, one would expect a similar effect. In general, although we cannot guarantee that there are no specific attributes that lead the participants to self-selection, the demographics indicate that, if they exist, they are well hidden.

An important aspect regarding the demographics of the experiment concerns the fraction of junior scientists. Since, as a rule, the regular panel members serving on the OPC are required to have a minimum seniority level (typically starting with scientists at their second postdoc onward), this establishes a significant difference between the two pools of reviewers. In the case of the OPC, the distribution is heavily skewed towards senior members (88%), with a small fraction of postdocs (12%) and no students (Patat, 2016), while the postdoc and student reviewers reach about 18% in the case of the DPR sample (Figure 1).

Most DPR participants were relatively experienced in submitting proposals (Figure 2), although almost 60% of them had never served on a time allocation commitee before (Figure 3). Although

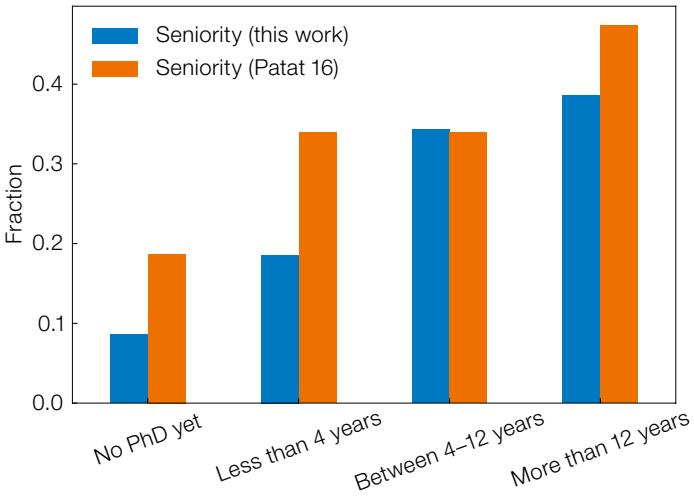

Figure 1. Scientific seniority distribution of the DPR sample (blue) and the OPC sample (orange). From Patat (2016).

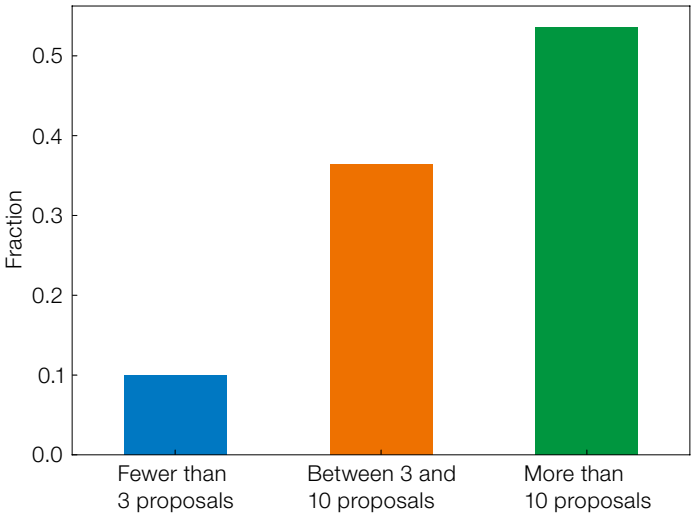

Figure 2. Distribution of the number of proposals submitted to ESO by the DPR participants.

there are published studies that indicate reviewers who self-report higher levels of expertise tend to be less generous in assigning the top grades (Gallo et al., 2016), the differences seen between the grade distributions of senior and junior DPR participants are not statistically significant.

Referee-Proposal matching

In the regular OPC process, the panel members are recruited to cover the widest possible range of astrophysical areas. Each of the selected reviewers is asked to declare her/his expertise by providing sub-categories from the same list used by the applicants to categorise their proposal. While the PI is allowed to indicate one single proposal sub-category (within a given scientific category), the panel members are requested to identify three sub-categories, ranking them in order of expertise. This information is then used to compose review panels in such a way that the expertise coverage within each of them is as broad as possible. This is required by any schema in which physical panels exist, which is in turn a constraint stemming from the fact that the panels have to meet face-to-face and discuss the same set of proposals. This introduces a certain rigidity, which is also related to the relatively small number of available reviewers.

Since DPR has the advantage of involving a much larger number of reviewers, it allows a significantly more flexible and

more objective approach in which, for each proposal, an ad hoc, optimised panel can be formed. A key ingredient in this approach is the proposal-referee matching, which should work without the need for human supervision, especially when the turnaround has to be fast.

For this purpose, the DT algorithm used in the DPR experiment was designed to predict what we call domain expertise, which in this context can be considered to be the objective ability of a given scientist to review a given proposal. Before we discuss its reliability, we examine how referees assessed their own ability to review each proposal assigned to them. As anticipated in the introduction, during the refereeing process each participant was asked to express their self-perceived expertise level for each of the assigned proposals, resulting in about 1200 evaluations. The distribution of participants' self-evaluated ability to review the assigned proposals is presented in Figure 4, where we have used different colours for the different classes of scientific seniority. As expected, junior scientists tend to perceive themselves as experts less often than senior scientists do. Also, they often indicate that they have limited knowledge of a given field. We take this is an indication that the self-evaluated ability of a referee to review the assigned proposals is a useful proxy of the more objective (albeit more abstract) concept of domain knowledge.

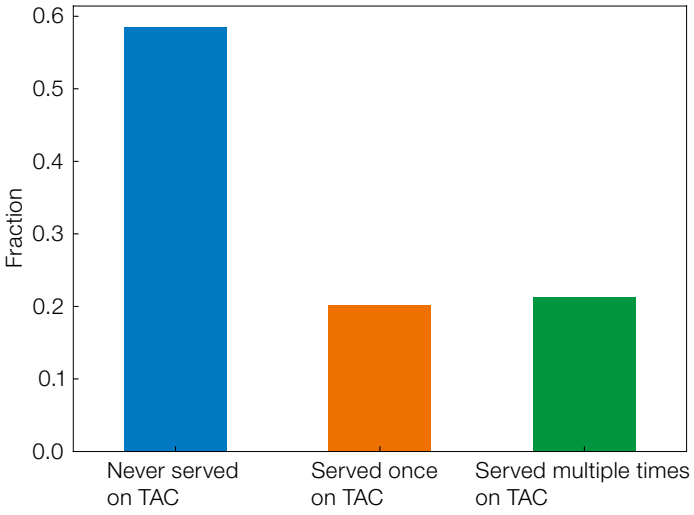

Figure 3. Distribution of expertise in serving on Time Allocation Committees (TAC) for the DPR participants.

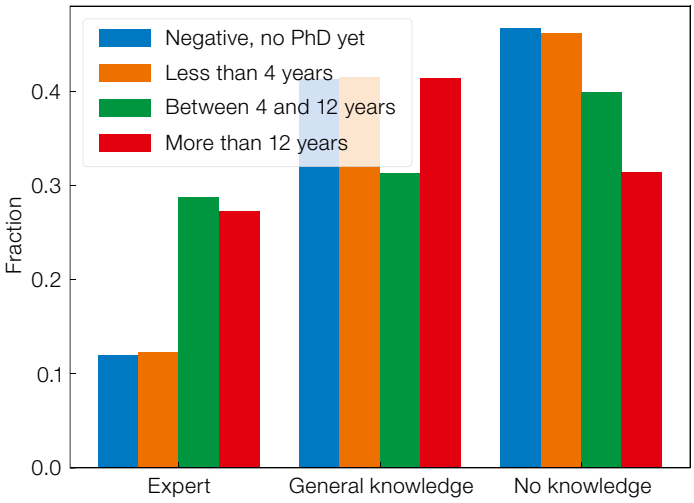

Figure 4. Distribution of self-reported domain knowledge for the different scientific seniority of the DPR participants.

The data collected in the DPR experiment enable an additional analysis of a possible gender dependence on the above self-evaluation. This has been reported, for instance, by Huang (2013), who concluded that females tend to under-predict their performance in certain STEM fields. Our data suggest that, at least for postgraduates in the domain of astrophysics, there is no statistically significant gender difference.

Since the DT is designed to predict the expertise of a referee with respect to a given proposal, the first question one should ask is how reliable the algorithm is. Obviously, there is no absolute reference; the DT is one possible objective estimate of this quality. Therefore, as a first exploratory test, one can check the DT results against the self-evaluation of expertise, which can be considered as a reasonable first approximation to the underlying domain knowledge. From a statistical point of view, this is equivalent to computing the Bayesian conditional probability P (self-reported | DT) of having a certain self-reported expertise level, given the DT-inferred level. In simpler words, one checks how the self-reported and DT-inferred levels correlate. The result is presented in Figure 5, which shows an encouragingly high correlation. For instance, the probability that the DT considers a match as the worst which the referee believes is the best, is less then 1%. At the other extreme, it is very likely (78%) that if the DT estimates the match is poor, the referee is of the same opinion. The agreement on the best matches is at the level of 50%, while for 81% of the best DT matches, the referees perceive them to be the top and intermediate classes. As shown in Figure 5, the correlation in the intermediate cases becomes fuzzier. With the available data it is impossible to tell which of the two estimators is responsible for the observed noise. If on the one hand we can argue that the DT approach has obvious limitations (which is certainly true), on the other hand the self-reported levels are affected by a significant level of uncertainty, as they are related to subjective perceptions rather than to objective criteria.

Another aspect is the importance of proper proposal-referee matching. Our direct experience, accumulated over many years of managing the review process at ESO, shows that, in addition to the obvious problem related to excessively large numbers of proposals, panel

members report a general uneasiness when dealing with proposals in areas in which they feel they are not experts. For a more quantitative assessment, DPR participants were asked to express their level of confidence, using a four-point scale, when asked to evaluate those cases; the corresponding distribution is presented in Figure 6. In about 60% of the cases, the reviewers were not comfortable with this situation. This implies that better matching of expertise gives the reviewers a better experience, an aspect which should not be underestimated.

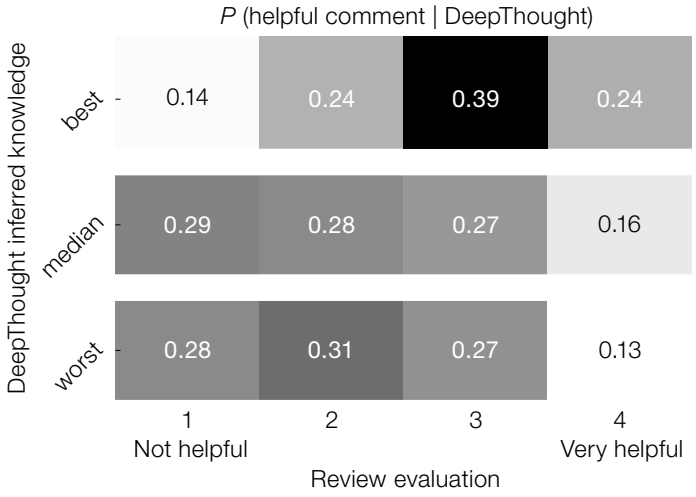

Figure 5. Conditional probability for the various combinations of self-reported and DT-inferred knowledge level.

Feedback quality

In the classical review concept, the feedback provided by the panel to the PI is supposed to reflect the consensus opinion. This paradigm has at least two obvious limitations: (a) proposals that are triaged out (i.e., the bottom ~ 30%) are not discussed, and the feedback is based on the opinion of the primary referee; (b) for proposals that are discussed during the face-to-face meeting the primary referee tries to capture the main points of the discussion and produces a single comment. There is simply not enough time for the panel members to review all the feedback and to make sure it reflects all the aspects of the discussion. In the current implementation at ESO, the comments are formally supervised by panel chairs, who are responsible for the integrity of the feedback (particularly as it relates to the language used). The net effect, possibly coupled with a sub-optimal matching between proposal and referee, is a high level of dissatisfaction in the community, which is consistently reported by the Users Committee; the dissatisfaction reported is about 30% for all of ESO and exceeds 50% for ALMA[3].

Since the TAWG recommended the use of DPR for a FTC, no attempt was made to produce consensus feedback and/or to edit/check individual comments, which were distributed to the PIs in their original form. The purpose of this implementation was two-fold: (a) to get feedback on the concept itself, and (b) to detect possible problems (for example, inappropriate language) generated by the unedited/unfiltered text.

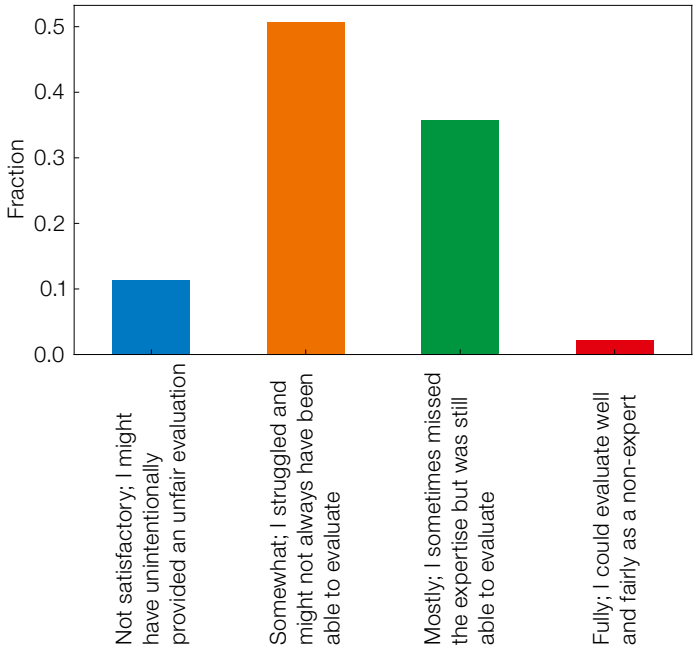

Figure 6. Distribution of the answers to the question: "How satisfactorily were you able to evaluate the proposals for which you were not an expert?".

The participants were asked to rate each of the comments they received for their proposal, based on its helpfulness. It is important to stress that they were not asked whether the comments were good or bad, or whether they liked them or not, but whether they were useful for improving the quality of their proposal. The general response was very satisfactory, as shown in Figure 7, with more than 60% of the comments judged as being useful, and about 5% not useful. One of the questions also concerned the comparison with the edited OPC comments received by the PIs in previous semesters (99% of the sub-sample that responded). In about 40% of the cases the DPR was reported to have provided better comments, while the fraction of comments with quality similar to, or better than the OPC reaches about 85%.

The analysis of comment helpfulness as a function of the reviewer's expertise (either self-reported or DT-inferred) shows that the dependence is mild in the central regions; the experts very rarely gave unhelpful comments and, conversely, non-experts rarely gave very helpful comments. A similar analysis as a function of

the reviewer's scientific seniority reveals a flat distribution (within the noise), with one remarkable exception: graduate students seem to be unable to provide very useful comments. This may signal a training issue, which can probably be addressed by exposing the students to schemes like the DPR. Finally, no statistically significant difference is seen between the helpfulness of comments written by female and male referees.

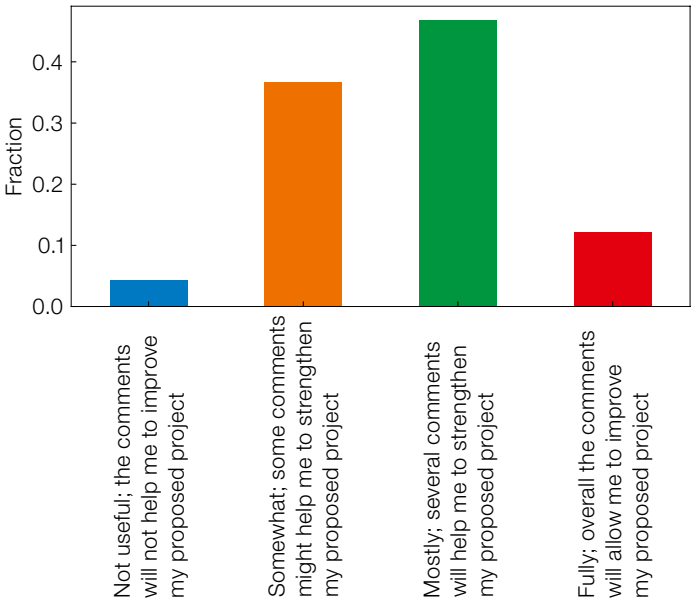

Figure 7. Distribution of the "helpfulness" ratings of the referee comments for the entire DPR sample.

A brief primer on subjectivity

Before we proceed with the comparison between the final OPC and DPR outcomes, a digression on the subjectivity inherent in the process is necessary. Although it is common knowledge that two different panels reviewing the same set of proposals would provide different rankings (and this is often used to compare time allocation committees to roulette), quantitative statements are very rare. This matter is addressed in great detail in an extensive study based on about 15 000 ESO proposals (Patat, 2018b; hereafter P18). The interested reader is referred to the paper for a thorough discussion, while here we will focus only on the concepts relevant to the present discussion.

One way of quantitatively describing the reproducibility of a review process is the correlation between the grades attributed to the same set of applications by two distinct bodies. These bodies can be composed of a single individual or of several members. We will be talking about referee–referee (r–r) and panel–panel (p–p) correlations. In the first instance, one simply considers all the distinct grade pairs attributed by referee #1 and referee #2 to the same set of proposals, placing them in a diagram in which the grades are used as coordinates, so that each single grade pair is represented by a point. One can then repeat the process for all possible referee pairs, plotting all the corresponding points on the r–r plane. Since the same proposal is graded by many reviewers, each single proposal is represented on the r–r plane by a cloud of points.

In the simplifying assumption that each proposal is seen by N_r referees, the number of distinct grade pairs n_p for each proposal is $n_p = N_r\,(N_r–1)/2$. For instance, in the case of the DPR experiment, with typically $N_r = 7$, the above combinatorics formula yields 21 distinct pairs per proposal. Of course, the same operation can be repeated for all N_p proposals in the sample, which will populate the diagram with $N_p = 172$ clouds of points. In the case of the DPR experiment, this would yield 172 × 21 = 3612 points. In an ideal situation, all the clouds would be very small in size (meaning that all referees would provide very similar grades for the given proposal), and so the points would

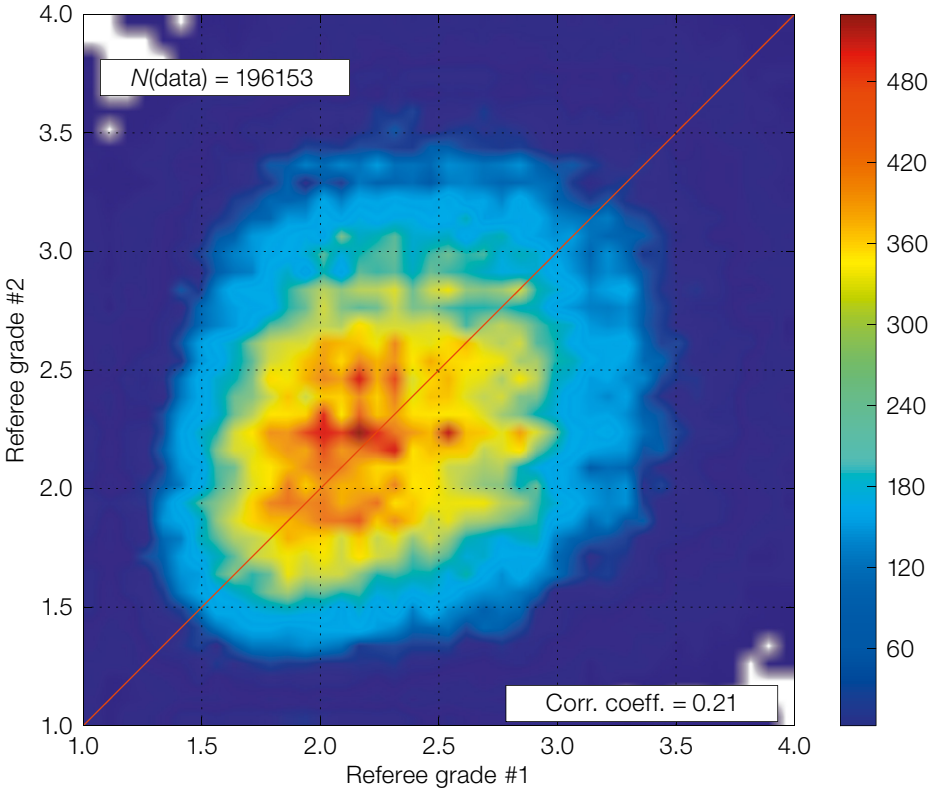

Figure 8 (below). Pre-meeting OPC referee–referee correlation. In this density diagram each point represents a pair of grades attributed to the same proposal by two distinct referees. The data are from the P18 sample.

be distributed very close to the straight-line $y = x$ on the r–r plane.

To illustrate what one is to expect in real life, we have constructed the r–r plane for the pre-meeting OPC P18 sample, from which we derived almost 200 000 grade pairs accumulated over 16 ESO cycles. The resulting diagram is presented in Figure 8. It is important to note that for a perfectly stochastic process, the points would be distributed within a circular area, with some radial, typically Gaussian, distribution. The fact that the real distribution is elongated along the diagonal direction signals that the process is not aleatory. This qualitative conclusion can be made more quantitative by computing the Pearson linear correlation coefficient, which ranges from –1 (complete anti-correlation) to 1 (complete correlation) and is null for complete uncorrelation. The value derived for the sample is 0.21. Given the very large number of points, this is a very robust estimate which can be reliably taken as a low correlation. For the same reason, however, this value reveals that there is a statistically significant signal indicating that the process is not completely aleatory. If on the one hand this may sound discouraging, it helps to put things in the correct context, as it characterises the subjectivity of the process in a more quantitative and objective way, as opposed to the common statements which are normally based on pure anecdotal evidence.

A different way of measuring the repeatability of the process, which we will use extensively in the next section, is the quartile agreement fraction (P18). The concept is as follows. When the same set of proposals is reviewed by two different bodies #1 and #2, one can compile the rankings for the two distinct reviews based on their distinct grades. The rankings are then used to derive a merit classification within the classical quartile scheme. For instance, the top 25% of proposals are ranked in the first quartile of the distribution of grades.

Once this is done, one can compute the fraction of applications ranked in the first quartile by review #1 which are also graded in the same quartile by review #2. For a complete agreement the fraction is 1, while a null value would signal a complete disagreement. The average agreement is expected to be 0.25 in case a fully stochastic process, i.e., when there is no correlation between the two bodies. The concept can be extended to all quartiles, including cross-quartile values, and the quartile agreement matrix (QAM) can be constructed. In statistical terms, the generic element M_{ij} of the QAM is the conditional probability that a proposal ranked in the i-th quartile by referee #1 is ranked in the j-th quartile by referee #2.

The application of this concept to the P18 pre-meeting sample shows that, on average, the ranking lists produced by two distinct referees have about 33% of the proposals in common in their first and last quartiles. In the central quartiles the intersection is compatible with a purely random selection (25%). This extends to the mixed cases ($i \neq j$), with the exception of the extreme quartiles; the fraction of proposals ranked in the first quartile by referee #1 and in the fourth quartile by referee #2 is ~ 17%, which deviates in a statistically significant way from the random value. As in the case of the r–r correlation introduced above, the r–r agreement fraction gives a quantitative estimate of the high level of subjectivity that characterises the process, providing a precise indication of what one should expect.

The reason why the applications are usually evaluated by more than one reviewer is to reduce the inherent "noise" which, as we have just seen, is quite substantial. For this purpose, the grades attributed by different referees to the same proposal (typically grouped in panels) are aggregated to form one single figure of merit. In the ESO implementation (and this is a common recipe), this is achieved simply taking the average, with no weights and/or rejection. The effect of increasing the number of reviews is diffusely discussed in P18; here it suffices to say that for $N_r = 3$ the first quartile agreement fraction grows to 43% and 30% in the first and second quartiles, respectively.

Armed with these terms of reference we can now discuss the results of the DPR experiment.

Comparing the OPC and DPR outcomes

The first test we apply to the DPR data concerns the subjectivity level characterising the typical participant. For this purpose, we have computed the average r–r QAM that we introduced in the previous section. Because of the DPR setup, the ranking list for each referee includes at most eight proposals, so each quartile contains no more than two proposals. Also, at variance with the classical panel scheme, the number of proposals in common between two reviewers is typically very small. As a direct comparison between ranks is not possible, we use a bootstrap approach. Very briefly, for each of the 172 proposals we randomly extract one grade pair and form two ranking lists, which are used to compute the quartile agreement fractions. The process is repeated a large number of times and the average values and standard deviations are derived for each of the QAM elements. The result is presented in Table 1. A direct comparison with the values derived from the P18 sample reveals that the two results are statistically indistinguishable. No meaningful difference is seen in the QAMs computed for the OE and DT sub-samples.

In a further test, we have investigated the possible dependence on the scientific seniority level introduced above. Of the 167 reviewers, 136 provided this information, which we used to sub-divide the reviewers into two classes: junior (groups 0 and 1) and senior (groups 2 and 3). These classes roughly correspond to PhD students plus junior postdocs (37), and advanced postdocs plus senior scientists (99), respectively. We then computed the r–r QAM for the two classes; the first quartile terms are 0.22 and 0.32, respectively. At face value this indicates a larger agreement between senior reviewers. However, the small size of the

Table 1. Bootstrapped r–r Quartile Agreement Matrix for the DPR experiment.

Referee #1 quartile	Referee #2 quartile 1	2	3	4
1	0.33	0.26	0.24	0.18
2	0.26	0.26	0.25	0.23
3	0.24	0.25	0.25	0.26
4	0.18	0.23	0.26	0.34

junior class produces a significant scatter, so the difference may not be significant.

One can extend the above bootstrapping procedure to subsets with a number of referees $N_r > 1$. The case of $N_r = 3$ is particularly interesting as this is directly comparable to the results presented in P18. The procedure is as follows: we first make a selection of the proposals having at least 6 reviews (164); for each of these we randomly select two distinct (i.e., non-intersecting) subsets of $N_r = 3$ grades each, from which two average grades are derived; the subsequent steps are identical to the r–r procedure, and lead to what we will call the p–p QAM.

The first-quartile agreement turns out to be 41%, while for the second and third quartiles this is 30%. The top-bottom quartile agreement is 10%. These values are very similar to those presented in P18 for the OPC process for $N_r = 3$ subpanels. As for the r–r case, the OE and DT sub-samples yield statistically indistinguishable values. The conclusion is that, in terms of self-consistency, the DPR review behaves in the same way as the pre-meeting OPC process.

We now come to what is perhaps one of the most interesting aspects. As anticipated, the proposals used in the DPR experiment were also subject to the regular OPC review. This enables the comparison between the outcomes of the two selections, with the caveats outlined above about their inherent differences.

For a first test we used a bootstrap procedure in which, for each proposal included in the DPR, we randomly extracted one evaluation from the DPR (typically one out of 7) and one from the OPC (one out of 3), forming two ranking lists from which a r–r QAM was computed. The operation was repeated a large number of times and the average and standard deviation matrices were constructed. This approach provides a direct indication of the DPR-OPC agreement at the r–r level and overcomes the problem that the two reviews have a different number of evaluations per proposal (see below). The result is presented in Table 2. The typical standard deviation of single realisations from the average is 0.06.

Table 2. Average DPR–OPC (pre-meeting) r–r Quartile Agreement Matrix.

DPR referee quartile	OPC referee quartile 1	2	3	4
1	0.31	0.26	0.24	0.18
2	0.24	0.27	0.25	0.24
3	0.24	0.23	0.26	0.26
4	0.20	0.23	0.25	0.31

This matrix is very similar to that derived within the DPR reviews (see Table 1), possibly indicating a DPR–OPC r–r agreement slightly lower than the corresponding DPR–DPR. A check performed on the two sub-samples for the junior and senior DPR reviewers (according to the classification described above) has given statistically indistinguishable results.

As explained in the introduction, the proposals were reviewed by $N_r = 3$ OPC referees in the pre-meeting phase. This constitutes a significant difference, in that the DPR ranking is typically based on ~ 7 grades, whereas the pre-meeting OPC ranking rests on 3 grades only. With this caveat in mind, one can nevertheless compute the QAM for the two overall ranking lists. The result is presented in Table 3. At face value, about 37% of the proposals ranked in the 1st quartile by the DPR were ranked in the same quartile by the OPC, with a similar fraction for the bottom quartile. When looking at these values, one needs to consider that this is only one realisation, which is affected by large scatter, as can be deduced from the comparatively large fluctuations in the QAM. These are evident when compared to, for instance, the average values obtained from the bootstrapping procedures described above. The numerical simulations show that the standard deviation of a single realisation is ~ 0.1.

Using the model presented in P18, one can predict that, on average, the top and bottom quartile agreement between the DPR and the pre-meeting OPC should be around 0.5 (see Kerzendorf, 2019 for more detail). The observed value (0.37) differs at the 1.3-σ level from the average value. For the central quartiles the difference is at the ~ 1.5-σ level. Therefore, although lower than expected on average, the observed DPR–OPC agreement is statistically consistent with that expected from the statistical description

Table 3. DPR–OPC (pre-meeting) p–p Quartile Agreement Matrix.

DPR quartile	OPC (pre-meeting) quartile 1	2	3	4
1	0.37	0.26	0.28	0.09
2	0.28	0.16	0.28	0.28
3	0.16	0.40	0.19	0.26
4	0.19	0.19	0.26	0.37

of the pre-meeting OPC process (P18). Note that, given the large noise inherent in the process, a much larger data set (or more realisations of the experiment) would be required to reach a sufficiently high statistical significance and to make robust claims about possible systematic deviations.

The fact that in the real OPC process there is a face-to-face meeting constitutes the most pronounced difference between the two review schemes. In the meeting, the opinions of single reviewers are changed during the discussion, so that grades assigned by individual referees are not completely independent from each other (as opposed to in the pre-meeting phase, in which any significant correlation should depend only on the intrinsic merits of the proposal). The effects of the meeting can be quantified in terms of the quartile agreement fractions between the pre- and post-meeting outcomes, as outlined in Patat (in preparation; hereafter called P19). Based on the P18 sample, P19 concludes that the change is significant; on average, only 75% of the proposals ranked in the top quartile before the meeting remain in the top quartile after the discussion (about 20% are demoted to the second quartile, and 5% to the third quartile). P19 characterises this effect by introducing the Quartile Migration Matrix (QMM). For the specific case of Period 103, the QMM is reported in Table 4 for the subset of the DPR experiment. Of the initial 172 proposals included in the DPR sample, 36 were triaged out in the OPC process and are therefore not considered.

As anticipated, the effect is very marked; the meeting does have a strong effect on the final outcome. In light of these facts, we can finally inspect the QAM between the DPR and the final outcome of the OPC process. This is presented in Table 5. With the only possible exception of $M_{4,4}$, which indicates a relatively

Table 4. OPC Quartile Migration Matrix for the DPR sub-sample (N = 136).

OPC pre-meeting quartile	OPC post-meeting quartile 1	2	3	4
1	0.56	0.32	0.12	0.00
2	0.32	0.32	0.29	0.06
3	0.12	0.26	0.38	0.24
4	0.00	0.09	0.21	0.71

Table 5. DPR–OPC (post-meeting) Quartile Agreement Fraction.

DPR quartile	OPC post-meeting quartile 1	2	3	4
1	0.26	0.38	0.24	0.12
2	0.24	0.35	0.24	0.18
3	0.32	0.12	0.29	0.26
4	0.19	0.15	0.24	0.44

marked agreement for the proposals in the bottom quartile, the two reviews appear to be almost completely uncorrelated. By means of simple Monte-Carlo calculations one can show that for two fully aleatory panels, the standard deviation of a single realisation around the average value (0.25) is 0.10. We conclude the majority of the $M_{i,j}$ elements in Table 5 are consistent with a stochastic process at the 1-σ level.

The main conclusion of this analysis is that, while the pre-meeting agreement is consistent, with the DPR and OPC reviewers behaving in a very similar way (in terms of r–r and p–p agreements), the face-to-face meeting has the effect of significantly increasing the discrepancy between the two processes. However, we caution that the sample is relatively small, and therefore the results are significantly affected by noise.

That the DPR–OPC agreement is smaller than the internal DPR–DPR agreement is not unexpected, as there are intrinsic differences between the two setups, the largest one being the absence of a face-to-face meeting, which is potentially the weakest aspect of the DPR. However, it remains unclear whether panel discussions lead to the selection of better science. In this respect, it is important to note that several studies have shown that panel meetings can increase the differences between two panels with respect to the pre-meeting agreement. In other words, while the meeting increases the internal consensus by polarising different opinions within the panels, it does not lead to a better panel-panel agreement (see Obrecht et al., 2007 and references therein). One would expect the discussions to bring judgment closer to identifying the best science; however, these studies indicate that a face-to-face meeting does not necessarily make the process better.

Conclusions and outlook

Gemini has already implemented a variant of this mechanism successfully over the past few years for their Fast Turnaround (Andersen et al., 2019). The approach presented here enhances this process, using better review-proposal matching based on natural language processing. The next logical step is to expand this experiment and distribute a fraction of observing time using DPR at more facilities. More than 95% of the participants suggest an implementation of such a scheme for some part of the ESO proposal types, with 75% support for the short programmes (time requests < 20 hours). Fewer than 5% of the responses were against implementing DPR for any of the programme types. In particular, about 70% of the responses are in favour of deploying DPR for the Fast Track Channel, while only about 15% are against it (the remaining 15% is indifferent). We take this as a clear indication of support.

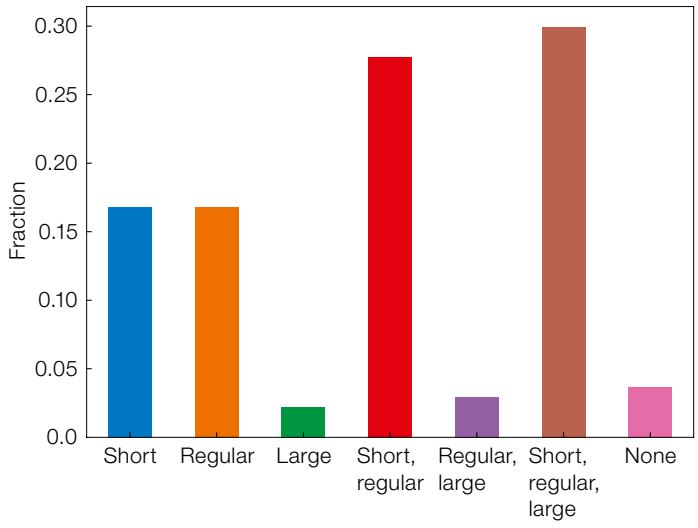

Figure 9. Distribution of the answers to the question: "For which types of proposals do you think distributed peer review would be beneficial?" in the DPR survey.

One of the objections that is typically made to the DPR concept is that, by distributing the proposals to a larger number of unselected scientists, it increases the chances of information leakage and plagiarism. In the specific case of the DPR experiment, the proposals were distributed to 172 reviewers, while in the OPC process the applications were seen by 78 individuals. However, while in the OPC implementation each reviewer has access to all proposals assigned within her/his panel (typically 70–80), the DPR reviewer sees a factor of ~ 10 fewer proposals. Therefore, under the reasonable hypothesis that the fraction of "malevolent" scientists is the same in both review bodies (which are selected from the same community), one would actually expect that the DPR is less prone to confidentiality issues on average. To get a direct opinion from DPR participants, the questionnaire contained an explicit question about this aspect. The distribution of the responses is shown in Figure 10. Excluding the "no strong opinion" cases, 66% of the users declared themselves to be equally or more confident in the DPR process, resulting in about a third of the users placing more trust in the classical scheme.

Another concern that is often heard when discussing DPR is the possible presence of biases. Again, the specific question put to the participants regarding this point does not support this concern; 74% of the respondents believe DPR is equally or more robust against biases (Figure 11).

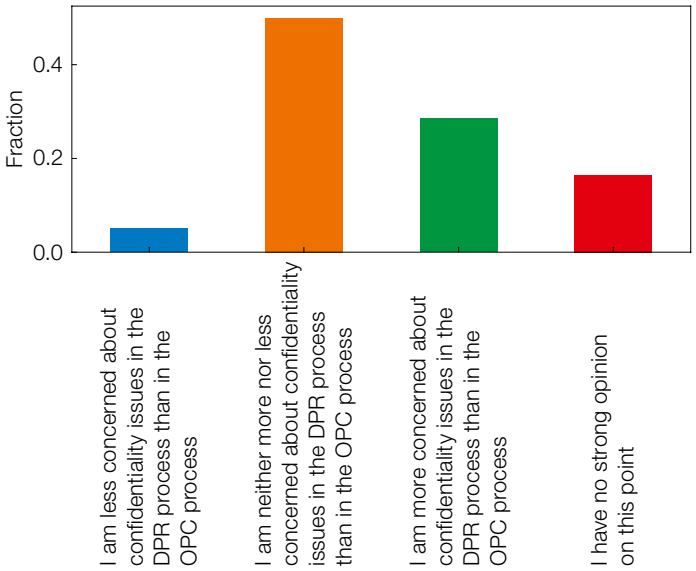

Figure 10. Distribution of answers to a question about how secure the participants felt about confidentiality issues.

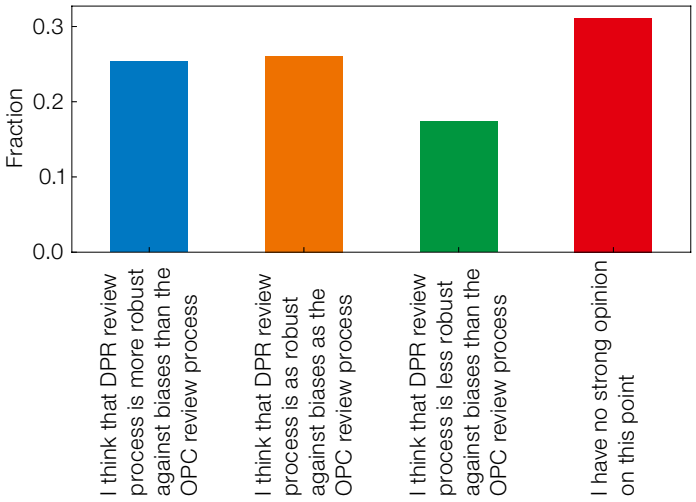

Figure 11. Distribution of answers to a question about the robustness of the process against biases.

The main conclusions drawn from the DPR experiment can be summarised as follows:

– The DeepThought-enhanced DPR experiment was very well received by the participants.
– The mechanism allows an optimal referee-proposal matching.
– The DPR process is as subjective as the OPC process.
– The participants do not see the confidentiality and bias issues as being more severe than in the classical scheme.
– ESO should consider deploying DPR for regular proposals below a certain time request, while leaving the classical review for larger time requests.

To these aspects, which come directly from the data, other positive facts can be added. DPR allows a much larger statistical basis enabling robust outlier rejection (the number of proposals per referee can be easily brought to 10–12) and it removes possible biases generated by panel member nominations. The larger pool of scientists allows much better coverage in terms of proposal expertise matching, and the smaller number of proposals per reviewer allows more careful work and more useful feedback.

Another aspect of the DeepThought approach to proposal-referee matching is that it can be semi-automated; it also gives an objective criterion to assign a particular expertise, eliminating biases in self-reporting. DPR implicitly removes the concept of panel, which adds rigidity to the process. For instance, it maximises the overlap in evaluations, which is a typical issue in pre-allocated panels. The lack of a face-to-face meeting prevents strong personal opinions from having a pivotal influence on the process. Also, DPR involves a larger part of the community, increasing its democratic breadth and exposing all applicants to the typical quality of the proposals. This allows them to better understand if their request is not allocated time by placing it in a wider context, which will help to improve their proposal-writing skills, training the members of the community without additional effort.

We acknowledge that the lack of a meeting does not allow the exchange of opinions and the possibility of asking and answering questions to/from the peers. Despite the fact that its effectiveness remains to be demonstrated and quantified (see above), it is clear that the social, educational and networking aspects of the face-to-face meeting should not be undervalued. In this respect, we note that the resources freed by the DPR approach can be used by the organisations for education and community networking (training on proposal writing, fostering collaborations, etc.).

In April and May 2019, results of the DPR experiment were presented to the ESO governing bodies most closely concerned with the Peer Review process (i.e., the Scientific Technical Committee, the Users Committee and the Observing Programmes Committee). The ensuing discussions have resulted in a wealth of useful feedback that is being discussed internally. We would like to conclude by pointing out that these kinds of studies are crucial if we are to progress from a situation in which the classical peer review process is adopted notwithstanding its limitations simply due to the lack of better alternatives. As scientists, we firmly believe in experiments, including those that address the selection of the experiments themselves.

Acknowledgements

The authors wish to express their gratitude to the 167 volunteers who participated in the DPR experiment, for their work and enthusiasm. The authors are also grateful to Markus Kissler-Patig for passionately promoting the DPR experiment following his experience at Gemini; to ESO's Director General Xavier Barcons and ESO's Director for Science Rob Ivison for their support; and to Hinrich Schütze for several suggestions on the NLP process.

Links

[1] Gemini Observatory Fast Turnaround Observing Mode webpage: http://www.gemini.edu/sciops/observing-gemini/proposal-routes-and-observing-modes/fast-turnaround

[2] Distributed Peer Review Pilot in Foundational Program: https://nifa.usda.gov/resource/distributed-peer-review-pilot-foundational-program

[3] Report from ESO Users Committee No. 42 (2018): https://www.eso.org/public/about-eso/committees/uc/uc-42nd/UCreport2018.pdf

References

Andersen, M. et al. 2019, AAS, 233, 455.03
Ardabili, P. N. & Liu, M. 2013, CoRR, arxiv:1307.6528
Brinks, E. et al. 2012, The Messenger, 150, 20
Gallo, S. A., Sullivan, J. H. & Glisson, S. R. 2016, PLoS ONE, 11, e0165147
Huang, C. 2013, European Journal of Psychology of Education, 28, 1
Kerzendorf, W. E. 2017, Journal of Astrophysics and Astronomy, arxiv:1705.05840
Kerzendorf, W. E. et al. 2019, submitted to Nature Astronomy
Merrifield, M. R. & Saari, D. G. 2009, Astronomy and Geophysics, 50, 4.16
Mervis, J. 2014a, Science, 344, 1328
Mervis, J. 2014b, Science, 345, 248
Obrecht, M., Tibelius, K. & D'Aloisio, G. 2007, Research Evaluation, 16 (2), 79
Patat, F. 2016, The Messenger, 165, 2
Patat, F. et al. 2017, The Messenger, 169, 5
Patat, F. 2018a, The Messenger, 173, 7
Patat, F. 2018b, PASP, 130, 084501
Strolger, L.-G. et al. 2017, AJ, 153, 181

ESO/G. Hüdepohl (atacamaphoto.com)

Snowfall at Paranal is a rare phenomenon that serves to utterly transform the surroundings of the VLT/I into an otherworldly landscape.