F1000Research

RESEARCH ARTICLE

REVISED Community review: a robust and scalable selection system for resource allocation within open science and innovation communities [version 2; peer review: 2 approved]

Chris L.B. Graham [iD] [1,2]*, Thomas E. Landrain[1]*, Amber Vjestica[3], Camille Masselot[4], Elliot Lawton[1], Leo Blondel[1], Luca Haenal[1], Bastian Greshake Tzovaras[4], Marc Santolini [iD] [1,4]*

[1]Just One Giant Lab, Paris, France
[2]University of Warwick, Coventry, UK
[3]University of Nottingham, Nottingham, UK
[4]Learning Planet Institute, Université de Paris, Paris, France

* Equal contributors

Abstract

Resource allocation is essential to the selection and implementation of innovative projects in science and technology. With large stakes involved in concentrating large fundings over a few promising projects, current "winner-take-all" models for grant applications are time-intensive endeavours that mobilise significant researcher time in writing extensive project proposals, and rely on the availability of a few time-saturated volunteer experts. Such processes usually carry over several months, resulting in high effective costs compared to expected benefits. Faced with the need for a rapid response to the COVID-19 pandemic in 2020, we devised an agile "community review" system, similar to distributed peer review (DPR) systems, to allocate micro-grants for the fast prototyping of innovative solutions. Here we describe and evaluate the implementation of this community review across 147 projects from the "Just One Giant Lab's OpenCOVID19 initiative" and "Helpful Engineering" open research communities. The community review process uses granular review forms and requires the participation of grant applicants in the review process. We show that this system is fast, with a median duration of 10 days, scalable, with a median of 4 reviewers per project independent of the total number of projects, and fair, with project rankings highly preserved after the synthetic removal of reviewers. We investigate potential bias introduced by involving applicants in the process, and find that review scores from both applicants and non-applicants have a similar correlation of r=0.28 with other reviews within a project, matching previous observations using traditional approaches. Finally, we find

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 18 Apr 2023	✓ view	✓ view
version 1 06 Dec 2022	? view	

1. **Ferdinando Patat** [iD], European Southern Observatory, München, Germany

2. **Dimitrios Kouis**, University of West Attica, Athens, Greece

Any reports and responses or comments on the article can be found at the end of the article.

that the ability of projects to apply to several rounds allows to both foster the further implementation of successful early prototypes, as well as provide a pathway to constructively improve an initially failing proposal in an agile manner. This study quantitatively highlights the benefits of a frugal community review system for agile resource allocation.

Keywords
community science, decision making, research funding, research and innovation

This article is included in the Research on
Research, Policy & Culture gateway.

This article is included in the Citizen Science
Association collection.

Corresponding authors: Chris L.B. Graham (christopherlbgraham@live.co.uk), Marc Santolini (marc@jogl.io)

Author roles: Graham CLB: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Landrain TE**: Conceptualization, Data Curation, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vjestica A**: Data Curation, Formal Analysis, Investigation; **Masselot C**: Methodology, Project Administration; **Lawton E**: Data Curation, Project Administration; **Blondel L**: Data Curation, Methodology, Software; **Haenal L**: Software; **Greshake Tzovaras B**: Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Santolini M**: Data Curation, Formal Analysis, Investigation, Methodology, Supervision, Visualization, Writing – Review & Editing

> **REVISED** **Amendments from Version 1**
>
> In our revision to the paper, as recommended by reviewers we have included a small summary on distributed peer review, and how this played a part in our community review process which are very similar. Along with a small summary on why it should not be overlooked.
>
> We hope that in light of this literature, a balanced approach to new peer review approaches can be taken by future programmes which incorporate the applicants as reviewers, much like we have and others have in distributed peer review, as well as including a community of non-applicants to take part in the review process, despite the difficulties in recruiting these experts. We also hope that our data, on the applicant vs non applicant bias in such programmes as reviewers is informative. The first reviewer for this article was well informed on distributed peer review and we thank him for improving our literature review in the introduction.
>
> **Any further responses from the reviewers can be found at the end of the article**

Introduction

The distribution of scientific funding through grants requires the identification of novel, feasible and potentially impactful projects. However, the traditional scientific grant allocation system involving a closed panel of experts in the field, or in similar fields,[1] is notoriously slow,[2] time consuming and expensive, often taking months and occurring in timescales of yearly rounds or grant calls. In extreme cases, the grant review program can be more costly than simply allocating small grants to each applicant, as in the case of the NSERC grant system of 2008.[3] In addition, the allocation of grants has shown to suffer from various biases, such as the composition of the grant panel,[4] gender and geographical location,[5] group based dynamics personality triumphing over other qualitative factors,[6–8] and socio-psychological factors such as group dynamics and personality traits triumphing over other qualitative factors.[8,9] Overall, selection results are only weakly predictive of future performance.[10]

Often, the reason to conduct grant allocations in a 'closed' setting is to protect the intellectual property of the grant applicants. As a result, the majority of unsuccessful grant applications, which contain a large amount of research effort, are inevitably lost, unavailable to the public after the fact.[11] The recent emergence of the open science movement[12–14] has reversed this incentive, with open access practices and early sharing of results such as pre-registration now becoming normalised by institutions and journals.[15]

Beyond the allocation of funding, the review of early-stage, unpublished work by community peers has been leveraged to allocate other types of resources. For example, conferences often need to allocate time for their participants to showcase their work to other members of the community during a usually short amount of time, thereby providing a platform for promoting the work, building novel collaborations, and getting feedback to improve a manuscript. In such cases, peer reviewing is needed to decide in a collegial fashion whether a work is worth a full oral presentation, a shorter lightning talk, a poster, or is not of a high enough standard to be showcased to participants. For example, the EasyChair online platform has been used by close to 100k conferences for handling such review processes.[16] Often, participants to a conference are also part of the "program committee" reviewing the proposed abstracts and papers of peer applicants, alongside external members of the scientific community. This allows for a rapid process usually lasting less than a few weeks. Digital review score and review sorting algorithms have also been implemented for a more data-driven approach to peer review.[17]

These examples suggest there is a potential for a more agile route for community-driven grant allocation that would bypass pre-selected grant panels that handle funds and introduce barriers,[6] and rely instead on peer applicants to handle a large-scale application process in a short timescale. A similar concept, the distributed peer review (DPR) system, was previously introduced within some academic fields. In DPR, applicants act as reviewers in a distributed fashion. The first paper based on this idea was published in 2009 in the field of Astronomy.[18] The concept then became popular for astronomical facilities, and it was further used to select conference papers in Data Science.[19] Recently, even the US National Science Foundation has toyed around with the idea of leveraging applicants as reviewers, or a community as reviewers.[20] NSF's use of the protocol was found to lead to a disadvantage for groups applying to funds with controversial projects, and to increase the potential for collusion between teams, as applicants could see the projects from other applying teams.[21] Yet, it is unclear how these findings generalize to collaborative contexts, where inter-team interactions are incentivized, and where project resubmission is made possible in an agile and regular manner, allowing for constructive feedback to be immediately integrated.

In this study, we present the design, implementation, and results of a community-driven, open peer-review system similar to DPR, to support two open research communities during the coronavirus disease 2019 (COVID-19) pandemic across

seven selection rounds (Figure 1): the "OpenCOVID19" initiative from Just One Giant Lab (JOGL)[14,22] and the COVID relief charity Helpful Engineering.[23] Part way through our community review process we also used distributed peer review to form a hybrid process of applicants and non applicants reviewing projects, here termed as 'Community review'. We show that this system is robust (unaffected by reviewer removal), agile (fast timeline), iterative (covering multiple grant rounds), decentralised (driven by the community), and scalable. Finally, we discuss these results and the perspectives they offer for the design of future community-driven review systems that may differ from the current model of grant panel review.
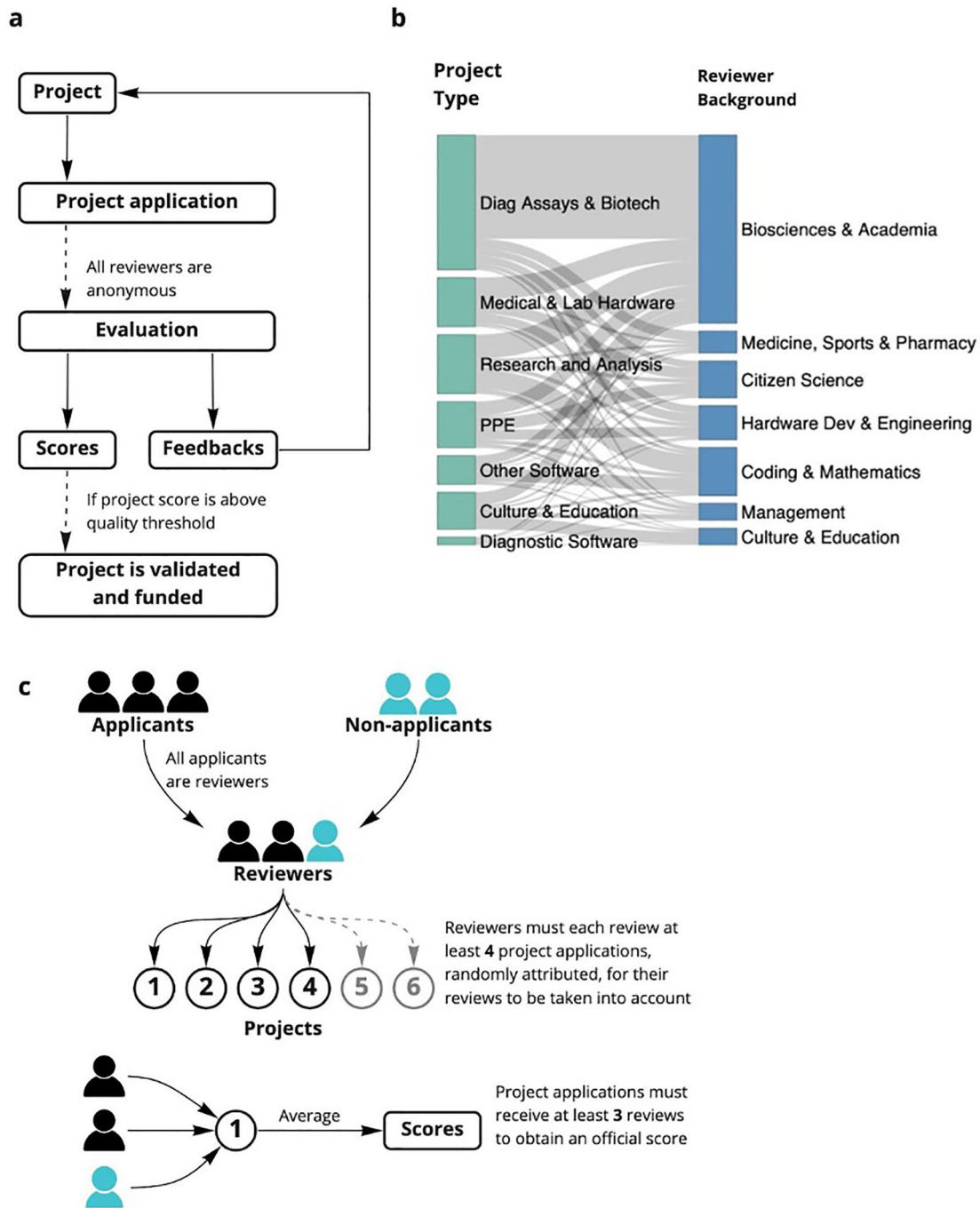

Figure 1. Overview of the open peer review process. (a) Stages of the open peer review process JOGL rounds 3-5. The online review forms and templates are found in supplementary data. (b) community review method JOGL rounds 3-5 (c) distribution of project type to expertise across rounds.

Methods

Context

The implementation of a crowd-based, open peer-review system followed the need to support two nascent community efforts, first by allocating volunteers to projects in the COVID relief charity Helpful Engineering,[23] then by allocating funding to projects in the JOGL "OpenCOVID19" initiative.[22] The method was developed as an open access grant review and funding allocation system, meaning that it was open to anyone willing to review. It was implemented using the Just One Giant Lab platform (app.jogl.io) as the project proposal host, and free-to-use tools and forms to conduct the review process (Extended Data:FigS2[24]). The implementation was applied and refined over 7 rounds across 1 year.

General process of review

The peer review system was conducted on early phase projects within both JOGL and Helpful Engineering. These projects were submitted by project leaders to a grant review process in order to allocate volunteers in the case of Helpful Engineering, and funding in the context of OpenCovid19. Reviews of these projects (see Figure 1b) were initially conducted by members of the community and included members of other projects who also submitted their project for review.

As a consequence of the process being experimental and serving an urgent need, the process was altered over time. However, it followed the same general pattern (Figure 1, Extended Data:FigS1[24]). First, a template for the grant proposal was created by the community and was iteratively edited (Extended Data[24]) template followed typical grant application templates,[25] with sections on team composition, the project general hypothesis and its timeline. The proposal was then submitted using a google form, which requested an email address and required only one application per project (Extended Data:FigS2a[24]). In Helpful Engineering rounds this included a link to their proposal hosted in editable google documents, while in JOGL rounds this included instead a link to their open access JOGL page proposal. The project links were manually formatted into a google sheet with a link to a review form for convenience, along with descriptions of desirable reviewer skills by the applicants in the proposal submission form to help reviewers find relevant projects (Extended Data: FigS2B[24]). A technical evaluation form scoring various criteria (e.g.: proposal efficacy, team composition, impact) on a scale from 1-5 (Extended Data[24]) was created by the designers of the program and iteratively changed following feedback from the community (Extended Data:FigS2c[24]). This form separated questions on projects into two areas centred around Impact and Feasibility for ease of identifying the problems and/or strengths in their grant application.

As part of the community aspect of the review, a message with a link to the reviewer form for use in review, along with a nested google sheet containing project proposal links was spread among the community through announcements and email. In later rounds (JOGL 3-5) all applicants were asked to review at least three other projects and the process was randomised, removing the need for a sheet and reducing the possibility of collusion. The review process was given between 4 days HE 1, 8 days HE 2, 7 Days - JOGL 1, 10 days - JOGL 2, 16 days - JOGL 3, 21 days - JOGL 4 and 28 days JOGL 5, (Extended Data:FigS1b[24]) to allow reviews to occur and be collected via a google form into a google sheet automatically (Extended Data:FigS2d[24]). No reviewer selection was performed, however usernames (Slack handles or JOGL user names depending on the round) and emails were collected for conducting further analyses. The average reviewer scores were then composed into a presentation to the community, and those projects with a score above a given impact/feasibility threshold (Extended Data:FigS2e[24]) were chosen for grant funding. Due to the community aspect of our study, members from the JOGL HQ participated in the process, and their removal from the analysis does not change the observations (Extended Data:FigS10[24]), we therefore retain these in our analysis.

Iterative changes to the review process

As mentioned in the previous section, the method of review was iteratively changed throughout the programme, elongating from an initial "emergency style" four day period of review and allocation (HE round 1) to 21 and 28 days in JOGL rounds 4 and 5 as the need for rapid response reduced, with an overall median average of 10 days per round (Extended Data:FigS1b[24]). As such, the design of the general process described in Figure 1 and Extended Data:FigS1[24] had some variations. For example, initially applicants were not required to review applications (Figure 1b). Upon scaling up of the programme, the process was adapted to be less dependent on volunteer reviewers, (Extended Data:Fig S1b,A-D[24]) and more dependent on the applicant's reviews of their competing peers (Figure 1c). In JOGL rounds 3, 4 and 5 (Extended Data:FigS1b[24]) teams depositing a proposal could only be eligible after having reviewed at least three other teams. The changes in the process and differences in the rounds are summarised in Extended Data:FigS1c.[24] The major changes between Helpful Engineering (HE) and JOGL rounds (Extended Data:FigS1c[24]) occurred through changes in the nature of proposal submission from google document links to an online project repository. In addition, HE rounds offered no grants, but instead publicity and allocation of members to projects, while JOGL offered microgrants worth up to 4000 euros per team (Extended Data:FigS2c[24]).

Final selection process

In Helpful Engineering, this review method allowed 54 projects to be reviewed and ranked by score for community recruitment purposes, with no official threshold, but instead an arbitrary set of "Highlighted projects". Within JOGL this grant system reviewed 96 eligible applications (Figure 2) and allocated requested funds to 36 of these. Once the review process had taken place, the cut-off threshold of scores given by reviewers to projects for funding by JOGL was decided by an absolute threshold (above 3.5/5 average reviewed score) rather than a rejection rate. The absolute 3.5/5 threshold was chosen due to the gap in project scores in the first JOGL round, and maintained at this standard for consistency. Those with a score above the threshold were funded.

Detection of fraudulent reviewer behaviour

The results of each round, and number of reviews per reviewer were closely monitored through simple email handle tracking by a data handling administrator. If a number of emails were found to be grading a particular project and not others this was suggestive of fraudulent behaviour and self-grading. These reviews were then removed, and teams that were found responsible for this bad behaviour were removed from the review process and their project removed from the review pool, as described in grant round participation rules. This was performed only one time across all rounds prior to the rule of each reviewer having a minimum review count for their scores to be counted, which was created after this event.

Computation of inter-review correlations

In order to compute the correlation between reviews within a project, we first proceeded with data cleaning. Indeed, in several rounds, reviewers had to answer only a subset of questions from the review form that corresponded to the topic of the project (e.g. data project vs bio project). However, in some cases, projects were assigned to one or the other category by the different reviewers, leading them to answer to different sets of questions, making the correlation only partial. To mitigate this effect, for each project we kept only the reviews that corresponded to the choice of topic that was most expressed among reviewers. If no majority could be found, the project was removed from analysis. We then converted review scores into vectors of length the number of grades in the form. A Spearman's rho correlation was then computed between all pairs of reviews within a project. Finally, for each review we computed the average correlation with the other reviews in the project. This number was then associated with the features of the reviewer who produced the review (Figure 4 and Extended Data:FigS7[24]).
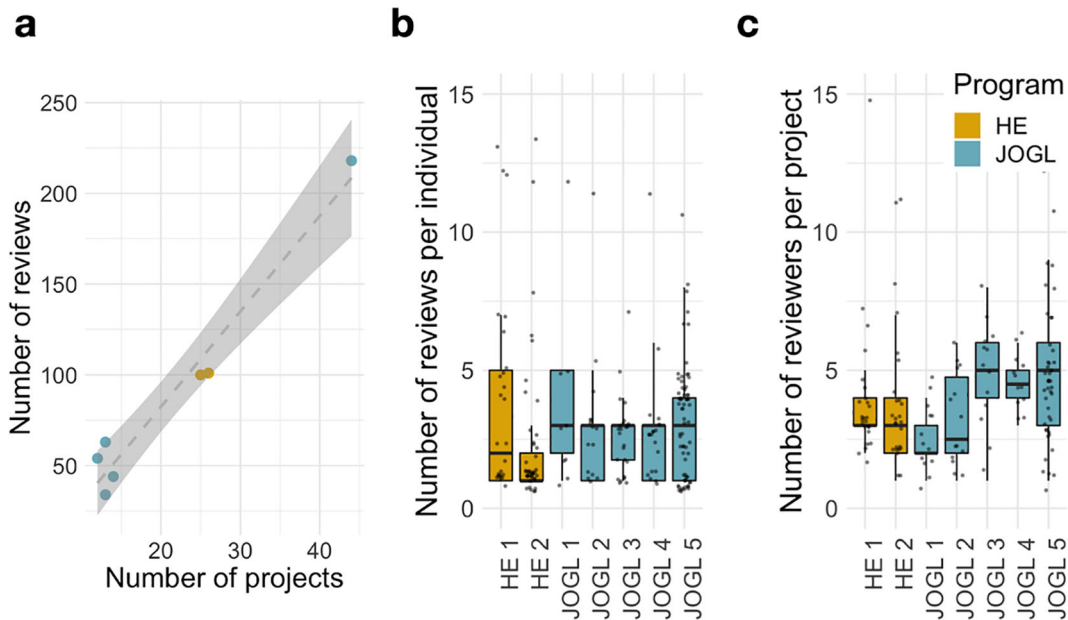

Figure 2. Scalability of the community review methodology. (a) Number of Reviewers and projects during each round of peer/grant review. HE- Helpful Engineering Crowd reviews, JOGL- Just One Giant Lab funded projects. (b) Number of reviews per individual reviewer. (c) Number of reviewers per project. Despite a scale-up in the number of projects, the number of reviews per round scales linearly with the number of projects applying.

Reviewer feasibility and impact scores

For JOGL rounds 1-5, we categorised the 23 to 29 questions from the review forms into either impact or feasibility related questions (see Underlying Data Review forms). The feasibility and impact categories were used to provide two dimensional projections of project scores during the result presentation.

Reviewer professions and project types

For all JOGL rounds, reviewer responses of the "What is your expertise relevant to this project" question were manually coded into simple categories per review (see Table S1 in the Extended Data[24]). This data was then used as a proxy for expertise distribution across rounds (Figure 1b).

In addition, reviewer responses to the "Which category would you say the project falls under?" question were manually coded into a set of simple categories, representing a summary of the project types across rounds per review (see Extended Data conversion table[26]). The data, due to suggested categories provided by the form, needed little manual coding, but was formatted into a list, then concatenated into similar project types for simplicity. This data was used to assess project type distribution across rounds (Figure 1b).

Bootstrap analysis

In order to perform the bootstrap analysis of Figure 3d, we first ranked all projects using their average review score across reviewers. We then selected a review at random. If the corresponding project had at least another review, we removed the selected review and recomputed the average scores and final ranking. We then computed the Spearman correlation between the obtained scores and the original scores. This process was repeated until each project had only one review. Finally, we reiterated this analysis 50 times. The analysis code can be found as Extended Data.[24]

Ethics/Consent

We confirm all ethical guidelines have been followed, using the same ethical procedures described in Commission Nationale de l'Informatique et des Libertes registration number "2221728" for user research and "2227764" for grant administration. Consent was granted by a user agreement on JOGL's website upon signup (https://app.jogl.io/data), and on the google forms used during the study.

Results

Scalability of the review process

We describe in Figure 2 the reviewing activity across the seven rounds implemented. Despite the large differences in number of projects between rounds, we find that the number of reviews per round scales linearly with the number of projects applying (Figure 2a) therefore indicating the scalability promoted through the use of DPR or community. In addition, the number of reviews per individual and number of reviewers per project have relatively stable distributions across rounds, independent of scale (Figure 2b-a). For example, despite the substantial growth in reviewers and projects in JOGL round 5, we find that the distributions of number of reviews per reviewer and number of reviewers per project are comparable to those observed in the previous rounds, highlighting the scalability of this review system to different systems. Finally, we note that the number of reviewers per project show a sustained increase from JOGL round 3 onwards, corresponding to the change in review process from a crowd sourced review to a DPR style applicant driven review, where applicants were required to review at least 3 other projects (see Methods). This highlights the benefits of this requirement in promoting sustained engagement.

Robustness of the final project ranking

In order to obtain a granular score for each project, the reviewers had to grade between 23 (JOGL 1-2) and 29 (JOGL 3-5) criteria in the review form.[26] We first investigate whether these questions would cover different dimensions of project quality. We show in Figure 3a a heatmap of reviewer scores in JOGL round 4 across 20 questions (removing questions only representing a minority of projects), visually showing a greater inter-review variability (rows) than inter-questions variability (columns). As such, respondents seem to assign a project with either low scores or high scores throughout their review. To quantify the number of dimensions of variation across grades, we conduct a Principal Component Analysis (PCA) on the questions correlation matrix, i.e correlations between pairs of questions across reviews (see Extended Data: Fig S2a[24]). We find that the first principal component (PC1) explains most of the variance (53%), with the next largest PC explaining less than 6% of the variance (Extended Data:Fig S3[24]). When examining the weights of the various questions in PC1, we find that they all contribute to a similar level (Figure 3b), meaning that the PC1 is close to the average over all questions, confirming the visual insight from Figure 3a. This shows that scores are highly correlated, and that the average score across the review form is a reasonable operationalisation of project quality. In addition, we find that the top 10 PCs explain ~90% of the variance, indicating that review forms could be reduced in complexity using only half of the number of questions to obtain a similar outcome.
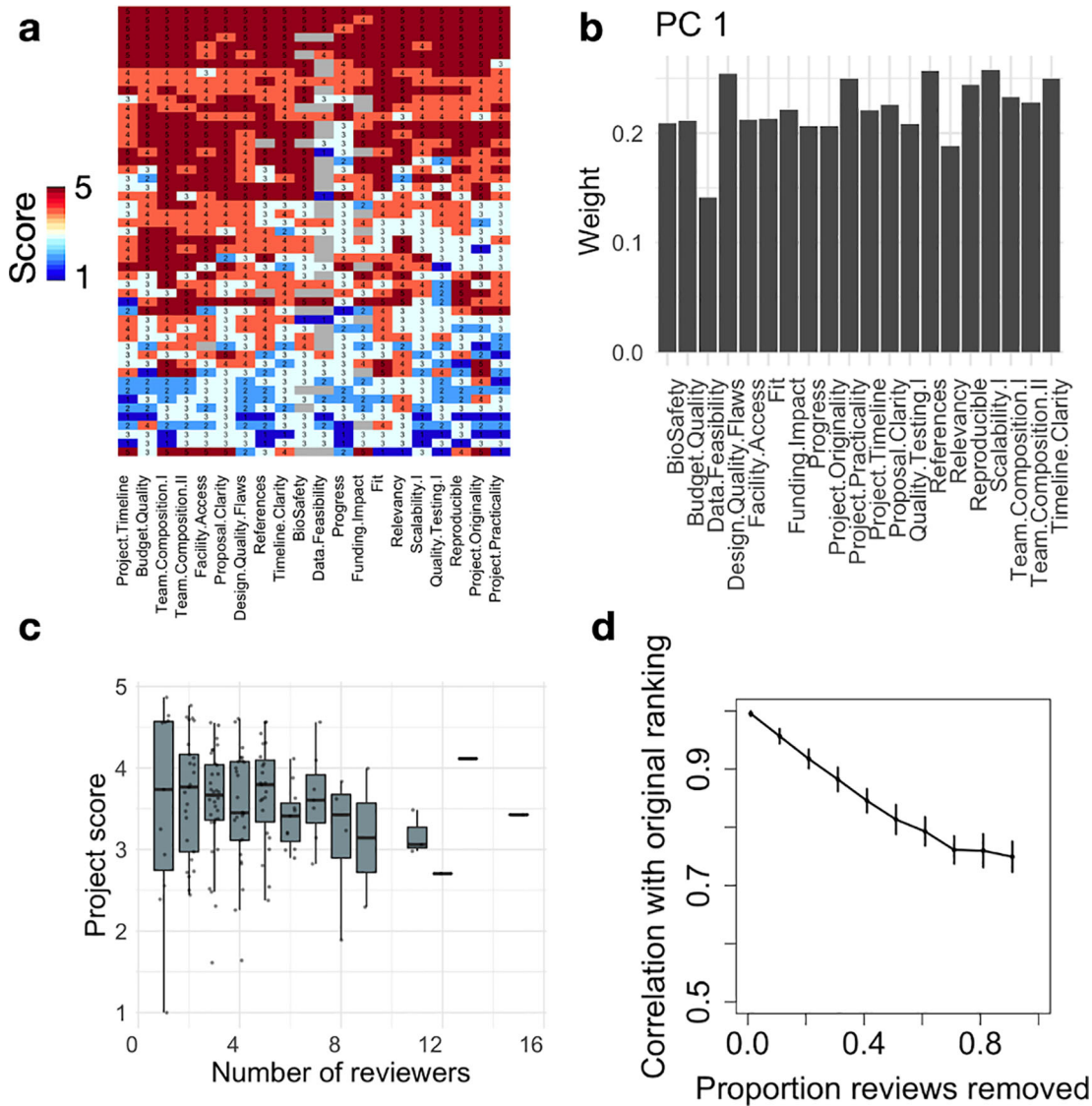
Figure 3. Robustness of the Community review process. (a) Heatmap showing review scores (rows) across questions (columns) for the JOGL round 4. Row and column clustering was performed using correlation distance and average linkage. (b) We show for PC1 (53% variance) the weights of the questions from the original question space. PC1 has near uniform weights across dimensions, indicating that it corresponds to an average score. (c) Project average score across reviewers as a function of number of reviewers. (d) Bootstrap analysis showing the Spearman correlation between the final project ranking and simulated project rankings with increasing proportion of reviews removed from the analysis (see Methods).

We next investigate the reliability of the review scores obtained across reviewers. As suggested by the previous section, for each review we compute the average score across all criteria from the review form. In the following, we refer to this average score as the review score. We observe a generally good discrimination of review scores between projects, with intra-project variation smaller than inter-project variation (Extended Data:FigS4[24]).

Finally, we investigate the robustness of the final project ranking as a function of the number of reviews performed using a bootstrap analysis (see Methods). For each project, a project score is computed by averaging its review scores, and projects are then ranked by decreasing score. We show in Figure 3d the Spearman correlation between the original project ranking and the ranking obtained when removing a certain proportion of reviews. We find that even with only one review per project, the final ranking is strongly conserved (rho=0.75 and see Extended Data:FigS5[24]), confirming that intra-project variability is much smaller than the range of inter-project variability. This supports our design strategy, showing that the use of a granular form allows us to differentiate between projects whilst minimising the impact of individual reviewers' variability.

Measuring reviewer biases

The previous results show the existence of variability between reviews from different reviewers, yet with limited impact on final rankings (Figure 3d). Here we investigate the source of review variability: is it due to inherent grading variability between individuals, or can it be attributed to other factors? To evaluate this question, we analyse how review score varies with reviewer attributes. We explore in particular two possible sources of bias for which we could gather data: expertise and application status. First, reviewer expertise might be important in determining an accurate project score. This feature is operationalised using the self-reported expertise grade (1 to 5) present in the review forms of JOGL rounds. Second, a majority of reviewers (65%) were applicants of other competing projects, which could lead to a negative bias when reviewing other competing projects.
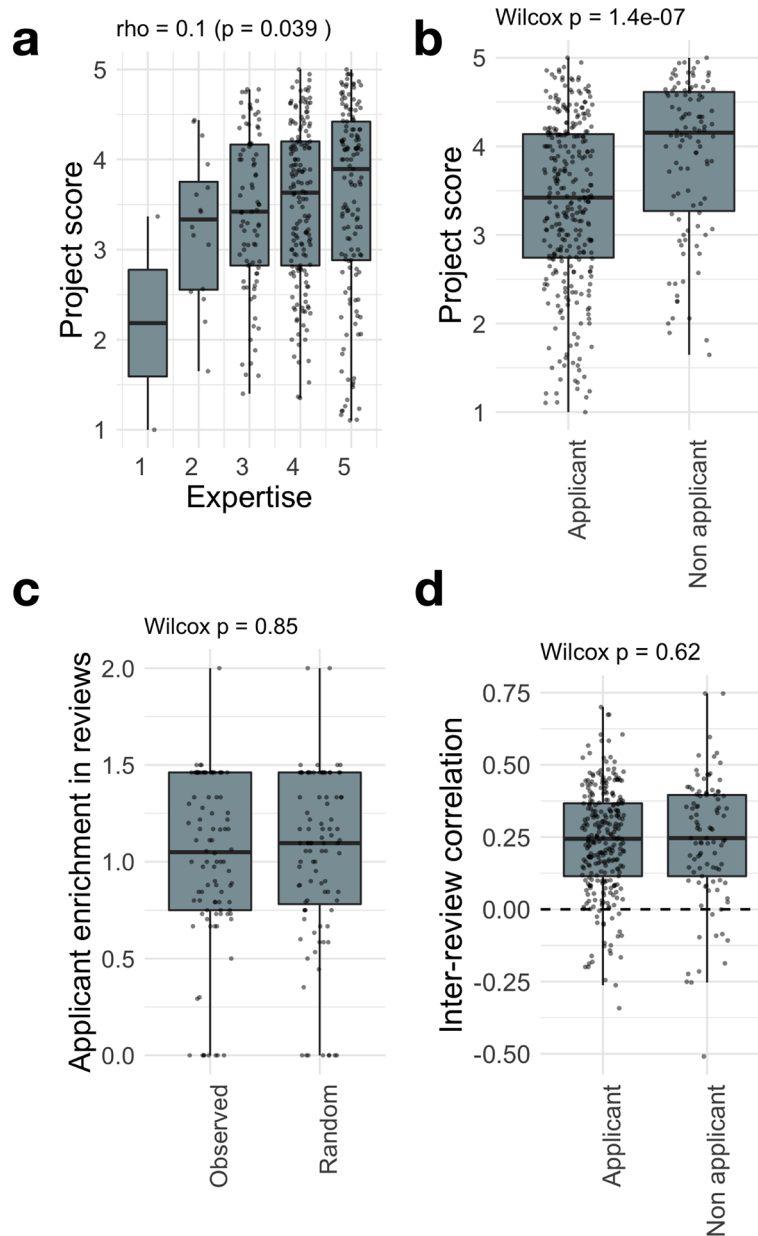

Figure 4. Questionnaire granularity allows to measure and mitigate reviewer biases. Breakdown of project score as a function of (a) self-assessed expertise, (b) applicant status (i.e. the reviewer is also an applicant in the round). See Fig S4 for a breakdown by review round. (c) The inter-review correlation or agreement between reviewers for a project is compared among applicants and non-applicants. (d) For each project, we compute the ratio between the proportion of applicant reviewers to the average proportion of applicant reviewers observed in the round. The boxplot compares the computed enrichments to the ones obtained for randomly assigned reviewers to projects, showing that applicants are evenly distributed across projects.
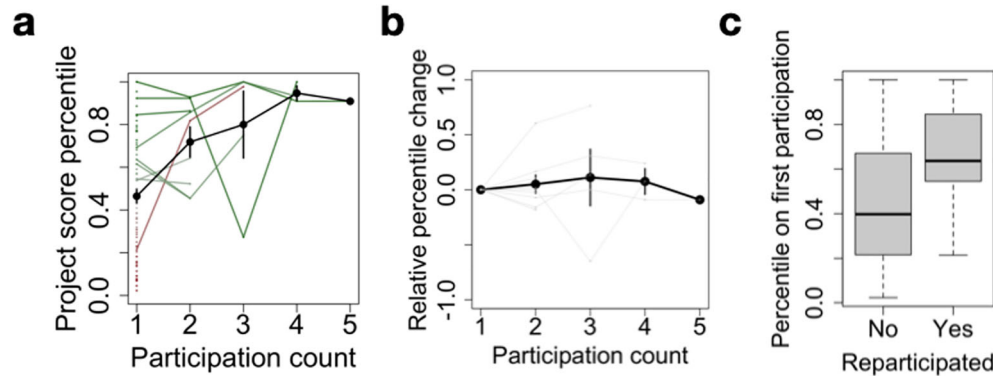
Figure 5. Multiple participations foster long-term project sustainability. (a) Project score percentile as a function of participation count. For each project, a score percentile is computed to quantify their relative rank within a specific application round, allowing to compare multiple projects across rounds. Participation count refers to the successive number of rounds a project has applied to. The black line denotes the average across projects, error bars represent standard error. Dots correspond to projects with only one participation, and lines to re-participating projects. Finally, the color gradient indicates relative score at first participation, from red (low) to green (high). (b) Same as a., after subtracting the percentile at first participation. (c) Score percentile at first participation as a function of whether or not a project has re-participated.

We show in Figure 4 how the review score varies as a function of these reviewer characteristics. We find that review score increases slightly with expertise (Figure 4a, Spearman's rho=0.1, p=0.039). However, the strongest effect is found when looking at applicant bias: review scores from applicants are significantly lower than those from non-applicants (Figure 4b, p=1.4e-7). Given the fact that in JOGL rounds 3-5, applicants were required to score at least 3 projects, they are found to have a lower expertise towards other projects (Extended Data:Fig S6[24]), which could explain the lower scores as suggested by Figure 4a. Yet, when controlling for review expertise, we find that application status is the main contributing factor, with a score difference between applicants and non-applicants of -0.52 points (p=1.61e-6, Extended Data: Supplementary Table 1[24]). This supports that application status is a significant source of bias in the final score.

Such differences could be due to unfair grading, with reviewers from a certain category (applicants or non-applicants) grading more "randomly" than others. To analyse this effect, we need to look beyond average score into correlations. Indeed, two similar average scores could stem from highly different fine-grain score vectors. Imagine two reviewers grading 3 questions from 1 to 5. The first reviewer gives the grades 1, 2, and 5, while the second gives 5, 1, and 2. These reviews produce the same average score (2.67). However, their fine-grain structure is anti-correlated, with a Pearson correlation r = -0.69. In our context, we find that review scores are positively correlated, with a median Pearson correlation between their reviews of r = 0.28 across rounds (Figure 4d), in line with previous observations in traditional funding schemes [35]. More importantly, we find no difference between applicants and non-applicants in their correlation with other project reviews (Figure 4c). This indicates that the variability between grades within a review form are conserved across reviewer characteristics (see Fig S7 and Extended Data:Fig S9 for the other characteristics[24]). As such, if applicants are uniformly distributed across projects, one will not expect a difference in the final rankings. Whilst this is true for our own analysis, the data for our review is also available for independent analysis.

A framework for iterative project implementations
In the JOGL implementation of the community review system, projects can apply to any number of rounds, irrespective of whether or not they have already successfully obtained funding in a previous round. We found 9 projects that applied to multiple rounds. On average, the relative performance of the projects in a grant round increases as a function of the number of participations (Figure 5a). We find that this effect is explained by re-participation being associated with early success, with initially lower performing projects eventually dropping out (Figure 5b-c). As such, the multiple round scheme supports projects with a high initial potential in the long-term through repeated micro-funding allocations. We also note that in the case of 2 projects, re-participation after an initial failure allowed them to pass the acceptance threshold. This highlights how constructive feedback allows for a rapid improvement of a project and its successful re-application in the process.

Discussion
In this manuscript we describe the "community review" method for the identification of novel, feasible and potentially impactful projects within two communities of open innovation: Helpful Engineering and OpenCovid19. This process was leveraged for the attribution of volunteers as well as micro-grants to projects over a year, in an agile and iterative framework.

Key to the system is the requirement of applicants to take part in the reviewing process, ensuring its scalability. As such, the number of reviews is proportional to the number of projects applying (Figure 2), with a fast median process duration of 10 days. This requirement comes at a risk, since applicants might be negatively biased towards other projects they are competing against. Accordingly, we found that applicants consistently give a lower score to projects when compared to non-applicants (-0.52 points). This bias cannot be explained solely by the lower expertise of applicants towards the randomly assigned projects. Indeed, we found that self-reported expertise has only a limited impact on the final score (Figure 4c). The effect is most stringent for rare cases of self-reported expertise of 1 and 2 out of 5, suggesting that a threshold of 3 might be implemented to remove non-expert bias. It is on the other hand possible that non-applicants are positively biased towards projects from which they might have personally been invited to review. We however noted no such report in the conflict of interest question in the review form.

Despite these biases, we found that applicants and non-applicants have a similar behaviour when grading questions in the form, with a stable Pearson correlation between their reviews of r = 0.28 (Figure 4/Extended Data:Fig S8[24]). This is slightly higher than the correlation of 0.2 observed in an analysis of the ESRC's existing peer review metrics,[27] suggesting comparable outcomes when compared to existing institutional methods. The similarity of their correlation profiles means that such biases contribute a similar "noise" to the system: they might change the overall average scores, but not their ranking as long as applicants are well distributed across projects. Accordingly, we found that the community review system is robust to the removal of reviewers, with an average ranking Spearman correlation of 0.7 in the extreme case of one reviewer per project.

Finally, we showed that some projects apply multiple times to the application rounds. While the number of such projects of this type is small (9 projects), we find that it had two benefits. First, we found two projects that re-applied after an unsuccessful application, allowing them to pass the acceptance threshold on the second application. This showcases the ability of the feedback system to benefit projects in constructively improving their application. Furthermore, we found that the number of applications of a project is strongly dependent on its performance on the first application. This means that the iterative process allows to select highly promising projects and sustain their implementation in the mid- to long-term. This is of particular importance when considering traditional hackathon systems, where promising projects are usually not supported over longer periods of time.

The speed and cost-efficiency of the community review process, through a distributed peer review like system, has allowed for a reactive response to the high-pressure environment created by the pandemic. This agility has meant that within the short time frame given, projects have been able to produce literature, methods and hardware and put them to use.[28–33] Overall, the community review system allows for a rapid, agile, iterative, distributed and scalable review process for volunteer action and micro-grant attribution. It is particularly suited for open research and innovation communities collaborating in a decentralized manner and looking for ways to distribute common resources fairly and swiftly. Finally, community review offers a robust alternative to institutional frameworks for building trust within a network and paves the way for the installation of community-driven decentralized laboratories.

Data availability
Underlying data
Open Science Framework: DATA FOR: Community review: a robust and scalable selection system for resource allocation within open science and innovation communities. https://doi.org/10.17605/OSF.IO/CAZ4N.[26]

This project contains the following underlying data:

- Review Data (the raw responses of the review rounds analysed by the paper, and the raw data used in the study.)

- Project round progress.csv (aggregated data and is based on data post analysis for our final figure, and the scores of each project over time, however we have aggregated this for ease of viewing.)

- Grant Review forms (the forms used to assess each proposal)

- Peer Review protocol (the protocol used to analyse the raw data, giving the correlation values we refer to in the paper)

- Coded expertise (the simplified version of project and reviewer type collected during review)

Extended data

Open Science Framework: EXTENDED DATA FOR: Community review: a robust and scalable selection system for resource allocation within open science and innovation communities, https://doi.org/10.17605/OSF.IO/W5Q9B.[24]

This project contains the following extended data:

- Supplementary figures 1-10

- Supplementary table 1

- Analysis code

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgements

References

1. Langfeldt L: **The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome.** *Soc. Stud. Sci.* 2001 Dec 1; **31**(6): 820–841.
Publisher Full Text

2. Herbert DL, Barnett AG, Graves N: **Australia's grant system wastes time.** *Nature.* 2013 Mar; **495**(7441): 314–314.
PubMed Abstract | **Publisher Full Text**

3. Gordon R, Poulin BJ: **Cost of the NSERC Science Grant Peer Review System Exceeds the Cost of Giving Every Qualified Researcher a Baseline Grant.** *Account. Res.* 2009 Feb 27; **16**(1): 13–40.
PubMed Abstract | **Publisher Full Text**

4. Roumbanis L: **Peer Review or Lottery? A Critical Analysis of Two Different Forms of Decision-making Mechanisms for Allocation of Research Grants.** *Sci. Technol. Hum. Values.* 2019 Nov 1; **44**(6): 994–1019.
Publisher Full Text

5. Severin A, Martins J, Delavy F, *et al.*: **Gender and other potential biases in peer review: Analysis of 38,250 external peer review reports.** *PeerJ. Inc.* 2019 Jun [cited 2021 Jul 9]. Report No.: e27587v3.
Reference Source

6. Coveney J, Herbert DL, Hill K, *et al.*: **'Are you siding with a personality or the grant proposal?': observations on how peer review panels function.** *Res. Integr. Peer. Rev.* 2017 Dec 4; **2**(1): 19.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Pier EL, Raclaw J, Kaatz A, *et al.*: **'Your comments are meaner than your score': score calibration talk influences intra- and inter-panel variability during scientific grant peer review.** *Res. Eval.* 2017 Jan; **26**(1): 1–14.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Fogelholm M, Leppinen S, Auvinen A, *et al.*: **Panel discussion does not improve reliability of peer review for medical research grant proposals.** *J. Clin. Epidemiol.* 2012 Jan; **65**(1): 47–52.
Publisher Full Text

9. Coveney J, Herbert DL, Hill K, *et al.*: **'Are you siding with a personality or the grant proposal?': observations on how peer review panels function.** *Res. Integr. Peer. Rev.* 2017; **2**: 19.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

10. Fang FC, Casadevall A: **NIH peer review reform--change we need, or lipstick on a pig?** *Infect. Immun.* 2009 Jan 21;2009 Mar; **77**(3): 929–932.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. von Hippel T, von Hippel C: **To apply or not to apply: a survey analysis of grant writing costs and benefits.** *PLoS One.* 2015; **10**(3): e0118494.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Landrain T, Meyer M, Perez AM, *et al.*: **Do-it-yourself biology: challenges and promises for an open science and technology movement.** *Syst. Synth. Biol.* 2013 Sep; **7**(3): 115–126.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. Masselot C, Tzovaras BG, Graham CLB, *et al.*: **Implementing the Co-Immune Open Innovation Program to Address Vaccination Hesitancy and Access to Vaccines: Retrospective Study.** *J. Particip. Med.* 2022 Jan 21; **14**(1): e32125.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

14. Santolini M: *Covid-19: the rise of a global collective intelligence?* The Conversation; 2020 [cited 2021 Apr 2].
Reference Source

15. eLife Journal Policy: **eLife.** 2022.
Reference Source

16. Easy Chair: 2021.
Reference Source

17. Flach PA, Spiegler S, Golénia B, *et al.*: **Novel tools to streamline the conference review process: experiences from SIGKDD'09.** *ACM SIGKDD Explor. Newsl.* 2010 May 27; **11**(2): 63–67.
Publisher Full Text

18. Merrifield MR, Saari DG: **Telescope time without tears: a distributed approach to peer review.** *Astron. Geophys.* 2009; **50**(4): 4.16–4.20.
Publisher Full Text

19. Shah NB: **Challenges, experiments, and computational solutions in peer review.** *Commun. ACM.* 2022 May 20; **65**(6): 76–87.
Publisher Full Text

20. Mervis J: **A radical change in peer review.** *Science.* 2014 Jul 18; **345**(6194): 248–249.
PubMed Abstract | **Publisher Full Text**

21. Naghizadeh P, Liu M: **Incentives, Quality, and Risks: A Look Into the NSF Proposal Review Pilot [Internet].** *arXiv.* 2013 [cited 2023 Mar 7].
Reference Source

22. Kokshagina O: **Open Covid-19: Organizing an extreme crowdsourcing campaign to tackle grand challenges.** *RD Manag.* 2021 Mar 23; **52**: 206–219.
Publisher Full Text

23. Helpful Engineering: **Helpful.** [cited 2021 Jul 9].
Reference Source

24. Graham CLB: **EXTENDED DATA FOR: Community review: a robust and scalable selection system for resource allocation within**

open science and innovation communities. 2022, November 26.
Publisher Full Text

25. Je-S electronic applications - Economic and Social Research Council: [cited 2021 Jul 9].
Reference Source

26. Graham CLB:DATA FOR: Community review: a robust and scalable selection system for resource allocation within open science and innovation communities. [Dataset]. 2022, November 26.
Publisher Full Text

27. Jerrim J, de Vries R: **Are peer-reviews of grant proposals reliable? An analysis of Economic and Social Research Council (ESRC) funding applications.** *Soc. Sci. J.* 2020 Mar 6; 1–19.
Publisher Full Text

28. Aidelberg G, Aranoff R, Javier Quero F, *et al.*: *Corona Detective: a simple, scalable, and robust SARS-CoV-2 detection method based on reverse transcription loop-mediated isothermal amplification.*ABRF; 2021.
Reference Source

29. Bektas A: **Accessible LAMP-Enabled Rapid Test (ALERT) for detecting SARS-CoV-2.** *Viruses.* 2021.
Publisher Full Text | **Reference Source**

30. Cheng C: **COVID-19 government response event dataset (CoronaNet v. 1.0).** *Nat. Hum. Behav.* 2020; **4**(7): 756–768.
PubMed Abstract | **Publisher Full Text**

31. Greshake Tzovaras B, Rera M, Wintermute EH, *et al.*: **Empowering grassroots innovation to accelerate biomedical research.** *PLoS Biol.* 2021 Aug 9; **19**(8): e3001349.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

32. Monaco C, Jorgensen E, Ware S: *The One Hour COVID Test: A Rapid Colorimetric Reverse-Transcription LAMP–Based COVID-19 Test Requiring Minimal Equipment.*ABRF; 2021 Sep.
Reference Source

33. Tzovaras BG: **Quantified Flu: an individual-centered approach to gaining sickness-related insights from wearable data.** *medRxiv.* 2021.

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 26 January 2024

https://doi.org/10.5256/f1000research.145384.r185935

✓ **Dimitrios Kouis**

University of West Attica, Athens, Greece

The article presents an interesting topic related to a different model/method of evaluating proposals for funding based on applying the distributed peer review method with specific variations and assumptions. Implementing the proposed model is suggested as an alternative approach to the traditional way of evaluating funding proposals, meaning the evaluation through peer review by a panel of experts. This particular topic increasingly engages the research community (indicative articles 1,2,3) in selecting proposals for funding and evaluating scientific publications and other aspects of research, mainly within the context of Open Science.

In terms of the current article, both the proposed evaluation flow and the workflow adaptations that were applied and tested at various stages and subsequently evaluated with an in-depth statistical analysis give the article the necessary elements of innovation and originality to be accepted for publication. Specifically, the methodological approach, the assumptions made, the data collected, and the further analysis and conclusions drawn are based on a correct and logical scientific approach and present great interest.

If there are any points of critique and thought for a potential future research effort (and not so much regarding corrections to the current article), these are summarized in the following two points:

○ To what extent can the scientific field of application for evaluating the proposed methodology affect the results? Will applying the proposed method to a different scientific area provide the same results? (As expected, different scientific topic means other evaluators' mentality, evaluation forms etc.)

○ To what extent can the proposed methodology be adopted by a governmental organization or an international funder, given that the evaluators might also be applicants for getting funded and thus exhibit an "expected behaviour of bias" and, as the authors observed in some cases, must be removed? Although such a process (i.e., the one proposed in the article) might ultimately yield better results (fairer evaluation), it may also be subjected to greater criticism due to the possible involvement of the participants themselves.

○ Concerning the previous observation, which might be the weak point for adopting the

proposed methodology on a broad scale, an alternative solution could be using the participants' reviews only to improve the quality of the proposals and not for grading purposes.

The article provides all the necessary elements, such as technical validity, methodological accuracy, results reproducibility, etc. It also opens the discussion within a correctly documented scientific framework for the revision/improvement of the evaluation framework for research proposals for funding in the context of Open Science.

References

1. Abma-Schouten R, Gijbels J, Reijmerink W, Meijer I: Evaluation of research proposals by peer review panels: broader panels for broader assessments?. *Science and Public Policy*. 2023. Publisher Full Text

2. Feliciani T, Morreau M, Luo J, Lucas P, et al.: Designing grant-review panels for better funding decisions: Lessons from an empirically calibrated simulation model. *Research Policy*. 2022; **51** (4). Publisher Full Text

3. Bendiscioli S: The troubles with peer review for allocating research funding: Funders need to experiment with versions of peer review and decision-making.*EMBO Rep*. 2019; **20** (12): e49472 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Library Networks with an emphasis on shared bibliographic data management systems and collaborative systems for managing digital collections; Digital Content Management with an emphasis on electronic scientific publications; Scholarly communication, and specifically models of access and dissemination of scientific content, on research data as well as on the procedures of measurement and evaluation of research activity

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 25 April 2023

✔️ **Ferdinando Patat** 🆔
European Southern Observatory, München, Germany

I am fully satisfied with the changes the authors have made following my recommendations.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Astrophysics: Supernovae. Peer-review: proposal selection and time allocation processes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

- -

Version 1

Reviewer Report 17 February 2023

https://doi.org/10.5256/f1000research.138239.r161990

❓

Ferdinando Patat [iD]
European Southern Observatory, München, Germany

The paper reports on the design, deployment and evaluation of what the authors describe as an 'agile community review system' for allocating micro-grants. As opposed to classical peer-review performed by pre-selected expert panels, the community review involves the participation of applicants themselves. It therefore provides as scalable solution, in the sense that the number of reviewers increases with the number of applications, hence keeping the review load manageable and the duty-cycle faster than in the classical panel case. The data show that the inter-reviewer correlation is similar to what has been reported for the classical paradigm. Also, the final rankings are shown to be robust against the randomised removal for reviewers.

The paper presents the case very clearly and accurately. One aspect which is weak (honestly this is probably the only weakness of the paper) is that the casual reader is left under the impression that the idea of a community review is new and proposed here for the first time. And that is not the case. This concept, sometimes indicated as *distributed peer review* (DPR), has been around for many years. As far as I know the idea appeared first in 2009 in a paper by Saari & Merrifield (2009)[1]. Although it was referring to applications for telescope time, the concept was of general interest. This mechanism has been implemented at least in three major, ground-based astronomical facilities (GEMINI, ALMA, ESO). Also, a similar distributed process has been deployed in the field of computer sciences for the selection of conference papers[5]. Other interesting examples are those of the US National Science Foundation[2,3] and the US National Institute of Food and Agriculture (NIFA), which has deployed DPR in 2016[6,7].

The authors should therefore provide a short recap of these initiatives and cite the applicable references. In addition to the papers published by the above organisations, there is an article on Physics Today[4] which gives a nice summary for the specific cases in astronomy.

Other minor points, which the authors should consider, are listed here:
1. Scalability. The authors correctly state that the approach is scalable, in the sense that increasing the number of projects linearly increases the number of reviewers. They show this in Fig. 2b. At a first read it looks like this is an unexpected results and, somehow, the data show that this is the case. However, I argue that this is completely expected, since the number of potential reviewers, on average, is directly proportional to the number of submitted projects. This would not be necessarily the case if the system would not force the applicants to review proposals. Maybe the authors should be more explicit on this.

2. In the section about fraudulent behaviour, the authors state that this was removed from the review process. However, they do not explain what this means in practice in terms of actions taken against those reviewers. Is their project simply rejected? Are they warned? Or is just their evaluation removed from the data?

3. The caption of Figure 4 for panels c) and d) does not match what is presented in the figure and also its description in the main text section *Measuring reviewer biases*:

"For each project, we compute the ratio between the proportion of applicant reviewers to the average proportion of applicant reviewers observed in the round. The boxplot compares the computed enrichments to the ones obtained for randomly assigned reviewers to projects, showing that applicants are evenly distributed across projects. (d) For each project, we compute the ratio between the proportion of applicant reviewers to the average proportion of applicant reviewers observed in the round. The boxplot compares the computed enrichments to the ones obtained for randomly assigned reviewers to projects, showing that applicants are evenly distributed across projects."

The authors never explain what they mean by "enrichment". Also, while panel c) may be showing a ratio, this is not the case for panel d), which on the y-axis has the label "inter-review correlation", which sounds like a Pearson correlation, and not a ratio. In general, I am rather confused about panels c) and d) in Fig. 4, which I reckon requires a better explanation also in the main text (which, by the way, never includes the word "enrichment").

4. As shown in Fig. 3a, there are two evaluation criteria related to the Team composition. Given that many organisations are now moving to a dual anonymisation of the applications, it would be interesting to see how the scores change if one removes these two indicators from the final score. I guess that, given the large number of indicators (and the results shown by the authors about the slight change when 50% of them are removed), no measurable effect is going to be seen. Also because, since the team's identity is known to the reviewers, there is certainly a cross-talk between this indicator and all others. It is probably therefore impossible to disentangle its effect from the available data. Nevertheless, I suggest the authors mention this aspect.

The study is well designed, the analysis was conducted in an appropriate way and included a satisfactory level of checks and controls. The methods are well explained and the necessary data are made available so that the analysis can be repeated and validated. The statistical analysis and its presentation meet the required standards.

The conclusions are well supported, interesting and useful for other organisations which may consider adopting a similar schema. I therefore recommend the article for indexing after the point on the missing citations and short description of existing cases is addressed.

References
1. Merrifield M, Saari D: Telescope time without tears: a distributed approach to peer review. *Astronomy & Geophysics*. 2009; **50** (4): 4.16-4.20 Publisher Full Text
2. Naghizadeh P, Liu M: Incentives, Quality, and Risks: A Look Into the NSF Proposal Review Pilot. *arXiv*. 2013. Publisher Full Text | Reference Source
3. Mervis J: A radical change in peer review. *Science*. 2014; **345** (6194): 248-249 Publisher Full Text
4. Singh Chawla D: Distributed peer review passes test for allocating telescope slots. *Physics Today*. 2020. Publisher Full Text | Reference Source
5. Shah N: Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*. 2022; **65** (6): 76-87 Publisher Full Text
6. NIFA - FAQs for Distributed Peer Review (DPR). *NIFA*. 2016. Reference Source
7. distributed peer-review (DPR). *NIFA*. 2016. Reference Source

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Astrophysics: Supernovae. Peer-review: proposal selection and time allocation processes.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 08 Mar 2023
Chris Graham

Dear Ferdinando,

Thank you for a thorough and well-rounded review of our article. We are glad that you found the article interesting and our methods to be robust and the data well described. You are correct, our method is very similar to distributed peer review (DPR) and we should have included citations and mentions of this in our literature review. Its use in astronomy, and especially by the NSF are particularly interesting, as are the articles which cite its drawbacks. Thank you for improving the paper with mention of this. We have incorporated the mention of our use of DPR throughout the paper.

We have now expanded the introduction to speak about distributed peer review, and throughout conclusions in the article have also made reference to distributed peer review, although we did not initially use DPR, and were instead adopting a crowd-sourcing approach, and then combined the two methods. These changes have been made especially to the Introduction, where we have incorporated your suggested references, as well as

others not mentioned

In terms of your specific comments, for Figure 2d, the wording has now been changed to reflect the reality that the community review method was expected to be scalable. The use of language was through the assumption of a null hypothesis and was innappropriate.

We have also changed the Figure legend of Figure 4c, as pointed out to be reflective of the figure. This was a copy-editing error. The use of the word ' enrichment' has also been specified for Figure 4d.

Your final point, on removing team composition questions is very interesting especially when thinking about double anonymisation of reviewers. We hope, based on the data of the PCAs which separated the questions by their power and difference to other questions (Figure 3b), that this was not a big influence on the final scores therefore would agree with your own assumption that this wouldn't effect the rankings greatly. Unfortunately although your hypothesis for the inclusion of a further figure to analyse this is sound, our team would prefer to keep the paper in its current state in terms of new hypotheses, but think this is an excellent question and in future publications will ask this very question if given the chance.

Overall your review has identified some important literature and potential issues, as well as a new hypothesis and we are very grateful for it.

Thank you for taking the time to analyse our article in the way you have, you've enlightened us specifically on the astronomy field's use of DPR and shown how there is precedent for such methods, with the example of the National Science Foundation's pilot study. Hopefully there will be further attempts with such large organisations in the future, perhaps with safeguarding measures to counteract any collusion, as well as controls studies for grant review with existing techniques.

Competing Interests: No competing interests.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research