Research Policy 51 (2022) 104467

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Designing grant-review panels for better funding decisions: Lessons from an empirically calibrated simulation model

Thomas Feliciani [a,*], Michael Morreau [b], Junwen Luo [c], Pablo Lucas [a], Kalpana Shankar [c]

[a] School of Sociology and Geary Institute for Public Policy, University College Dublin, Dublin, Ireland
[b] Department of Philosophy, UiT The Arctic University of Norway, Tromsø, Norway
[c] School of Information and Communication Studies and Geary Institute for Public Policy, University College Dublin, Dublin, Ireland

ARTICLE INFO

Keywords:
Research evaluation
Scoring
Inter-rater reliability
Social simulation
Peer review

ABSTRACT

Objectives: To explore how factors relating to grades and grading affect the correctness of choices that grant-review panels make among submitted proposals. To identify interventions in panel design that may be expected to increase the correctness of choices.

Method: Experimentation with an empirically-calibrated computer simulation model of panel review. Model parameters are set in accordance with procedures at a national science funding agency. Correctness of choices among research proposals is operationalized as agreement with the choices of an elite panel.

Conclusions: The simulation model generates several hypotheses to guide further research. Increasing the number of grades used by panel members increases the correctness of simulated choices among submitted proposals. Collective decision procedures giving panels a greater capacity for discriminating among proposals also increase correctness. Surprisingly, differences in grading standards among panel members do not appreciably decrease correctness.

1. Introduction

Evaluation plays a critical role in research management and policy making. Many evaluative processes make use of *grading languages*. These are vocabularies of evaluative expressions that come in a “top” to “bottom” order. Representative examples of grading languages used in grant review are the numerical scale from ‘5’ (the top score) to ‘1’ (the bottom score) used by Science Foundation Ireland (SFI), and the scale from ‘excellent’ down to ‘poor’ used by National Science Foundation (NSF) panels in the United States.

Several factors relating to grading languages can affect a grant-review panel’s choices. One is the number of scores or grades: the more there are, the more finely the panel can discriminate better proposals from worse ones. Another factor is differences in understanding of grades from one panelist to the next, which can result in equivocation, as well as spurious agreements and disagreements. A third factor is just how the panel moves from individual panelists’ grades for proposals to collective decisions. Additional factors unrelated to grades and grading include the number of panelists, their individual expertise, how much effort they put into studying proposals, how many proposals can be chosen, and how much the proposals differ in their merits.

This article has two main objectives. The first is to explore how features of grading languages together with other factors determine the correctness of a grant-review panel’s choices. The second is to draw lessons about panel designs that can help grant-review panels make correct choices.

Grant-review panels are complex social entities. Their decisions emerge in the interplay of many factors, and behavioral studies of grant-review panels aimed at measuring the effect of each factor *in situ*, while controlling for the others, will require complex experimental designs and very large sample sizes. Some guidance on what hypotheses to test first would be helpful. This study uses experiments with a computer model to tease apart the contributions of different factors. Computer modeling is a well-established method for studying complexity in the sciences; recently, it has been used to study peer review and science funding (Squazzoni and Takács, 2011; Roebber and Schultz, 2011), and to explore potential effects of implementing science policies (Ahrweiler et al., 2012, 2015). While computer modeling cannot replace experiments with real grant-review panels, it can help us identify plausible and interesting hypotheses.

Many social simulation models lack empirical calibration, which limits their realism and relevance to target phenomena (Hassan et al.,

[*] Corresponding author.
E-mail address: thomas.feliciani@ucd.ie (T. Feliciani).

https://doi.org/10.1016/j.respol.2021.104467
Received 22 December 2020; Received in revised form 20 December 2021; Accepted 21 December 2021
Available online 22 January 2022

2010). Simulation studies of peer review have similarly suffered from lack of relevant data (Feliciani et al., 2019; Gurwitz et al., 2014; Squazzoni et al., 2020). The present simulation model has been calibrated using empirical data on grant panels from Science Foundation Ireland (SFI), the largest national science-funding agency in Ireland, as well as publicly available SFI grant call documents, and results from a survey and interviews that we conducted with SFI reviewers.

The next section introduces our notion of correct choice and discusses our research questions. Section 3 presents a model of grant review and the parameters of our simulation. Section 4 presents the results of our simulation experiments. Section 5 draws conclusions and discusses implications for the design and management of grant-review panels.

2. Conceptual framework and research questions

This article identifies conditions under which grant-review panels may be expected to make correct choices among research proposals. Here, correctness is understood not in procedural but in *epistemic* terms.[1] That is, whether a panel's choice is correct is, in the present study, not a matter of how this choice *comes about*—of whether applicable rules and guidelines are adhered to, or anything that happens while the panel is making its choice. Rather, it is a matter of the *outcome* of the choice—the set of proposals that the panel settles on in the end. A panel's choice is epistemically correct if the chosen proposals are, intuitively speaking, the right ones to choose.

Some studies, following early and still influential models of peer review, assume that research proposals have an "objective" or "true" merit (Thurner and Hanel, 2011; Squazzoni and Gandelli, 2013; Roebber and Schultz, 2011). Thinking in this way, an epistemically correct choice might be conceived as one that picks out from among all submissions those proposals whose true merit is sufficiently high, or whose true merit is higher than that of the rejected proposals. Such an understanding of correctness, though, is dubious. Evaluating a research proposal is, intuitively speaking, quite unlike measuring, say, the length of something with a ruler. There is no "gold standard" instrument for measuring true merit in research proposals.

We take the choices of an elite *reference panel* as our standard of correctness. An ordinary or *field* panel makes a correct choice if it chooses the same set of proposals as the reference panel chooses. In an experiment with real panels, a reference panel would be empaneled using elite reviewers: people with great expertise, not only in the academic domains of the submitted proposals but also in interpreting evaluation criteria, setting weights or priorities, and in other aspects of panel review. The reference panels in our simulation experiments, like the field panels, are simulated; but the idea is the same: correctness of a simulated field panel's choice is agreement with a simulated reference panel's choice.[2]

We study how various factors affect the epistemic correctness of a field panel's choices. One factor is the extent to which panels are able, by assigning different grades, to discriminate better proposals from worse ones. We investigate the consequences of providing panelists with fine-grained grading languages, and of using judgment-aggregation procedures that give panels greater powers of discrimination than their individual panelists. Another factor is differences in grading standards among panelists. We investigate the consequences for panel decisions by manipulating panelists' thresholds for awarding grades. These and other factors including panel size, the individual expertise of panelists and targeted funding rates are systematically manipulated in the simulation experiments of Section 4 to determine the consequences for the correctness of a panel's choices.

Our simulation experiments suggest hypotheses for behavioral testing using real panels, with potential implications for the practice of grant review. First, judgment-aggregation rules that boost discrimination can contribute to the correctness of choices. Switching to these procedures would be straightforward, with some panel designs. Second, differences in panelists' grade thresholds have little effect on panel correctness. We suspect that training panelists to coordinate understandings of grades might, with some panel designs, often not be worth associated costs. We return in Section 5 to these matters, and to implications for crowdsourced review and funding research by partial lotteries.

3. A model of panel review

This section introduces a general model of panel review and main parameters of the simulation model. The simulation model is discussed in detail in Appendix A.

We model a grant-review panel that is given some proposals for review and tasked with choosing some of them for funding. The panel's choice is based on evaluation criteria such as novelty, methodology and impact. The funding agency might specify precisely how these criteria should be understood, and what their weights or priorities should be; or it might allow the panel to decide this itself.

The review panel uses a *grading language* to evaluate proposals. This is a finite list of *grade expressions* that come in a "top" to "bottom" order. Typically they are natural-language predicates (such as 'outstanding', 'very good',…) or numerals ('5', '4', …); but they could be almost any marks or signs. A typical grading language used by review panels has around five grades. There could be as few as two or upwards of ten.

Training materials and other guidelines typically explain in natural language the intended meanings of grades. They may be expected to constrain panelists' use of grades to some extent (Sattler et al., 2015). It is nevertheless common for people to have different thresholds for assigning grades—even experts with similar training and experience, such as members of science panels (Wallsten et al., 1986; Wardekker et al., 2008; Morgan, 2014).

The panel makes its choice as follows. First, the submitted proposals are divided among the panelists for individual review. Then, panelists assign to the proposals grades from the panel's grading language. Finally, on the basis of the individually assigned grades, a decision is made about which proposals to fund. The process leading from individual inputs to the whole panel's funding decision can include deliberation, voting or another form of judgment aggregation, or both.

Section 3.1 through 3.5 discuss components of our model of grant-review panels, summarized in a flowchart in 3.6.

3.1. Underlying scales and categorization

To grade proposals is, necessarily, to sort them from "top" to "bottom." The top category contains proposals that receive the highest grade; the next are proposals with the next-highest grade, and so on. Any such linear series of categories we call a *category scale*.

The category scale corresponding to a grading language obviously has the same number of categories as there are grades, typically about five. Most reviewers, though, can make finer distinctions than this. They may distinguish, among proposals that they give the same grade, some they think are better and others they think worse.

Reviewers, in our model of grading, judge proposals in the first instance on category scales that are comparatively fine-grained. Only then do they express their judgments as inputs to the panel, using the grading language specified by the funding agency. These fine-grained category scales we call *underlying scales*.

For example, suppose a reviewer gives any given proposal up to 10 points for each of 10 evaluation criteria, and then tallies up all points as

[1] Other social institutions analyzed in epistemic terms include economic systems (Hayek, 1945), democracies (Goodin and Spiekermann, 2018) and think tanks (Claveau and Veillette, 2020).

[2] Reference standards are similarly used to evaluate diagnostic methods in medicine, when "gold standard" diagnostic tests are unavailable (Bertens et al., 2013).

a basis for awarding a grade. This reviewer's underlying scale has 101 categories, from 0 to 100 total points.

Our model does not represent mechanisms of *underlying categorization*, the process by which reviewers place proposals on their underlying scales. In the example, that is done by assigning and then adding up criterial scores, all given the same weight. This is just one of many possibilities, though. The particulars of the categorization process can be different for different reviewers and, for any particular reviewer, from one proposal to the next.[3]

Random variation in a reviewer's underlying categorization process we call *underlying noise*. It can cause inter-rater unreliability, which must be reckoned with in peer review (Bornmann et al., 2010; Guthrie et al., 2018; Nicolai et al., 2015). The reviewers' average underlying noise is a parameter of the simulation model, λ.

Deficiencies in reviewers' capabilities, such as missing domain expertise, are one source of underlying noise. Another is a lack of time or other resources. Suppose furthermore that the funding agency leaves it to panelists to settle the meanings of evaluation criteria. Then where a panelist puts some given proposal on their underlying scale can turn on just how this reviewer understands, for instance, *impact*, and the importance given to this criterion relative to other criteria such as *scholarship*. Underspecification of the evaluation criteria is a further source of underlying noise.

3.2. Reference distributions

The submission of proposals for review is represented, in our model, as sampling from a population of possible submissions. Suppose an elite panel considers all proposals in this population, and categorizes them on some given underlying scale—perhaps from 0 to 100 criterial points, as in the example in Section 3.1. It settles all questions left open by the specification of the evaluation criteria, resolving ambiguities and setting weights or priorities. Then, we assume, the elite panel is able to assign to each possible submission a unique underlying category. There is a frequency distribution of underlying categories to which proposals are in this way assigned. This is called the *reference distribution*, for these possible submissions and this resolution of criterial indeterminacies.

The shape of reference distributions represents, in our model, the combined effects of several aspects of the writing, submission and evaluation of proposals. One determinant of the shape, already discussed, is the evaluation criteria and their weights or priorities. Holding fixed any given population of proposals and their relevant properties, a change in the interpretation, weights or priorities of evaluation criteria can result in their being assigned to different underlying categories.

A second determinant of the shape of reference distributions is properties of the proposals themselves. Holding the evaluation criteria and their weights and priorities fixed, but changing properties of the proposals relevant to how they measure up by these criteria, will in general also result in a reference distribution with a different shape.

Relevant properties of research proposals submitted for review can depend on the incentive structure in grant applications. Where researchers have a high degree of trust that funding decisions really do depend on the merits of submissions, they are incentivized to submit high quality proposals. Rules such as the European grant program Horizon 2020′s (H2020) "quarantine" for rejected proposals presumably have a similar effect. Under such conditions, we suppose, expert reviewers will by and large find submitted proposals to be reasonably good, resulting in a "high" distribution, like that on the left of Fig. 1. This seems to be the case for SFI funding schemes: several experienced SFI reviewers we interviewed reported that most proposals they reviewed were of high quality.

Lower expectations about the extent to which funding decisions reflect the merits of proposals might on the other hand incentivize writing proposals quickly, resulting in lower quality submissions. A "low" reference distribution can represent this condition (Fig. 1, middle panel). A bimodal distribution, finally, can represent the condition in which there are outstanding researchers who make every effort to submit proposals that are as good as can be, but there is also a distinct group of applicants whose submissions are poor (Fig. 1, right panel).

The shape of reference distributions could affect the correctness of choices. Where most proposals are in top categories, for instance, it might be hard to distinguish the truly outstanding ones in a strong pool. The reference distribution is a parameter of our simulation model. In the simulation experiments of Section 4, it is switched between high, low and bimodal values in order to test the robustness of our findings.

3.3. Grading languages

The panel's grading language is modeled as a finite vocabulary of grading expressions in a fixed order from "top" to "bottom". The number L of grading expressions is a parameter of the simulation model.

Panelists are assumed to have some understanding of what the different available grades mean. We model panelists' understanding of grades using the technical device of an *interpretation.* An interpretation of a grading language, on any given underlying scale, is an ordered partition of this scale, specifying which underlying categories are covered by which grade expressions. The *threshold* of a grade expression is the lowest underlying category that this expression covers. An interpretation of a grading language amounts to a specification of thresholds for each of its grade expressions (see Fig. 2).

Funding agencies typically explain the intended meanings of grades in rubrics provided to reviewers. We model these guidelines as delimiting a class of preferred interpretations, which we call *reference* interpretations. Intuitively, a reference interpretation models a correct understanding of the funding agency's grades.[4] For any given underlying scale, there are in general several reference interpretations, some giving the grades thresholds that are a bit higher, and others setting lower thresholds. This is how our model represents indeterminacy in the meanings of grades, due to vagueness and ambiguity in the natural language with which grades are defined.

The more precise and unambiguous the grade definitions, the fewer reference interpretations there are. Formal calibration training and the informal culture of a funding agency, established over time, can further limit the class of reference interpretations. Similarities and differences among the particular proposals under consideration might also be relevant: thresholds between grades should be set so as to lump similar cases together and to avoid blowing small differences out of proportion (Maudlin, 2008).

Our model allows reviewers' understandings to deviate from correct understandings, modelled as reference interpretations, and to deviate also from each other's understandings. In addition to underlying noise, these variations can reduce inter-rater reliability. Indeed, reviewers might assign different grades to the same proposal solely because of a difference in their interpretations of grades (Pier et al., 2018; Sattler et al., 2015). The extent of random variation in reviewers' interpretations is another parameter ϑ of the simulation model, the panel's *grade-threshold noise.* By varying values of this parameter, we model cases in which random factors affect panelists' grading standards to various extents.

To coordinate grading thresholds among reviewers, funding agencies can adopt interpretations of grading languages that agree with

[3] Another categorization mechanism involves comparing similarities. See Hampton (2001).

[4] Similarly, the Intergovernmental Panel on Climate Change, in order to avoid miscommunication, has stipulated a single correct interpretation of a scale of seven probability expressions used in its publications, ranging from 'virtually certain' at the top to 'exceptionally unlikely' at the bottom (Mastrandrea et al., 2011).

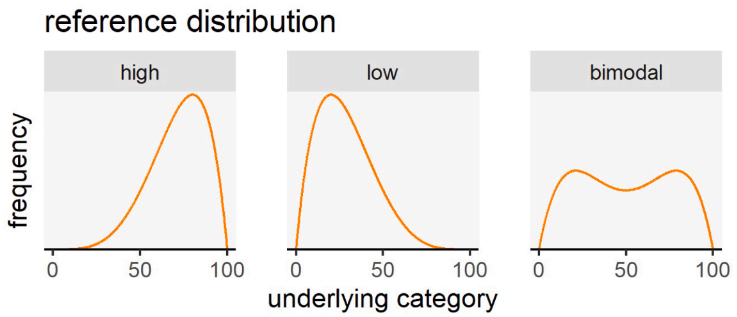

Fig. 1. Three reference distributions model different cultures surrounding the writing, submission and reviewing of grant proposals.

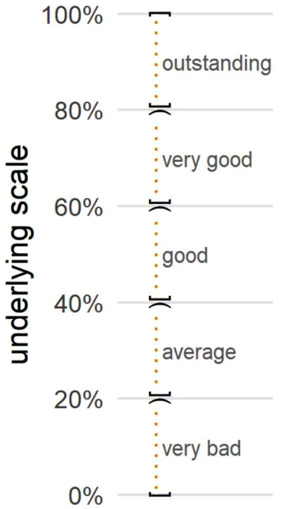

Fig. 2. An interpretation of the SFI grading scale within an underlying category scale from 0 to 100%. The top expression 'outstanding', according to this interpretation, applies to any proposal with an underlying score of 80% or above. The threshold for the next grade expression 'very good' is 60%, and so on down. The bottom SFI grade 'very bad', according to this interpretation, covers proposals whose underlying score is between 0 and 19%.

reviewers' natural and intuitive understandings.[5] They can also recruit experienced reviewers and provide calibration training. By manipulating the parameter of grade-threshold noise, our model can be used to study the likely consequences of coordinating grading thresholds for the correctness of a panel's choices.

3.4. *Grade decisions and panel rankings*

An individual panelist arrives at a grade for a proposal by first positioning it on their own underlying scale and then "looking up" the corresponding grade, according to their own interpretation of the grading language.

Reviewers could in principle change their grading standards from one proposal to the next. Our model accommodates this possibility by allowing reviewers to look up corresponding grades using different interpretations for different proposals. We suppose, though, that a reviewer's interpretation of the grading language is a more-or-less fixed attribute of them, at least for the duration of their work on any given panel. In our simulation model, accordingly, any given reviewer uses the same interpretation for grading all proposals allocated to them for review.

A panel arrives at a grade for a proposal on the basis of the grades assigned to it by the individual panelists. On so-called sitting panels, this involves deliberation among panelists (Fogelholm et al., 2012; Langfeldt, 2001; Obrecht et al., 2007). On SFI's postal review panels, on the other hand, there is no deliberation. A panel's score for a proposal is simply calculated by taking the mean of individually assigned scores.

Our simulation model explicitly represents judgment aggregation by taking the mean as well as several other judgment-aggregation rules. It does not explicitly represent non-algorithmic deliberation among panelists. Notice, though, that judgment-aggregation rules even so can model deliberation to some extent. For example, aggregating individually assigned scores by taking their mean may be taken to represent, in an idealized way, a deliberative process in which panelists tend to seek the middle ground.

How a panel aggregates inputs from panelists can have consequences for its capacity to discriminate better from worse. Consider a panel which, like an SFI postal panel, has five panelists who score proposals on a numerical scale from 1 to 5. Unlike the SFI panel, though, this other panel does not take the mean of individually assigned scores. Instead, it takes the median.[6] Such a panel sorts proposals into at most five classes, because the median of five numerals between 1 and 5 is itself a numeral between 1 and 5. The SFI panel on the other hand can make many more distinctions than this, since there are 21 different means of five numerals from 1 to 5. Taking the mean of individually assigned scores is an example of what we call a *discrimination boosting* aggregation rule, or *booster* for short. The simulation experiments of Section 4 explore the contribution of boosters to the quality of panel decisions.

The aggregation rule used by field panels is a further parameter R of the simulation model. The values studied in Section 4 are given in Table 1.

3.5. *Choice and correctness*

Once a panel has graded all proposals submitted for review, it ranks them in the order of the collectively assigned grades. The result is this panel's *panel ranking*. The field panel and the reference panel each choose some number of proposals from the top of their own panel rankings, guided by a target set by the funding agency. The *choice rate K*, or targeted percentage of proposals to choose, is another parameter.

Section 2 introduces the idea that a field panel's choice is correct if it agrees with a reference panel's choice. Notice two implications. First, correctness is a matter of degree. The field panel's choice set overlaps to

[5] Basing official interpretations of natural-language probability expressions on people's actual understandings, instead of stipulating them without regard to linguistic evidence, has been shown to increase the consistency of understanding among trained intelligence analysts (Ho et al., 2015).

[6] It might be thought that this is better when the numerals on the scale have a merely ordinal significance, and adding them up is not strictly speaking meaningful.

Table 1
Judgment-aggregation rules.

Boosters
Mean. The aggregate score for a given proposal is the arithmetic mean of its individually assigned scores. **Hypermean**. A weighted mean with weights that dampen the contributions of panel members who tend to disagree with the ordinary mean and amplify contributions that tend to agree. The idea is that, due to the wisdom of crowds, members who tend to disagree are likely to be more wrong than those who tend to agree.[1] **Majority judgment**. A ranking method based on the median, proposed by Balinski and Laraki (2010). When two proposals have the same median grade, the 50th percentile is removed from the grade profiles[2] and the median is calculated once again.[3] If there is still a tie this step is repeated, until either the tie is broken or there is only one grade left.
Non-boosters
Median. The aggregate score for a proposal is the middlemost (50th percentile) when all individual grade judgments are put in their “top” to “bottom” order. With an even number of judgments, it is the average of the two middle ones. Used by some H2020 panels (European Commission, 2020). **Lowest score**. The aggregate score is the lowest of the individually assigned scores. This might be expected to minimize type I errors, where proposals that should be rejected are accepted.[4] **Highest score**. The mirror image of **Lowest score**. Might be expected to minimize type II errors, where proposals that should be accepted are rejected.

[1] The hypermean represents a wider class of weighting rules including the trimmed mean (Jose and Winkler 2008) and a proposal by Budescu & Chen (2015) in which individual weights are based on past performance. For implementation details, see Appendix A.

[2] The *grade profile* of a proposal is the (multi)set of grades assigned to it by all of its reviewers.

[3] With an even number of reviewers, the grade to be removed is the lowest of the two grades around the 50th percentile.

[4] Esarey (2017) considers an equivalent rule. Intuitively, the Lowest score rule corresponds to the practice of rejecting proposals just because of a single negative review.

a greater or lesser extent with a reference panel’s choice set.[7] Second, correctness is plural: there are in general several sets of proposals that it is correct to choose. This is because there is often some room for panels to interpret evaluation criteria and to decide the importance of the different ones (Lee et al., 2013; Abdoul et al., 2012; Langfeldt, 2001). A reference panel may choose any of several sets of proposals, depending on which interpretations, weights and priorities it settles on, and a field panel’s choice is completely correct if the outcome is identical to any of these. Where it is not identical to any of them, we understand the degree of correctness of the field panel’s choice to be the *maximum* extent of agreement between the set it chooses and any set that the reference panel can choose.[8]

It is instructive to compare the correctness of a panel’s choices with the correctness of choices by a single member, chosen at random. This *control*, as we call it, represents the case in which a single reviewer decides, alone. It also represents the case in which the panel tends to defer to the judgment of one of its members — perhaps a forceful personality, or someone who, just by *happening* to speak first, anchors the grade decisions of the others. We count a panel as exhibiting the wisdom of crowds to the extent that it outperforms the control.

3.6. Model overview

Fig. 3 shows a flow diagram of our model of individual grading, grade aggregation, panel ranking, choice and correctness.

4. Simulation experiments

This section presents results of a series of experiments with simulated grant-review panels. The simulated panels are purely aggregative. That is, they arrive at their panel rankings without any deliberation: panelists’ grades for proposals are simply aggregated using a suitable rule from Table 1. Table 2 gives an overview of all model parameters and indicates settings defining the parameter space explored. Implementational details of the simulation model are in Appendix A. Code and documentation are available at https://github.com/thomasfeliciani/wisdom-of-expert-crowds/.

The *baseline configuration* is a privileged point within the parameter space from which our experiments venture out by systematically changing parameter settings. It represents the review process that the case-study funding agency (SFI) established for some of its funding programs. Baseline parameter settings chosen using empirical data include: $N = 5$ reviews per proposal, choice rate $K = 20$, $L = 5$ grades, judgment aggregation by averaging scores, and a reference distribution skewed towards the top (“high”). Non-extreme baseline values were stipulated for underlying noise and grade-threshold noise: λ=0.2, and ϑ=0.5.[9] In Table 2, all baseline values are underlined.

Appendices A and B have details on the implementation of the model parameters and on how baseline settings were derived from empirical data.

In the rest of this section, we first report on the correctness of choices by a panel in the baseline configuration. Then we report the results of varying some parameters around this configuration, while keeping others fixed. We ran 500 independent simulations for each parameter configuration.

4.1. Performance in the baseline

Fig. 4 plots the correctness of the choice by simulated panels in the baseline configuration (orange).[10] It also shows the performance of just one reviewer, the control, chosen at random from among the panel’s members (dark gray). The two boxplots show descriptive statistics of the distributions of correctness in choices. The middle line of each boxplot indicates median correctness; its vertical span indicates the interquartile range. In the background, violins (light gray) show the distributions underlying the boxplots. This plotting and coloring convention holds throughout Section 4.

Fig. 4 shows that the panel in the baseline configuration greatly outperforms the control. The median correctness of its choices is clearly higher than the median correctness of the control, and the correctness of its choices is somewhat less variable. The simulated SFI panel displays a clear wisdom of crowds.

Fig. 5 compares performance of the baseline panel and control of Fig. 4 with that of other panels of varying sizes (with all parameters other than panel size set to the baseline values). The performance of larger panels is clearly better, other things being equal, than that of smaller panels, with higher median correctness and less variability.

Notice that, for each subsequent new member on the panel, median correctness increases by a smaller amount. It seems that, under the conditions of grant review, there is a limit to the wisdom that can be created just by increasing the size of crowds. We turn now to factors

[7] The degree of agreement can be quantified using any of several measures of inter-rater reliability. In the simulation experiments in Section 4, we use Cohen’s kappa.

[8] Notice that on this accounting of the degree of correctness, any choice that a reference panel actually makes is bound to be completely correct.

[9] The stipulated baseline values for λ and ϑ are different because of the different implementations of the two corresponding sources of noise in the simulation model (see Appendix A for details). In both cases, the baseline values represent moderate levels of noise.

[10] Correctness, scored using Cohen’s kappa, ranges between -1 and 1. A field panel scores 1 when it chooses the same proposals as the reference panel. It scores 0 by doing no better than it would by choosing at random. It scores below 0 if it perversely mainly accepts proposals it ought to reject or rejects proposals it ought to accept.

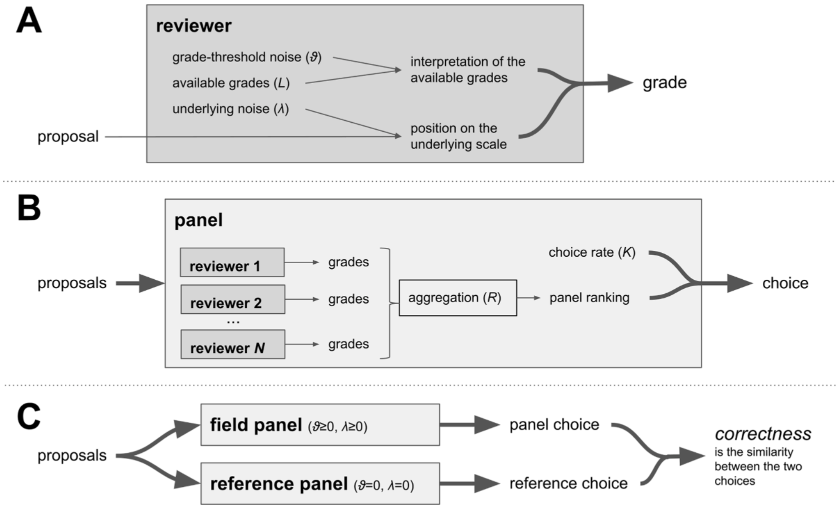

Fig. 3. Model overview. A reviewer grades one of the submitted proposals (A). A review panel collects all grades submitted by all reviewers for all proposals, aggregates them, ranks the proposals according to the aggregate grades and chooses some predetermined number of them from the top of its panel ranking (B). The correctness of the field panel's choice is the similarity of the set of proposals chosen by the field panel to the set chosen by the reference panel (C).

Table 2
Overview of model parameters and their explored values. Baseline settings are underlined.

Parameter	Values	Represents
N	2 through 13, baseline 5	number of reviewers on the field panel
K	5, 10, 20, 50	choice rate: the target percentage of proposals to be chosen
L	2, 3, 4, 5, 7, 10	number of grades available to reviewers for evaluating proposals
R	mean, hypermean, majority judgment, median, lowest score, highest score	rule for aggregating individual judgments to arrive at panel judgments
Reference distribution	high, low, bimodal	probability distribution of reference categorizations
λ	0, 0.1, 0.2, 0.4	underlying noise
ϑ	0, 0.05, 0.1	grade-threshold noise

connected to grading languages, and how they affect the correctness of choices by individuals and panels.

4.2. Grading languages: discrimination

Discrimination in grading is the capacity of a panelist or a panel to give higher grades to better proposals than to worse ones. The number of available grades is a limiting factor. Now we equip the simulated panels with different grading languages and observe the consequences for panel performance.

4.2.1. The number of grades

Fig. 6 shows variation in correctness of individual and panel choices with changes in *L*. As might be expected, individuals (the controls) and panels make better choices when using a finer-grained grading language. As with panel size, there are diminishing marginal returns in correctness as the number of grades increases.

4.2.2. Aggregation rules

Another factor affecting discrimination is the method by which the panel aggregates grades contributed by panelists. With some

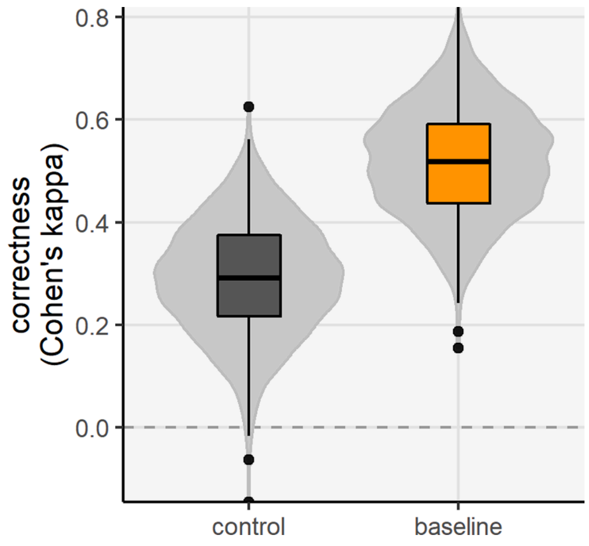

Fig. 4. Baseline configuration compared with control condition.

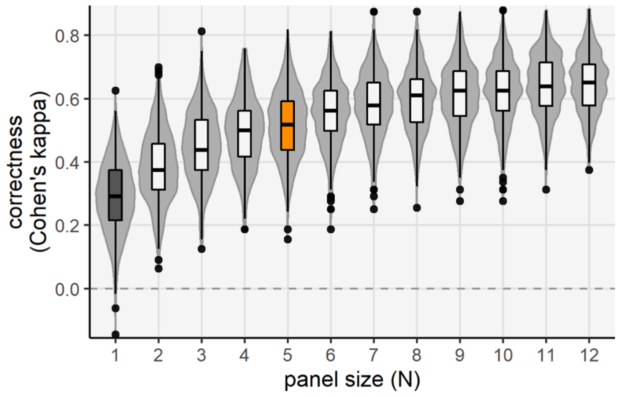

Fig. 5. Performance of panels of different sizes. Other than *N*, all parameters are set to baseline values.

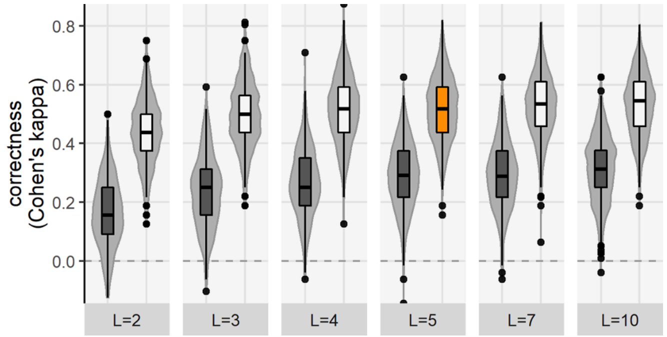

Fig. 6. Number *L* of available grades and correctness of panel choices. Other than *L*, model parameters are set to baseline values.

aggregation rules, such as the median rule, the panel makes the same number of distinctions as its panelists. Booster rules, such as averaging individually assigned numerical scores, push the panel's discrimination above theirs. In this subsection, we equip simulated panels with the aggregation rules defined in Table 1, and observe the consequences for the correctness of their choices.

Fig. 7 shows equal or higher correctness when panels use one of these booster rules rather than one of the non-boosters. We now single out certain rules for further discussion.

Aggregation by taking the mean is of special interest. This is common in practice and standard for SFI postal panels. Under baseline conditions it also results in a high level of correctness. Nevertheless, this might often not be the best rule. Sometimes it is outperformed by the hypermean, as seen in Fig. 7. Furthermore, it allows any individual reviewer by grading strategically to skew the whole panel's decision for or against any given proposal. Rules like the hypermean and majority judgment are more robust to bad faith manipulation (see Section 5.2).

Median aggregation stands out in Fig. 7 among the non-boosting rules. Correctness using this rule is on a par with that obtained using any of the boosters (it is slightly lower), and higher than with the other non-boosters we considered. We suggest a division of aggregation rules into those that are centrally tending (median, mean, hypermean, majority judgment) and those that are extremes seeking (lowest and highest score). Fig. 7 shows that all centrally tending rules we considered outperform the extremes seeking rules. In general, we hypothesize, a centrally tending booster rule is most conducive to correct choice.

Majority judgment might be considered a booster version of median aggregation. It is built on the median, but makes finer discriminations by using more of the information contained in panel members' inputs. Under the specific conditions of Fig. 7, though, majority judgment does not produce noticeably higher correctness than the median.

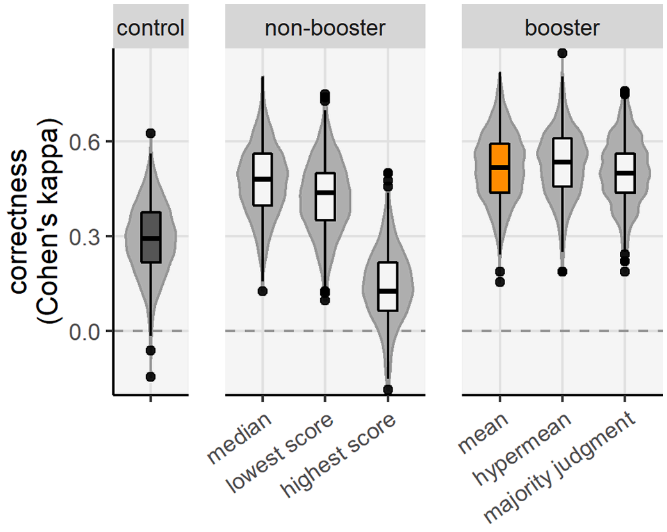

Fig. 7. Booster rules versus non-boosters. Model parameters other than the aggregation rule are set to baseline values.

4.3. Grading languages: different understandings

We turn now to consequences of differences in panelists' interpretations of grades for the correctness of panel choices. Fig. 8 shows the correctness of individual and panel choices at different levels of grade-threshold noise. This, recall, is random variation in panelists' interpretations around a reference interpretation.

Higher levels of grade-threshold noise might be expected to result in lower correctness of panel choices. Fig. 8 shows that this is hardly if at all the case: the differences in correctness among the different levels of ϑ are in the expected direction, but they are negligibly small.[11]

4.4. Underlying noise

Fig. 9 shows the correctness of individual and panel choices at different levels of underlying noise, λ. This, recall, is random variation in where panelists place proposals on their underlying scales. Like grade-threshold noise (ϑ), it is a source of inter-rater unreliability. High values of underlying noise represent in our model the presence of detrimental individual-level factors that are unrelated to how grades are understood: reviewers' lack of domain expertise, say, or their lack of care when studying proposals, or their inability to add up criterial scores consistently.

Fig. 9 shows that underlying noise (λ) has a strong effect on the correctness of choices by both individuals and panels. This contrasts strikingly with the results from Fig. 8, where grade-threshold noise (ϑ) is seen to be of little or no consequence. Figs. 8 and 9 together suggest that factors captured by underlying noise are much more critical to the correctness of a panel's choices than whether everyone is "on the same page" in the matter of grading standards.

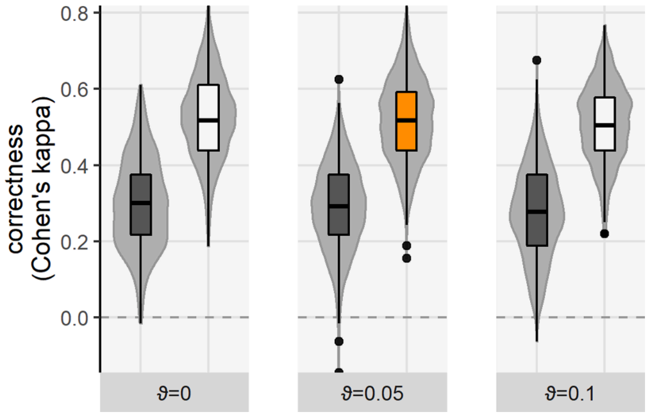

Fig. 8. Impact of grade-threshold noise, ϑ, on the correctness of a panel's choices. All parameters other than ϑ are set to baseline values.

[11] This result is anticipated in an earlier simulation experiment with a simpler model (Lyon & Morreau, 2018). An anonymous reviewer suggested that the result might be an artifact of allowing unrealistically little variation in thresholds. To test this, we repeated our experiment with much more variation, gotten by sampling thresholds from a uniform distribution. This did reduce correctness to some extent but not by much.

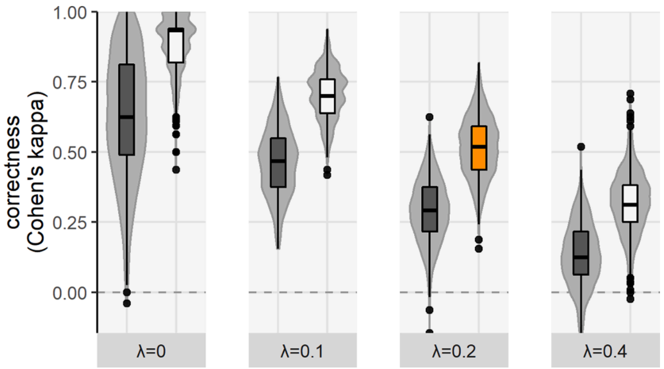

Fig. 9. Impact of underlying noise, λ, on the correctness of a panel's choices. All parameters other than λ are set to baseline values.

4.5. *Choice rate*

One might expect that the higher the targeted choice rate K, the less it matters in which order proposals are ranked, and the easier it is for a field panel to match the reference panel's choice.[12] Results from Fig. 10 confirm this. With higher choice rates, choices by individuals and panels are much more correct and also somewhat more reliable.

4.6. *Robustness: reference distributions*

The results reported in previous subsections are obtained with the "high" reference distribution of Fig. 1. Repeating the simulation experiments with instead the "low" and "bimodal" reference distributions, similar results are obtained: panels display a clear wisdom of crowds (compare 4.1); the use of discriminating grading languages and booster rules improves panel performance (4.2); any detrimental impact of inter-reviewer differences in grading thresholds is negligible (4.3); underlying noise due to random errors by reviewers greatly diminishes panel performance (4.4); and increasing the choice rate greatly improves it (4.5).

Changes in the reference distribution can result from changes in the interpretation or weighting of evaluation criteria, and from changes in properties of proposals that are relevant to criterial scores (see Section 3.2). That our main results hold for a range of reference distributions suggests that they do not depend on any special assumptions about the evaluation criteria, or about the distribution of relevant properties among submitted proposals.

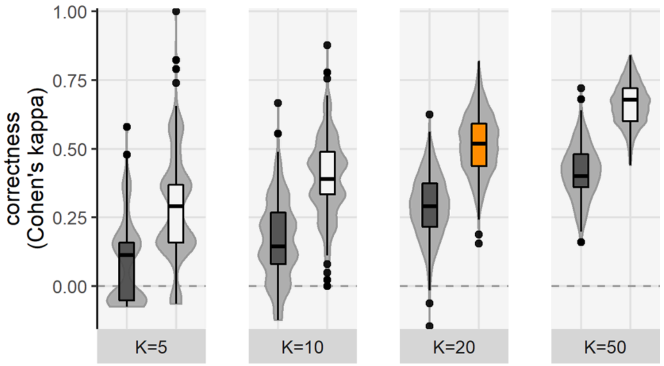

Fig. 10. Impact of choice rate K on the correctness of a panel's choices. All parameters other than K are set to baseline values.

5. Conclusions

The results in Section 4 point to several practical suggestions for the design of grant-review panels. We emphasize that these results have not yet been tested in trials with real panels. Section 5.1 states these suggestions and our expectations about their effectiveness. Section 5.2 discusses caveats and Section 5.3 considers their feasibility.

5.1. *Suggestions*

We begin with grading:

- *Use fine-grained evaluation scales*. Grading languages often have around five grades. Using a larger number of grades generally increases the correctness of a panel's choices, although with diminishing returns.[13]
- *Use "discrimination boosting" judgment aggregation*. Methods such as the mean rule, which give panels a greater capacity to distinguish better from worse than their members, generally give higher correctness than non-boosting rules, such as the median rule. Among the boosters we studied, the hypermean did best.
- *Allow diverse grading standards among panel members*. Our model suggests that reviewer training to calibrate grading thresholds has little impact on the correctness of a panel's choices.

These suggestions concerning grades can lead to inexpensive improvements in current practice. The model corroborates other and more familiar advice that might, however, be less actionable:

- *Increase the quality of reviews*. This can be achieved by recruiting more-capable reviewers, offering better training in recognizing the underlying merits of proposals, or allowing reviewers more time to study proposals before grading them.
- *Collect many reviews for each proposal*. This means recruiting more reviewers or allocating more proposals to each one. If the result is lower review quality this can easily be counterproductive.
- *Increase the target choice rate*. While funding agencies and panel designers might often not be able to do much about the funding rate, there are indirect ways for them to control the target choice rate. See Section 5.3.

5.2. *Caveats*

Like all models in science, our simulation simplifies reality. We turn now to certain simplifications that could affect some of our results and suggestions.

On our simulated panels, all panelists review all submitted proposals. On many real panels, on the other hand, review work is divided among subpanels and results are merged: the whole panel's grade for any given proposal is the grade assigned by the subpanel that reviewed it (perhaps with some adjustment during discussion in the whole panel). This creates scope for differences in grading standards among subpanels to depress correctness. High grades from subpanels with low grading standards can promote inferior proposals in the panel's ranking, displacing superior proposals from the top and resulting in incorrect choices. This cannot occur with the simulated panels of our experiments, simply because there are no results from different subpanels to merge. Thus our simulation experiments set aside a mechanism by which

[12] At the limit where all proposals can be chosen ($K = 100\%$), the rank order is irrelevant: choosing all the proposals is the only option and necessarily it is correct (Cohen's kappa = 1).

[13] Evidence suggests that finer-grained scales are effective if they include "anchoring points" at discrete intervals (G. Derrick, personal communication, 27 August 2021).

diverse grading standards could decrease the correctness of panel choices.

There is another such mechanism that we have set aside. The target of our simulation model is SFI's postal review, which is a purely aggregative design: proposals are graded and ranked, and a choice is made among them, without any deliberation among panelists. This matters because panelists in a deliberative phase would communicate among themselves about the relative merits of proposals, discussing preliminary grades they have assigned. If they happen to have different grading standards, some panelists being "tougher" graders than others, then spurious agreements and disagreements could arise. Resulting miscommunication may be expected to introduce noise and error, increasing the variability of panel choices and reducing their correctness.

A third simplification is that we have excluded individual biases from our simulation model.[14] Notice, though, that these need not reduce the correctness of a panel's choices. First, individual biases often cancel out when judgments are aggregated. This is a basic mechanism of the wisdom of crowds. Second, some biases such as overly "tough" grading can improve discrimination and thus have the potential to increase panel correctness, under circumstances that regularly arise in grant review.[15] Furthermore, some aggregation rules are not so vulnerable to individual biases: a single biased reviewer can shift the mean,[16] for instance, but not in general the median[17] score of any given proposal — especially with purely aggregative designs, where there is little opportunity to influence inputs from other panelists.

Each simplification indicates a direction for future research with practical consequences for the design, management or everyday running of grant-review panels. Extending the simulation model to include a division of review work among subpanels will make it possible to study realistic review networks and also to study the impact of differences in grading standards among (sub)disciplines. Realistic modeling of social deliberation among panelists will lead to a better understanding of how social structure and processes can affect the quality of panel review, both for better and for worse. Elaborating the simulation model to include individual biases will enable an investigation of the response of deliberative and aggregative decision procedures to the diverse preferences and values of reviewers, as well as their robustness to favoritism and prejudice.

5.3. *Feasibility*

We conclude with remarks on the feasibility of the suggestions of 5.1. Much depends on the burdens they impose on stakeholders.

The first suggestion is that panelists start using finer-grained grading languages. While this may be expected to increase the correctness of panel choices, it entails a heavier cognitive load on panelists. At some point, depending on levels of experience and domain expertise, this will become excessive.

We suggest also using discrimination-boosting decision procedures for moving from the individual inputs of panelists to rankings and choices made by whole panels. This we see as a promising intervention. With panel designs with an aggregative component, such as postal review, the judgment-aggregation rule can be changed practically for free.

We have suggested that training to coordinate grading standards has little impact on the correctness of a panel's decisions. This might be found very counterintuitive. If it is confirmed in behavioral experiments there will be obvious advantages for several stakeholders. Funding agencies will not have to set up and run calibration training sessions, and panelists will be spared the time and effort of attending them.

Among the interventions that we suggest, increasing the quality of panelists' reviews is predicted to be especially effective. Top experts are scarce and busy, and sometimes hard to identify (Callaham and Tercier, 2007); but there are other paths to improvement. One is to improve reviewer training. This will place extra demands on reviewers' time and attention. Training might have to be ongoing, in light of evidence that its benefits do not last long (Schroter et al., 2004). There are grounds for optimism, though, that reviewers will be willing to put in extra effort: evidence suggests that they appreciate the training they receive (Freda et al. 2009; Derrick and Samuel 2017).

Another path to improving reviews is to reduce workloads, allowing panelists more time to study and evaluate each proposal. Funding agencies could do this without recruiting more panelists. The number of proposals requiring review can be limited by tightening eligibility requirements for instance, or by introducing some form of "desk rejection" for submitted proposals.

Increasing the number of reviews for each proposal is an intervention that will, we expect, often not be feasible. Overworked panelists might balk at reviewing more proposals, and recruiting more panelists is logistically difficult and also expensive, if financial compensation is used as an incentive. Introducing "open peer review" (also known as "crowdsourced" or "intelligent crowd" review) might be proposed as a cost effective way to get more reviews. We expect it to be counterproductive, though. Reviews from large and self-selected crowds presumably will be of lower quality than those from small, handpicked panels of experts. Our simulations suggest that low review quality puts a very strong downward pressure on panel correctness.

Increasing the targeted choice rate we expect to be both effective and feasible. This could be done by increasing the funding rate, or proportion of submissions that are funded. While budgetary constraints usually set an upper limit to the number of funded proposals, both rates can be increased by reducing the number of submissions, for instance by changing eligibility requirements. The choice rate can also be increased independently of the funding rate. One way to do this is desk rejection, which reduces the proportion of all submissions passed to panels for review. Another is partial lotteries: a review panel is tasked with choosing a set of fundable proposals, from which some are selected at random for funding (Avin, 2019).

The consequences of the suggested interventions are different for different stakeholders. The organizational and administrative costs of recruiting more reviewers fall on funding agencies and review managers. Savings of various kinds are enjoyed by agencies and reviewers, should calibration training sometimes be unnecessary. Additional loads on reviewers include time and cognitive effort required for more training, studying proposals more closely, writing more reviews and using more finely-grained grading scales. The entire enterprise of research evaluation and funding as we know it depends on informal understandings between academics, university departments and funding agencies about what can be expected from whom, and for what in return. When it comes to innovation, maintaining this unwritten social contract is an important consideration.

[14] For an overview of different forms of bias in peer review, see Lee et al. (2013).

[15] Say there are some proposals to choose among, all of them excellent, but some very slightly better than others. Unbiased reviewers are bound to give all of them the same high grade. An overly "tough" reviewer on the other hand can give the slightly worse proposals lower grades than they really deserve, thus better communicating differences in merit.

[16] The same is true for Borda counting, another discrimination boosting aggregation method that generally produces high choice performance. It is named after Jean-Charles de Borda, a prominent figure in naval and scientific circles in the French Enlightenment. When the vulnerability of this method to manipulation was pointed out to him, Borda is said to have replied in indignation that "*mon scrutin n'est fait que pour d'honnêtes gens*" — "my voting method is only intended for honest people" (Mascart, 2000, p. 130).

[17] Sir Francis Galton noted similarly that the median or "middlemost" of many estimates is preferable because judgment aggregation by the mean gives "a voting power to 'cranks' in proportion to their crankiness" (Galton, 1907, p. 414). Such vulnerabilities are studied in the theory of social choice, in connection with strategic manipulation (Barberà, 2011).

CRediT authorship contribution statement

Thomas Feliciani: Conceptualization, Writing – original draft, Investigation, Methodology, Software, Visualization. **Michael Morreau:** Conceptualization, Writing – original draft, Investigation, Methodology. **Junwen Luo:** Conceptualization, Writing – review & editing, Investigation, Data curation. **Pablo Lucas:** Conceptualization, Software, Writing – review & editing. **Kalpana Shankar:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Science Foundation Ireland is both the case study in this article and the funder of the research project that contributed to this article. Other than inviting the pool of reviewers and providing access to the survey, Science Foundation Ireland had no involvement in the study, e.g. in its design, analyses, writing, or the decision to submit the article for publication.

Acknowledgments

We thank Rida Laraki, Aidan Lyon, Lai Ma, Rebekka Lisøy, Gerit Pfuhl, Gemma Derrick and three anonymous referees for their comments and suggestions. This material is based upon works supported by the Science Foundation Ireland under Grant No.17/SPR/5319.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.respol.2021.104467.

References

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., Durand-Zaleski, I., Alberti, C., 2012. Peer review of grant applications: criteria used and qualitative study of reviewer practices. PLoS One 7 (9), e46054. https://doi.org/10.1371/journal.pone.0046054.

Ahrweiler, P., Gilbert, N., Pyka, A., 2012. Final Report: Using Network Analysis to Monitor and Track Effect resulting from Changes in Policy Intervention and Instruments. SMART 2010/0025. European Commission, DG Information Society & Media.

Ahrweiler, P., Schilperoord, M., Pyka, A., Gilbert, N., 2015. Modelling research policy: ex-ante evaluation of complex policy instruments. J. Artif. Soc. Soc. Simul. 18 (4), 5. https://doi.org/10.18564/jasss.2927.

Avin, S., 2019. Mavericks and lotteries. Stud. Hist. Philos. Sci. Part A 76, 13–23. https://doi.org/10.1016/j.shpsa.2018.11.006.

Balinski, M.L., Laraki, R., 2010. Majority Judgment: Measuring, Ranking, and Electing. MIT Press.

Barberà, S., 2011. Strategyproof social choice. Handbook of Social Choice and Welfare. Elsevier, pp. 731–831. https://doi.org/10.1016/S0169-7218(10)00025-0. Vol. 2.

Bertens, L.C.M., Broekhuizen, B.D.L., Naaktgeboren, C.A., Rutten, F.H., Hoes, A.W., van Mourik, Y., Moons, K.G.M., Reitsma, J.B., 2013. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. PLoS Med. 10 (10), e1001531 https://doi.org/10.1371/journal.pmed.1001531.

Bornmann, L., Mutz, R., Daniel, H.D., 2010. A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. PLoS One 5 (12), e14331. https://doi.org/10.1371/journal.pone.0014331.

Budescu, D.V., Chen, E., 2015. Identifying expertise to extract the wisdom of crowds. Manag. Sci. 61 (2), 267–280. https://doi.org/10.1287/mnsc.2014.1909.

Callaham, M.L., Tercier, J., 2007. The relationship of previous training and experience of journal peer reviewers to subsequent review quality. PLoS Med. 4 (1), e40. https://doi.org/10.1371/journal.pmed.0040040.

Claveau, F., Veillette, A., 2020. Appraising the epistemic performance of social systems: the case of think tank evaluations. Episteme 1–19. https://doi.org/10.1017/epi.2020.16.

Derrick, G., Samuel, G., 2017. The future of societal impact assessment using peer review: pre-evaluation training, consensus building and inter-reviewer reliability. Palgrave Commun. 3 (1), 17040. https://doi.org/10.1057/palcomms.2017.40.

Esarey, J., 2017. Does peer review identify the best papers? A simulation study of editors, reviewers, and the scientific publication process. PS Political Sci. Politics 50 (04), 963–969. https://doi.org/10.1017/S1049096517001081.

European Commission, 2020. Evaluation Process and Results. European Commission. H2020 Online Manual. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/grants/from-evaluation-to-grant-signature/evaluation-of-proposals/eval_process_results_en.htm.

Feliciani, T., Luo, J., Ma, L., Lucas, P., Squazzoni, F., Marušić, A., Shankar, K., 2019. A scoping review of simulation models of peer review. Scientometrics 121 (1), 555–594. https://doi.org/10.1007/s11192-019-03205-w.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., Väänänen, K., 2012. Panel discussion does not improve reliability of peer review for medical research grant proposals. J. Clin. Epidemiol. 65 (1), 47–52. https://doi.org/10.1016/j.jclinepi.2011.05.001.

Freda, M.C., Kearney, M.H., Baggs, J.G., Broome, M.E., Dougherty, M., 2009. Peer reviewer training and editor support: results from an international survey of nursing peer reviewers. J. Prof. Nurs. 25 (2), 101–108. https://doi.org/10.1016/j.profnurs.2008.08.007.

Galton, F., 1907. One vote, one value. Nature 75 (1948). https://doi.org/10.1038/075414a0, 414–414.

Goodin, R.E., Spiekermann, K., 2018. An Epistemic Theory of Democracy. Oxford University Press.

Gurwitz, D., Milanesi, E., Koenig, T., 2014. Grant application review: the case of transparency. PLoS Biol. 12 (12), e1002010 https://doi.org/10.1371/journal.pbio.1002010.

Guthrie, S., Ghiga, I., Wooding, S., 2018. What do we know about grant peer review in the health sciences? F1000Research 6, 1335. https://doi.org/10.12688/f1000research.11917.2.

Hampton, J.A., 2001. The role of similarity in natural categorization. In: Hahn, U., Ramscar, M. (Eds.), Similarity and Categorization. Oxford University Press, pp. 13–28. https://doi.org/10.1093/acprof:oso/9780198506287.003.0002.

Hassan, S., Pavón, J., Antunes, L., Gilbert, N., 2010. Injecting data into agent-based simulation. In: Takadama, K., Cioffi-Revilla, C., Deffuant, G. (Eds.), Simulating Interacting Agents and Social Phenomena. Springer, Japan, pp. 177–191. https://doi.org/10.1007/978-4-431-99781-8_13.

Hayek, F.A., 1945. The use of knowledge in society. Am. Econ. Rev. 35 (4), 519–530.

Ho, E.H., Budescu, D.V., Dhami, M.K., Mandel, D.R., 2015. Improving the communication of uncertainty in climate science and intelligence analysis. Behav. Sci. Policy 1 (2), 43–55. https://doi.org/10.1353/bsp.2015.0015.

Jose, V.R.R., Winkler, R.L., 2008. Simple robust averages of forecasts: some empirical results. Int. J. Forecast 24 (1), 163–169. https://doi.org/10.1016/j.ijforecast.2007.06.001.

Langfeldt, L., 2001. The decision-making constraints and processes of grant peer review, and their effects on the review outcome. Soc. Stud. Sci. 31 (6), 820–841. https://doi.org/10.1177/030631201031006002.

Lee, C.J., Sugimoto, C.R., Zhang, G., Cronin, B., 2013. Bias in peer review. J. Am. Soc. Inf. Sci. Technol. 64 (1), 2–17. https://doi.org/10.1002/asi.22784.

Lyon, A., Morreau, M., 2018. The wisdom of collective grading and the effects of epistemic and semantic diversity. Theory and Decision 85 (1), 99–116. https://doi.org/10.1007/s11238-017-9643-7.

Mascart, J., 2000. La Vie Et Les Travaux Du Chevalier Jean-Charles de Borda: (1733-1799) ; épisodes de La Vie Scientifique Au XVIIIe Siècle. Presses de l'Univ. de Paris-Sorbonne.

Mastrandrea, M.D., Mach, K.J., Plattner, G.K., Edenhofer, O., Stocker, T.F., Field, C.B., Ebi, K.L., Matschoss, P.R., 2011. The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. Clim. Change 108 (4), 675–691. https://doi.org/10.1007/s10584-011-0178-6.

Maudlin, T., 2008. Grading, sorting, and the sorites. Midwest Stud. Philos. 32 (1), 141–168. https://doi.org/10.1111/j.1475-4975.2008.00170.x.

Morgan, M.G., 2014. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc. Natl. Acad. Sci. 111 (20), 7176–7184. https://doi.org/10.1073/pnas.1319946111.

Nicolai, A.T., Schmal, S., Schuster, C.L., 2015. Interrater reliability of the peer review process in management journals. In: Welpe, I.M., Wollersheim, J., Ringelhan, S., Osterloh, M. (Eds.), Incentives and Performance. Springer International Publishing, pp. 107–119. https://doi.org/10.1007/978-3-319-09785-5_7.

Obrecht, M., Tibelius, K., D'Aloisio, G, 2007. Examining the value added by committee discussion in the review of applications for research awards. Res. Eval. 16 (2), 70–91. https://doi.org/10.3152/095820207X223785.

Pier, E.L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M.J., Ford, C.E., Carnes, M., 2018. Low agreement among reviewers evaluating the same NIH grant applications. Proc. Natl. Acad. Sci. 115 (12), 2952–2957. https://doi.org/10.1073/pnas.1714379115.

Roebber, P.J., Schultz, D.M., 2011. Peer review, program officers and science funding. PLoS One 6 (4), e18680. https://doi.org/10.1371/journal.pone.0018680.

Sattler, D.N., McKnight, P.E., Naney, L., Mathis, R., 2015. Grant peer review: improving inter-rater reliability with training. PLoS One 10 (6), e0130450. https://doi.org/10.1371/journal.pone.0130450.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F., Smith, R., 2004. Effects of training on quality of peer review: randomised controlled trial. BMJ 328 (7441), 673. https://doi.org/10.1136/bmj.38023.700775.AE.

Squazzoni, F., Ahrweiler, P., Barros, T., Bianchi, F., Birukou, A., Blom, H.J.J., Bravo, G., Cowley, S., Dignum, V., Dondio, P., Grimaldo, F., Haire, L., Hoyt, J., Hurst, P., Lammey, R., MacCallum, C., Marušić, A., Mehmani, B., Murray, H., Willis, M., 2020. Unlock ways to share data on peer review. Nature 578 (7796), 512–514. https://doi.org/10.1038/d41586-020-00500-y.

Squazzoni, F., Gandelli, C., 2013. Opening the black-box of peer review: an agent-based model of scientist behaviour. J. Artif. Soc. Soc. Simul. 16 (2) https://doi.org/10.18564/jasss.2128.

Squazzoni, F., Takács, K., 2011. Social simulation that "peers into peer review. J. Artif. Soc. Soc. Simul. 14 (4) https://doi.org/10.18564/jasss.1821.

Thurner, S., Hanel, R., 2011. Peer-review in a world with rational scientists: toward selection of the average. Eur. Phys. J. B 84 (4), 707–711. https://doi.org/10.1140/epjb/e2011-20545-7.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., Forsyth, B., 1986. Measuring the vague meanings of probability terms. J. Exp. Psychol. Gen. 115 (4), 348–365.

Wardekker, J.A., van der Sluijs, J.P., Janssen, P.H.M., Kloprogge, P., Petersen, A.C., 2008. Uncertainty communication in environmental assessments: views from the Dutch science-policy interface. Environ. Sci. Policy 11 (7), 627–641. https://doi.org/10.1016/j.envsci.2008.05.005.