Research Evaluation, 2024, **33**, rvae025
https://doi.org/10.1093/reseval/rvae025
Article

OXFORD

Funding lotteries for research grant allocation: An extended taxonomy and evaluation of their fairness

Thomas Feliciani [1,2,*], Junwen Luo [3,4] and Kalpana Shankar [3,*]

[1]School of Sociology, University College Dublin, Belfield, Dublin 4, Dublin, Ireland
[2]Now with, Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4/b, 20156 Milan, Italy
[3]School of Information and Communication Studies, University College Dublin, Belfield, Dublin 4, Dublin, Ireland
[4]Boole Library, University College Cork, College Rd, Cork, Ireland

*Corresponding authors. Email: thomas.feliciani@polimi.it; kalpana.shankar@ucd.ie

Abstract

Some research funding organizations (*funders*) are experimenting with random allocation of funding (*funding lotteries*), whereby funding is awarded to a random subset of eligible applicants evaluated positively by review panels. There is no consensus on which allocation rule is fairer—traditional peer review or funding lotteries—partly because there exist different ways of implementing funding lotteries, and partly because different selection procedures satisfy different ideas of fairness (*desiderata*). Here we focus on two desiderata: that funding be allocated by 'merit' (epistemic correctness) versus following ethical considerations, for example without perpetuating biases (unbiased fairness) and without concentrating resources in the hands of a few (distributive fairness). We contribute to the debate first by differentiating among different existing lottery types in an extended taxonomy of selection procedures; and second, by evaluating (via Monte Carlo simulations) how these different selection procedures meet the different desiderata under different conditions. The extended taxonomy distinguishes "Types" of selection procedures by the role of randomness in guiding funding decisions, from null (traditional peer review), to minimal and extensive (various types of funding lotteries). Simulations show that low-randomness Types (e.g. 'tie-breaking' lotteries) do not differ meaningfully from traditional peer review in the way they prioritize epistemic correctness at the cost of lower unbiased and distributive fairness. Probably unbeknownst to funders, another common lottery Type (lotteries where some favorably-evaluated proposals bypass the lottery) displays marked variation in epistemic correctness and fairness depending on the specific bypass implementation. We discuss implications for funders who run funding lotteries or are considering doing so.

Keywords: funding lottery; randomization; research evaluation; science funding; social simulation.

1. Introduction

The allocation of grants by research funding bodies (hereafter: *funders*) relies extensively on peer review by experts who are expected to evaluate proposals for their scientific merit, potential for impact, and other factors, depending on the goals of the funding agency and the particular grant program. However, researchers and funders have expressed doubts about the effectiveness of peer review in the process of grant allocation, and trust in grant peer review is dwindling (Langfeldt, Reymert and Svartefoss 2023). Biased outcomes; the time and labor required of applicants, reviewers, and program officers; conservatism in outcomes; the lack of reliability of the peer review process; and poor correlation of funded research with desired impacts are some of the criticisms raised by stakeholders (Luukkonen 2012; Lee et al. 2013; Bol, de Vaan and van de Rijt 2018; Pier et al. 2018; Aczel, Szaszi and Holcombe 2021; Conix et al. 2021, among others).

In light of these critiques, some funders[1] have introduced an element of randomness into their selection procedures as an attempt to reduce some of the identified problems (Greenberg 1998; Fang and Casadevall 2016; Adam 2019; Piper 2019; Liu et al. 2020). In these selection procedures, some of the proposals recommended by the review panel are funded directly and some are put in a lottery pool whence grant winners are chosen at random following a predetermined protocol. Such selection procedures are generally called *funding lotteries*.

Funding lotteries are controversial: for instance, some question their capacity to address the flaws of traditional peer review (Barnett 2016; Avin 2019; Lee, Grant and Erosheva 2020; Reinhart and Schendzielorz 2020; Philipps 2021; De Peuter and Conix 2022). Epistemic and ethical considerations are also central in this debate. Some scholars put epistemic considerations first: they argue that a fair selection procedure is one that allocates funding to the 'best' – meaning most meritorious, most promising or most significant—project proposals (e.g. Reinhart and Schendzielorz 2020). These scholars prefer traditional peer review over funding lotteries. Other scholars put ethical considerations first. For instance, selection procedures are fair to the extent that they are unbiased—that is if, ceteris paribus, they give equal opportunity to male and female applicants; early career and senior applicants; applicants from renowned and less-renowned institutions; (mono)disciplinary and interdisciplinary research; novel research ideas and approaches and more traditional ones; and so forth (e.g. Gillies 2014; Fang and Casadevall 2016; Adam 2019; Zhuang and Acuna 2019; Osterloh and Frey 2020). Another ethical consideration is the even distribution of research funding. Selection procedures would be fair to the extent that they distribute funding evenly across researchers, labs, universities, or geographic regions, thus avoiding the accumulation of funding in the hands of a few (e.g. Gildenhuys 2020). Ethical considerations are typically brough forth as arguments in favor of funding lotteries.

Adding nuance to the discussion is the fact that there is considerable and largely unacknowledged variation in how funding lotteries can be implemented (Philipps 2022; Shaw 2022). At one end of the spectrum are funding lotteries where peer review panels are minimally (if at all) involved in the funding decision. At the opposite extreme are tie-breaking lotteries, where most funding decisions are made by traditional peer review panels—and funders draw lots only to decide among equally funding-worthy proposals that cannot all be funded. This variation in lottery implementation holds significance, as funding allocation might be more or less epistemically correct or fair depending on how exactly randomization is incorporated in the selection procedure. However, a critical gap in the literature is the lack of a systematic classification of selection procedures that distinguishes existing *types* of funding lotteries.

In this article we do not take a stance on the desirability of funding lotteries or on which considerations should be privileged in funding. Rather, we aim to contribute to an informed debate on funding lotteries in two ways: (1) by expanding an existing taxonomy of selection procedures for research funding to account for variation among existing funding lotteries; (2) by using Monte Carlo simulations to evaluate the fairness of various lottery types versus traditional peer review capturing both epistemic and ethical considerations.

The article is structured as follows. Section 2 reviews some prominent epistemic and ethical considerations on the fairness of funding lotteries that emerge from published opinions in the literature. Section 3 extends a taxonomy of selection procedures to differentiate among alternative types of funding lotteries. Then we focus on evaluating the epistemic correctness and fairness of funding allocation: Section 4 presents the formal and computational model of selection procedures from the extended taxonomy; Section 5 presents the results of numerical simulations. Our findings suggest possible ramifications for funders' decisions on whether—and how exactly—to implement a funding lottery: these are discussed in Section 6 along with the generalizability and limitations of our study.

2. Fairness in science funding allocation: epistemic and ethical considerations

Epistemic and ethical considerations feature prominently in the debate around funding lotteries—and they can lead to potentially conflicting expectations as to which are fair ways to allocate research funding. We provide an outline of these concepts in Table 1.

Epistemic considerations are about ensuring that funding is allocated to the 'best' proposals. Selecting the best proposals (and thus determining which is an epistemically correct distribution of funding across applicants) is not easy. To begin with, one has to define what is meant by 'a good proposal'. A proposal might be 'good' by virtue of having a sound research plan, the aim to fill an important gap in the literature; or a good potential for, for example, societal impact; several of these things together, or other things entirely. There is not one recipe, though usually there is a list of evaluation criteria and reviewer instructions issued by funders to guide panels in their assessment. Secondly, the determination of which proposals are the best is to be made 'ex-ante', meaning prior to the research being done. Review panels tasked with evaluating grant proposals ex-ante have limited information: they may evaluate the feasibility of a research project, judge the fit of the applicants and supporting institutions with the proposed research, and they may rely on applicants' prior research experience, publication record and other performance indicators as proxies to guess their future achievements. Ultimately, there is not one objective way for reviewers to select the best proposals; and the determination of which proposals are the best hinges on funders and review panels selecting some relevant indicators of what constitutes 'good'.

Because so many decisions come into play in a panel's funding recommendations, we remain agnostic as to how exactly reviewers and review panels select the 'best proposals': instead we define *epistemic correctness* as the panel's ability to direct funding to what *they* find to be the 'best' (and thus most funding-worthy) proposals. For some scholars, funding lotteries risk compromising the epistemic correctness of funding decisions by potentially allocating funding to proposals that are not the 'best proposals' in reviewers' eyes (see e.g. Bedessem 2020; Vindin 2020; Philipps 2021). Arguably, traditional peer review, despite being imperfect, is the most epistemically correct method we currently have to identify the 'best' science.

While epistemic considerations seem to be better handled by traditional peer review, ethical considerations are often cited as reasons for drawing lots in many domains (Broome 1984), including science funding (Woods and Wilsdon 2021). We distinguish between unbiased fairness and distributive fairness as two prominent examples of ethical considerations.

First of all, funding decisions are *unbiasedly fair* if they are blind to characteristics of applicants or proposals that ought not matter in the evaluation and thus bias, or skew, panel recommendations. In peer review, evaluations can be biased by various characteristics of applicants or proposals (Lee et al. 2013), but here we only use gender as an illustrative example. Applicants' gender should not matter for funding decisions; and yet, ceteris paribus and with few reported exceptions (i.e. Mutz, Bornmann and Daniel 2012; Yip et al. 2020), funding rates and grant sizes are higher for male applicants than for female applicants (Bornmann, Mutz, and Daniel 2007; McAllister, Juillerat and Hunter 2016; Burns et al. 2019; Witteman et al. 2019). It is argued that randomization can help prevent bad reasons (such as bias on demographic characteristics) from informing funding decisions (Stone 2009). Because lotteries are blind to these characteristics of applicants and proposals, they should correct—or help minimize—the influence of bias on funding decisions. Therefore, proponents of unbiased fairness in funding decisions

Table 1. Summary of the alternative notions of fairness as emerging from the literature.

Notion of fairness	Normative definition	Chosen selection procedure
Epistemic correctness	Funding should go to the 'best' proposals.	Traditional peer review
Ethical consideration: unbiased fairness	Funding should be distributed without gender, racial, geographic, or other biases.	Funding lotteries
Ethical consideration: distributive fairness	Funding should be distributed evenly.	Funding lotteries

generally expect funding lotteries to be fairer than traditional peer review (Gillies 2014; Adam 2019; Gildenhuys 2020; Osterloh and Frey 2020; Horbach, Tijdink and Bouter 2022).

The other ethical consideration examined here is *distributive fairness*: the idea that funding should be evenly distributed across equally eligible applicants. Distributive fairness can be motivated by the fact that it can be difficult to tell which of two projects is more deserving (Bornmann, Leydesdorff and Van den Besselaar 2010; van den Besselaar, Sandström and Schiffbaenker 2018). In these cases an even distribution of resources may feel 'fairer' than an arbitrary decision to fund one project and not the other. Another reason to pursue distribute fairness is that scholars who have won funding in the past are more likely to receive funding moving forward (Bol, de Vaan and van de Rijt 2018; Zhang et al. 2020). This is because reviewer assessments skew in favor of previously successful applicants, causing a progressive, self-perpetuating concentration of funding in the hands of a few researchers, labs, institutions which have already won funding in the past or are well-networked. Pursuing distributive fairness means trying to prevent the concentration of funding in the hands of a few. Distributive fairness is presumed to favor funding lotteries because, over the long run, lotteries will tend to distribute funding more evenly across researchers regardless of their past grants, networks, affiliations or location (Martin 2000; Gildenhuys 2020).[2]

In principle, epistemic correctness on the one hand, and unbiased and distributive fairness on the other, need not be at odds, in the sense that funding can theoretically be distributed in a way that simultaneously meets all desiderata. However, reviewer evaluations tend to skew in favor of some groups of applicants—for example male researchers, or researchers who have already won funding in the past (Lee et al. 2013; Bol, de Vaan and van de Rijt 2018). These and other biases can give rise to conflicts between epistemic correctness and ethical considerations such that pursuing one can mean sacrificing the others. For a panel with strong gender bias, for example, an epistemically correct allocation will favor male applicants because applications by male applicants are found to be the best; such allocation would of course lack unbiased fairness and be ethically questionable. Likewise, a panel with a strong bias in favor of past grant winners will evaluate more positively proposals from past grant winners: thus, a recommendation to award funding to those past winners will be epistemically correct, but distributively unfair. This makes it particularly important to tease out these various dimensions of fairness and understand how different selection procedures navigate epistemic and ethical considerations under various levels of reviewer biases and various other conditions.

Ultimately, the decision to prioritize some considerations over others is strategic. Usually, this decision around priorities aligns with the mandates of funders and the specific objectives of the funding calls. Ideally, priorities would also be agreed upon by the scholarly community and, for public funders, the taxpayers. But our focus lies downstream of this decision: in light of the established priorities, what selection procedures should be preferred, under what conditions, and at what costs?

3. Extended taxonomy: five types of selection procedures

The contemporary debate on funding lotteries started with Greenberg (1998). Greenberg noticed that peer reviewers rate most applicants as meritorious and there is often little distinction between proposals that fall just above versus just below the funding line. This intuition has gained empirical support: comparing peer review reports and citation metrics of grant application just above and below the funding line reveals that proposals around the funding line are often indistinguishably good (Bornmann, Leydesdorff and Van den Besselaar 2010; van den Besselaar, Sandström and Schiffbaenker 2018). Indistinguishability matters: the fact of being above or below the funding line, however aleatory, is shown to have important ramifications for science and science careers (Bol, de Vaan and van de Rijt 2018).

For Greenberg indistinguishability around the funding line indicated that we should consider distributing research funding at random among all qualified applicants and do away with expensive peer review panels (Greenberg 1998).[3] Greenberg proposed what today we call a 'non-partial', or *full* lottery: 'full' because grant winners are chosen by lot, without any involvement of peer reviewers. To our knowledge no funder has ever run a full lottery.

Others after Greenberg proposed instead that randomization should support funding decisions only where distinctions must be made among indistinguishably good proposals (Adam 2019; Osterloh and Frey 2020; Heyard et al. 2022). So, for example, if indistinguishability mainly affects proposals around the funding line (e.g. when the 'exceptionally good' and the 'exceptionally bad' proposals are easier to tell apart from the rest), then only decisions about proposals in the gray middle around the funding line should be made by drawing lots. This solution is called a *partial lottery*.

In essence, partial lotteries are selection procedures that combine recommendations by a peer review panel and randomization to make decisions among proposals around the funding line that are indistinguishably funding-worthy. Partial lotteries have historically been used as sortition mechanisms in some voting systems[4]; and they are the kind of funding lotteries adopted by some funders in recent years.

The distinction between full and partial lotteries is one of the three criteria on which Shaw constructed a preliminary taxonomy of funding lotteries (Shaw 2022):

1) Partiality—or whether a traditional peer review panel is involved in setting up the lottery pool: partial lotteries require a peer review panel that ranks proposals and a cut-off rule to determine which portion of the ranking constitutes the lottery pool. Conversely, no peer review panel is involved in full lotteries.
2) Weighting—or whether the probability of winning is uniform in the lottery pool: uniform in non-weighted lotteries; not uniform in weighted lotteries.
3) Tiering—that is, the granularity of the ranking of proposals. A ranking with minimal granularity is one that only has two valid ranking positions (e.g. eligible proposals at the top and ineligible proposals at the bottom); a ranking with maximal granularity is one where there are as many ranking positions as there are proposals, virtually without any ties (i.e. a strict ordering of proposals, from 'best' to 'worst').

To the best of our knowledge, all funding lotteries that were ever implemented are partial, non-weighted lotteries. Therefore, our first goal is to extend Shaw's taxonomy to account for the existing range of variation within this class of

Table 2. Taxonomy of selection procedures populated with empirical examples.

Type	Description	Is a lottery	Has focal randomization	Adopted in some funding schemes[a]
0	Traditional peer review	No	No	Most research funding organizations
1	Tie-breaking partial lottery	Yes	Yes	• Swiss National Science Foundation; • Natural Environment Research Council (UK); • Research Council of Norway; • Science Foundation Ireland.
2	Partial lottery with bypass	Yes	Yes	• Volkswagen Foundation (Germany); • Austrian Science Fund; • Novo Nordisk Foundation (Denmark).
3	Partial lottery of fundable proposals	Yes	No	• Health Research Council of New Zealand; • Science for Technological Innovation (New Zealand); • New Frontiers in Research Fund (Canada); • The British Academy.
4	Full lottery	Yes	No	N/A

[a]The list of funders in Table 2 only includes those funders for which we could determine with certainty the position in the taxonomy by examining publicly available call documents. Therefore, Table 2 excludes the Foundational Questions Institute (U.S.), for which insufficient information is publicly available, and several (unnamed) funders who are currently considering setting up a lottery but who, at the time of writing, have not yet to published call documents that describe in detail how the lottery is to be implemented.

funding lotteries. We do so by making finer distinctions along the partiality criterion. We note that the criteria of weighting and tiering are orthogonal to the criterion of partiality—in other words, weighting and tiering need to be specified for all selection procedures, including partial and full lotteries. To stay within scope we do not thoroughly examine these two variations here; instead, we compare selection procedures assuming the same weighting and tiering conditions.[5] We come back to this point in our conclusion and discussion (Section 6).

In the extended taxonomy we classify funding lotteries by the role of randomization in the funding decisions, from no-randomness (Type 0—traditional peer review) to partial lotteries (Types 1–3), to complete randomness (Type 4 full lotteries).[6] This extension allows the taxonomy to include non-lotteries (Type 0) and to distinguish between various existing implementations of partial lotteries sitting at different levels of 'partiality' (Types 1–3). Table 2 provides an overview along with pointers to funders that adopt or have adopted one or the other selection procedures in some of their calls.

Peer reviewers and review panels make their assessments of the submitted proposals following some protocol—for example evaluating each applicant or proposal against some predefined criteria. The taxonomy of selection procedures is agnostic as to which criteria are adopted (e.g. whether citation metrics or other bibliometric indicators are used). What matters in the taxonomy is that the review panel is able, at the end of the assessment, to construct a ranking of all eligible proposals. This ranking is more properly called a weak ordering, which means that it may have ties: two or more proposals may be found to be equally worthy of funding. Some Types of selection procedures are characterized by the way these ties in the panel ranking are treated.

Type 0 denotes a traditional selection procedure, where funding is allocated following the recommendation by peer review panels. Proposals at the top of the ranking are recommended for funding directly; those at the bottom are rejected. Crucially, there may be ties around the funding line: that is, the panel deems equally funding-worthy two or more proposals that cannot all be funded. In Type 0 selection procedures, these ties need to be arbitrarily broken, generally by invoking a further discussion by the review panel.

By contrast, Type 1 'tie-breaking' lotteries resolve ties that need breaking by choosing at random from the tied proposals the one(s) that will be funded. Type 1 are *partial lotteries* with *focal randomization*. The label 'partial' signifies some form of involvement of a peer review panel in the funding decision; and 'focal randomization' signifies that only some of the funding decisions are made by lot (the remaining funding decisions are made directly by the review panel). The concept of 'focal randomization' is owed to Brezis (2007), who describes partial lotteries where the review panel divides the submitted proposals into three sets. The first set comprises the most promising proposals to be directly recommended for funding—we call this the 'bypass set' because it comprises proposals to be directly funded bypassing the lottery. The second is the focal randomization set (or 'lottery pool'). The lottery pool comprises all funding-worthy proposals that are somewhat 'less good' than the bypass set. The third set are proposals that are not recommended for funding. In selection procedures with focal randomization, the lottery pool is a device that allows review panels to recommend some proposals directly for funding, and some other proposals to be eligible for entering a funding lottery.

Type 2 selection procedures are a different flavor of partial lotteries with focal randomization. While in Type 1 the review panel *must* come up with a bypass set and *may* create a lottery pool to resolve ties, in Type 2 the panel *must* set up a lottery pool.[7] The lottery pool typically consists of all eligible proposals deemed to be funding-worthy. Partial lotteries with bypass (Type 2) aim to ensure that excellent proposals are not being missed to bad luck (thanks to the bypass), and also promote a chance for those transformative, innovative or riskier research projects that may be disliked by some reviewers (thanks to the lottery).

Note that Type 2 can be further distinguished by the size of the lottery pool. For example, a funder using a Type 2 procedure may mandate that a number of funding decisions (e.g. 10% of them) is to be determined by lottery (and 90% chosen directly via the bypass): in this case, we can say that the Type 2 lottery is running with a 10% *lottery choice rate*. A different funder may mandate a 90% lottery choice rate, meaning that 90% of funding awards are chosen from the

lottery pool, and only 10% via the bypass. Such lottery choice rates will become important in our simulation experiment and recommendations for funders.

Type 3 selection procedures are partial lotteries without focal randomization: the lottery pool comprises all eligible proposals deemed funding-worthy by review panels; and there is no bypass set. This is equivalent to a Type 2 procedure with a 100% lottery choice rate.

Lastly, Type 4 selection procedures are Greenberg-like *full lotteries* where funding decisions are guided solely by the lottery and no peer review is involved. In a Type 4 lottery all eligible proposals enter the lottery pool and winners are selected at random uniform. Although to our knowledge no research funding organization has ever adopted a Type 4 selection procedure, we include Type 4 in our study because it is discussed in the literature and has some advocates (see e.g. Roumbanis 2019; Shaw 2023).

These different Types of selection procedures can in theory be combined to create novel hybrids. For example, in a recent exploration of alternative selection procedures, Luebber et al. (2023) propose one consisting of two steps. The first step is a Type 4 full lottery, the winners of which move on to the second step in the selection. The second step consists of a traditional peer review panel ranking the proposals (Type 0). While other hybrids are theoretically possible, all existing implementations of funding lottery we could find conform to the fundamental Types outlined here.

4. Simulation model and experiment

Our study follows a growing body of research that studies selection procedures in grant funding (and beyond) using computer simulation experiments (e.g. Brezis 2007; Zollman 2009; Avin 2015; Gross and Bergstrom 2019; Sakamoto 2023). Formal and computational models of science evaluation are growing in use, mainly because they allow to arbitrarily manipulate conditions and refine interventions 'in silico' in ways that ethics, costs, data availability or practicality prohibit to directly do '*in vivo*' (Feliciani et al. 2019).

Here, simulation experiments allow us to observe how different selection procedures would allocate grants over the same sets of candidate proposals; and to gauge the degree to which each funding allocation is epistemically correct or unbiasedly/distributively fair. Each run of the simulation model simulates the selection process for a hypothetical funding call, from the submission of proposals through the funding decisions. We repeat independent simulation runs multiple times, systematically varying all relevant conditions including the Type of selection procedure adopted. At the end of each simulation run we measure the degree of epistemic correctness, unbiased and distributive fairness of the funding decisions.

In the following subsections we define the components of a simulation run in the order in which they become relevant during the execution of the simulation. All model parameters and notations are summarized in Section 4.4. Reproducible code for R 4.3.2 (R Core Team 2023) is available in a public repository at https://github.com/thomasfeliciani/mixed-lotteries-for-science-funding/.

4.1. Proposal evaluation and standard of correctness

Each simulation run starts with the creation of a review panel. The task of a review panel, in essence, is to examine a set of competing proposals in the light of various evaluation criteria, and then to synthesize the opinions of different panel members on the different proposals and criteria into a one-dimensional ranking of proposals from the 'best' (shorthand for the most funding-worthy) to the 'worst'. It is important here to define what is the standard of correctness against which to measure how well panels did their evaluations—how epistemically correct they were.

Some of the empirical (Lindsey 1988; Lee et al. 2013) and most of the theoretical/modeling (Feliciani et al. 2019) literature on peer review assumes that submissions (in our case, funding proposals) have some intrinsic, objective value, sometimes referred to as 'true quality'. This literature takes submissions' true quality as the standard of correctness that reviewers try to adhere to in their assessments. This standard assumption is challenged in more recent work (e.g. Feliciani et al. 2022): first, because it makes epistemic correctness difficult to measure (as we cannot observe true quality directly); and second, because the very notion that research quality can be assessed objectively is questionable.

Therefore, we assume instead that reviewers and review panels have *subjective standards of correctness*—that is, ideal evaluations that might differ between reviewers or between review panels. The subjective ideal evaluation can be thought of as the evaluation each reviewer (or panel) would make under ideal conditions, for example with infinite knowledge, time and resources. This is a more realistic assumption and makes correctness somewhat easier to measurable empirically.

To implement these assumptions we model proposals in the simulation as 'entities' with three attributes: their *reference evaluation*, the actual panel *evaluation*, and a *bias factor*. Assume there exists a review panel that can work under ideal conditions and can rank all proposals from best to worst with perfect accuracy. This is our *reference panel*; its evaluation is our standard of correctness and we call it the *reference evaluation*. To the extent that different reviewers and panels will evaluate merit differently, typical panels will each have their own 'reference' version—this is what makes the notion of reference evaluation a *subjective* standard of correctness (Feliciani et al. 2022).

Compared to the reference panel, *evaluations* of a 'typical' peer review panel are: (1) prone to errors for example due to lack of knowledge, time or other resources; (2) biased—that is, evaluations are skewed by some attributes that ought not matter, for example the gender of the applicants (the *bias factor*); (3) coarse, meaning that sometimes two or more proposals are indistinguishably good, and so are assigned the very same grade or ranking position. More formally, we assume that each review panel evaluates N funding proposals, and that the evaluation x of a proposal i is:

$$x_i = \hat{x}_i + \varepsilon_i + (b_i \times \lambda)$$

Where $\hat{x}_i$ is the reference evaluation of i, ε_i is random error in the evaluation of i, and $(b_i \times \lambda)$ is a quantity that represents the effect of bias on the evaluation. Reference evaluations $\hat{x}$ range in [0,1], where higher values represent a more positive evaluation. They follow a beta distribution with parameters α and β. Generally we assume this distribution to be bell-shaped ($\alpha > 1$ and $\beta > 1$), and we allow for it to be skewed (when $\alpha \neq \beta$). From the literature and from previous research we know that reviewer evaluations tend in fact to be negatively skewed—meaning that most proposals are found to be

good, and few proposals are found to be bad (Feliciani et al. 2022). To reflect this we will assume that $\alpha > \beta > 1$.

The random error ε is assumed to be normally distributed ($\mu = 0$; $\sigma = \sigma_\varepsilon$). Its standard deviation σ_ε is a key model parameter representing the size of the errors made by the typical review panel compared to their reference panel.

The effect of bias is the product of two terms: b_i and λ. The former represents some attribute (or the combination of multiple attributes) of a proposal i that may bias the panel evaluation, though it should not. This attribute b_i is modeled as a real number drawn from a uniform distribution in [0,1]. Whether and to what degree this attribute influences the panel evaluation depends on the evaluation bias (λ): this is a model parameter also ranging in [0,1]. By setting $\lambda = 0$ we model a review panel whose judgment is not at all affected by the potential bias towards the attribute b. Conversely, higher λ models biased panels whose evaluation will be unduly steered by proposals' attribute b.

The resulting value of x_i needs to be truncated to ensure its range is in [0,1]. We do so as follows:

$$x_i = \begin{cases} 0, & if\ x_i < 0 \\ x_i, & if\ 0 \le x_i \le 1 \\ 1, & if\ x_i > 1 \end{cases}$$

Lastly, the typical panel evaluation is converted into a grade on an appropriate grading scale. The preliminary taxonomy of funding lotteries by Shaw (2022) highlights the granularity of the evaluation, or tiering, as an important degree of freedom in designing a funding lottery. Generally, coarse evaluations imply proposal rankings with fewer and larger equivalence classes: this means that many proposals will receive the same grade resulting in many ranking ties. Therefore, we introduce G as an additional model parameter that captures the granularity of the evaluation by the typical peer review panel.

The value of parameter G can be any natural number larger than 1 and can be thought of as the number of grades available for a peer review panel to choose from. Peer review panels usually aggregate the grades assigned by their individual reviewers, for example by averaging them. Therefore, G depends on how many reviewers there are on the panel (denoted R), and on the granularity of the evaluation scale used by the individual reviewers (denoted γ). G then is:

$$G = \lambda \times R - (R - 1)$$

So, for example, if there are $R = 2$ reviewers on the typical peer review panel who issue a pass/fail grade (thus, $\lambda = 2$), the panel as a whole will have $G = 2 \times 2\text{-}(2\text{-}1) = 3$ available grades: the top grade if both reviewers choose 'pass'; the middle grade if one reviewers chooses 'pass' and the other 'fail'; and the bottom grade if both choose 'fail'.

Therefore, let us denote g_i the final grade assigned by the typical peer review panel to a proposal i. To determine g_i we first discretize the interval of x_i (which is [0,1]) into G evenly-sized partitions. Numbering the panel grades from 1 (worst available grade) to G (best available grade), we can define the chosen grade g_i as the index number of the partition corresponding to x_i. Following the example where $G = 3$, suppose that $xi = 0.6$. The interval [0,1] can be partitioned into: [0,1/3], (1/3, 2/3], (2/3, 1]. The evaluation of 0.6 falls within the second interval: thus, on our scale from 1 to $G = 3$, the final grade is $g_i = 2$.

In the end, the evaluations by the typical peer review panel are used to rank proposals from worst-rated to best-rated.[8] Generally, there will be more ranking positions and fewer ties when G is high. This ranking constitutes the panel recommendations and is used as input for a Type 0–4 selection procedure, which in turn determines the final funding decisions.

4.2. Types of selection procedures and funding decisions

In our model, funding decisions are binary decisions ('fund'/'not fund') made by a review panel on the basis of the ranking of proposals. We assume that generally, the review panel is tasked with identifying T proposals for funding: thus, there will be N funding decisions to be made, T of which will be 'fund', and N—T will be 'not fund'. To increase model realism, we also assume that in selection procedures where a review panel is involved (i.e. Types 0–3), proposals must receive from the panel a sufficiently positive grade in order for them to be considered for funding. How high is the bar for sufficiency is determined by a model parameter denoted θ with range [0,1]. Formally, proposal i can only be considered for funding if $g_i \ge \theta \times (G\text{-}1)$.[9]

How exactly the ranking is converted into funding decisions varies depending on the Type of selection procedure being run.

In Type 0 selection procedures—traditional peer review—the panel recommends for funding the T proposals at the top of the ranking that received a sufficiently positive evaluation ($g_i \ge \theta \times (G - 1)$). It can happen that two or more proposals tie for T^{th} position because they receive the same final grade from the panel. The tie is broken by re-evaluating the tied proposals, following the same equation as above:

$$x_i = \hat{x}_i + \varepsilon_i + (b_i \times \lambda)$$

Just like before, the new evaluation approximates the reference evaluation $\hat{x}$, plus the random error ε and the effect of bias ($b_i \times \lambda$). Of the tied proposals, the panel recommends funding the one(s) with the highest evaluation x.

Type 1 selection procedures—tie-breaking partial lotteries—work similarly to Type 0, in that they also recommend T proposals moving from the top of the ranking downwards, as long as $g_i \ge \theta \times (G - 1)$. If there are no ties for the T^{th} position, a panel following a Type 1 procedure will choose the same proposals as a panel following Type 0. The difference between Types 1 and 0 comes about when there are ties for T^{th} position that need to be broken. Instead of re-evaluating the tied proposals (as in Type 0), Type 1 panels break the panel at random uniform.

Type 2 selection procedures—partial lotteries with bypass –, unlike Types 0–1, always have a lottery element, even if there are no ties in the ranking. More precisely, Type 2 panels choose (1) a *bypass set* of proposals from the top of the ranking (these are funded directly); and then (2) a *lottery pool* from which some of the winners are drawn. Let S be the number of proposals that must be chosen by lottery. Then, the bypass set will consist of the T-S proposals chosen from the top of the ranking.[10] The remaining S proposals to be funded will be randomly drawn from the lottery pool with uniform probability. The lottery pool consists of all proposals that have

received a sufficiently high grade and that are not in the bypass set.

In Type 3 selection procedures—partial lotteries of fundable proposals—all proposals with a sufficiently high grade enter the lottery pool, and T grant winners are drawn at random uniform.

In Type 4 selection procedures—full lotteries—the lottery pool consists of all proposals. All T funding winners are drawn at random from the lottery pool.

4.3. Operationalization of epistemic correctness, unbiased fairness and distributive fairness

Epistemic correctness is operationalized as the rank correlation (Spearman's ρ) between funding recommendations of typical peer review panels and their reference evaluation. A strong positive correlation indicates that funding recommendations align with the subjective evaluation that the review panel could make under ideal conditions—in other words, funding is being allocated to those proposals that the panel should identify as the best, or most funding-worthy.

Unbiased fairness is the degree to which a selection procedure can compensate for the biased evaluations by the typical review panels. To operationalize unbiased fairness we take the rank correlation (Spearman's ρ) between funding recommendations and proposals' *bias factor*. The bias factor, recall, is a real number that defines, for each proposal, the size of the skew in panel evaluations. It is meant to represent the (combined) effect of proposal attributes that should not matter in the evaluation, but which do—for example applicants' gender, ethnicity, career stage, etc. Technically, we define unbiased fairness as 1 - $|\rho|$. Panels whose decisions are strongly influenced by these biases will tend to rank proposals in a way that matches or resembles the ordering of proposals by bias factors (e.g. senior male applicants at the top, and early-career female applicants at the bottom). The Spearman ρ between funding recommendations and proposals' bias factor will therefore be high, and unbiased fairness low. Conversely, unbiased fairness will score high when there is no correlation between the irrelevant attributes and the funding decisions.

To measure distributive fairness we simulate Type 0–4 selection procedures multiple times over the same set of proposals. We keep a tally of how many times each proposal has won funding. Distributive fairness is operationalized as 1—the Gini coefficient calculated over the tallies.[11] As before, a high score indicates a distributively-fairer decision where, over the long run, funding is being distributed evenly across all proposals/applicants. Conversely, a low score indicates that, over the long run, most of the funding goes to a few proposals/applicants, and most others receive little funding or no funding at all.

4.4. Outline of the simulation experiment

We conduct our experiment by simulating a number of selection procedures and systematically varying the conditions under which the selection takes place. Even with an abstract model of selection such as ours there are many conditions—degrees of freedom—to consider and explore. This is an overview of the main ones:

- *Evaluation bias* (denoted λ). Correlation between the evaluation by the typical panel and the bias factor—that is the proposal attribute, like for example gender, that should not correlate with the evaluation. A high value λ models a scenario where reviewer evaluations are strongly biased. This represents the bias of reviewers that an unbiasedly fair selection procedure should mitigate. An unbiasedly fair selection procedure will compensate for this bias by distributing funding evenly across proposals that were favorably and unfavorably discriminated by evaluation bias.
- *Size of the review panel* (denoted R). Number of reviewers in the typical panel. This parameter, too, influences the number of ties in the panel ranking. Ceteris paribus, a large review panel as a whole is able to make finer distinctions (and thus have fewer ties in the ranking) than a small panel.
- *Granularity of the evaluation scale* (denoted γ). Number of grades available for reviewers to choose from. Generally, more granularity helps reduce the number of ties in the panel ranking, because reviewers can make finer distinctions among proposals.
- *Funding rate* (denoted T). Indicates what percentage of all submissions is to be recommended for funding. For example, a high funding rate (e.g. 75%) characterizes a very generous grant scheme with relatively low competition, where three in four submissions are awarded funding.
- *Sufficient merit* (denoted θ). Funders may elect never to fund proposals that were rated poorly. Proposals that do not receive a sufficiently high grade cannot be chosen by the review panel, cannot enter the lottery pool and generally cannot be recommended for funding. The parameter 'sufficient merit' determines what is the minimum sufficient grade a submission must be given by the typical panel for it to be considered 'fundable'. Higher sufficient merit implies stricter quality requirements.
- *Lottery choice rate* (for Type 2 partial lotteries with bypass; denoted S). This parameter is inversely proportional to the size of the bypass and is more technically defined as the percentage of proposals that a Type 2 panel will choose by lottery. The remainder will be directly chosen by the panel.

We show below a summary table of model parameters along with the values we have explored in the simulation experiment (Table 3). We have run 100 independent repetitions for each unique combination of parameters.[12] Each repetition was executed as follows. First, we generated a new set of proposals and reference evaluations; then, we repeated each Type of selection procedure 100 times. Our outcome measures (epistemic correctness, unbiased and distributive fairness) are measured separately each time, for each Type and for each repetition.

5. Results

Figure 1 compares Types of selection procedures under a parameter configuration that represents typical, non-extreme conditions. This configuration is characterized by, among others: a little evaluation bias ($\lambda = 0.2$); a review panel consisting of 5 panel members ($R = 5$), a grading scale with 5 available grades ($\gamma = 5$, e.g. as in a five-point Likert scale where 1 = 'bad'; 5 = 'excellent'); a generous funding rate ($T = 50\%$, meaning that half of all proposals will be funded); and a fairly low threshold for sufficient merit ($\theta = 0.5$) of the submitted proposals.

In Figure 1, the different Types of selection procedures are shown on the Y-axis and are ordered by role of randomness, from Type 0 to Type 4. This includes the different flavors of Type 2 partial lotteries based on the lottery choice rate adopted ($S \in \{10, 50, 90\}\%$). The plot is divided into three panels, each dedicated to one of the three measures of fairness (X-axis). Violins (gray) and boxplots[13] (colored) show distributions across simulation runs.

At a glance, the effect of randomness on fairness is monotonic: increasing the role of randomness (i.e. selecting a higher-numbered Type) generally degrades epistemic correctness while improving unbiased fairness and distributive fairness. As can be expected, traditional peer review (Type 0) is better performing in terms of epistemic correctness, whereas the various lottery Types (1–4) do better with regard to ethical considerations: they better compensate for evaluation bias (higher unbiased fairness) and distribute funding more evenly (higher distributive fairness).

However, Figure 1 also shows some less predictable results. Across epistemic and ethical considerations, Type 0 performs very similarly to Type 2 and, to a lesser degree, to Type 2 (when $S = 10$). Types 0 and 1 have nearly identical scores on epistemic correctness and unbiased fairness; only in distributive fairness they come apart, with Type 0 coming at the bottom. This suggests that, at least under the conditions explored in Figure 1 and for these metrics of fairness, traditional peer review may not be distinguishably—or meaningfully—different from those Types of lotteries where chance plays a limited role.

We observe a similar pattern for Type 3 (which we labeled 'partial lotteries of fundable proposals') and Type 4 (full lotteries). They have very similarly high scores in unbiased and distributive fairness. They only differ in epistemic correctness, where Type 4 performs more poorly.

Another unanticipated result concerns Type 2 partial lotteries with bypass. Across all notions of fairness, the performance of Type 2 procedures seems to greatly depend on the lottery choice rate (parameter S). When the lottery choice rate is low (e.g. Type 2, $S = 10$), Type 2 resembles Type 0–1 in having very high epistemic correctness and very low unbiased and distributive fairness. Conversely, when the lottery choice rate is high (Type 2, $S = 90$), Type 2 resembles Types 3–4, with very low epistemic correctness and very high unbiased and distributive fairness. Considering that Type 2 procedures are in use by several funders (see Section 3, Table 2) and that we could find no literature that describes, motivates or attempts to evaluate the lottery choice rate in these selection procedures, the role of the lottery choice rate in Type 2 lotteries seems to be an underappreciated consequential aspect of lottery design.

All these things considered, we may synthesize these findings by grouping selection procedures into *low randomness* and *strong randomness* Types. Low randomness Types are procedures where lottery plays little to no role. These include Types 0–1 (traditional peer review and tie-breaking partial lotteries), as well as Type 2 partial lotteries with bypass and low lottery choice rate (e.g. Type 2, $S = 10$). Conversely, strong randomness Types are procedures where funding

Table 3. Summary table of model parameters and their explored values.

Parameter	Value	Description
N	100	Number of proposals.
α, β	$\alpha = 10; \beta = 3$	Parameters of the beta distribution of reference evaluations.
σ_ε	0.1	Size of evaluation errors.
λ	{0.2, 0.5}	Size of evaluation bias.
G, R, γ	$R \in \{3, 5, 7\}$ $\gamma \in \{3, 5, 10\}$ This results in: $G \in \{7, 11, 13, 15, 21, 28, 29, 46, 64\}$	Granularity of the ranking of proposals by the typical peer review panel. G is the number of available ranking positions, and is determined by R (the number of reviewers on the panel) and by γ (the number of grades available for each reviewer to choose from).
T	{25, 50, 75}%	Funding rate.
θ	{0.25, 0.5, 0.75}	In Type 0–3 funding procedures, θ is the 'sufficiency bar': a threshold that determines whether grade g is sufficiently high for a proposal to be funded. Proposals with a grade lower than $\theta \times (G\text{-}1)$ cannot be funded.
S	{10, 50, 90}%	Lottery choice rate. In Type 2 funding procedures, this parameter denotes the percentage of proposals to be chosen by lot.

Underscored values represent what we call a 'non-extreme' condition.

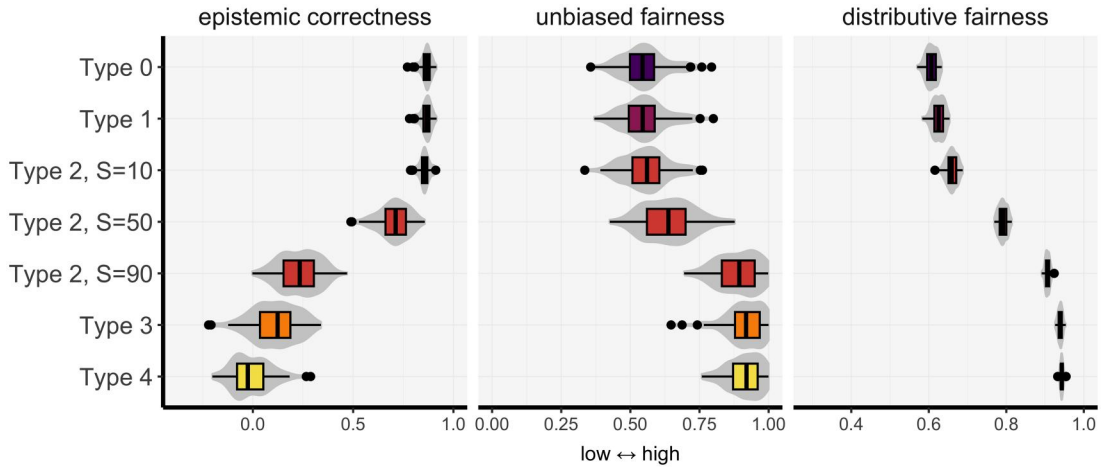

Figure 1. Comparison of the three types of fairness of various grant selection procedures under non-extreme conditions (technically, $\lambda = 0.2$, $R = 5$, $\gamma = 5$, $T = 50\%$, $h = 0.5$).

decisions are mainly or solely determined by chance. These include Type 2 with high lottery choice rate (e.g. Type 2, $S = 90$) and Types 3–4 (partial lotteries of fundable proposals and full lotteries).

With this grouping of selection procedures in mind, we can say that low randomness Types are better at allocating funding consistently with the evaluation of the peer review panel (high epistemic correctness). Conversely, strong randomness Types are better at correcting evaluation biases by the panel (high unbiased fairness) and at awarding grants more evenly across applicants and proposals (high distributive fairness).

The Supplementary appendix presents more fully the results from other regions of the parameter space. The findings presented thus far persist under many of the explored conditions; but there are noteworthy exceptions that we report here.

These exceptions emerge in a few circumstances, particularly if we assume stronger competition among proposals through, for example, a lower funding rate or tougher 'sufficient merit' requirements.[14] This scenario may be particularly relevant because these conditions of stronger competition are more representative of typical funding calls in the real world where submissions are many but funding is scarce. Figure 2 illustrates this scenario by showing simulation results where all conditions are the same as in Figure 1 except for the funding rate (here set to a lower level, $T = 25\%$) and a higher-than-earlier bar for sufficient merit ($\theta = 0.75$).

The first emerging exception we can see in Figure 2 concerns distributive fairness: under stronger competition, the effect of randomness on distributive fairness is not monotonic. Here, tie-breaking partial lotteries (Type 1) show better distributive fairness than both traditional peer review (Type 0) and low randomness partial lotteries with bypass (Type 2, $S = 10$); and this is different than in Figure 1 where we had the monotonic effect on distributive fairness: Type 0 < Type 1 < Type 2. The reason why Type 1 can outperform Type 2 ($S = 10$) under strong competition lies with the composition of the lottery pools in these selection procedures. Recall that in Type 1 the lottery pool only comprises proposals that tie for the last fundable ranking position; whereas in Type 2 the pool comprises all proposals that received a sufficient grade. With strong competition the selection is being made among many proposals with a very positive evaluation, and many of these might tie for the last fundable ranking position. Thus, the lottery pool in this condition may be larger for Type 1 than Type 2 ($S = 10$). In turn, a larger pool can result in awards being distributed more evenly among proposals, eventually leading to higher distributive fairness for Type 1 than for Type 2 with $S = 10$. Furthermore, as can be seen in Figure 2, while strong competition affects the relative performance of Type 1 and Type 2 in terms of distributive fairness, it does not affect their relative performance in terms of epistemic correctness and unbiased fairness. Ultimately, this is because the lottery pool in Type 1—however larger—tends to comprise higher-rated proposals than in Type 2. In sum, this suggests that, according to all definitions of fairness examined here, when there is strong competition a tie-breaking partial lottery might be the selection procedure with the highest distributive fairness among all low randomness Types.

The second noteworthy new result that emerges under strong competition (Figure 2) concerns the differences among strong randomness Types. These differences were subtle and mostly limited to epistemic correctness in Figure 1; by contrast, under strong competition these differences between partial lotteries of fundable proposals (Type 3) and full lotteries (Type 4) are wider across all measures. Specifically, Type 4 is markedly less epistemically correct than Type 3 and the rest, and markedly more unbiasedly and distributively fair.

This is an important finding about the difference between partial and full lotteries. Whereas Type 4 full lotteries do not involve any peer review panel, Type 3 partial lotteries do involve a panel, albeit to a limited capacity (the panel is only tasked with determining which proposals are 'fundable'). The differences between Type 3 and 4 in Figure 2 demonstrate that, however minimal, the involvement of peer reviewers can make a big difference, under the right conditions. Peer review panels can steer funding decisions in the directions of reference panel evaluations (higher epistemic correctness); inject bias in funding decisions (lower unbiased fairness); and/or lead to the concentration of resources in the hands of a few (lower distributive fairness).

6. Conclusion and discussion

In this study we introduced an expanded taxonomy of selection procedures in science funding systems. Our expansion

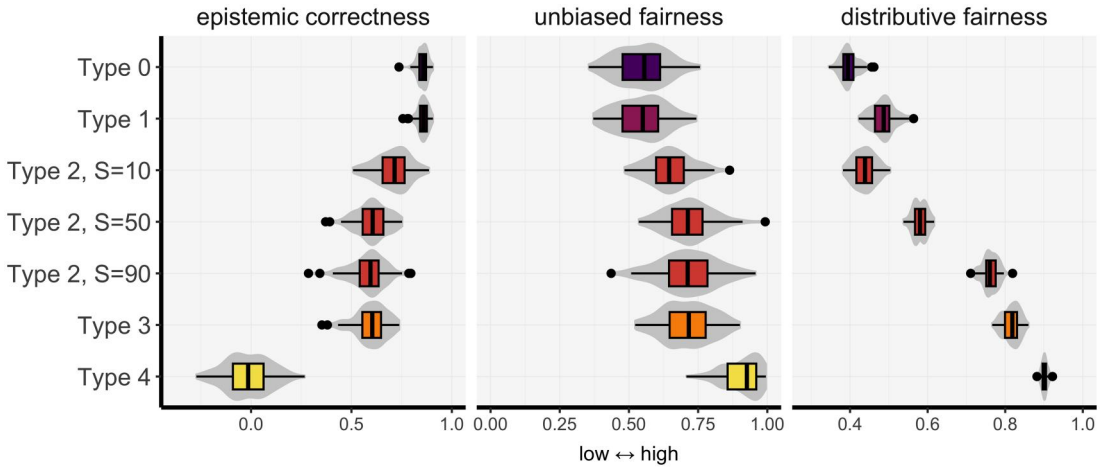

Figure 2. Comparison of the three types of fairness of various grant selection procedures under a stronger competition condition: lower funding rate T and higher θ ($\lambda = 0.2$, $R = 5$, $\gamma = 5$, $T = 25\%$, $\theta = 0.75$).

allows us to further distinguish among 'Types' of selection procedures adopted by research funding organizations that were all previously classified as 'partial lotteries'—despite their large differences in implementation. In the expanded taxonomy we rank selection procedures by the role of randomness from Type 0 (traditional peer review), where there is no randomness involved, to Type 4 (full lotteries), where all grant winners are determined exclusively by chance.

Then, we compared these Types of selection procedures in terms of two often-conflicting desired properties, or 'desiderata': epistemic correctness and ethical considerations including unbiased fairness and distributive fairness. We base our comparison of Types 0–4 on simulated counterfactuals run in numerical (Monte Carlo) experiments. Numerical simulations allow us to quantify how exactly each Type performs in terms of epistemic and ethical considerations, and how the performance varies under different conditions (e.g. for different funding rates, different review panel sizes, etc).

The simulation results show that the big distinction in epistemic and ethical considerations of fairness may not distinguish between traditional peer review and funding lotteries overall, but rather between low- and strong-randomness Types. Under most conditions, we find that 'low-randomness' Types (i.e. Type 0, traditional peer review, and Type 1, tie-breaking lotteries) with small distinctions similarly prioritize epistemic correctness at the expense of ethical considerations. The distinction among strong-randomness Types such as Type 3 (partial lotteries of fundable proposals) and Type 4 (full lotteries) is bigger than among low-randomness types but they also similarly prioritize ethical considerations at the cost of lower epistemic correctness.

The similarity among low-randomness and among strong-randomness Types can be important for funders to consider. For example, if Types 0 and 1 are similarly fair, what is the point of adopting a controversial Type 1 in the first place? The decision to choose a higher-randomness Type (e.g. Type 1 over 0, or Type 4 over 3) should be based on some other criteria such as their costs and public acceptance.

Another unexpected finding that deserves discussion is the large variation in the performance of partial lotteries with bypass (Type 2) which are being run by several funders. The performance of Type 2 strongly depends on the lottery choice rate: when only a few of the awards are determined by lot, Type 2 procedures balance epistemic and ethical considerations in the way low-randomness Types 0 and 1 do. When most of the awards are determined by lot, Type 2 is at the opposite end of the spectrum, coming closer to high-randomness Types 3 and 4. Funders adopting a Type 2 selection procedure should therefore be aware that the lottery rate can be very consequential, tune it carefully, and perhaps most importantly, make it known to applicants and the public.

More broadly, our results exemplify the usefulness of our simulation model—and of simulation experiments in general—as tools to ask new research questions, inform the debate on science funding allocation and, eventually, advise policy. Re-elaborating the results from Figure 1 in another way, for example, we can highlight the balance between epistemic and ethical considerations in a two-way plot that shows which areas of the plot are covered by each Type (see Figure 3). Given the priorities set by different funders, plots such as those proposed in Figure 3 can be generated showing the expected performance of each Type for the conditions set by the funder for a given funding call (from the expected funding rate, to the number of reviewers available, the evaluation scales adopted, etc). Such plots can inform funders as to which Type of selection procedure best meets the mix of desiderata they set for specific calls.

Our research has several limitations, the first of which is the very narrow domain of validity of our results. On a methodological level, our findings follow deductively from a series of assumptions about the grant selection process: how proposals are submitted, how review panels make their evaluation and how selection procedures are implemented. We tried to surface these assumptions and to balance model realism and plausibility with experimental needs to the best of our

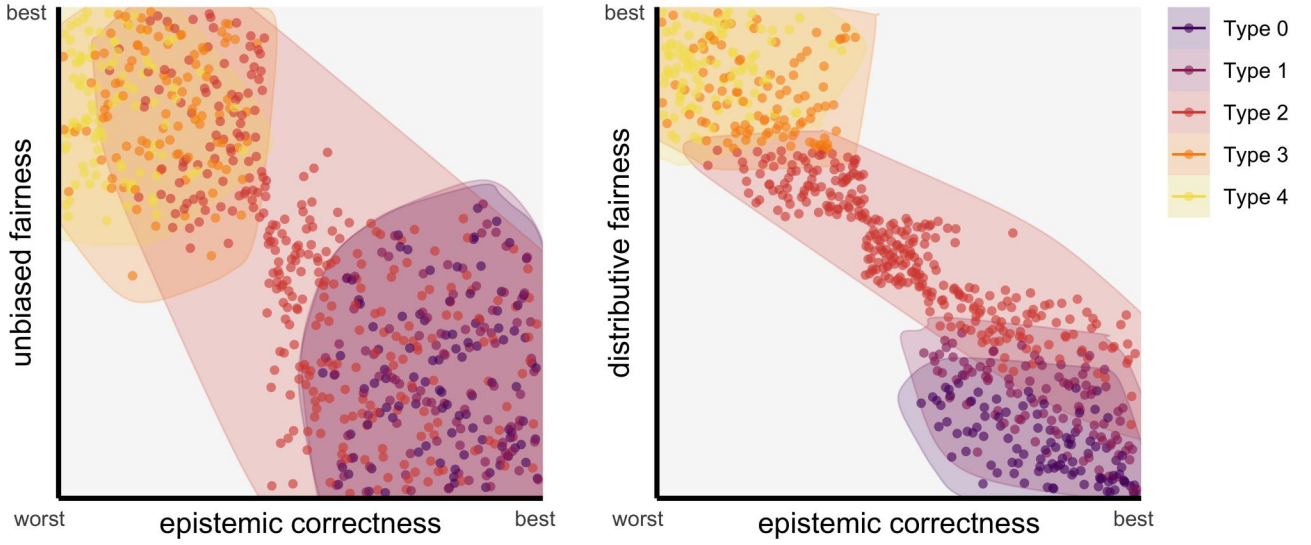

Figure 3. Re-elaboration of the results presented in Figure 1. Here, points are individual simulation runs under non-extreme conditions ($\lambda = 0.2$, $R = 5$, $\gamma = 5$, $T = 50\%$, $\theta = 0.5$). Plot axes show the desiderata, epistemic correctness (X-axis) and ethical considerations (unbiased fairness on the Y-axis of the left-hand panel; distributive fairness on the Y-axis of the right-hand panel). Color represents the Type of selection procedure. Shaded areas show the convex hull of each Type, with some padding for visual clarity. The distribution of convex hulls on the plane shows how each Type balances the two desiderata under the specified conditions.

ability. However, our findings are only as good as the underlying assumptions are sound.

On a more substantive level, here we only considered some aspects of fairness in grant allocation. Fairness is of course an important consideration for funders who are considering adopting lotteries, but there are other considerations being discussed in the literature that may be of equal or greater importance—mainly cost, time savings and workload for applicants and reviewers (Avin 2015; Fang and Casadevall 2016; Gross and Bergstrom 2019; Conix et al. 2021; Dresler et al. 2023). Another is diversity, often mentioned as an argument in favor of implementing funding lotteries (Horbach, Tijdink and Bouter 2022); and while here we did examine the related concept of distributive fairness, diversity is a more general, multifaceted idea. Other considerations concern potential unintended consequences of lottery systems on the production of science (see e.g. Reinhart and Schendzielorz 2020; Reinhart 2021). One last consideration is the public acceptance of funding lotteries as alternative to traditional peer review: for instance, Type 4 selection procedures seem not welcome by scholars, though opinions may be more nuanced when it comes to Type 1–3 partial lotteries (Liu et al. 2020; Reinhart and Schendzielorz 2020; Philipps 2021, 2022; Barlösius and Philipps 2022; Barlösius, Paruschke and Philipps 2023). Acceptability by funders and by the wider public is also under discussion (Barnett 2016; Bouacida and Foucard 2020).

Lastly, we remind the reader that ours is just a simulation experiment. While it does offer insights into what we might expect of funding lotteries, these expectations will need to be tested empirically. This might be possible in the near future when funders who are running lotteries today will have gathered sufficient data for researchers to retrospectively compare selection procedures. Data needed for studying lottery-based selection procedures primarily entail microdata on applicants, ranging from their demographic characteristics to the calls they participated in, to the outcomes of their applications—from review panel evaluations to final funding decisions. Second, data on the selection procedure adopted for each call are also needed—including a description of the peer review process and the adopted procedure for drawing lots. We particularly encourage researchers to re-use or extend our simulation model in order to explore relevant questions or test their empirical assumptions.

Ultimately, funders who are implementing lottery-based grants are in a unique position to also enable and pioneer empirical research on funding lotteries. Involving these funders in future research on funding lotteries—first and foremost via data sharing agreements—seems the only way forward for investigating funding lotteries empirically.

Future research should also examine more closely the interplay between selection procedures Types 0–4 and the other aspects of the design of selection procedures (Fiss 2007). Two particularly relevant aspects are, for instance, weighting and tiering (Shaw 2022). In short, weighting refers to the idea of assigning a different number of 'lottery tickets' to different proposals, as a way to proportionally favor disadvantaged or otherwise underrepresented groups of applicants; tiering refers to a panel's capacity to construct finer-grained rankings of proposals. These two aspects are orthogonal to the aspect of partiality that we captured by teasing apart Types 0–4 and they may be consequential for the epistemic correctness and unbiased/distributive fairness: weights are conceived as a means to better distribute research resources among applicants; and tiering affects the performance of all Types of selection procedures. Research on the interplay between the role of randomness (i.e. from Type 0 to Type 4), weighting and tiering of selection procedures is the natural next step in this line of research.

Acknowledgements

We have several people to thank for sharing valuable input on different aspects and at various stages of our work: Martin Reinhart, Michael Morreau, Gemma Derrick, Diana Hicks, Jamie Shaw, Kristin Oxley, Chiara Franzoni, and our SPRING (Studying Peer Review IN Grant Funding) team at University College Dublin. We are also grateful for the thoughtful suggestions we received from three anonymous referees. Any remaining errors and omissions are our own.

Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

Conflict of interest statement. None declared.

Funding

This work was supported by the Science Foundation Ireland [17/SPR/5319].

Authors' contributions

Thomas Feliciani: Conceptualization, Writing—Original draft, Investigation, Methodology, Software, Visualization. **Junwen Luo**: Conceptualization, Investigation, Writing—Review & Editing. **Kalpana Shankar**: Conceptualization, Investigation, Writing—Review & Editing, Project administration.

Data availability

To ensure the reproducibility of simulations and analyses, all relevant code and resources have been made publicly available in a dedicated code repository accessible at https://github.com/thomasfeliciani/mixed-lotteries-for-science-funding/. The simulation scripts are designed to run in R 4.3.2.

Notes

1. A tentative list of funders who have introduced a funding lottery is provided below in Section 3, Table 2.
2. We set aside the distinction between a 'more even' distribution of funding and the notion of 'perfect equality'. While funding lotteries might in principle distribute funding more evenly than traditional peer review, they are not ideal method of distributing funding perfectly evenly across applicants. By definition, the ideal method would be an 'egalitarian sharing' of government research funding among eligible researchers, e.g. by replacing grants with stipends, the same for all researchers, and possibly adjusted to compensate for differences in costs (e.g. of equipment) between research fields. Although the proposal of egalitarian sharing has been considered in the literature (Martin 2000; Sakamoto 2023; Vaesen and Katzav 2017), egalitarian sharing does not belong in the extended taxonomy of selection procedures because it requires no selection at all and therefore, in the strict sense of the term, it is not a selection procedure.

3. Here we discuss 'indistinguishability' around the funding line. Admittedly, other authors have found that the 'area of indistinguishability' may not be (limited to) near the funding line (for a discussion, see Shaw 2022, Section 4.2.4). Greenberg's argument extends to these cases, too.
4. Historically, the idea of partial lotteries originates in an electoral system in Renaissance Florence known as *scrutinio e tratta* ('scrutiny and lot'). *Scrutinio e tratta* is a form of sortition where several candidates were elected or nominated for the same public office; the winner among them was then determined by lot. It was proposed as a method to preserve political balance in the city, preventing all political offices from being assigned to one faction, the powerful Medici family and their allies (Najemy 1982).
5. Concerning weighting, we will assume that no weighting is applied (all proposals in a lottery pool have the same chance of winning—ie uniform win probability). Concerning tiering, we will assume that the number of possible ranking positions ('granularity') depends solely on (1) the number of reviewers on the peer review panel, and (2) the number of available grades that each reviewer can choose from. We will set both to plausible values. Implementation details are discussed in Section 4.1.
6. This generally follows Shaw's distinction between non-lotteries (which we denote Type 0), 'partial' lotteries (Types 1–3) and 'non-partial' lotteries (Type 4) (Shaw 2022).
7. Practically, the bypass in Type 2 selection procedures can be done via 'wildcards' or 'unanimous approval'. Wildcards are the opportunity given by the funder to individual reviewers to arbitrarily choose one or a few proposals to be funded, even though the rest of the review panel would disagree. By contrast, bypass by unanimous approval means that all reviewers must agree that a given proposal should be funded, for it to bypass the lottery. Despite this implementation differences, in both cases it is optional for the panel to choose a bypass set (whereas it is a requirement for it to choose a lottery set). We do not reflect this distinction in the taxonomy because the role of chance in the final funding decision is the same in both cases.
8. Because evaluations are coarse, the ranking will generally be a weak ordering of the evaluated proposals—that is, two or more proposals might tie for the same ranking position.
9. This implies that, by setting $\theta = 0$, the model reproduces a selection procedure where there are no 'sufficiency of grade' requirements. Conversely, $\theta = 1$ reproduces a selection procedure where only proposals that obtained the best possible grade are considered for funding.
10. If there are ties for the T-S^{th} position, the tie is broken at random uniform. If, among the T-S highest-ranking proposals there are some that received an insufficient grade, these are removed from the bypass set, and the lottery pool will be empty—this is a degenerate case where Type 2 is equivalent to Type 1.
11. For a similar operationalization, see Squazzoni and Gandelli (2013) and follow-up work.
12. In other words, we ran a full-factorial experiment design. In the case of parameters G, R and γ we disregarded R and γ in the factorial experimental design and only considered G instead. This is for simplicity: of the three, G is the only parameter of interest and R and γ have a predictable, positive linear effect on G and therefore on the simulation results.
13. Boxplots display the quartiles of the distributions (first: outside of the box, on the left; second: left part of the box; third: right part of the box; fourth: outside of the box, on the right). The vertical line inside the boxplots is therefore median; and the whiskers outside of the boxplots extend horizontally 1.5 times the interquartile range.
14. These exceptions are observed not only when we model stronger competition (lower T and higher θ, but also when we increase the evaluation bias by the typical panel, λ; and if we reduce the granularity of the panel ranking, e.g. by lowering R and/or γ. Results from these other conditions are presented in the Supplementary Data.

References

Aczel, B., Szaszi, B., and Holcombe, A. O. (2021) 'A Billion-Dollar Donation: estimating the Cost of Researchers' Time Spent on Peer Review', *Research Integrity and Peer Review*, 6: 14.

Adam, D. (2019) 'Science Funders Gamble on Grant Lotteries', *Nature*, 575: 574–5.

Avin, S. (2015). 'Funding Science by Lottery'. in U. Mäki, I. Votsis, S. Ruphy, and G. Schurz (eds) *Recent Developments in the Philosophy of Science: EPSA13 Helsinki, European Studies in Philosophy of Science*, Vol. 1, pp. 111–26. Cham: Springer International Publishing.

Avin, S. (2019) 'Mavericks and Lotteries', *Studies in History and Philosophy of Science Part A*, 76: 13–23.

Barlösius, E., and Philipps, A. (2022) 'Random Grant Allocation from the Researchers' Perspective: Introducing the Distinction into Legitimate and Illegitimate Problems in Bourdieu's Field Theory', *Social Science Information*, 61: 154–78.

Barlösius, E., Paruschke, L., and Philipps, A. (2023) 'Peer Review's Irremediable Flaws: Scientists' Perspectives on Grant Evaluation in Germany', *Research Evaluation*, 32: 623–34.

Barnett, A. G. (2016) 'Funding by Lottery: Political Problems and Research Opportunities', *mBio*, 7: e01369-16.

Bedessem, B. (2020) 'Should we Fund Research Randomly? An Epistemological Criticism of the Lottery Model as an Alternative to Peer Review for the Funding of Science', *Research Evaluation*, 29: 150–7.

Bol, T., de Vaan, M., and van de Rijt, A. (2018) 'The Matthew Effect in Science Funding', *Proceedings of the National Academy of Sciences*, 115: 4887–90.

Bornmann, L., Leydesdorff, L., and Van den Besselaar, P. (2010) 'A Meta-Evaluation of Scientific Research Proposals: Different Ways of Comparing Rejected to Awarded Applications', *Journal of Informetrics*, 4: 211–20.

Bornmann, L., Mutz, R., and Daniel, H. (2007) 'Gender Differences in Grant Peer Review: A Meta-Analysis', *Journal of Informetrics*, 1: 226–38.

Bouacida, E., and Foucard, R. (2020). *The Acceptability of Lotteries in Allocation Problems:A Choice-Based Approach*. Lancaster: Lancaster University, Department of Economics,.

Brezis, E. S. (2007) 'Focal Randomisation: An Optimal Mechanism for the Evaluation of R&D Projects', *Science and Public Policy*, 34: 691–8.

Broome, J. (1984) 'Selecting People Randomly', *Ethics*, 95: 38–55.

Burns, K. E. A., Straus, S. E., Liu, K., Rizvi, L., and Guyatt, G. (2019) 'Gender Differences in Grant and Personnel Award Funding Rates at the Canadian Institutes of Health Research Based on Research Content Area: A Retrospective Analysis' *PLOS Medicine*, 16: e1002935.

Conix, S., De Block, A., and Vaesen, K. (2021) 'Grant Writing and Grant Peer Review as Questionable Research Practices', *F1000Research*, 10: 1126.

De Peuter, S., and Conix, S. (2022) 'The Modified Lottery: Formalizing the Intrinsic Randomness of Research Funding', *Accountability in Research*, 29: 324–45.

Dresler, M., Buddeberg, E., Endesfelder, U., Haaker, J., Hof, C., Kretschmer, R., Pflüger, D., and Schmidt, F. (2023) 'Effective or Predatory Funding? Evaluating the Hidden Costs of Grant Applications', *Immunology & Cell Biology*, 101: 104–11.

Fang, F. C., and Casadevall, A. (2016) 'Research Funding: The Case for a Modified Lottery', *mBio*, 7: e00422-16.

Feliciani, T., Luo, J., Ma, L., Lucas, P., Squazzoni, F., Marušić, A., and Shankar, K. (2019) 'A Scoping Review of Simulation Models of Peer Review', *Scientometrics*, 121: 555–94.

Feliciani, T., Morreau, M., Luo, J., Lucas, P., and Shankar, K. (2022) 'Designing Grant-Review Panels for Better Funding Decisions: Lessons from an Empirically Calibrated Simulation Model', *Research Policy*, 51: 104467.

Fiss, P. C. (2007) 'A Set-Theoretic Approach to Organizational Configurations', *Academy of Management Review*, 32: 1180–98.

Gildenhuys, P. (2020) 'Lotteries Make Science Fairer', *Journal of Responsible Innovation*, 7: S30–43.

Gillies, D. (2014) 'Selecting Applications for Funding: why Random Choice is Better than Peer Review', *RT. A Journal on Research Policy and Evaluation*, 2: 1.

Greenberg, D. S. (1998) 'Chance and Grants', *The Lancet*, 351: 686.

Gross, K., and Bergstrom, C. T. (2019) 'Contest Models Highlight Inherent Inefficiencies of Scientific Funding Competitions', *PLOS Biology*, 17: e3000065.

Heyard, R., Ott, M., Salanti, G., and Egger, M. (2022) 'Rethinking the Funding Line at the Swiss National Science Foundation: Bayesian Ranking and Lottery', *Statistics and Public Policy*, 9: 110–21.
Horbach, S. P. J. M., Tijdink, J. K., and Bouter, L. M. (2022) 'Partial Lottery Can Make Grant Allocation More Fair, More Efficient, and More Diverse', *Science and Public Policy*, 49: 580–2.
Langfeldt, L., Reymert, I., and Svartefoss, S. M. (2023) 'Distrust in Grant Peer Review—Reasons and Remedies', *Science and Public Policy*, 51: 28–41.
Lee, C. J., Grant, S., and Erosheva, E. A. (2020) 'Alternative Grant Models Might Perpetuate Black–White Funding Gaps', *The Lancet*, 396: 955–6.
Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013) 'Bias in Peer Review', *Journal of the American Society for Information Science and Technology*, 64: 2–17.
Lindsey, D. (1988) 'Assessing Precision in the Manuscript Review Process: A Little Better than a Dice Roll', *Scientometrics*, 14: 75–82.
Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely, T., and Pomeroy, L. (2020) 'The Acceptability of Using a Lottery to Allocate Research Funding: A Survey of Applicants', *Research Integrity and Peer Review*, 5: 3.
Luebber, F., Krach, S., Martinez Mateo, M., Paulus, F. M., Rademacher, L., Rahal, R.-M., and Specht, J. (2023) 'Rethink Funding by Putting the Lottery First', *Nature Human Behaviour*, 7: 1031–3.
Luukkonen, T. (2012) 'Conservatism and Risk-Taking in Peer Review: Emerging ERC Practices', *Research Evaluation*, 21: 48–60.
Martin B. (2000) 'Reserach Grants: problems and Options', *The Australian Universities' Review*, 43: 17–22.
McAllister, D., Juillerat, J., and Hunter, J. (2016) 'What Stops Women Getting More Grants?', *Nature*, 529: 466.
Mutz, R., Bornmann, L., and Daniel, H.-D. (2012) 'Does Gender Matter in Grant Peer Review?: An Empirical Investigation Using the Example of the Austrian Science Fund', *Zeitschrift Für Psychologie*, 220: 121–9.
Najemy, J. M. (1982). *Corporatism and Consensus in Florentine Electoral Politics, 1280-1400*. Chapel Hill: University of North Carolina Press.
Osterloh, M., and Frey, B. S. (2020) 'How to Avoid Borrowed Plumes in Academia', *Research Policy*, 49: 103831.
Philipps, A. (2021) 'Science Rules! A Qualitative Study of Scientists' Approaches to Grant Lottery', *Research Evaluation*, 30: 102–11.
Philipps, A. (2022) 'Research Funding Randomly Allocated? A Survey of Scientists' Views on Peer Review and Lottery', *Science and Public Policy*, 49: 365–77.
Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., and Carnes, M. (2018) 'Low Agreement among Reviewers Evaluating the Same NIH Grant Applications', *Proceedings of the National Academy of Sciences*, 115: 2952–7.
Piper, K. (2019). 'Science funding is a mess. Could grant lotteries make it better?'. *Vox*. <https://www.vox.com/future-perfect/2019/1/18/18183939/science-funding-grant-lotteries-research> accessed 18 June 2024.
R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Reinhart, M. (2021). Open Science as an Engine of Anxiety: How Scientists Promote and Defend the Visibility of Their Digital Selves, while Becoming Fatalistic about Academic Careers (preprint). SocArXiv. <https://osf.io/2vr7j> accessed 27 April 2022. DOI: 10.31235/osf.io/2vr7j
Reinhart, M., and Schendzielorz, C. (2020) 'The Lottery in Babylon—On the Role of Chance in Scientific Success', *Journal of Responsible Innovation*, 7: S25–9.
Roumbanis, L. (2019) 'Peer Review or Lottery? A Critical Analysis of Two Different Forms of Decision-Making Mechanisms for Allocation of Research Grants', *Science, Technology, & Human Values*, 44: 994–1019.
Sakamoto, T. (2023) 'Is Lottery a Better Way of Resource Distribution Than Baseline Funding?', *Philosophy of Science*, 90: 603–28.
Shaw, J. (2022) 'Peer Review in Funding-by-Lottery: A Systematic Overview and Expansion', *Research Evaluation*, 32: 86–100.
Shaw, J. (2023) 'Peer Review, Innovation, and Predicting the Future of Science: The Scope of Lotteries in Science Funding Policy', *Philosophy of Science*, 90: 1297–306.
Squazzoni, F., and Gandelli, C. (2013) 'Opening the Black-Box of Peer Review: An Agent-Based Model of Scientist Behaviour', *Journal of Artificial Societies and Social Simulation*, 16: 3.
Stone, P. (2009) 'The Logic of Random Selection', *Political Theory*, 37: 375–97.
Vaesen, K., and Katzav, J. (2017) 'How Much Would Each Researcher Receive If Competitive Government Research Funding Were Distributed Equally among Researchers?', *Plos One*, 12: e0183967.
van den Besselaar, P., Sandström, U., and Schiffbaenker, H. (2018) 'Studying Grant Decision-Making: A Linguistic Analysis of Review Reports', *Scientometrics*, 117: 313–29.
Vindin, H. (2020) 'Grant Lottery: don't Stall Ideas and Careers', *Nature*, 577: 472.
Witteman, H. O., Hendricks, M., Straus, S., and Tannenbaum, C. (2019) 'Are Gender Gaps Due to Evaluations of the Applicant or the Science? A Natural Experiment at a National Funding Agency', *The Lancet*, 393: 531–40.
Woods, H. B., and Wilsdon, J. (2021). *Why draw lots? Funder motivations for using partial randomisation to allocate research grants (RoRI Working Paper No.7)*., p. 2603617 Bytes. Research on Research Institute. <https://rori.figshare.com/articles/report/Why_draw_lots_Funder_motivations_for_using_partial_randomisation_to_allocate_research_grants_RoRI_Working_Paper_No_7_/17102495/2> accessed 10 February 2023.
Yip, P. S. F., Xiao, Y., Wong, C. L. H., and Au, T. K. F. (2020) 'Is There Gender Bias in Research Grant Success in Social Sciences?: Hong Kong as a Case Study', *Humanities and Social Sciences Communications*, 7: 173.
Zhang, G., Xiong, L., Wang, X., Dong, J., and Duan, H. (2020) 'Artificial Selection versus Natural Selection: Which Causes the Matthew Effect of Science Funding Allocation in China?', *Science and Public Policy*, 47: 434–45.
Zhuang, H., and Acuna, D. E. (2019). 'The effect of novelty on the future impact of scientific grants', arXiv. DOI: 10.48550/ARXIV.1911.02712
Zollman, K. J. S. (2009) 'Optimal Publishing Strategies', *Episteme*, 6: 185–99.

Research Evaluation, 2024, 33, 1–13
https://doi.org/10.1093/reseval/rvae025
Article