Addressing conservatism in funding decisions via enhanced selection mechanisms

Thomas Feliciani [a]*, Rikke Nørding Christensen [b], Chiara Franzoni [a]

[a] School of Management, Politecnico di Milano, Via Lambruschini 4/B, 20156 Milan, Italy
[b] Novo Nordisk Foundation, Tuborg Havnevej 19, 2900 Hellerup, Denmark
* Corresponding author: thomas.feliciani@polimi.it

Abstract

In recent years, concerns that peer review of grant proposals may disfavor innovative research – often referred to as *conservatism* in science funding – have led to proposed interventions on the selection mechanisms to counteract this conservatism. While several funders have begun implementing some of these interventions, their effectiveness remains unknown. We evaluate the impact of these interventions on conservatism using data from the Novo Nordisk Foundation on 273 calls for proposals and more than 46,000 peer-review evaluations collected over a decade. Our study relies on a natural experiment and on the simulation of counterfactuals with computer models. The interventions examined include: (1) the creation of novelty-dedicated funding streams designed to attract and reward innovative, interdisciplinary, and high-risk high-gain research proposals. (2) The use of aggregation rules to replace panel discussions (ranking by mean, median or best score). (3) The use of "golden tickets" ("wild cards"), which allow each reviewer to unilaterally promote one proposal of their choice. (4) Funding lotteries, where reviewers select a broad set of fundable proposals from which grant winners are randomly chosen.

Results show that the effectiveness of these interventions depends on what aspect of novelty are prioritized. Novelty-dedicated calls effectively attract and promote science projects that recombine theories, concepts, methods, and tools in novel ways (stronger recombination). In contrast, golden tickets and funding lotteries are more effective in supporting projects that develop entirely new theories, concepts, methods, or tools (newer vintage).

Keywords: Science funding; peer review; creativity; aggregation; golden tickets; lotteries.

Acknowledgements: This work was supported by the Alfred P. Sloan Foundation (G-2021-14189) and by the Novo Nordisk Foundation (NNF21SA0071711). We are grateful for feedback and suggestions provided at various stages of this work by: Jerome Adda, Ruggero Barzon, Sofie Cairo, Christoph Carnehl, Stijn Conix, Henrik Barslund Fosse, Jared Gars, Daniel Gross, Massimiliano Guerini, Martina Iori, Nicola Melluso, Marco Ottaviani, Kristin Oxley, Cem Peker, Justus Preusser, Bhaven Sampat, Jesper W. Schneider, Sotaro Shibayama, Daniel Souza, Andola Stanaj, and Paula Stephan. Any remaining errors and omissions are our own. We thank the NIH National Library of Medicine for providing the infrastructure we relied upon for extrapolating the MeSH terms and metadata from research proposals.

1. Introduction

Many research funding organizations (*funders*) pledge to fund innovative science. However, a widespread concern is that funding decisions tend to minimize risk by supporting projects that have a higher likelihood of success, based on established methodologies and past achievements. As a result, more traditional and incremental research proposals would be more likely to be funded over those that develop new methods and theories, draw on interdisciplinary approaches, or outline uncertain outcomes (Azoulay et al., 2011; Franzoni et al., 2022; Luukkonen, 2012; Packalen & Bhattacharya, 2020; Veugelers et al., 2022). This creates a paradox whereby research articles from funded projects may be *less* innovative than those that do not report any funding (Shin et al., 2022). The literature refers to the purported bias against innovative research – and its consequences – as conservatism (Lee et al., 2013; Wessely, 1998).

There is an active scholarly debate on possible interventions to curb conservativism in science funding. Proposed interventions generally seek either to increase the awareness and receptiveness of individual reviewers toward novel or innovative aspects of research proposals, or to prevent panel discussions from amplifying any novelty-averse tendencies among peer reviewers. While some funders are currently implementing or trialing these interventions, it remains uncertain (1) whether they are effective, and (2) to what degree. Given the nature of the trials, the small samples involved in these and the time lag with which scientific results materialize, it is unlikely that ongoing trials will lead to results in the next few years. In the meantime, we aim to answer these questions by leveraging a natural experiment and computer-based counterfactual simulations conducted on real-world data on funding applications.

We use data from 13,106 proposals and 46,228 peer review evaluations related to 273 funding calls administered between 2012-2022 by Novo Nordisk Foundation (NNF). NNF is a major funder for research in the life sciences, with an annual research funding budget ranging from DKK 960 million in 2012 (awarded grants; DKK 484 million paid out) to DKK 7.4 billion in 2022 (DKK 5.8 billion paid out). Project evaluation and selection at NNF is done by panels of internationally renowned scholars who first review projects separately and then meet to discuss and select winners collectively (sitting panels). We examine four families of selection mechanisms (*interventions*) alternative to the traditional sitting panels meetings. The interventions variously affect how reviewer evaluation are processed into final funding recommendations. They are: 1) Novelty-dedicated calls, 2) Aggregation rules, 3) Golden tickets and 4) Funding lotteries (Table 1). We test the effectiveness of these interventions in reducing or neutralizing the purported conservatism in research funding by comparing the novelty of proposals that have been chosen using the traditional mechanism with those that would have been selected if the intervention had been in place.

Novelty is a multifaceted concept and several operationalizations have been proposed in the literature. Two dimensions of novelty are important here: *vintage* and *recombination*. Research of newer vintage is one that introduces or develops theories, concepts, methods or tools never previously seen in the published literature (Arts et al., 2023; Azoulay et al., 2011; Mishra & Torvik, 2016). Research uses *recombination* if it combines two or more streams of literature, theories, methods or tools that are rarely put together in the published literature (Shibayama et al., 2021; Uzzi et al., 2013; Wang et al., 2017). Vintage and recombination capture two key dimensions of novelty valued by funders, and we operationalize them using adapting several indicators from the literature to bolster the validity and robustness of our findings.

Our objectives are twofold. First, we search for evidence of conservatism in funding decisions. Whereas earlier work finds mixed results both at NNF (Keldke et al., 2023; Peng et al., 2024) and other funding institutions (Boudreau et al., 2016; Packalen & Bhattacharya, 2020; Veugelers et al., 2022), our results show robust evidence for conservatism in regular sitting panels, suggesting that more novel proposals are significantly less likely to be funded.

Our second and primary objective is to gauge the potential efficacy of the four interventions. For the first intervention, novelty-dedicated calls, NNF offers a natural experiment: of the 273 calls in our dataset, 13 were novelty-dedicated ("Synergy") calls. These explicitly invite the submission of novel ideas, high-risk high-gain projects and interdisciplinary approaches. Comparing the average novelty of funded proposals between regular and novelty-dedicated calls we find ambivalent results. We find evidence for a novelty penalty (conservatism) in funding decisions in regular calls; but not in novelty-dedicated calls. This seems to place them as an effective intervention to curb conservativism. However, reducing conservatism in funding decisions does not guarantee that projects funded via novelty-dedicated calls are on average more novel than those funded by regular calls, as novelty-dedicated calls may induce applicants to submit proposals that are different than those they would submit to regular calls. We indeed find that the novelty-dedicated calls elicit submissions that involve more recombination than regular calls, but do not necessarily elicit projects of newer vintage. Therefore, the overall efficacy of this intervention may be limited to some dimensions of novelty (in the case of NNF, recombination), at the expense of other dimensions (vintage).

We then analyze the other three interventions – aggregation rules, golden tickets and funding lotteries. Underling these interventions is the premise that conservatism may emerge, at least partly, from the interpersonal dynamics during panel meetings. During these meetings, reviewers must collectively choose among proposals of comparable quality (Bornmann et al., 2010; van den Besselaar et al., 2018). To make these difficult decisions, panelists may go in search for reasons *to reject* proposals, giving excessive weight to negative opinions, which emerge more frequently for novel approaches (Lane et al., 2022), or they may rely on bibliometric indicators – which, again, put novel proposals at a disadvantage (Stephan et al., 2017). Because panel meetings may be a source of conservatism, these three interventions are alternative ways to reduce or altogether prevent the influence of panel meetings on final funding decisions.

Although our data does not include calls that were run using aggregation rules, golden tickets or funding lotteries, we study the efficacy of these interventions through Monte Carlo simulation experiments (Feliciani et al., 2019; Fortunato et al., 2018). In essence, for each NNF call in our dataset we create multiple virtual counterparts (called counterfactual calls) where funding decisions are generated according to the intervention protocol. We then compare the degree of novelty of funded projects between counterfactual and actual calls: this allows us to determine which conditions and selection mechanisms work better at curbing conservatism and enabling novel science. Surprisingly, the results predict that the use of partial randomization and certain aggregation methods can achieve only modest gains in novelty, these gains are only limited to some dimensions of novelty, and they are of smaller magnitude compared to what is already achieved by adopting novelty-dedicated calls.

By evaluating the effectiveness of interventions to the selection mechanisms in curbing conservatism, our study identifies those most promising ones for enhancing innovation. It also moves first steps and prompts further inquiry into the differential impact these interventions might have on different dimensions of novelty.

The next Section 2 elaborates on the four families of interventions. Section 3 focuses on data and operationalization. Sections 4 presents the results, and Section 5 discusses their implications for science policy and identifies avenues for further research.

2. Four interventions

We choose four prominent families of interventions either because they are frequently debated by scholars and practitioners (Bieri et al., 2021; Fang & Casadevall, 2016; Greenberg, 1998; Franzoni et al., 2022) or because they are being trialed by funders (Feliciani et al., 2024; Singh Chawla, 2023b).

Family of interventions	Description	Method of investigation
Novelty-dedicated calls (high-risk high-gain schemes and similar)	Funding calls specifically targeted at high-novelty proposals. Novelty is a prominent selection criterion.	Natural experiment with NNF data.
Aggregation rules (mathematical formulas)	Panel discussions are replaced by the mathematical aggregation of the scores issued by independent reviewers. We explore aggregation by *mean*, *median* or *best score*. Funding is allocated to proposals with the best aggregate score.	Counterfactual simulations with NNF data.
Golden tickets (a.k.a. wild cards)	Once per call, each reviewer is allowed to unilaterally promote one proposal they particularly like. Funding is allocated to the best-rated proposals and to proposals on whom reviewers spent their golden ticket.	Counterfactual simulations with NNF data.
Funding lotteries (partial randomization)	Review scores are used to select a pool of proposals of fundable quality. Funding is allocated to a random subset of these.	Counterfactual simulations with NNF data.

Table 1. Interventions targeting the selection mechanism of grant peer review.

2.1. Novelty-dedicated calls

Novelty-dedicated calls are any stream of research funding that funders explicitly dedicate to promote highly innovative science. By attracting highly novel applications and by instructing reviewers to recognize and prioritize novelty, it is hoped that novelty-dedicated calls do better at funding novel science than other regular calls. Novelty-dedicated calls represent a broader trend among funders to explicitly support innovative and interdisciplinary research: examples include some funding programs run by the European Research Council ("Starting Grants"), and several programs run by the NIH (the "NIH Director's Pioneer Award", the "NIH Director's New Innovator Award", or the "NIH Director's Transformative Research Award"). In our dataset, 13 of the 273 funding calls were novelty-dedicated. These are called "Interdisciplinary Synergy Programme", and invite submissions "*based on a novel research idea, where an interdisciplinary, high-risk/high-gain challenge is at the centre*". Furthermore, the assessment criteria used by reviewers included the "*creativity and novelty of the exploratory idea*", and the "*interdisciplinary approach of the project and its relevance*".

2.2. Aggregation rules

This is a family of interventions that use as input the assessments of individual reviewers and apply rules or mathematical formulas to combine them to find some aggregate, overall score for each proposal. A rule for combining the opinions may be as simple as averaging the individually assigned scores. Then, proposals are ranked by their aggregate score, and funding is allocated to proposals in a descending order, from the best aggregate score to the worst, until the budget threshold is met.

Funders may formally prescribe a specific protocol for this step. In other cases, an aggregation rule may be used informally during face-to-face panel meetings to facilitate the deliberation. The intervention we examine here concerns the use of an aggregation rule not to assist, but to replace the deliberations of panel meetings, so that a program officer may rank proposals and draw funding decisions without the need to call in a panel meeting. Avoiding a panel meeting may be a way to prevent the negativity bias, or bibliometric-driven preferences that can induce conservatism in funding recommendations.

We examine three prominent aggregation rules from among those that may be applicable to the context of grant peer review (Csaszar & Eggers, 2013; Feliciani et al., 2022; Hastie & Kameda, 2005).

i) *Aggregation by mean* involves averaging the scores given by reviewers. Previous research has found that aggregation by mean is in fact a good approximation of panel deliberations (Bieri et al., 2021; Pina et al., 2021), and the simplicity of this rule makes it likely that review panels already use this rule informally to guide panel meetings. All these factors make using the mean score aggregation a strong initial candidate for an aggregation rule to replace panel meetings entirely. It does have shortcomings, though: as Sir Francis Galton put it, averages give "a voting power to 'cranks' in proportion to their crankiness" (Galton, 1907). In other words, aggregate evaluations based on averages may be steered by the conservativism of any panel member.

ii) *Aggregation by median* involves aggregating individually-assigned scores using the median. This addresses the shortcoming of aggregation by mean: for panels with more than two reviewers, individual risk-averse reviewers cannot steer the aggregate evaluation. However, just like the mean, the median is a measure of central tendency of a distribution. Thus, scores aggregated by median may be about as conservative as the typical reviewer on the panel.

iii) *Aggregation by best score* involves choosing the most positive score given to a proposal by any of the panel members. This rule can in principle minimize type II errors, i.e. the erroneous rejection of good proposals. Further, it does not seek some potentially conservative central tendency in the distribution of scores and ignores the contribution of particularly conservative panel members. It will allow a good, highly novel proposal to be at the top of the ranking as long as at least one panel member is not biased against novelty. However, there are downsides explaining why we do not know of any funder that uses this aggregation rule. One possible downside is that reviewers who give the best score could be mistaken, e.g. they might have missed some critical flaws in the proposal. If so, other reviewers on the panel who might have noticed the flaws would be dismissed, leading to funding being allocated to flawed proposals. Furthermore, by ignoring all but the most positive grades, many proposals will have the same aggregate score – the best score – creating a high number of ties for the first ranking position, which may be impractical. Despite these flaws, it is worth considering this aggregation rule as a thought experiment to gauge the efficacy of extreme interventions such as replacing panel meetings with aggregation by best score.

2.3. Golden tickets

Golden tickets, or "wild cards", are the option given to each reviewer – only once within a funding call – to unilaterally promote a single proposal, ensuring it gets funded regardless of potential disagreement with other reviewers on the panel. If used as replacement for sitting panels, reviewers choose to spend their golden ticket upon submitting their review. If used during panel meetings, reviewers may spend their ticket during the deliberation. Either way, golden tickets may be a way to prevent or mitigate the potentially conservative influence of panel meetings on funding recommendations. In a way, golden tickets are a pragmatic alternative to aggregation by best score: if a reviewer is certain a proposal deserves funding, they can guarantee its success. However, unlike aggregation by best score, golden tickets come with a limit on reviewers' discretion. First, reviewers may only use this privilege once. Second, they do so deliberately and publicly: this puts their reputation at stake, deterring the misuse of golden tickets, but also potentially make reviewers afraid of losing status for unpopular preferences (Ottaviani & Sørensen, 2001).

To our knowledge, the list of funders currently adopting selection mechanisms based on golden tickets includes the U.S. National Science Foundation, Volkswagen Foundation, and Villum Foundation (Singh Chawla, 2023b, 2023a). We rely on the implementation and use of golden tickets at Villum Foundation as a blueprint for our counterfactual simulations. At Villum Foundation there are no panel meetings, meaning reviewers choose if and how to spend their golden ticket without knowing the opinions of their peers. Unsurprisingly, reviewers tend to spend their golden tickets on the proposal they rate the highest: these are often proposals that would have been funded regardless, as they typically receive excellent scores from other reviewers as well. Finally, roughly half of the reviewers at Villum Foundation use their golden ticket within a call. Our simulations of golden tickets are designed to adhere to these observations.

2.4. Funding lotteries

The third intervention is the partial randomization of funding decisions through so-called funding lotteries (Fang & Casadevall, 2016; Shaw, 2023). There are several ways of implementing a funding lottery (Feliciani et al., 2024). One of these, the so-called "partial lottery of fundable proposals", has been shown to be particularly effective at minimizing biased funding decisions, and has been adopted by several funders (Health Research Council of New Zealand; Science for Technological Innovation of New Zealand; the New Frontiers in Research Fund in Canada, and The British Academy). Reviewer scores are aggregated, most commonly by averaging, and a proposal enters the lottery pool if its aggregate score meets a certain threshold (in our simulation we set this threshold at the midpoint of the evaluation scale). All grants are awarded by drawing lots from the pool of eligible, positively evaluated proposals, using uniform probabilities.

3. Methods

3.1. Data

We use comprehensive data from the Novo Nordisk Foundation (NNF) covering more than a decade of activities, from fall 2012 to spring 2022. We exclude all funding instruments for infrastructures, fellowships, scholarships, incubation, product innovation and any non-research funding to focus only on calls that strictly fund scientific research grants for principal investigators. After eliminating calls that received fewer than two submissions, that awarded no grant or had incomplete or missing review scores, we retain a sample of 273 calls for scientific research grants, comprising full-text proposals, applicant information, and final funding decisions for 13,106 proposals and 46,228 reviewers' evaluations related to those proposals.

3.2. Operationalization of novelty

We operationalize two dimensions of novelty and measure them for each submitted proposal: *vintage*, capturing the recency of the content investigated in the focal proposal, and the *recombination*, capturing the degree of combinatorial creativity involved in the focal proposal. Here we provide an overview of our indicators; a detailed explanation is provided in Section (a) of the Supplementary Information.

To compute our indicator of *vintage*, we extrapolated from the text of each proposal the related Medical Subject Headings (MeSH). MeSH are metadata expressed in the form of terms, or keywords, from a curated taxonomy that maps the fields of life sciences – the target domain of NNF grants. Borrowing from previous work that operationalized novelty (vintage) similarly, we constructed a binary indicator of vintage that counts proposals as novel if they present at least one new MeSH term – that is, a MeSH term that did not yet appear in the published literature at the time the proposal was submitted to NNF (Azoulay et al., 2011; Boudreau et al., 2016; Mishra & Torvik, 2016).

To compute our indicator of *recombination* we instead relied on the indicator of "recombinant novelty" proposed by Shibayama et al. (2021). In essence, this indicator captures the semantic distance among the papers cited by proposals: novel proposals are those with semantically distant references. To measure the recombination of a given proposal, an algorithm first extracts the titles of the references cited in the reference list; then, word embeddings are used to calculate the semantic distance between these titles. It should be noted that not all proposals in our dataset contained identifiable references, and we could thus only calculate this indicator for 9,115 out of the 13,106 proposals.

As efforts to validate and systematize the landscape of novelty indicators are still underway (Arts et al., 2023), in the Supplementary Information we repeat our analyses with a broader set of measures which give a more robust picture of the results and show all prominent differences between the novelty dimensions of vintage and recombination.

The alternative measures of vintage and recombination (Section (a) of the Supplementary Information) include:

i) An indicator of vintage based on Arts et al. (2023) that uses N-grams in place of MeSH keywords. N-grams include all single words (e.g. "CRISPR"), pairs of words (e.g. "neural network") or three-word strings (e.g. "large language model"), that appear in a proposal's text. Vintage is the tally of N-grams not previously appearing in the published literature.

ii) An indicator of recombination based on the semantic distance between the *abstracts* of the publications found in the reference list. This indicator was proposed and tested as an alternative metric by Shibayama et al. (2021).

iii) An indicator to simultaneously capture vintage and recombination. Inspired by Azoulay et al. (2011), we identify all new *pairs* of MeSH terms related to a proposal that have not appeared together in the published literature prior to the proposal's date of submission. This indicator seeks to capture the first time that any pair of ideas, concepts or methods were adopted together in a research project.

3.3. Empirical estimates

To achieve our first objective of assessing whether there is evidence of conservatism in funding decisions we compare the novelty of submitted proposals with the novelty of awarded proposals. In Section (b) of the Supplementary Information we also show, more formally, the estimates for a probit model with a binary variable of "funding success" as dependent variable, and novelty (vintage or recombination) as independent variable, and interaction term to capture the moderating role of the type of call (regular vs novelty-dedicated). We expect a negative correlation between novelty and funding success, constituting evidence of conservatism in funding decisions, and we expect a moderating role of the type of call, such that conservatism is more pronounced in regular calls than in novelty-dedicated calls. We control for several confounders: applicants' gender (Bornmann et al., 2007; Hofstra et al., 2020; Jørgensen & Christensen, 2022); their age (treated as binary, with a cut-off threshold at 35 years of age); year of submission; rejection rate within the funding call; average score received from the reviewers; grant size (i.e. the inflation-adjusted applied amount). A more complete description of these control variables is provided in Section (b) of Supplementary Information.

3.4. Simulation of counterfactuals

To test whether interventions based on aggregation rules, golden tickets and funding lotteries may curb conservatism, we cannot rely on real-world data, but we can project the outcomes by running computer-based counterfactual simulations. This entails simulating which projects would have received funding if one of the alternative selection mechanisms was used instead of traditional panels, holding all other elements constant, i.e. the number of submissions, the scores they received, and the funding rate. Counterfactual funding decisions are obtained from the scores provided by reviewers, following the protocol prescribed by each intervention.

After running 100 independent Monte Carlo simulations for each regular NNF call, we compute the average novelty of each projected funded set, obtaining a distribution of the levels of novelty projected from each intervention.

In our simulations we also track the overlap between counterfactual decisions and the funding decisions made by the real panels. This gives us a sense of how impactful each simulated intervention can be on the decisions of real-world panels: some interventions may hardly differ from the funding decisions made in regular, real-world panel meetings, whereas other interventions may lead to the funding of proposals that are largely different from the proposals funded by real-world panels. We measure the similarity between the decisions by counterfactual and real-world panels using Cohen's Kappa.

Simulating counterfactuals is certainly convenient in domains where data is scarce and experiments impossible or impractical to run, which explains their growing popularity in the field of science of science (Feliciani et al., 2019; Fortunato et al., 2018). However, simulations also pose challenges. One of the challenges is the handling of ties in the ranking. To ensure that our counterfactual simulations have the exact same funding rate as the actual NNF call they replicate, our simulated panels must sometimes face decisions among equally strong proposals. In real panel meetings these ties are resolved by real reviewers who re-examine the tied proposals and make a final determination. In our simulations, we need to assume that counterfactual panels will instead choose at random uniform.

There is also a theoretical issue with counterfactual simulations that deserves attention. NNF applicants participated in calls that used the standard NNF selection mechanism. Assuming that applicants learn from past decisions, the adoption of a particular selection mechanism seen as more lenient towards novel research may cause applicants to adjust to the new incentives and adapt the quality of the proposals by submitting a different, perhaps more daring proposal than they would otherwise. In Section 5 we will discuss the significance of this issue and its remedies.

4. Results

4.1. Evidence of conservatism

The first question we address is whether funding decision at NNF are conservative. Regarding regular NNF calls, Figure 1 shows that granted proposals tend to be less novel (exhibiting older vintage and lower recombination) than the proposals that were declined. This holds true in our formal hypothesis test: the effect is statistically significant at conventional confidence levels, after controlling for a broad set of potential confounders (Tables 2-3 in the Supplementary Information). In sum, we find robust evidence of conservatism in the funding decisions for regular calls in our sample.

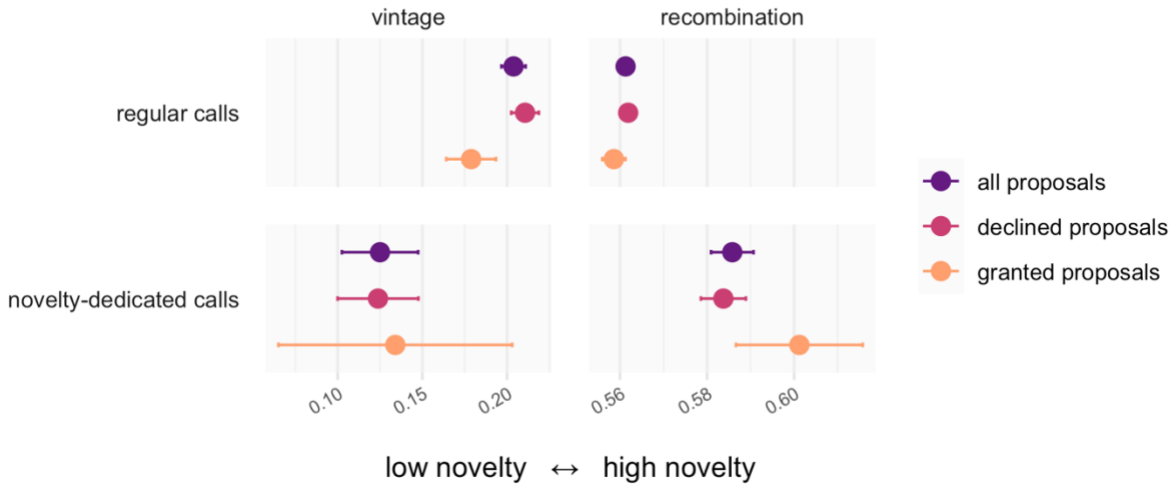


Figure 1. The X-axis shows the average novelty (vintage, on the left, and recombination, on the right) of: submitted proposals (dark, purple points); declined proposals (medium, magenta); and granted proposals (light, orange). Horizontal bars show 95% confidence intervals. The Y-axis distinguishes between proposals submitted to regular NNF calls versus novelty-dedicated NNF calls.

Some nuance emerges if we also consider novelty-dedicated calls. Figure 1 shows that, in these calls, the novelty (both vintage and recombination) of funded projects is on average higher than, or at least within the confidence bands of, the average novelty of declined proposals. This already suggests that review panels in novelty-dedicated calls may be less conservative (if at all) than panels in regular calls.

We also note from Figure 1 that novelty-dedicated calls, compared to regular calls, receive submissions that score on average *higher* on the recombination dimension of novelty, but *lower* on the vintage dimension. The explanation for this difference is that the novelty-dedicated calls run by NNF target explicitly interdisciplinary research. The interdisciplinary flavor of these calls makes it so applicants prioritize combining ideas or methods from different disciplines in novel ways – as is reflected in Figure 1 by the higher level of recombination. On the other hand, there is a high variability in the vintage of the topics investigated, with a good share of proposals recombining knowledge from old topics. One explanation is that working in interdisciplinary teams may be easier in mature fields of study: collaborations in teams that rarely work together may make it more difficult, or at least not a priority, to contribute brand new ideas, theories or methods that are novel to all fields involved.

4.2. Evaluating interventions

Our second objective is to assess the effectiveness of the four interventions to the selection mechanisms with regard to mitigating conservatism. Figure 2 presents a comparison of the average novelty of projects funded in regular NNF calls (blue), in novelty-dedicated calls (light blue) and in simulated counterfactual regular calls (white). To evaluate the efficacy of each intervention we compare the distribution of average funded novelty with that of regular NNF calls (i.e. our baseline).

Starting with novelty-dedicated calls, we find that this intervention is effective in boosting recombination, as confirmed by a significant T-test result ($p<0.001$). It is not as effective, though, in promoting new topics, as shown by older average vintages.

On the other hand, counterfactual calls simulated using aggregation rules in place of traditional panels show no meaningful improvements on either dimension of novelty. It is also noteworthy that there is minimal variation between the different rules, with aggregation by best score performing only marginally better than aggregation by mean and median. A possible explanation for the consistent behavior across different aggregation rules is in the similarity between the assessments from different reviewers. Peer review is often criticized for its poor inter-rater reliability, meaning that reviewers frequently disagree when assessing the same proposals (Graves et al., 2011; Pier et al., 2018). Despite disagreements, reviews still tend to correlate somewhat, and so the choice of using the mean, median, or highest score will likely yield similar outcomes.

Simulated golden tickets yield an average level of novelty that differs very little from the average novelty of projects funded through regular calls. They only improve the average vintage of funded projects, though to a limited degree. Finally, simulated funding lotteries improve the average vintage of funded projects, with no perceivable change in the degree of recombination compared to regular calls.

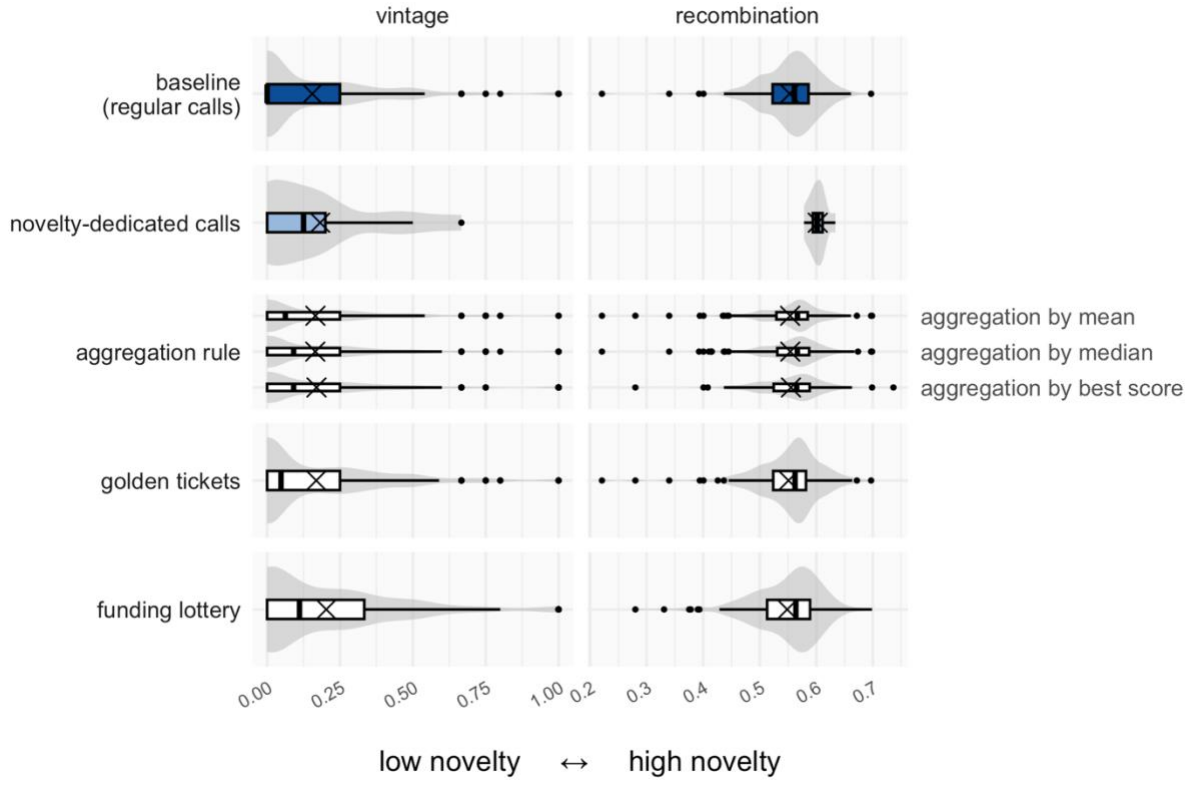


Figure 2. Violins (gray) and boxplots, on the X-axis, display the distribution average novelty of funded proposals. The left-hand panel shows the novelty dimension of vintage, whereas the right-hand panel shows recombination. The Y-axis distinguishes between regular NNF calls (denoted "baseline", blue boxplot), novelty-dedicated NNF calls (light blue boxplot), and counterfactual calls (white boxplots). The middle line within each boxplot indicates the median of the underlying distribution; a cross represents the mean.

Overall, novelty-dedicated calls stand out as the most effective intervention to improve recombination over the standard selection mechanisms of regular NNF calls. Novelty-dedicated calls, however, do not necessarily boost newer topics (more recent vintage). Funding lotteries are the most effective approach to increase funding for newer topics, followed by golden tickets, but, in turn, neither funding lotteries nor golden tickets have a meaningful impact on recombination.

Regardless of the selection mechanisms adopted, we can expect that a share of the proposals of poor quality would always be rejected, and a share of the proposals of very high quality would always be accepted. Before closing, we additionally want to consider the size of the differential impact that these interventions would have on the funding decisions. We estimate what proportion of proposals would change status – from *funded* to *declined* or vice versa – under the alternative selection mechanisms compared to the traditional sitting panel. Figure 3 shows, on the X-axis, the degree of similarity between the decisions by simulated panels and the decisions by their real-world counterpart. Different counterfactual interventions have a different impact on panel decisions. Aggregation by mean, median and golden tickets would lead to funding decisions that results in funding decisions that closely align with those made by traditional panel meetings (high Cohen's Kappa). In contrast, funding lotteries and, to a more limited degree, aggregation by best score, lead to very different funding decisions (low Cohen's Kappa). This means that they enable the funding (or, conversely, the rejection) of many projects that would not be funded (or rejected) by traditional panel meetings, which could be considered a disadvantage of funding lotteries and aggregation by best score.

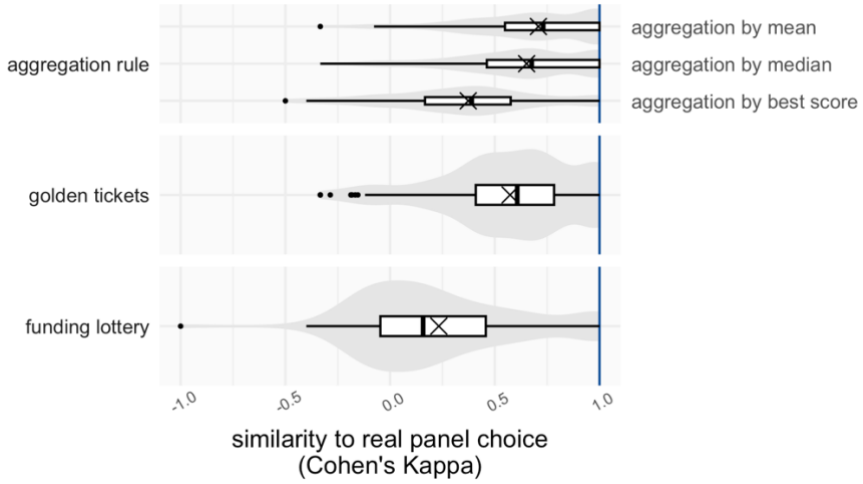


Figure 3. Violins (gray) and boxplots (white) illustrate the distribution of similarity between decisions made by counterfactual panels and real funding decisions, for each intervention listed on the Y-axis. The vertical line at X=1 (blue) represents the maximum theoretical level of similarity, achieved when counterfactual panels make the exact same funding decisions as their real counterpart. The middle line within each boxplot indicates the median of the underlying distribution; a cross indicates the mean.

5. Conclusion and discussion

We have examined four interventions to the selection mechanism for grant proposals and have evaluated their effectiveness in countering conservatism in funding decisions. Our results indicate that none of these interventions acts as a universal solution: interventions that excel in certain areas appear to perform poorly in other areas.

One of these interventions is the institution of novelty-dedicated calls. These calls are expected to mitigate conservatism by encouraging the submission of more novel proposals and by instructing reviewers to especially prioritize novel research. Our results from a natural experiment at NNF support these expectations only partially. One the one hand, compared to regular calls, novelty-dedicated calls attract more creative submissions, and conservatism appears less influential in funding decisions. On the other hand, the effectiveness of novelty-dedicated seems to be limited to a specific dimension of novelty, which we refer to as *recombination*. For another dimension of novelty, *vintage*, this intervention appears ineffective; in fact, regular calls tend to fund projects of newer vintage more frequently than novelty-dedicated calls.

The opposite trend is observed for other interventions, such as golden tickets and funding lotteries. These interventions aim to reduce the potentially risk- and novelty-averse panel discussions by limiting their role in final funding decisions. Results from our computer simulation experiments suggest that golden tickets and funding lotteries are effective in promoting research projects of newer vintage; however, they do not enhance recombination in funded projects. A potential drawback, particularly with funding lotteries, is that they often lead to funding decisions that differ significantly from those made by real panels. Many projects that were declined (or, conversely, awarded) by real panels would have received opposite outcomes under a lottery-based selection mechanism.

Lastly, the intervention based on aggregation rules – where panel deliberation is bypassed entirely by using mathematical formulas to rank proposals – proved ineffective overall. In simulations none of the simple aggregation rules we experimented with had any meaningful impact on conservatism.

In sum, to promote novel science, funders might consider implementing novelty-dedicated calls, utilizing golden tickets, or running funding lotteries. Since these interventions seem effective towards specific dimensions of novelty, the choice of selection mechanism should be made to match the mandate and objectives of the funder and the funding scheme.

While novelty is the primary focus of this article, it is important to point out that it is not only aspect funders should consider when designing the selection mechanism for a funding scheme. Other important considerations may include, for example, the costs of these interventions, their vulnerability to gaming or misuse by applicants or reviewers, and their potential direct or indirect effects on the distribution of funding across disciplines, institutions, or social groups. How these and other important considerations relate to one another seems an important area for follow-up research.

We also need to point out again an important limitation of our counterfactual simulation experiments. Implementing one or the other selection mechanism will influence the incentive structure of applicants and reviewers, ultimately changing their behavior. For example, applicants might be more willing to submit more – and more novel – proposals if a funder adopts funding lotteries. This might in turn reflect on the workload of reviewers, and thus the quality of their

assessments and, ultimately, the funding decisions. In our simulations we have momentarily put aside these and similar issues arising from changed incentives. We have instead experimented with "ceteris paribus" counterfactual worlds, where interventions on the selection mechanism do not have upstream effects on the behavior of applicants and individual reviewers. Fortunately, simulations allow to explore scenarios to account for these effects. Exploring these is the natural next step for our research.

In conclusion, we call on those funders who adopt – or intend to adopt – these or similar selection mechanisms to monitor the implementations and outcomes of their interventions. Furthermore, it is crucial that these data be made accessible to science-of-science researchers, as this is the only way to enable further research on these interventions.

Supplementary Information

a. Vintage and recombination: alternative measures

In this section we report the implementation details of our measurements of novelty (vintage and recombination) and replicate the figures presented in the article (Figures 1-3) with the alternative measurements.

These are the labels and descriptions of the indicators we measured:

i. *Vintage – MeSH terms*. This is the main indicator of vintage that we used in the article. It is a binary variable that identifies all proposals that can be assigned a new MeSH term. The taxonomy of MeSH terms is developed and curated by the NIH National Library of Medicine, which also provides the tools and infrastructures (e-utils and PubMed) we used for generating MeSH metadata. We assigned MeSH terms to proposals algorithmically based on the proposal's title and description, leveraging the NIH-NLM MeSH-on-Demand APIs. Then, we queried the NIH-NLM MeSH descriptors databases to find the date in which each of the MeSH terms was added to the taxonomy. If any of the MeSH terms assigned to a proposal was created after the date of submission of the proposal, we considered the proposal 'novel', coding vintage = "1" (and "0" otherwise). For some proposals we only knew the year of submission to a NNF call instead of the exact date. In these cases, we assumed the date to be January 1st of the corresponding year.

ii. *Recombination – title*. This is our main indicator of recombination. To calculate the recombination of a given proposal we first extracted the titles of the references cited in the reference list. Then we used pre-trained word embeddings from scispaCy (Neumann et al., 2019), a specialized NLP (natural language processing) model for scientific text and biomedical text specifically. Word embeddings are used to represent titles in high-dimensional vectors which capture their semantic meaning. Recombination was measured as the cosine distance among the vector representations of the reference titles. It should be noted we could calculate this indicator for 9,115 out of 13,106 proposals, as not all proposals had an identifiable reference.

iii. *Vintage – N-grams*. An indicator of vintage based off Arts et al. (2023). We calculated it by identifying all single words, two-word strings (bigrams), and three-word strings (trigrams) – collectively referred to as N-grams – that appear in a proposal's text. We queried OpenAlex to search for these N-grams in the literature published prior to the date of submission of the proposal. If we found no prior occurrences, then we marked the N-gram as "new". Upon close inspection we found that the identification of new N-grams is prone to false positives: many N-grams marked as new are in fact common utterances in academic writing, e.g. "work package investigate[s]". We sought to mitigate this issue through a mix of strategies. First, we manually cleaned the most extreme outliers (false positive that occurred in many proposals). Second, we devised ad-hoc regular expressions to filter out common false positive. Third, we dichotomized this variable to code as "novel" all proposals that had at least *two* new N-grams, so that one false new N-gram could not wrongly lead us to mark a proposal as novel.

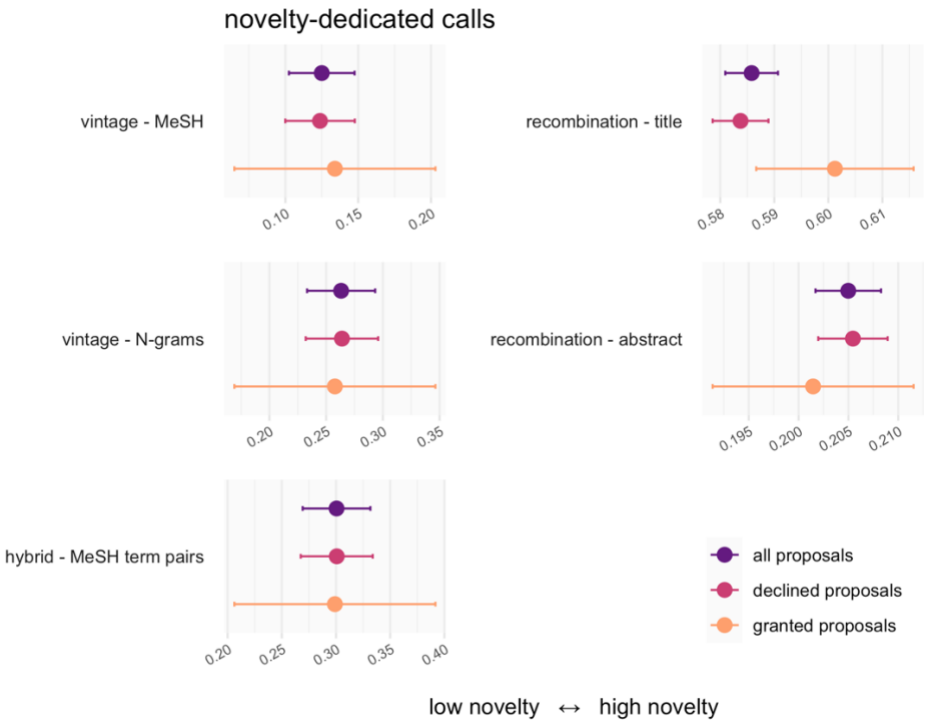
iv. *Recombination – abstract*. This is calculated in the same way as "recombination – title", the only difference being the input text (Shibayama et al., 2021). Here, the semantic distance between references is not based on the title of cited references, but their abstract.

v. *Hybrid – MeSH term pairs.* This indicator captures both vintage and recombination by identifying if a proposal has any new *pair* of MeSH terms. We queried the PubMed APIs by the NIH to determine which pairs of MeSH terms were new. This hybrid measure of novelty is binary: we assigned "1" to propels that had at least one new pair of MeSH terms.

We replicated the results shown in the main text using all five operationalizations of novelty. Appendix Figures 1 and 2 replicate Figure 1 in the main article. These figures show, across all measures, evidence of conservatism in regular calls, whereby granted proposals are on average less novel than declined proposals (Appendix Figure 1). The same cannot be said for novelty-dedicated calls (Appendix Figure 2). For novelty-dedicated calls, some novelty indicators suggest a mild conservative tendency (recombination - abstract); some suggest that novelty is rewarded (recombination - title), and the rest show no meaningful difference at all. In sum, for novelty-dedicated calls we do not find evidence for conservatism.

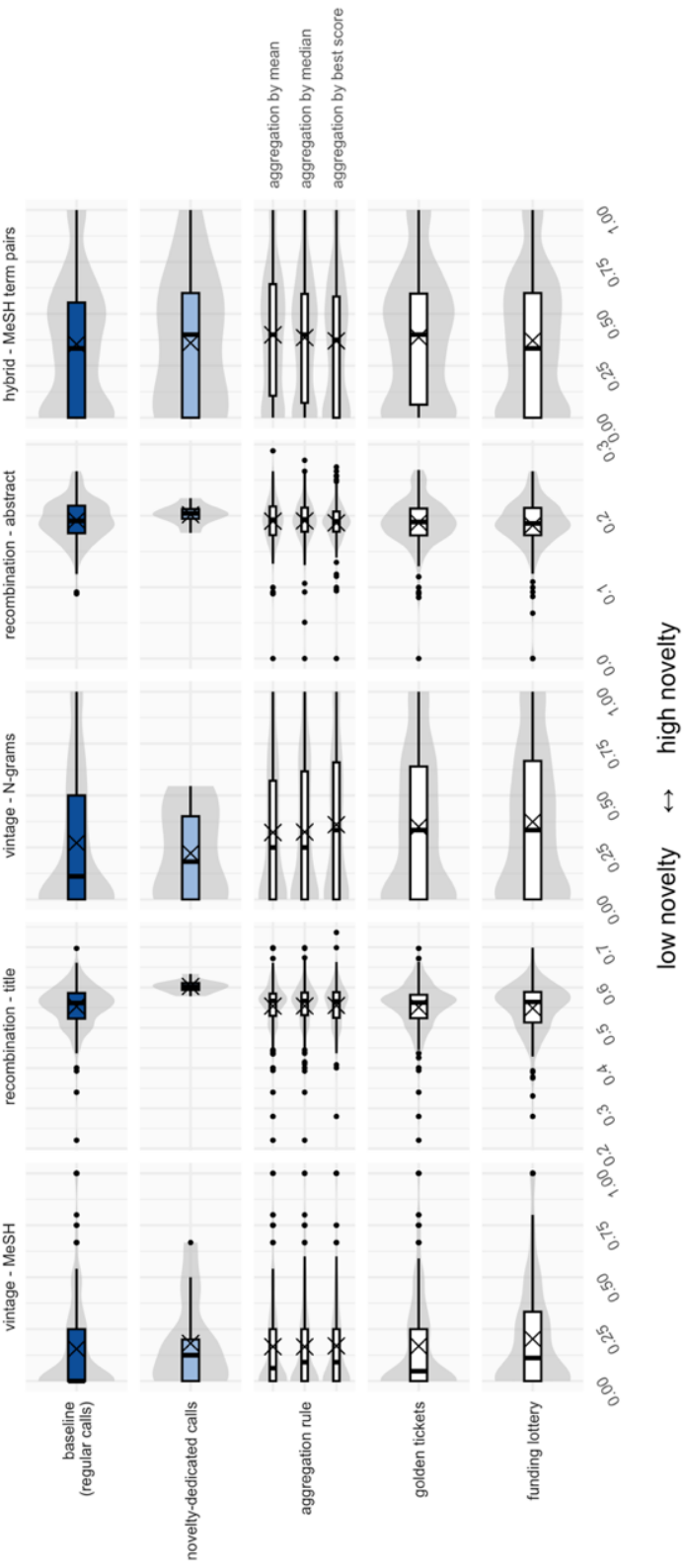


Appendix Figure 1. Average novelty of submitted proposals (dark, purple points); declined proposals (medium, magenta); and granted proposals (light, orange) in regular NNF calls. Different plots report different operationalizations of novelty. Horizontal bars show 95% confidence intervals.



Appendix Figure 2. Average novelty of submitted proposals (dark, purple points); declined proposals (medium, magenta); and granted proposals (light, orange) in novelty-dedicated NNF calls. Different plots report different operationalizations of novelty. Horizontal bars show 95% confidence intervals.

Appendix Figure 3 replicates Figure 2 of the main article using all available operationalizations of novelty. This wider array of measures confirms the main results we reported: novelty-dedicated calls improve recombination in funded projects but has no meaningful impact on vintage. Conversely, golden tickets and funding lotteries from counterfactual simulations do lead to the funding of projects of newer vintage but have no effect on recombination. Lastly, aggregation rules seem to have little impact overall.



Appendix Figure 3. Violins (gray) and boxplots, on the X-axis, display the distribution average novelty of funded proposals for different operationalization of novelty (columns). Rows distinguish between regular NNF calls (denoted "baseline", blue boxplot), novelty-dedicated NNF calls (light blue boxplot), and counterfactual calls (white boxplots). The middle line within each boxplot indicates the median of the underlying distribution; a cross represents the mean.

b. Evidence of conservatism

We fitted a probit model to formally test for conservatism in NNF panels. Here we outline dependent and independent variables, the fitted model and the resulting regression coefficients.

This is the list of variables we included, in addition to the main independent variables "vintage - MeSH" and "recombination – title".

- Funding success is our dependent variable. We treat it as binary, coding 1 all proposals that were awarded a grant (even if applicants subsequently declined the grant), and 0 otherwise (rejection).
- Male PI (i.e. Principal Investigator). In our data, gender is self-reported and has some missing values. In the light of this, we constructed a binary variable coding "1" the advantaged group, i.e. proposals whose main applicant indicated "male". Non-male main applicants and unknowns are coded "0".
- Young PI. We constructed a binary variable to identify young main applicants. We define young applicants those with an age of 34 or younger at the time of submission.
- Competition. The "competition" faced by a given application is the rejection rate in the funding call where the application competed – in other words, the number of declined applications divided by the number of all submitted applications.
- Average score received. This is the arithmetic mean of the grade issued by individual reviewers. Reviewers used Likert scales where low values signify better grades. Up to circa 2018 NNF reviewers were provided with five-point Likert scales and moved to using six-point scales in 2018. We observed that the worst grade, 6, was very seldomly used, so we made no particular adjustment to correct for this change of scale.
- Applied amount. This captures the size of the grant. Many proposals in our dataset do not report any applied amount, and some simply report "0", which we also treat as missing values. To reduce the number of missing values we adopted the following imputation strategies, in this order. First we tried to replace the missing applied amount with the awarded amount. When this was not available, we used call-level average requested amount. And when this was also not available, we used the average amount requested in similar calls in that same year. These imputation approaches brought down the number of missing values from 1,341 to 579. Lastly, the applied amounts were in Danish krone (DKK). We thus adjusted the applied amounts for inflation in Denmark, and log-transformed the amounts to compensate for kurtosis.
- Year. We treat each year as a dummy variable and include all dummies as fixed effects.

We fitted three models using vintage as the main independent variable (Table 2) and three additional models using recombination instead (Table 3). In all models, the dependent variable is funding success.

In Model (1), the independent variables include fixed effects for novelty (vintage or recombination, as the case may be) and a dummy variable for call type (0 = regular, 1 = novelty-dedicated). In Model (2) we test the moderating role of call type by interacting novelty with call type, distinguishing between novelty in regular calls (e.g., "Regular call × vintage") and in novelty-dedicated calls ("Novelty-dedicated call × vintage"). Model (3) extends this by adding the control variables.

Overall, we find evidence of conservatism in regular calls, as indicated by the significant, negative coefficient of "Regular call × vintage" (Table 2) and "Regular call × recombination" (Table 3). This effect appears robust to various confounders. In contrast, evidence of conservatism in novelty-dedicated calls is somewhat more ambiguous (see "Novelty-dedicated call × vintage" in Table 2 and "Novelty-dedicated call × recombination" in Table 3).

Table 2. Fitted probit model with vintage (vintage – MeSH) as independent variable.

	Dependent variable:		
	funding success (granted)		
	(1)	(2)	(3)
Constant	-0.772***	-0.771***	-2.089***
	(0.014)	(0.014)	(0.106)
Vintage	-0.111***		
	(0.032)		
Novelty-dedicated call	-0.407***	-0.428***	0.486***
	(0.058)	(0.062)	(0.117)
Regular call × vintage		-0.117***	-0.125**
		(0.032)	(0.048)
Novelty-dedicated call × vintage		0.048	-0.142
		(0.169)	(0.282)
Male PI			-0.069·
			(0.039)
Young PI			0.008
			(0.088)
Competition			-0.878***
			(0.036)
Average score received			-1.744***
			(0.036)
Applied amount			0.071**
			(0.022)
Year 2013			0.232*
			(0.116)
Year 2014			0.352**
			(0.114)
…			…
Year 2021			0.521***
			(0.109)
Year 2022			1.463***
			(0.183)
Observations	13,106	13,106	12,121
Log Likelihood	-6,661.414	-6,660.967	-2,865.159
Akaike Inf. Crit.	13,328.830	13,329.930	5,768.318
Note:	· p<0.1; * p<0.05; ** p<0.01; *** p<0.001		

Table 3. Fitted probit model with recombination (recombination – title) as independent variable. Due to missing reference lists in proposals submitted before 2015 and in 2022, recombination could not be calculated for these years, resulting in some year fixed effects not being computed in Model 3.

	Dependent variable:		
	funding success (granted)		
	(1)	(2)	(3)
Constant	-0.920***	-0.920***	-1.728***
	(0.016)	(0.016)	(0.084)
Recombination	-0.016		
	(0.015)		
Novelty-dedicated call	-0.262***	-0.319***	0.283·
	(0.064)	(0.070)	(0.149)
Regular call × recombination		-0.025	-0.073**
		(0.016)	(0.024)
Novelty-dedicated call × recombination		0.136*	-0.009
		(0.066)	(0.154)
Male PI			-0.208***
			(0.048)
Young PI			-0.015
			(0.118)
Competition			-0.794***
			(0.055)
Average score received			-1.853***
			(0.045)
Applied amount			0.046
			(0.030)
Year 2016			-0.069
			(0.097)
Year 2017			-0.001
			(0.094)
…			…
Year 2021			0.136
			(0.090)
Observations	9,296	9,296	8,565
Log Likelihood	-4,291.874	-4,288.927	-1,912.546
Akaike Inf. Crit.	8,589.748	8,585.854	3,855.092
Note:	· p<0.1; * p<0.05; ** p<0.01; *** p<0.001		

References

Arts, S., Melluso, N., & Veugelers, R. (2023). *Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text*. https://doi.org/10.48550/ARXIV.2309.16437

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: Evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

Bieri, M., Roser, K., Heyard, R., & Egger, M. (2021). Face-to-face panel meetings versus remote evaluation of fellowship applications: Simulation study at the Swiss National Science Foundation. *BMJ Open*, *11*(5), e047386. https://doi.org/10.1136/bmjopen-2020-047386

Bornmann, L., Leydesdorff, L., & Van den Besselaar, P. (2010). A meta-evaluation of scientific research proposals: Different ways of comparing rejected to awarded applications. *Journal of Informetrics*, *4*(3), 211–220. https://doi.org/10.1016/j.joi.2009.10.004

Bornmann, L., Mutz, R., & Daniel, H. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238. https://doi.org/10.1016/j.joi.2007.03.001

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. *Management Science*, *62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Csaszar, F. A., & Eggers, J. P. (2013). Organizational Decision Making: An Information Aggregation View. *Management Science*, *59*(10), 2257–2277. https://doi.org/10.1287/mnsc.1120.1698

Fang, F. C., & Casadevall, A. (2016). Research Funding: The Case for a Modified Lottery. *mBio*, *7*(2), e00422-16, /mbio/7/2/e00422-16.atom. https://doi.org/10.1128/mBio.00422-16

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). https://doi.org/10.1007/s11192-019-03205-w

Feliciani, T., Luo, J., & Shankar, K. (2024). Funding lotteries for research grant allocation: An extended taxonomy and evaluation of their fairness. *Research Evaluation*, *33*(1), rvae025. https://doi.org/10.1093/reseval/rvae025

Feliciani, T., Morreau, M., Luo, J., Lucas, P., & Shankar, K. (2022). Designing grant-review panels for better funding decisions: Lessons from an empirically calibrated simulation model. *Research Policy*, *51*(4), 104467. https://doi.org/10.1016/j.respol.2021.104467

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science*, *359*(6379), eaao0185. https://doi.org/10.1126/science.aao0185

Franzoni, C., Stephan, P., & Veugelers, R. (2022). Funding Risky Research. *Entrepreneurship and Innovation Policy and the Economy*, *1*, 103–133. https://doi.org/10.1086/719252

Galton, F. (1907). One Vote, One Value. *Nature*, *75*(1948), 414–414. https://doi.org/10.1038/075414a0

Graves, N., Barnett, A. G., & Clarke, P. (2011). Funding grant proposals for scientific research: Retrospective analysis of scores by members of grant review panel. *BMJ*, *343*(sep27 1), d4797–d4797. https://doi.org/10.1136/bmj.d4797

Greenberg, D. S. (1998). Chance and grants. *The Lancet*, *351*(9103), 686. https://doi.org/10.1016/S0140-6736(05)78485-3

Hastie, R., & Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. *Psychological Review*, *112*(2), 494–508. https://doi.org/10.1037/0033-295X.112.2.494

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The Diversity–Innovation Paradox in Science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291. https://doi.org/10.1073/pnas.1915378117

Jørgensen, M. A., & Christensen, R. N. (2022). *Gender Composition in NNF Funded Publications and Scientific Excellence*.

Keldke, T. M., Fosse, H. B., & Christensen, R. N. (2023). *Bias Against Novelty* (Working Paper Series). The Novo Nordisk Foundation.

Lane, J. N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E. C., & Lakhani, K. R. (2022). Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation. *Management Science*, *68*(6), 4478–4495. https://doi.org/10.1287/mnsc.2021.4107

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation*, *21*(1), 48–60. https://doi.org/10.1093/reseval/rvs001

Mishra, S., & Torvik, V. I. (2016). Quantifying Conceptual Novelty in the Biomedical Literature. *D-Lib Magazine*, *22*(9/10). https://doi.org/10.1045/september2016-mishra

Ottaviani, M., & Sørensen, P. (2001). Information aggregation in debate: Who should speak first? *Journal of Public Economics*, *81*(3), 393–421. https://doi.org/10.1016/S0047-2727(00)00119-5

Packalen, M., & Bhattacharya, J. (2020). NIH funding and the pursuit of edge science. *Proceedings of the National Academy of Sciences*, *117*(22), 12011–12016. https://doi.org/10.1073/pnas.1910160117

Peng, H., Qiu, H. S., Fosse, H. B., & Uzzi, B. (2024). Promotional language and the adoption of innovative ideas in science. *Proceedings of the National Academy of Sciences*, *121*(25), e2320066121. https://doi.org/10.1073/pnas.2320066121

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. *Proceedings of the National Academy of Sciences*, *115*(12), 2952–2957. https://doi.org/10.1073/pnas.1714379115

Pina, D. G., Buljan, I., Hren, D., & Marušić, A. (2021). A retrospective analysis of the peer review of more than 75,000 Marie Curie proposals between 2007 and 2018. *eLife*, *10*, e59338. https://doi.org/10.7554/eLife.59338

Shaw, J. (2023). Peer Review, Innovation, and Predicting the Future of Science: The Scope of Lotteries in Science Funding Policy. *Philosophy of Science*, 1–15. https://doi.org/10.1017/psa.2023.35

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PLOS ONE*, *16*(7), e0254034. https://doi.org/10.1371/journal.pone.0254034

Shin, H., Kim, K., & Kogler, D. F. (2022). Scientific collaboration, research funding, and novelty in scientific knowledge. *PLOS ONE*, *17*(7), e0271678. https://doi.org/10.1371/journal.pone.0271678

Singh Chawla, D. (2023a). 'Golden tickets' on the cards for NSF grant reviewers. *Nature*, d41586-023-00579-z. https://doi.org/10.1038/d41586-023-00579-z

Singh Chawla, D. (2023b). US agency launches experiments to find innovative ways to fund research. *Nature*, d41586-023-03933–3. https://doi.org/10.1038/d41586-023-03933-3

Stephan, P., Veugelers, R., & Wang, J. (2017). Reviewers are blinkered by bibliometrics. *Nature*, *544*(7651), 411–412. https://doi.org/10.1038/544411a

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474

van den Besselaar, P., Sandström, U., & Schiffbaenker, H. (2018). Studying grant decision-making: A linguistic analysis of review reports. *Scientometrics*, *117*(1), 313–329. https://doi.org/10.1007/s11192-018-2848-x

Veugelers, R., Wang, J., & Stephan, P. (2022). *Do Funding Agencies Select and Enable Risky Research: Evidence from ERC Using Novelty as a Proxy of Risk Taking* (w30320; p. w30320). National Bureau of Economic Research. https://doi.org/10.3386/w30320

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Wessely, S. (1998). Peer review of grant applications: What do we know? *The Lancet*, *352*(9124), 301–305. https://doi.org/10.1016/S0140-6736(97)11129-1