NBER WORKING PAPER SERIES

CHOOSE YOUR MOMENTS: PEER REVIEW AND SCIENTIFIC RISK TAKING

Richard T. Carson
Joshua S. Graff Zivin
Jeffrey G. Shrader

Working Paper 31409
http://www.nber.org/papers/w31409

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023

We acknowledge the financial support of the National Science Foundation through its SciSIP Program (Award SBE-1561257). We are grateful for comments from Peter Muennig, Matthew Neidell, and Bhaven Sampat, as well as from seminar participants at Columbia University. Special thanks to Kyle Myers and Pierre Azoulay for their help with NIH grant data as well as their valuable comments. Stephanie Khoury and Tarikua Erda provided excellent research assistance. We particularly want to acknowledge the late Jordan Louviere, who shared sage advice on experimental design issues before his untimely death. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Choose Your Moments: Peer Review and Scientific Risk Taking
Richard T. Carson, Joshua S. Graff Zivin, and Jeffrey G. Shrader
NBER Working Paper No. 31409
June 2023
JEL No. H40,O3,O38

ABSTRACT

Science funding agencies such as the NIH, NSF, and their counterparts around the world are often criticized for being too conservative, funding incremental innovations over more radical but riskier projects. One explanation for their conservatism is the way the agencies use peer review of scientific proposals. Peer review is the cornerstone of research allocation decisions, but agencies typically base decisions on a simple average of peer review scores. More novel ideas are less likely to gain consistently high ratings across evaluators and are less likely to be funded. Using a discrete choice experiment conducted with a large sample of active biomedical researchers, we find that—in contrast to funding agencies—scientists systematically prefer to fund projects with more reviewer dissensus. Rather than purely focusing on the first moment of the distribution of reviewer scores, they also value the second moment. Further, scientists with the greatest domain expertise on a proposal are more enthusiastic about dissensus, and while appetite for dissensus shrinks as budgets become tighter, it does not disappear completely. Applying our estimates to prior studies mimicking NIH's review process shows that incorporating expert scientists' preferences for dissensus would change marginal funding decisions for ten percent of projects worth billions of dollars per year.

Richard T. Carson
Department of Economics
University of California, San Diego
ECON 323
9500 Gilman Dr. #0508
La Jolla, CA 92093-0508
rcarson@ucsd.edu

Jeffrey G. Shrader
Columbia University
420 W 118th St.
New York, NY 10027
jgs2103@columbia.edu

Joshua S. Graff Zivin
University of California, San Diego
9500 Gilman Drive, MC 0519
La Jolla, CA 92093-0519
and NBER
jgraffzivin@ucsd.edu

A data appendix is available at http://www.nber.org/data-appendix/w31409

I. Introduction

Fundamental scientific knowledge and the technologies built on it significantly contribute to aggregate income and economic growth (Nelson and Phelps 1966; Lucas 1988; Romer 1990; Aghion and Howitt 1992; Mokyr 1992). The public good nature of basic scientific discovery implies that the government should play a prominent role in its funding, which should, in turn, catalyze private-sector investments in applied science (Arrow 1972; Nelson 1959; Bush 2020). Indeed, the U.S. federal government invested more than $95 billion into science funding in 2021 alone (National Science Foundation (NSF) 2023),[1] with the vast majority allocated based on some form of peer review process. Peer review is likewise the cornerstone of governmental research allocation decisions across the globe (Whitley and Gläser 2007), as well as grant awards from science-based philanthropic organizations[2] and firms' internal R&D decisions (Miller 1995).

Despite the ubiquity of peer review, previous research has left open important questions about the best way to incorporate the outputs from peer review into decisions about the allocation of scarce resources (Franzoni and Stephan 2022). This is especially true if the goal is to produce novel or transformative science (Sen 2014; Boudreau et al. 2016), with science agencies having long been criticized for being too conservative in their research funding decisions (Nicholson and Ioannidis 2012). In this paper, we study the aggregation of individual peer review evaluations and the implications of translating those evaluations into decisions about which projects in a given area get funded.

The specific focus of our work is the U.S. National Institutes of Health (NIH). NIH is the world's largest funder of research in the life sciences, distributing more than $30B in funding each year, with most spent on basic research (Moses et al. 2005). Virtually all of NIH's funding decisions are based on the results of a highly structured peer review process that can be broken down into three parts: (1) allocation of funding across broad research areas, where Congress and the Executive Branch play a large role (Science News Staff 2022), (2) a peer review process for projects within areas (Lee et al. 2013; Li 2017; Pier et al. 2018), and (3) the mechanism for using these peer reviews to inform funding decisions.

Our focus here is on point (3), which has received little attention. NIH, like many organizations around the world (Guthrie, Ghiga, and Wooding 2018), makes decisions based primarily on the first moment of the distribution of scores from peer reviewers. Specifically, NIH elicits reviews from the panel of peer reviewers, calculates the average score across the

[1] This amount is larger than the GDP of two-thirds of the world's economies. Including intramural funding for research conducted inside the federal government adds a further $34 billion.

[2] Note that philanthropic grants are distinct from (targeted) gifts made to universities, which often reflect different decision-making process such as naming rights or familial experience with a disease (Murray 2013).

reviewers, then ranks and funds projects based on that average until the budget is exhausted (NIH 2008; Azoulay, Graff Zivin, and Manso 2012; Lauer 2023). This mechanism tends to result in granting funding to projects with consistent high marks across evaluators. It is widely thought to favor incremental innovation over more radical ideas that could yield high payoffs but are less likely to yield consensus in the review process and, in particular, is vulnerable when there is a sharp division in the field on the best way forward (March 1991; Manso 2011; Azoulay, Graff Zivin, and Manso 2011; Nijstad, Berger-Selman, and De Dreu 2014).

Assessing whether the information from peer review scores could be aggregated to better effect would ideally entail a large randomized experiment that allocates grant applications to two or more different aggregation approaches and then tracks the outcomes that arise from those awards over a long time horizon.[3] Since such an experiment is likely to be politically infeasible, it is important to explore alternative options. One such approach could make use of variation in peer reviewer scores across funded projects to examine whether projects with greater levels of dissensus generated more pathbreaking scientific discoveries. Unfortunately, NIH has not been willing to provide researchers with access to individual reviewer scores. NSF is similarly guarded about sharing individual review scores, and comparable data on corporate and foundation R&D decisions is even more elusive. Because this makes it impossible to explore the implications of alternative approaches to synthesizing scores using programmatic data, a simulacrum is required.

We used discrete choice experiments to effectively ask scientists what they think the aggregation function should look like when evaluating grant proposals (Louviere, Hensher, and Swait 2000).[4] The participants in the experiments were active biomedical researchers with a track record of successful NIH funding, and the experiments simulated the research funding process NIH uses. Participants were presented with real (but anonymized) project abstracts and a set of experimentally assigned peer review scores for those projects. They were then asked to choose which projects they would fund with their allocated budget.

The distribution of peer review scores was randomly drawn from an experimental design that allowed us to examine the weight participants placed on various moments of the score distributions.[5] The core idea is that there is a clean null hypothesis: do participants place

[3]It is worth noting that Chiara Franzoni of Polytechnic University of Milan and Paula Stephan of Georgia State University are currently running a related experiment that explores project funding decision rules with the Novo Nordisk Foundation.

[4]Discrete choice experiments have previously been used to study R&D decisions in a private firm context (Carson et al. 2022). Recent work has also used preference elicitation to study how scientists trade off grant length versus grant size, comparing the preferences of scientists to the preferences of granting agencies (Myers and Tham 2023).

[5]In addition to filling an important data gap, the experiment afforded us reasonable power to detect

all of their weight on the mean value of scores, which is the decision rule that mirrors the current NIH approach? Our experimental design allows for reasonably powerful tests of the two alternatives that the researchers are either risk averse with respect to dissensus in reviewer ratings or risk seeking with respect to such dissensus.[6] The latter is consistent with the notion that some level of dissensus may indicate more promising but radical ideas (Ackermann 1986; Goldstein and Kearney 2017; Krieger et al. 2022). We are also able to look at other suggested deviations from NIH's mean-based funding rule.[7]

The results show that our samples of experienced biomedical scientists, on average, do not share the same objective function as the NIH. In addition to the average peer review scores of projects, they also placed value on other moments of the project score distribution. Specifically, participants were willing to trade-off a project with lower average score for one with more variance. Participants were willing to accept an average score 0.1 points lower in exchange for an increase in score variance of 1. This effect holds true even when accounting for other characteristics of the project score distribution. On average, scientists also preferred projects that had a higher skew, indicating that they preferred the presence of more right-tail scores, even at the expense of good but not great overall scores. At the same time, controlling for skewness did not eliminate scientists' preference for pure dissensus in the form of higher variance.

Armed with data from our scientists' preferences, we explore heterogeneity in their preferences and the robustness of our findings to a range of potentially important features for shaping the relationship between risk-taking and the peer review process. We first assess whether the scientists in our sample weighted negative reviews more strongly than positive reviews—a trend that has been documented in previous research and which has motivated calls for reforms which would allow individual reviewers with strong preferences to overrule potential naysayers. In contrast to previous work, we find little heterogeneity in the effects of positive versus negative reviews, suggesting that decisions based on the full set of project scores can be effectively used to support riskier projects, as long as the process places positive weights on the variance of proposal scores.

Second, we ask whether scientists in the sample preferred projects that had bimodal

preferences for these attributes. It also allowed us to investigate some of the hypothesized mechanisms that may be driving conservatism within the NIH peer review system.

[6]It is important to recognize that peer reviewers are assessing risky projects and NIH's mean score-based funding rule in that sense incorporates reviewer risk preferences. What it does not do is take into account the extra information contained in the distribution of reviewer scores. It is this extra information that participants in this experiment see in making choices concerning funding.

[7]Our use of the main types of NIH grants, which essentially have the same fixed budget within type, allows us to focus on the distribution of reviewer ratings because those reviewers do not need to make a simultaneous judgement about likely outcomes against grant cost. This is atypical in many granting agencies, although some of them also run fixed grant size competitions (e.g., NSF graduate research fellowships).

scores, a particularly extreme form of dissensus in scoring. We find that scientists did not prefer such projects relative to a model that simply accounts for high variance in scores.

Third, we use randomization in the proximity between a scientist's own research area and the research area of the projects we showed them to assess the dissensus preferences of relative experts versus outsiders (noting that the entire sample consisted of experts on relevant biomedical research). When acting as peer reviewers, previous research shows that scientists judge proposals inside their area of expertise relatively more harshly than proposals outside their area (Boudreau et al. 2016). Expert evaluators have also been found to focus first on feasibility of R&D proposals inside their own domain of expertise even at the cost of more innovative solutions (Lane et al. 2022a). These results raise concerns with review processes like those at the NIH, because expert peer reviewers might be especially unwilling to take risks on novel proposals in their area. Contrary to this concern, we find that participants who were in the best position to understand the proposal were substantially more tolerant of dissensus. The closer a proposal was to a researchers' own area of expertise, the stronger was the preference for project score variance. This novel finding on the risk-taking of insiders has important implications for the calculus that underlies the recently documented tensions between expertise and bias in the peer review process itself (Li 2017).

Fourth, we leverage results from an additional choice experiment, with an independent sample from our study population, to assess whether tighter funding budgets lead to lower dissensus tolerance. In this second study, participants were asked to construct portfolios of projects that they were willing to fund. We then administered a budget change shock by either tightening or relaxing the budget and asked them which projects they would cut from or add to the portfolio, in order to assess the characteristics of the marginal proposal. As expected, tightening the budget led participants to cut higher variance projects (those with greater dissensus in scores). The effect was not symmetric with relaxed budgets, however: a larger budget did not cause participants to notably add higher variance projects to the portfolio.

Putting things together, we assess the implications of the scientists' preferences for project funding. Using the project scores from this study, as well as three sets of expert-generated project scores repurposed from two previous studies (Pier et al. 2018; Lane et al. 2022b), we find that the funding rule based on the overall mean score and variance preferences of our successful biomedical scientists substantively alters which projects would get funded relative to the standard, mean-only NIH approach. On average across the four sets of project scores, fifty-eight percent of projects change their ranking when using the scientists' preferences versus the NIH rule. When funding constraints are tight, a ranking that incorporates both the average and variance of project scores can lead to changes in funding decisions for up

to twenty percent of projects in some settings, with an average funding reversal rate of ten percent when using the preferences from scientists with relatively greater domain-specific expertise and five percent when we expand to include the full sample of scientists.

The rest of the paper proceeds as follows. Section II describes the NIH review process and context. Section III lays out the experimental design, randomization, and recruitment procedures. Section IV gives details on the econometric model. Section V provides the results from fitting that model to the experimental data. Section VI concludes.

II. Background on NIH Peer Review

The existing NIH peer review process occurs in several stages. In the first phase, applications are assigned to study sections based on the proposal's scientific focus. These study sections are comprised of approximately twenty peer reviewers charged with assessing the quality of applications. Not all reviewers play the same role.

Each application is assigned two to five reviewers, chosen based on the relevance of their expertise, who thoroughly review the proposal before the study section meeting. These reviewers write a critique and assign preliminary scores for five distinct review criteria (significance, investigator, innovation, approach, and environment) as well as an initial overall impact score. The significance and innovation criteria are meant to assess a project's importance, while the approach, investigator, and environment criteria assess the project's feasibility and likelihood of success. Scores are based on a 9-point scale, where 1 is exceptional and 9 is poor. Reviewers are explicitly told that the overall impact score should reflect all the criteria but should capture an integrated assessment of the proposal, not simply be a mathematical sum of the parts.

The evaluations by the assigned reviewers are then used to launch a broader discussion of the proposal by the entire study section, after which the assigned reviewers can revise their preliminary scores. The remainder of the review committee is also asked to provide an overall impact score. These scores are then averaged, rounded mathematically to one decimal place, and multiplied by ten to create a final priority score. The NIH advisory council uses this priority score and the written critiques to make funding recommendations to the director of the institute or center that awards the funding. In practice, proposals are funded in descending order of score until the budget is exhausted, with out-of-order funding based on fit with the institutional mission and subjective judgments regarding application quality. Out-of-order funding occurs roughly five percent of the time (Jacob and Lefgren 2011). Since priority scores are a simple average of individual scores, they cannot reflect the intensity of individual reviewer preferences or convey any information about variation in

assessments across reviewers.

III. Study Design

The scientists in our study took part in discrete choice experiments that involved ranking research projects in terms of their priority for being funded. The first study involved choice scenarios where funding priority across four projects was decided. The experiment also contained a randomized intervention that altered the match between the participant's research area and the subject of the presented projects.

Projects were assigned a randomized set of scores from a hypothetical expert review panel. The exact scores were shown, along with the average and standard deviation of the scores. This intervention allowed us to identify the preferences of participants for different features of the score distribution. In particular, it allowed us to test whether participants preferred projects with higher average scores or had preferences for other features of the score distribution, like dissensus.

Project titles and abstracts were shown above the scores (see Section D). The titles and abstracts came from real NIH grants and were chosen to span a range of biomedical research fields. Participants were randomized into an experiment where they saw projects from either inside or outside their specialty field. This allowed us to test for differences in behavior between insiders and outsiders.

A separate set of participants was randomized into a second study experiment that involved forming portfolios under different budget constraints. That alternative experiment is described in Section B..

A. Design of Study 1: Estimating Preferences for Project Attributes

To estimate participant preferences for different distributions of project scores—particularly their preferences for consensus versus dissensus—we used a discrete choice experiment. In the experiment, each choice scenario involved ranking four proposals that had different distributions of scores from a hypothetical expert review panel. In this way, the participants were placed in the role of a NIH Scientific Review Officer, the individual who runs a study section and ultimately chooses projects to fund based on rating inputs from their study section's reviewers.

Participants were asked to complete four choice scenarios during the experiment. The choice scenarios were designed so that participants would be asked to rank projects with different average scores and score variances. Score variance was one of the main attributes of interest in the experiment because higher variance indicates greater dissensus among the

project reviewers.

Score distributions were generated using a balanced incomplete block design (BIBD), following Louviere, Flynn, and Marley (2015). BIBD designs are a type of fractional factorial design where preferences for combinations of different attributes or attribute levels are identified using a sparse matrix of choice options. We designed the BIBD to provide reasonably high power when estimating preferences over the average and standard deviation of project scores, while also allowing for estimation of preferences for other attributes of the project score distribution (e.g., number of top scores, number of bottom scores, score skewness). The BIBD did so by generating scores for ten hypothetical raters using nine different score levels. Following standard NIH practice, ratings were on a 1 to 9 scale with 1 indicating the best possible score. These ratings were reverse coded for the statistical analysis (described in Section IV.). The ratings from the ten reviewers were duplicated twice to yield thirty scores for each project. From the set of all resulting possible score distributions, fifty-four orthogonal combinations of average scores and score standard deviations were used to create the projects shown to the participants.

For each question, the participant was provided with thirty reviewer ratings, along with the computed average and variance of those ratings, for four distinct proposals. They were then asked to rank the four projects in terms of funding priority using a best/worst experiment design (Louviere, Flynn, and Marley 2015). The four projects in each of these choice scenarios were chosen to maximize power to identify preferences across the fifty-four attribute combinations. This grouping yielded 344 blocks of four projects each. See Section D for examples of the questions that participants saw. The choice scenarios were further grouped into sets of four scenarios to create eighty-six survey versions. Participants were uniformly randomized into receiving one of the versions.

A.i. Project Title and Abstract Randomization

Participants were also randomized into receiving projects whose description (title and abstract) fell inside or outside their direct area of expertise. This randomization was done independently of the randomization into different survey versions described above. The purpose of this second randomization was to assess the effect of subject area expertise or insider status on the types of projects chosen.

All individuals recruited for the study had a background in biomedical research and were part of at least one of the five NIH study sections.[8] Project titles and abstracts were selected from historical NIH R01, R35, or F32 grants listed on the Research Portfolio Online Reporting Tools (RePORTER) website in 2016. From the set of all potential grants, we

[8]See Section C. for details on recruitment.

kept those that were in one of the five study sections from which we recruited participants, and which had a project abstract length between 300 and 400 words, so the abstract would display consistently. All grants that could be tied to one of our study participants were dropped.

In total, sixteen title and abstract pairs were selected and were assigned to the discrete choice experiment projects. That assignment was done so that study participants would see either zero or one project(s) that matched their area of expertise. The matching was done based on the integrated review group (IRG) codes of the participants and the NIH proposal. Based on the randomization, thirty percent of the participants did not see any projects from their own IRGs. The remaining seventy percent of participants saw one choice scenario where all of the projects matched their IRGs and three choice scenarios where none of the projects matched their IRGs.

The IRG randomization was conducted at the study participant level. To identify the effect of proximity between a presented project and the participant's own research, while also including participant-level fixed effects, we constructed a more granular measure of research proximity using NIH Medical Subject Heading (MeSH) terms. The NIH maintains a structured dictionary of terms used for indexing research on PubMed, and all medical research can be assigned MeSH terms by passing it through an NIH indexing tool. We passed the titles and abstracts shown to participants and the grants received by participants through this tool, then calculated the proximity of a participant to a shown project by counting the unique, matching MeSH terms between the project and all of the participant's NIH-funded projects between 2012 to 2016, divided by the number of MeSH terms associated with the project.[9]

During each choice scenario, the project titles were shown above the project scores (see Section D for an example). All participants saw the project titles. If they hovered their mouse cursor over the title, they could also see the project abstract. Since not everyone chose to hover, we exploit this feature to further assess the veracity of our results on intellectual proximity. If an individual did not hover over the title to view the abstract, then the proximity of that abstract to the subject's research should be irrelevant to the project ranking.

B. Design of Study 2: The Effect of Budget Constraints

A second study was conducted with a separate set of scientists to assess the role of budget constraints on project funding preferences. The design utilized a similar discrete

[9]This measure of research proximity has been used in prior work on connections between researchers (Azoulay, Fons-Rosen, and Zivin 2019).

choice setup as Study 1, with two main differences. First, the participants were shown ten potential projects and asked to choose the four that they would most like to fund. This was presented as constructing a portfolio of projects (see Section D for an example of the choice scenario). The main goal of the study was to determine how individuals responded to tighter budgets, so after choosing their portfolio, participants were initially told that the budget had been cut, only allowing them to fund three projects. They were asked which project they would like to drop. Next, they were asked which project (of the six they did not select for funding) they would add if the budget were expanded to allow for the selection of five projects. This variation allowed us to identify the marginal project initially selected and rejected, to determine whether budgetary pressure affects preferences for project attributes. Each participant engaged in two of these choice scenarios.

C. Recruitment and Sample Construction

The initial sampling frame consisted of the set of all researchers who had received a R01, R35, or F32 NIH grant between 2012 and 2016, from any of the following IRGs: Brain Disorders and Clinical Neuroscience (BDCN), Cell Biology (CB); Molecular, Cellular, and Developmental Neuroscience (MDCN); Oncology-Basic Translational (OBT); or Oncology–Translational Clinical (OTC).[10] We further restricted the sample to individuals who were part of a study section that mapped to only one IRG code, to focus on individuals working within a single, albeit broad, scientific domain. The names and contact information for this set of potential participants was gathered from the NIH RePORTER database, yielding 6,678 total initial contacts.

These initial contacts were randomized into two groups. First, fifty percent (3,339) of the contacts were randomized into the group receiving the project ranking survey (Study 1). Second, the remaining fifty percent (3,339) of the contacts were randomized into the budget experiment (Study 2). Table A1 shows the summary statistics for contacts, broken down by randomization group.

Of the 6,678 scientists contacted by email, 590 either declined to participate or had an outdated email address (leading to the email bouncing), leading to a final contact sample of 6,088. Across the two studies, 563 participants completed all portions of the experiments, for a response rate of 9.2%.[11] 313 participants completed Study 1 and 250 participants completed Study 2. Attrition at each stage is assessed in Table 1. Across both studies, the

[10]The NIH Center for Scientific Review initially reviews grant submissions and assigns the submission to an IRG for assessment of scientific and technical merit.

[11]This response rate is consistent with, if not substantially higher than, other recent surveys of active scientists. For instance, Myers et al. (2020) had a 1.6% response rate and Myers and Tham (2023) had a 3.3% rate.

Table 1: Attrition

	Study 1			Study 2		
	(1)	(2)	(3)	(4)	(5)	(6)
	Attrited Mean [SD]	Finished Mean [SD]	Diff. Mean (SE)	Attrited Mean [SD]	Finished Mean [SD]	Diff. Mean (SE)
Fraction BDCN	0.25	0.26	-0.0071	0.27	0.30	-0.035
	[0.43]	[0.44]	(0.026)	[0.44]	[0.46]	(0.029)
Fraction CB	0.22	0.25	-0.024	0.20	0.20	-0.0020
	[0.42]	[0.43]	(0.025)	[0.40]	[0.40]	(0.026)
Fraction MDCN	0.17	0.22	-0.050	0.17	0.19	-0.019
	[0.38]	[0.42]	(0.023)	[0.38]	[0.39]	(0.025)
Fraction OBT/OTC	0.36	0.28	0.082	0.36	0.30	0.056
	[0.48]	[0.45]	(0.028)	[0.48]	[0.46]	(0.031)
Total funding	6.64	5.68	0.96	6.71	6.02	0.68
	[9.29]	[6.12]	(0.54)	[8.28]	[8.17]	(0.54)
Unique projects	4.23	4.12	0.11	4.32	4.05	0.27
	[2.92]	[2.73]	(0.54)	[2.93]	[2.57]	(0.19)
Total projects	16.3	15.0	1.27	16.6	15.7	0.82
	[15.6]	[15.2]	(0.93)	[15.8]	[13.3]	(1.00)
N	3,026	313		3,089	250	

This table shows statistics for the sample of individuals who did not complete the experiment (Column 1 for Study 1 and Column 4 for Study 2) versus those who completed the experiment (Column 2 for Study 1 and Column 5 for Study 2). Mean values are above and standard deviations are in the square braces below. Columns 3 and 6 show the difference in means between the two groups for Study 1 and Study 2, respectively. Standard errors are in parentheses below each value. "Total funding" is all NIH grant funding from 2012 to 2016. "Unique projects" counts unique NIH grants and "Total projects" is grants by years of grant funding from 2012 to 2016.

sample of completers versus attriters is comparable for the measures we can assess. The largest differences (compared to the standard errors) are that more participants with grants in the IRG code groups MDCN and OBT/OTC finished Study 1, compared to the group that did not complete the studies.

IV. Estimating Equation

Each participant's preferences for different project attributes were estimated by fitting a model for the probability that a participant would choose a given project. The baseline

results show the fit from the conditional logit model

$$\Pr(y_{ijk} = 1|\mathbf{x}_{ijk}) = F(\alpha_i + \beta_1 \text{avg}_{jk} + \beta_2 \text{var}_{jk} + \mathbf{z}_{ijk}\theta) \quad (1)$$

for participant i making a choice about project k as part of the choice scenario version and choice set j.[12] The function F is the cumulative logistic distribution.

The conditioning variables are indicated by x and fall into three groups: the mean and variance of project scores, other project attributes, and controls. The main right-hand-side variables of interest are project score attributes, with a particular focus on the mean and variance of project scores. If the scientists only cared about the average project scores, that would show up as a non-zero coefficient on average score and a zero coefficient on the score variance. In contrast, if they valued dissensus, they might still place a non-zero weight on the average score, but the coefficient on the score variance would be positive. Additional results allow for estimation of preferences around other project score attributes and project descriptions. For example, we assessed the effect of the count of individual project score levels, the effect of higher moments of the project score distribution (e.g., skew), participant expertise or experience, etc.

Control variables are fixed effects for each participant, α_i, such that all estimates reflect the average preferences of a given scientist, since the project attributes shown to that scientist were varied (average score, score variance, project description match with the scientist's research, etc.). Standard errors were clustered at the scientist level.

V. Results

A. Scientist Preferences for Dissensus

We first show models for scientist preferences over different attributes of project scores. The results from fitting Equation (1) are shown in Table 2. The dependent variable is equal to 1 if the participant chose a project in a given choice scenario. All right-hand-side variables are standardized so that the coefficient magnitudes are comparable.

Column 1 performs the simplest and most direct test of whether the scientists' preferences match the funding rule followed by NIH. The first coefficient shows that participants strongly

[12] We converted rankings into binary choices by considering each choice scenario to be composed of three different choice sets. In the first set, all projects are in the choice set and the chosen project is the top ranked one. The second choice set consists of all projects other than the top ranked one and the chosen project is the second ranked project. The third and final choice set consists of the remaining two projects and the chosen project is the third ranked project. Results are similar if we use a multinomial logit (see Table A3), but the conditional logit allows for more granular fixed effects controls. Inferential accuracy is maintained by clustering at the participant level.

Table 2: Scientist Preferences Over Project Scores

	(1) Project choice	(2) Project choice	(3) Project choice
Avg. score	0.82***	0.92***	0.79***
	(0.041)	(0.056)	(0.043)
Score variance	0.083***	0.11***	0.078**
	(0.027)	(0.033)	(0.036)
Score skew		0.10***	
		(0.035)	
Minimum score			0.034
			(0.030)
Maximum score			0.058*
			(0.030)
Clusters	313	313	313
N	11276	11094	11276

This table shows results from estimating Equation (1) on the baseline sample. The dependent variable is an indicator for whether the participant chose a given project. All right-hand side variables were standardized. The models include participant fixed effects. Standard errors are clustered at the participant level: * $p < .10$, ** $p < .05$, *** $p < .01$.

preferred projects with higher average ratings. For fixed effect values of 0 and other variables held at their mean, the model implies that a one standard deviation increase in average project score increased the chance that the project was chosen by eighteen percentage points, a fifty-five percent increase relative to the baseline probability (thirty-three percent) that a project was chosen.

The second column shows that the scientists also preferred projects with higher score variance. Conditional on the average score, a one standard deviation increase in the variance of project scores increased the chance that the project was chosen by 1.8 percentage points (a 5.4% increase). This effect shows that scientists were dissensus-seeking on average. It also runs counter to the mean-only scoring rules currently used by organizations such as NIH.

The subsequent columns test whether the preferences were for higher variance per se or for other correlated project attributes, some of which also indicate a preference for dissensus. Participants could appear to prefer high variance projects, for instance, if they simply preferred high scores and were relatively insensitive to the rest of the score distribution. Column 2 adds the skewness of project scores and Column 3 adds the minimum and maximum score assigned to the project, to see if these attributes explain the variance preferences. In both

cases, the preference for higher variance projects persists.[13]

In Column 2, for example, even after controlling for skewness and average score, participants still preferred higher variance projects. If anything, the preference for higher variance projects appears stronger. At the same time, participants also preferred projects with higher skew, with an effect size comparable to that of variance. The average project in our sample had negative skew, so an increase in skewness for that project resulted in a more symmetric distribution (while holding the mean and variance fixed).[14] This preference was consistent with scientists placing substantial value on high scores, particularly if the rest of the scores were concentrated near the middle of the range.

Column 3 adds the minimum and maximum scores to the estimating equation. Across all projects and after reverse coding, the minimum possible score was 1 and the maximum was 9, but different projects had different highest or lowest scores depending on their exact score distributions. A project with a maximum score below 9 or a minimum score above 1 often had a lower score variance than a project with scores across the full range, so Column 3 adds controls for the actual range. The results show that scientists preferred projects, on average, when both the minimum and maximum scores were higher, but these preferences were not as strong as the preference for variance. Neither coefficient is estimated precisely enough to reject at the five percent level that the preferences were zero.

A.i. Robustness and Sensitivity Checks

Further robustness checks are reported in Section B. Table A3 uses a multinomial logit model to estimate the effect of project scores on project rankings. The multinomial logit relaxes the assumption of homogeneous coefficients across the three choice sets involved in ranking projects, but at the cost of not including high-dimensional fixed effects. Estimating using the multinomial logit shows that the results are in line with the baseline conditional logit model with some notable heterogeneity across choice sets. When choosing the highest and second highest ranked projects, scientists preferred higher mean and higher variance projects. When ranking the third versus the fourth project, the scientists were indifferent between higher or lower variance. They also cared less about the average score. Similarly, Table A5 shows the results of estimating using a generalized multinomial logit model (G-MNL), which models subject-level heterogeneity and relaxes scale assumptions in the

[13]Table A2 adds further score statistics including kurtosis, interaction between mean and variance, the number of lowest or highest scores in the score distribution, and indicators for whether the project had at least one score of 1 (lowest possible score, after reverse coding) or 9 (highest possible score). In all cases, the estimated effect of score variance remains consistent.

[14]Heterogeneity analysis reveals that participants also preferred it when positively skewed distributions became even more positively skewed. See Table A6.

standard conditional logit model. On average, the preferences estimated using the G-MNL are similar to the primary results presented above.

Table A4 evaluates the sensitivity of the results to sample and control changes. As discussed in Section C., 313 participants completed the full study while 356 participants started the experiment and completed at least one ranking exercise. The results show that preferences were unchanged if all available data were included. The second column adds more granular fixed effects for the interaction of participant, question version, and choice scenario. Including these fixed effects, if anything, increases the magnitude of the estimated preference for dissensus.

B. Assessing Hypotheses About Dissensus Tolerance and Proposed Funding Reforms

Many commentators, including directors at NIH, have suggested that NIH is too cautious when funding research. The results above show that the average scientist in our sample agrees with that sentiment. Here we assess explanations that have been proffered for why funding decisions might be so dissensus-intolerant, and investigate scientist preferences for reforms that have been suggested to make the process less risk averse.

B.i. Are Positive and Negative Reviews Weighted Differently?

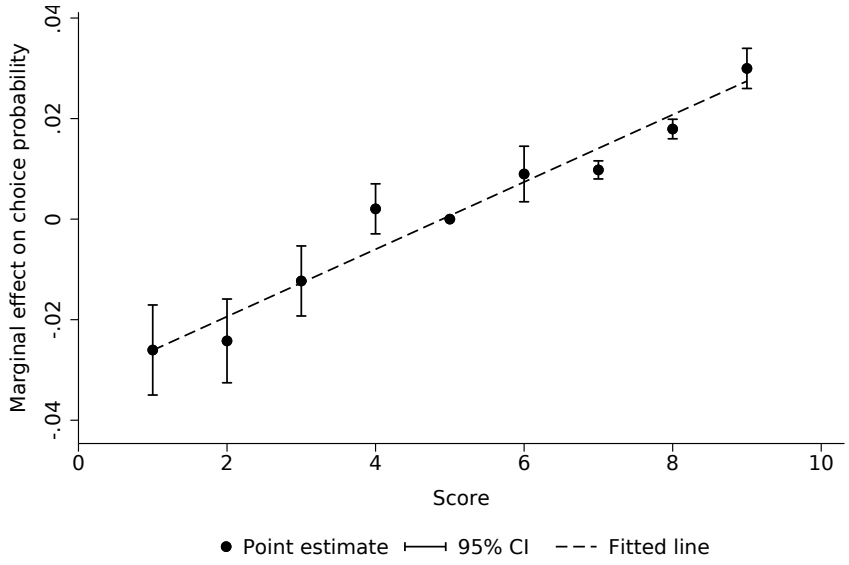
Testing for dissensus preferences using score variance treats low and high scores symmetrically. Previous studies of peer review for scientific grants have emphasized that negative reviews can have an oversized influence on the probability that a grant gets funded.[15] And consensus has been shown to emphasize the influence of negative scores (Lane et al. 2022b). In response, a variety of reforms have been proposed that would bypass some or all of the consensus-based peer review processes. For example, foundations have experimented with a so-called "golden ticket" that allows a reviewer to ensure that an application gets funded, even over the objections or low ratings of other reviewers (Sinkjaer 2018). A similar reform has also been proposed for Program Officers at NIH (Buck 2022).

Although we cannot directly test whether the scientists in our study would prefer a golden ticket-style selection procedure, we can test the underlying basis for that proposal—the idea that negative reviews exert an oversized influence. This hypothesis is assessed in Figure 1. The figure shows the effect, estimated from a conditional logit model, on the choice of project coming from the addition of one score from the range of possible scores. The omitted score is 5, the midpoint of the range from the best (reverse coded) score of 9 to the lowest score

[15] In particular, Jerrim and Vries (2020) found that "a single negative peer review is shown to reduce the chances of a proposal being funding from around 55% to around 25% (even when it has otherwise been rated highly)."

of 1. The coefficients can be interpreted as the effect of replacing a score of 5 with the score indicated on the x-axis. The dashed line shows a linear fit to the point estimates.

Figure 1: Marginal Effect of Each Score on Choice Probability



Notes: The figure shows the marginal effect of each possible project scores on the probability that the project was selected, relative to a score of 5. The estimates were generated by fitting a version of Equation (1) where the project attributes are the count of scores at each score level. The equation includes subject fixed effects. See Table A7 for the numerical coefficients. Whiskers are 95% confidence intervals based on standard errors clustered at the participant level.

The results show that choice probability was monotonically increasing in score, and that the effect of a low score was roughly symmetric with the effect of a high score. In particular, replacing a score of 5 with a score of 1 *reduced* the probability of a project being chosen by almost the same amount that replacing a score of 5 with a 9 *increased* the probability. A formal hypothesis test to see whether the sum of the coefficients is 0 yields a coefficient of 0.004 with a p-value of 0.25.

For less extreme scores, we did find some evidence for asymmetry. A score of 2 was penalized almost the same amount as a score of 1, while a score of 8 raised the probability of selection by less than would be expected based on the average slope of the marginal effects (as indicated by the dashed line). Even here, though, we cannot reject the hypothesis that the scores had marginal effects of the same magnitude. Overall, the results do not support the idea that negative scores disproportionately caused scientists to think poorly of a project. Instead, scores had a roughly uniform effect across the distribution of possible scores.

B.ii. Does Bimodality Better Capture Scientists' Preferences for Dissensus?

Buck (2022) proposes that projects with bimodal scores could receive higher funding priority as a way to reduce the conservatism of funding decisions. What preferences did the scientists in our experiment exhibit along this dimension? Table 3 shows estimated preferences for projects with bimodal scores (Column 1) and simultaneously for bimodality and higher variance (Column 2). In both cases, the average score was included as a control. Projects were classified as bimodal using the dip test from Hartigan and Hartigan (1985), as implemented in Stata by Cox (2016).[16]

Table 3: Preferences for Bimodality Versus Variance

	(1) Project choice	(2) Project choice
Avg. score	0.76***	0.82***
	(0.037)	(0.041)
Bimodal	-0.033	-0.018
	(0.13)	(0.12)
Score variance		0.083***
		(0.028)
Clusters	313	313
N	11276	11276

Notes: This table shows results from estimating Equation (1) on the baseline sample. The dependent variable is an indicator for whether the participant chose a given project. The average score and score variance variables were standardized. The variable "bimodal" is an indicator for the project scores exhibiting a dip statistic greater than 0.1 (Hartigan and Hartigan 1985). The models included participant fixed effects. Standard errors are clustered at the participant level: * $p < .10$, ** $p < .05$, *** $p < .01$.

The table shows that scientists did not prefer bimodal projects. Moreover, the preference for dissensus, as captured by project score variance, was unaffected by the inclusion of the bimodality measure. Bimodality is a particularly extreme form of dissensus that was not favored by the participants in our sample.

At the same time, bimodality is rare in both the scores we showed to participants and in

[16]In the table, the variable "bimodal" is an indicator for whether the dip statistic was above 0.1, although the results are robust to alternative cutoffs and available upon request.

current, real-world NIH scores. Over time, NIH has worked to avoid strategic behavior that results in bimodal scores.[17] Changes in the NIH scoring system have reduced the degree of bimodality, making it less relevant for judging dissensus.

B.iii. Does Expertise Decrease Dissensus Tolerance?

Lay observers, scientists, granting agencies, and previous research studies have debated whether expertise and experience increase or decrease the willingness of scientists to engage in high-risk research. The effect of expertise on dissensus tolerance could go in either direction. On one hand, greater expertise might increase a scientist's convictions about the correct direction of research, making them less subject to consensus-driven selection criteria. On the other hand, previous work has shown that the removal of incumbent researchers in a field can spur innovation (Azoulay, Fons-Rosen, and Zivin 2019), and recent work shows that the creativity of patents quickly declines with experience (Kalyani 2022). Arthur C. Clarke, in a quote that has come to be known as Clarke's Law, offered some additional nuance by arguing that the effect of experience is asymmetric: "When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong."

Understanding the direction of this effect is important because expert review is at the heart of nearly all scientific project evaluation, whether for funding or publication purposes. NIH in particular relies heavily on carefully matched peer evaluators when judging grant quality.

We assessed the effect of expertise and experience with the estimates shown in Table 4. Overall, we found that expertise *increased* dissensus tolerance. In other words, participants who were in the best position to understand the proposal had substantially stronger preferences for higher project score variance. Column 1 shows the effect of proximity between the shown project and the participant's research area, based on our measure of MeSH term overlap between the shown projects and the participant's NIH grants from 2012–2016. A stronger overlap in these terms indicates that the scientist was active in the area from which the project's description was drawn, and thus measures the degree to which the scientist was a relative insider for the specific field represented by the project. Recall that the projects shown to the participants were randomized to be either closer to or further from their field.

The results show that scientists modestly preferred projects that were more inside their research area. The "MeSH match" coefficient is positive and significant at the ten percent

[17] Ogden and Goldberg (2002) describes some reviewers as inflating the scores of projects that they like, while simultaneously lowering the scores of competing projects, so that the favored project would look even better by comparison.

Table 4: Preferences for Projects by Experts

	(1) Expertise	(2) Placebo Check	(3)
	Project choice	Project choice	Project choice
Avg. score	0.82***	0.70***	0.97***
	(0.041)	(0.052)	(0.064)
Score variance	0.081***	0.078**	0.088**
	(0.027)	(0.038)	(0.040)
MeSH match	0.038*	0.056*	0.015
	(0.021)	(0.030)	(0.031)
Avg. score × MeSH match	0.032	0.048	0.0042
	(0.029)	(0.042)	(0.039)
Score variance × MeSH match	0.051**	0.081**	0.012
	(0.024)	(0.040)	(0.029)
Hover subgroup	Full sample	Always	Never/rarely
Clusters	313	169	144
N	11276	6089	5187

Notes: This table shows results from estimating Equation (1) on the baseline sample. The dependent variable is an indicator for whether the participant chose a given project. All right-hand-side variables were standardized. The models include participant fixed effects. Standard errors are clustered at the participant level: * $p < .10$, ** $p < .05$, *** $p < .01$.

level, with an effect size that is about half the size of the effect of project score variance. The interactions between this measure of expertise and project score statistics shows that experts had a significantly stronger preference for dissensus, as indicated by the positive coefficient on the interaction between score variance and expertise, as measured by MeSH match. Given that all variables are standardized, the coefficient on "score variance" indicates the preference that a scientist with an average MeSH match had for a project with higher variance scores. The results show that this scientist preferred higher variance projects, on average, and that the preference was about one-tenth as strong as the preference for higher average score.

A scientist with a one standard deviation higher MeSH match showed little difference in their preferences over average scores but a substantially stronger preference for higher variance. In particular, the variance preferences were sixteen percent as strong as average score preferences for such individuals. Going the other direction, the results indicate that a scientist who was relatively far from the area of the shown project (one who has a 1 standard

deviation lower MeSH match) placed almost no weight on project score variance.

The second and third columns show the results of a placebo test that was built into the experiment to determine whether the results from Column 1 were driven by the study participants actually taking the time to understand the abstracts that were shown, instead of simply acting differently than individuals with lower match rates for reasons unrelated to project content. To see the abstracts of the projects included in the experiment, participants needed to hover over links. Column 2 shows the results for subjects who reported always hovering over the links. Column 3 reports results for subjects who said they rarely or never hovered to look at the abstracts. While the endogeneity of hovering means that these results should be interpreted with caution, one can see that the effect of expertise is substantially stronger for subjects who did report looking at the abstracts.[18]

B.iv. Do Tighter Budgets Decrease Dissensus Tolerance?

Francis Collins, NIH Director from 2009 to 2021, argued that budgetary pressures reduce scientific risk-taking, stating (emphasis ours): "Although the two-level NIH peer-review process is much admired and much copied around the world, its potential tendency toward conservatism is a chronic concern and *invariably worsens when funding is very tight*."[19]

Study 2 allowed us to test this hypothesis. The results are shown in Table 5. The first two columns show estimates for the attributes of the project that was dropped when scientists were told that the budget had been reduced. Column 1 shows the characteristics of the dropped project compared to the projects that were kept. Unsurprisingly, the dropped project had a lower average score compared to the projects that were kept in the portfolio. Lending support to Collins' statement, the dropped project also tended to have higher score variance. When faced with tighter budgets, participants preferentially dropped riskier projects characterized by higher dissensus.

Column 2 compares the dropped project to the projects that were originally not chosen for the portfolio of funded projects. Here, the average score clearly played an important role, but the variance of scores was no longer as important. The effect size is substantially smaller than when comparing the dropped project to projects that were kept in the portfolio, and the effect is not statistically significant.

Columns 3 and 4 show the characteristics of the projects that were added when budgets were expanded, with Column 3 showing the comparison to the four projects that were already

[18]Results using other subject-specific heterogeneity measures are shown in Table A8. In the sample, men were more dissensus-loving than women. An elicited measure of risk aversion did not strongly predict dissensus preference. And individuals with greater breadth in their research, as measured by the total number of unique MeSH terms, were more tolerant of dissensus.

[19]Quoted in Kolata (2009).

Table 5: Effect of Constrained or Relaxed Budgets

	(1)	(2)	(3)	(4)
	Tighter Budget		*Relaxed Budget*	
	Dropped proj. compared to kept	Dropped proj. compared to not chosen	Added proj. compared to kept	Added proj. compared to not chosen
Avg. score	-0.82***	1.88***	-1.80***	0.34***
	(0.10)	(0.14)	(0.14)	(0.056)
Score variance	0.17**	0.078	-0.045	-0.071
	(0.081)	(0.063)	(0.072)	(0.060)
Clusters	250	250	250	250
N	1983	3516	2483	3516

Notes: This table shows results from estimating Equation (1) on the baseline sample. The dependent variable is an indicator for whether the participant chose a given project. All right-hand side variables are standardized. The models include participant fixed effects. Standard errors are clustered at the participant level: * $p < .10$, ** $p < .05$, *** $p < .01$.

chosen and Column 4 showing the comparison with the projects that were not originally chosen. The variance of scores appeared to play little role in this choice.

Together, these results provide nuanced evidence for Collins' claim. Compared to projects that were kept in the portfolio, tighter budgets did cause scientists in our sample to cut higher-variance projects. But the reverse was not true for more expansive budgets, and the cut project was not substantially different than other non-chosen projects in terms of variance.

C. Implications for Project Funding

How large is the difference between the procedure NIH uses for funding (mean score) and the preferences possessed by the scientists in our study when it comes to actually ranking and funding projects? Although NIH does not maintain data on project scores and funding decisions that would allow us to test this question on historical NIH proposals, three datasets illuminate the scale of the difference. First, we calculated the changes in rankings for the fifty-four unique mean-variance combinations in the projects that we showed to participants in the first study. We also repurposed two prior experiments that closely replicated the NIH review process. The first of these was Pier et al. (2018), which carefully simulated the NIH review process using real NIH reviewers, former study section leaders, and proposals. The

second study, Lane et al. (2022b), conducted two experiments involving the evaluation of real submissions to a pair of research award opportunities. The data from Lane et al. (2022b) is especially revealing because it allowed us to assess whether the scientists' preferences would have resulted in different real-world funding decisions.

For the fifty-four different project score mean and variance combinations included in our Study 1, the overall ranking for half of them changed when ranked according to the mean and variance preferences given in Table 2, Column 1, versus a ranking purely based on mean score. The largest changes in overall rank occurred, naturally, for projects that had the highest variance. Given that project scores were bounded, these projects also tended to have average scores that were closer to the middle of the pack.

Thus, high variance caused two effects that drove a wedge between the NIH-style mean score ranking and the rankings that the scientists in our sample preferred. First, consider two projects with the same mean but different variances. The NIH procedure would give these two projects the same score, while the scientists gave the higher variance project a higher score. Thus, the NIH procedure gave the high variance project a relatively lower rank than the scientists. Second, consider two projects with different average scores. A higher average score was mechanically, positively correlated with lower variance given that scores were bounded. This caused the NIH procedure to rank a higher variance project lower (because of its lower mean score), while the scientists ranked the two projects closer together.

These two effects can be seen by comparing individual proposals drawn from our Study 1. To illustrate the first mechanism, we focus on two projects that had an average score of 6.3, but one had a low variance of 3.3 while the other had a high variance of 9.5. The NIH procedure would rank both of these projects right around the 50^{th} percentile across the entire set of projects in our study. Using scientists' preferences, however, would put the higher variance project at the 63^{rd} percentile and the low variance project at the 44^{th} percentile.

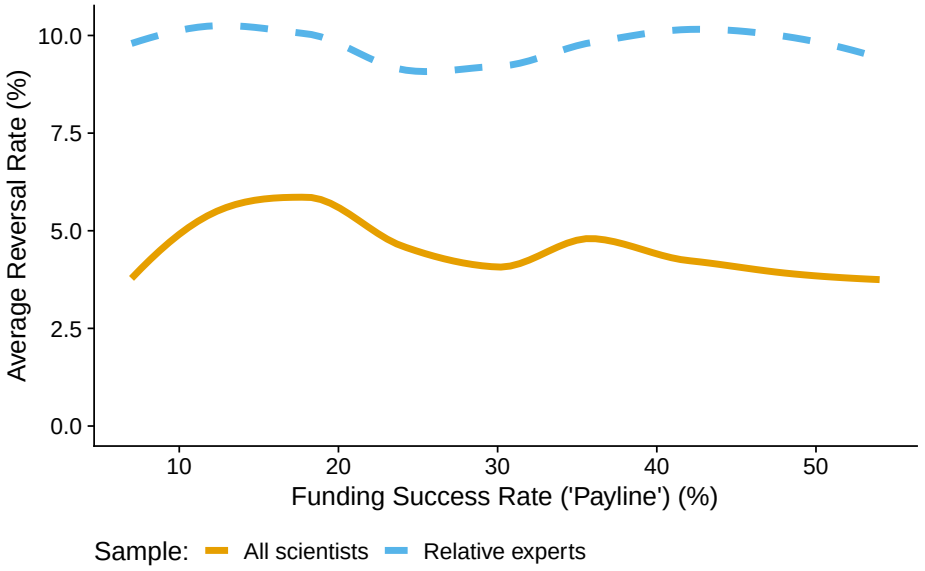
To illustrate the second mechanism, we can again consider the high variance project with an average score of 6.3 and a variance of 9.5. But this time we compare it to a project with an average score of 6.5 and a score variance of 2.3. The NIH procedure would rank the latter project in the 63^{rd} percentile of the overall project distribution, well ahead of the higher variance project (even though the difference in their means is only one-quarter of the standard deviation in average project scores across the experiment). If we ranked them according to the scientists' preferences, the lower variance project would drop down to the 55^{th} percentile, while the high variance project would again move up to the 63^{rd} percentile.

Using scores generated by Pier et al. (2018), which strove to closely replicate the NIH review process, we also found substantial differences in project ranking between the two

procedures. In the Pier et al. study, many projects near the top of the ranking received identical average scores. At the 80^{th} percentile, five studies were given the same average score of 7. Using variance, one can break three of these ties, with the highest variance project (variance of 4.7) being ranked first among the set, the lowest variance project (variance of 0.7) ranking last, and the remaining three projects with a variance of 1 being ranked in the middle.

This example from Pier et al. highlights an additional insight from our results. Taking variance into account can help break ties that often emerge when a small set of reviewers are judging each project or when reviewers (or the aggregation process) round their scores. It also shows that a useful approximation to the variance preferences of the scientists in our sample would be to use variance simply to break ties.

Figure 2: Funding Reversals Under NIH and Scientist Ranking Procedures



Notes: This figure shows the reversal rate for project funding as a function of the payline (fraction of projects that get funded) for four different sets of project scores (Study 1 from this paper, the two studies from Lane et al. (2022b), and Pier et al. (2018)). The reversal rate is the fraction of studies that changed whether they were funded under a mean-only ranking versus the mean and variance-based ranking. The lines are LOESS fits to reversal rates calculated at each payline percentile. The solid line shows the reversal rate when using the estimated preferences from the baseline results using the full sample of scientists. The dashed line shows the reversal rate when using the preferences of scientists who were relative experts (a MeSH match 1 standard deviation higher than average).

The data from the two experiments in Lane et al. (2022b) allow us to examine how actual funding decisions would have changed if variance had been taken into account. We did so

by first ranking projects by their average score.[20] Multiple reviewers rated each project, which allowed us to also calculate the variance of scores and re-rank the projects using the project score attribute preferences from our scientists. Importantly, we found that in both experiments, accounting for variance would have led to different projects being funded: the marginal projects funded would have been switched, with a higher-variance, unfunded project replacing a lower-variance project that actually did get funding.

We call such a change in funding a "reversal" of the project funding decision. For any given possible payline (the fraction of projects that get funded), we can calculate the reversal rate for the four sets of project ratings described in this section. The reversal rate is the fraction of projects funded at the payline that changes when we move from a mean-only to a mean-and-variance ranking. Figure 2 shows the average reversal rate across sets of projects from the different studies (this study, Pier et al., and two sets from Lane et al.) as a function of the payline.

Starting first with the scientists who were in the position to better understand the proposals (those with MeSH match values 1 standard deviation higher than average), we see from the dashed line that the reversal rate was around 10% for all paylines. Even when we expand to include the preferences for all scientists estimated in Table 2 Column 1, we still see reversal rates of 4 to 6%, depending on the payline. The highest reversal rates are near typical NIH paylines of 10 to 20%.[21] And this average reversal rate masks high rates that can appear for individual sets of project scores. Figure A1 shows the reversal rates separately for each of the four studies. Rates are as high as 20% for the proposals from Lane et al. (2022a). Together, these results underscore that variance preferences are not only statistically important, but can be consequential for funding decisions, and particularly so in cases that closely mimic real NIH grantmaking.

VI. Conclusion

Scientific research, through its influence on technological innovation, has long been recognized as an important contributor to aggregate income (Nelson and Phelps 1966) and a driver of economic growth (Lucas 1988; Romer 1990), yet the path from research to innovation is uncertain, requiring institutions that make substantial scientific investments to appropriately balance risk and return in the portfolio of projects they support. Research

[20] The main goal in Lane at al. is to study the effect of showing reviewers scores from other reviewers, to assess how exposure to others' scores affects one's own rankings. Thus, we only used the original, independent scores that participants provided for the exercises described here.

[21] For example, the National Institute of Allergy and Infectious Disease at NIH published annual information on paylines for grants. The payline for R01 grants in 2022 was twelve percent (sixteen percent for new PIs).

projects that closely build on existing scientific knowledge may be a relatively safe bet, but the incremental innovation they produce may have lesser social value. In contrast, research that eschews conventional wisdom for more speculative pursuits may be required to produce radical or paradigm-shifting innovations of enormous value, but it is also much more likely to end in failure (see Eric Lander as quoted in Fallows (2014); Manso (2011)). The design of public and private institutional structures employed to evaluate research projects plays a critical role in balancing the risk and rewards from research, which, in turn, informs future scientific frontiers.

The focus of this paper is on the peer review process and how NIH (and other science-based agencies) synthesizes the output of that process into resource allocation decisions. Of particular concern is that agencies base funding decisions on the average of peer review scores, ignoring higher moments of the score distribution that may confer valuable information about the radicality of a scientific proposal. Since data on individual scores from NIH is unavailable to the research community, we leveraged data from two novel discrete choice experiments, fielded in samples of active biomedical scientists with a successful NIH grant history, to assess their preferences for aggregating peer review evaluations into scientific funding decisions.

In contrast with current practice, we found that these scientists—the very scientists that NIH relies upon for expert evaluations of research proposals—preferred to fund projects where there was some disagreement among reviewers. This preference for higher-dissensus projects was not driven by lone wolf reviewers who were enamored with a project, nor was it driven by focus on an aberrant, critical review. Rather, it appears that our experts valued healthy disagreement over either middle-of-the-road reviews or more extreme forms of dissensus such as projects that received bimodal scores. While this appetite for risk shrank as budgets became tighter, it did not completely disappear. We also found that those scientists with relatively greater domain expertise on a proposal were consistently more enthusiastic about dissensus in their reviews than those asked to make decisions outside their specific area of expertise. Applying our estimates to prior studies that mimic the NIH review process suggests that incorporating preferences for dissensus would lead to changes in billions of dollars of research funding annually.

Our results should not be construed as a critique of the peer review process. Indeed, we believe the impartial review of proposals by experts in the field is essential for prioritizing scientific investments by both public and private agencies. The substance of our inquiry relates whether there is relevant information from that process beyond the simple mean of reviewer scores that should influence the funding decisions of a major government entity charged with undertaking risky R&D projects related to improving the public's health. While our findings have implications for funding rule reforms that could prove important,

many questions remain unanswered. Fundamental for the tasks ahead is a better understanding of the causal relationship between peer review scores and scientific impact. This will require a clever mix of experimental design and currently unavailable data from funding agencies containing individual reviewer scores on projects being evaluated. Prospective experimentation may offer additional insights and seems particularly well suited to the newly created Technology Innovation and Partnerships Directorate at the NSF. Shrinking research budgets, concerns about the technological competitiveness of the United States, and global declines in research productivity all underscore the need for more formal examinations of the policies and programs that ultimately shape research portfolios.

References

Ackermann, Robert. 1986. "Consensus and Dissensus in Science." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 88 (2):99–105.

Aghion, Philippe and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica* 60 (2):323–351.

Arrow, K. J. 1972. *Economic Welfare and the Allocation of Resources for Invention*. London: Macmillan Education UK, 219–236. URL https://doi.org/10.1007/978-1-349-15486-9_13.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin. 2019. "Does Science Advance One Funeral at a Time?" *American Economic Review* 109 (8):2889–2920.

Azoulay, Pierre, Joshua S Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence From the Academic Life Sciences." *RAND Journal of Economics* 42 (3):527–554.

———. 2012. "NIH Peer Review: Challenges and Avenues for Reform." *National Bureau of Economic Research Working Paper* 18116.

Boudreau, Kevin J, Eva C Guinan, Karim R Lakhani, and Christoph Riedl. 2016. "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science." *Management Science* 62 (10):2765–2783.

Buck, Stuart. 2022. "Reforming Peer Review at NIH." https://goodscienceproject.org/articles/reforming-peer-review-at-nih. Accessed: 2023-04-07.

Bush, Vannevar. 2020. *Science, the Endless Frontier*. Princeton: Princeton University Press. URL https://doi.org/10.1515/9780691201658.

Carson, Richard T, Joshua S Graff Zivin, Jordan J Louviere, Sally Sadoff, and Jeffrey G Shrader. 2022. "The Risk of Caution: Evidence From an Experiment." *Management Science* 68 (12):8515–9218.

Cox, Nicholas. 2016. "DIPTEST: Stata Module to Compute Dip Statistic to Test for Unimodality." https://ideas.repec.org/c/boc/bocode/s456998.html. Accessed: 2023-06-19.

Fallows, James. 2014. "When Will Genomics Cure Cancer?" https://www.theatlantic.com/magazine/archive/2014/01/when-will-genomics-cure-cancer/355739/.

Fiebig, Denzil G, Michael P Keane, Jordan Louviere, and Nada Wasi. 2010. "The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity." *Marketing Science* 29 (3):393–421.

Franzoni, Chiara and Paula Stephan. 2022. "Uncertainty and Risk-Taking in Science: Meaning, Measurement and Management in Peer Review of Research Proposals." *Research Policy* :104706.

Goldstein, Anna and Michael Kearney. 2017. "Uncertainty and Individual Discretion in Allocating Research Funds." *SSRN Working Paper* 3012169.

Gu, Yuanyuan, Arne Risa Hole, and Stephanie Knox. 2013. "Fitting the Generalized Multinomial Logit Model in Stata." *The Stata Journal* 13 (2):382–397.

Guthrie, Susan, Ioana Ghiga, and Steven Wooding. 2018. *What Do We Know About Grant Peer Review in the Health Sciences? An Updated Review of the Literature and Six Case Studies*. Santa Monica: RAND Corporation.

Hartigan, John A and Pamela M Hartigan. 1985. "The Dip Test of Unimodality." *Annals of Statistics* 13:70–84.

Jacob, Brian A and Lars Lefgren. 2011. "The Impact of Research Grant Funding on Scientific Productivity." *Journal of Public Economics* 95 (9-10):1168–1177.

Jerrim, John and Robert de Vries. 2020. "Are Peer-Reviews of Grant Proposals Reliable? An Analysis of Economic and Social Research Council (ESRC) Funding Applications." *The Social Science Journal* 60 (1):1–19.

Kalyani, Aakash. 2022. "The Creativity Decline: Evidence From US Patents." *SSRN Working Paper* 4318158.

Kolata, Gina. 2009. "Grant System Leads Cancer Researchers to Play It Safe." https://www.nytimes.com/2009/06/28/health/research/28cancer.html. Accessed: 2023-04-07.

Krieger, Joshua, Ramana Nanda, Josh Lerner, and Ahmed Tahoun. 2022. "Are Transformational Ideas Harder to Fund? Resource Allocation to R&D Projects at a Global Pharmaceutical Firm." *Havard Business School Working Paper* 23-014.

Lane, Jacqueline N, Zoe Szajnfarber, Jason Crusan, Michael Menietti, and Karim R Lakhani. 2022a. "Are Experts Blinded by Feasibility? Experimental Evidence From a NASA Robotics Challenge." *Working Paper* .

Lane, Jacqueline N, Misha Teplitskiy, Gary Gray, Hardeep Ranu, Michael Menietti, Eva C Guinan, and Karim R Lakhani. 2022b. "Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation." *Management Science* 68 (6):4478–4495.

Lauer, Mike. 2023. "FY 2022 by the Numbers: Extramural Grant Investments in Research." https://nexus.od.nih.gov/all/2023/03/01/fy-2022-by-the-numbers-extramural-grant-investments-in-research/. Accessed: 2023-04-07.

Lee, Carole J, Cassidy R Sugimoto, Guo Zhang, and Blaise Cronin. 2013. "Bias in Peer Review." *Journal of the American Society for Information Science and Technology* 64 (1):2–17.

Li, Danielle. 2017. "Expertise Versus Bias in Evaluation: Evidence From the NIH." *American Economic Journal: Applied Economics* 9 (2):60–92.

Louviere, Jordan J, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Louviere, Jordan J, David A Hensher, and Joffre D Swait. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press.

Lucas, Robert E. 1988. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22 (1):3–42.

Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance* 66 (5):1823–1860.

March, James G. 1991. "Exploration and Exploitation in Organizational Learning." *Organization Science* 2 (1):71–87.

Miller, Roger. 1995. "Applying Quality Practices to R&D." *Research-Technology Management* 38 (2):47–54.

Mokyr, Joel. 1992. *The Lever of Riches: Technological Creativity and Economic Progress*. Oxford University Press. URL https://doi.org/10.1093/acprof:oso/9780195074772.001.0001.

Moses, Hamilton, E Ray Dorsey, David HM Matheson, and Samuel O Thier. 2005. "Financial Anatomy of Biomedical Research." *Journal of the American Medical Association* 294 (11):1333–1342.

Murray, Fiona. 2013. "Evaluating the Role of Science Philanthropy in American Research Universities." *Innovation Policy and the Economy* 13 (1):23–60.

Myers, Kyle and Wei Yang Tham. 2023. "Money, Time, and Grant Design." *Working Paper* :33.

Myers, Kyle R, Wei Yang Tham, Yian Yin, Nina Cohodes, Jerry G Thursby, Marie C Thursby, Peter Schiffer, Joseph T Walsh, Karim R Lakhani, and Dashun Wang. 2020. "Unequal Effects of the COVID-19 Pandemic on Scientists." *Nature Human Behaviour* 4 (9):880–883.

National Science Foundation (NSF). 2023. "National Patterns of R&D Resources: 2020–21 Data Update." https://ncses.nsf.gov/pubs/nsf23321. Accessed: 2023-04-07.

Nelson, Richard R. 1959. "The Simple Economics of Basic Scientific Research." *Journal of Political Economy* 67 (3):297–306.

Nelson, Richard R and Edmund S Phelps. 1966. "Investment in Humans, Technological Diffusion, and Economic Growth." *American Economic Review* 56 (1/2):69–75.

Nicholson, Joshua M and John PA Ioannidis. 2012. "Conform and Be Funded." *Nature* 492 (7427):34–36.

NIH. 2008. "Enhancing Peer Review: The NIH Announces New Scoring Procedures for Evaluation of Research Applications Received for Potential FY2010 Funding." https://grants.nih.gov/grants/guide/notice-files/not-od-09-024.html. Accessed: 2023-04-07.

Nijstad, Bernard A, Floor Berger-Selman, and Carsten KW De Dreu. 2014. "Innovation in Top Management Teams: Minority Dissent, Transformational Leadership, and Radical Innovations." *European Journal of Work and Organizational Psychology* 23 (2):310–322.

Ogden, Thomas E and Israel A Goldberg. 2002. *Research Proposals: A Guide to Success*. San Diego, California: Academic Press.

Pier, Elizabeth L, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J Nathan, Cecilia E Ford, and Molly Carnes. 2018. "Low Agreement Among Reviewers Evaluating the Same NIH Grant Applications." *Proceedings of the National Academy of Sciences* 115 (12):2952–2957.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5, Part 2):S71–S102.

Science News Staff. 2022. "Research Gets a Boost in Final 2023 Spending Agreement." https://www.science.org/content/article/research-gets-boost-final-2023-spending-agreement. Accessed: 2023-06-19.

Sen, Avery. 2014. "Totally Radical: From Transformative Research to Transformative Innovation." *Science and Public Policy* 41 (3):344–358.

Sinkjaer, Thomas. 2018. "Fund Ideas, Not Pedigree, to Find Fresh Insight." *Nature* 555 (7697):143–144.

Whitley, Richard and Jochen Gläser, editors. 2007. *The Changing Governance of the Sciences*, vol. 26. Dordrecht: Springer.