Expert predictions and errors in research funding decisions

Chiara Franzoni[+*] Massimiliano Guerini[+] Andola Stanaj[+]

[+]School of Management, Politecnico di Milano, Piazza L. da Vinci 32, Milan 20123. *chiara.franzoni@polimi.it

Abstract

We analyze the predictiveness and accuracy of evaluators' assessments in a large sample of research proposals. We use a two-stage estimation that considers both the decision to fund and subsequent outcomes of the funded applications, measured by publications, citations, field-citation ratios, and altmetric influence. We find that evaluators' predictive ability is contingent on expertise. In particular, assessments by domain experts, defined as people with specific knowledge in the proposal's domain, are predictive of post-funding success, unlike those of evaluators with average or minimal domain expertise; they also are more influential in funding decisions, suggesting that their abilities are factored into the decision-making process. Despite this, domain experts are less accurate (i.e., make larger errors on average) in anticipating the level of outcome of proposals. Specifically, they are generally accurate when giving negative opinions but make large errors when giving positive opinions. The findings suggest that while domain expertise provides essential informational benefits, domain experts also tend to promote proposals that underperform in their area of expertise.

Keywords: grant peer review; expertise; predictive validity; forecasting; correctness; intellectual proximity

JEL: O38; O32; I28

Acknowledgements: This work was supported by the Alfred P. Sloan Foundation (G-2021-14189) and by the Novo Nordisk Foundation (NNF21SA0071711). The authors would like to express their gratitude to Henrik Barslund-Fosse and Rikke Nording-Christensen and the NNF personnel for essential support in retrieving and interpreting the data. They also thank Jared Gars, and Paula Stephan for their invaluable contributions to the research project. The authors have benefited from fruitful discussion with Jerome Adda, Stefano Baruffaldi, Cristoph Carnehl, Pietro Dedola, Thomas Feliciani, Chris Hesselbein, Marco Ottaviani, Evila Piva, Justus Preusser, and Daniel Souza.

1. INTRODUCTION

Decisions regarding whether to fund projects, be they for research or innovation, are challenging, but vital for scientific progress, societal wellbeing, and economic growth (Arrow 1962; Arrow and Lind 1970; Hall and Lerner 2010; Hall, Mairesse, and Mohnen 2010; Romer 1994; Stephan 2012). Research projects are intellectually sophisticated, highly specialized, and experimental in nature. Their understanding requires substantial domain-specific knowledge and a strong command of the technical language and methods. They also encompass some of the latest scientific advances, including approaches that may be novel, untested, or even disputed, requiring critical thinking about what is (Åstebro and Elhedhli 2006; Scott, Shu, and Lubynsky 2020).

Responsibility for evaluating these projects is largely entrusted to domain experts, defined as people with a strong intellectual understanding of research and with technical and scientific competencies on the specific domain knowledge of the proposal under evaluation (Chubin and Hackett 1990; Cole and Cole 1981; Lamont 2009). Reliance on domain expertise is a main tenet of all evaluation practices and is deeply rooted in the rules of peer review (Chubin and Hackett 1990; Polanyi 1962). It assumes that the possession of specific knowledge about the subject of the evaluation provides informational advantages in envisaging the possible future outcomes of a project (Henderson and Fredrickson 1996; Li 2017; Scott et al. 2020).

Despite the widespread belief in the importance of expertise, the few empirical tests on the predictive power of research evaluations conducted to date have been constrained by a lack of individual-level data linking the evaluations to the identity of the evaluators and, consequently, to their domain expertise (Danthi et al. 2014; Doyle et al. 2015; Fang, Bowen, and Casadevall 2016; Li 2017; Reinhart 2009). Studies conducted in other contexts, like consumers behavior, investment decisions, medical opinions, and the prediction of geopolitical events have generally not found a correlation between expertise and forecasting ability, showing instead that expert predictions are rarely better than those of basic algorithms and the few forecasters who systematically beat chance are generally not detectable by characteristics observed ex ante (Armstrong 1991; Broomell and Budescu 2009; Camerer and Johnson 1991; Clemen and Winkler 1985; DellaVigna and Pope 2018; Morgan 2014; Tetlock and Gardner 2015). These findings are based on contexts in which domain expertise is difficult to assess and is sometimes conflated with experience (length of exposure), decision-making authority

(holding top positions in organizations), or training (formal education in a subject), and so are not necessarily generalizable. However, they add to previous work documenting a number of problems with expert judgement, including the tendency for experts to be overconfident in their opinions (Ben-David, Graham, and Harvey 2013; Bradley 1981; Morgan 2014), to disagree with other experts and themselves over time (Cole, Cole, and Simon 1981; Litvinova et al. 2022; Mumpower and Stewart 1996), and to behave strategically to move decisions in desired directions (Krishna and Morgan 2001; Ottaviani and Sørensen 2006).

In this paper, we address the important and understudied topic of expert predictive ability in evaluations of research proposals, a context that is particularly suitable to the aim, because the knowledge of the experts is well-documented by publication records, allowing measuring domain expertise with a precision that would not be possible in other contexts. We use measures of domain expertise based on large language models to study empirically how well the opinions provided by evaluators with different degrees of domain expertise predict the actual outcomes of research proposals. We use a dataset comprising 16,636 unique evaluations of 5,769 research proposals (both accepted and rejected), submitted to the Novo Nordisk Foundation, the world largest private funder of scientific research, between 2012 and 2018. We supplement data with detailed information about the evaluators, the Principal Investigators (PIs) and the proposals. To be more specific, we collect individual-level information on the publications records of the evaluators, and compute a metric of domain expertise of each evaluator-proposal pair, defined as the intellectual proximity (cosine similarity) between the content of the proposal (based on title and summary) and the background of the evaluators at the time of the evaluation (based on their previous publications). We measure proposal outcomes for a minimum of 5 years after the date of funding (until the end of 2023), in terms of publications, citations, field-citation ratios, and influence metrics (altmetrics). We study the relationship between domain expertise and predictive ability in two ways. First, we model predictiveness, i.e., the degree to which scores provided by experts and non-experts prior to funding correlate to post-funding metrics of proposal success. Second, we model accuracy, i.e., the size and distribution of the prediction errors that experts and non-experts make when scoring proposals, by comparing the predicted (pre-funding) quantile rank to the actual (post-funding) quantile rank of the proposals with methods borrowed from the forecasting literature (Broomell and Budescu 2009; Larrick and Soll 2006; Lehmann and Casella 1998; Tetlock and Gardner 2015).

Our empirical exercise poses three major methodological challenges. First, measuring the outcomes of research projects is inherently difficult. We tackle this problem by collecting four different measures of outcomes, of which three (publications, citations, field citation ratios) measuring academic performance, and one (almetrics) measuring influence beyond academe. Second, although experts provide opinions on all proposals, the actual outcomes are observable only for the proposals that were selected for funding, inducing a sample selection bias. We tackle this challenge by using a Heckman two-stage estimation approach, in which the first stage models the probability of being funded, and the second stage models the ability of reviewers' evaluation scores to predict post-funding outcomes, or the accuracy of predictions, with a correction for sample selection. Reviewers are committee members, hired before the closure of the call and have limited control over the number of proposals assigned to them for evaluation. Accordingly, we use their workload (i.e., the total number of evaluations performed by a single person in the month of the review) as the exclusion restriction. We assume that reviewers' workload is exogenous to the unobservable variables (e.g., the quality of the proposal) in both the first and second stage, while it affects the choice of funding (higher workload makes evaluations noisier) but does not affect the research performed and its results, after funding. Third, applications are not randomly assigned to reviewers and expertise is one criteria of assignment, creating potential issues of identification regarding the effect of expertise on predictive power and accuracy. One source of concern is that proposals addressing more mainstream topics would be easier to evaluate, leading to better predictions, and also more likely to be assigned to a domain expert, because committee members are normally chosen to represent the major areas of a research field. Although we cannot eliminate this issue completely in our estimates, we try to minimize it, by using individual reviewer fixed effects in all estimates.

We provide four main findings: First, evaluation scores predict, quite logically, whether the proposal is funded, but they are not a good predictor of post-funding outcomes, however measured, suggesting that, in general, peer review opinions have, at best, weak predictive validity. Second, when we qualify evaluators by their domain knowledge on the topic of the proposals, we find that the predictive power of peer review is contingent on evaluators' domain expertise. Indeed, expert's scores are a strong predictor of post-funding outcomes, while non-expert's scores are not. This result is stable across all measures of post-funding outcome, suggesting that expertise is associated with important informational advantages. Third, evaluators with domain expertise are more influential towards the funding decision, suggesting

that their expertise carries weight. Moreover, supplemental analyses show that evaluators with prestigious research publications (articles in Science & Nature) are also more influential than non-prestigious evaluators, but their opinions were not more predictive of post-funding outcomes. This indicates that research excellence or prestige is not a substitute of domain expertise in peer review predictions. Fourth, experts are accurate (make minimal errors) when they give negative opinions (worst scores) but make large errors when they give positive opinions (best scores), suggesting that they tend to overestimate proposals in their domain. In supplemental analyses, we show that the overestimation bias of experts does not seem to be explained by a preference for novel research, which could have led to more volatile outcomes and larger average errors. Alternative, but untested, explanations for the large overestimation errors are that experts perceive the impact in their domain as exaggeratedly high, or are affected by confirmation biases, and/or that they behave strategically to promote underperforming proposals in their area of expertise.

We conclude that expertise is essential to ensure predictive evaluations in research funding. However, experts are prone to making large errors and their opinions, especially when very positive, should be taken with a grain of salt.

The paper is organized as follows. In section 2, we discuss the literature on expertise and forecasting and review the few studies that have looked at the predictiveness of peer review opinions. In section 3, we present data and methods. In Section 4, we show the results of the main and supplemental analyses and of robustness checks. In Section 5 we summarize the findings and discuss the implications for improving the predictiveness and accuracy of expert-based evaluations and for helping organizations to make more informed decisions regarding the funding of research.

2. BACKGROUND

2.1 The value of expertise

Decision-making on research and innovation is largely entrusted to experts, who have a strong intellectual understanding of research and specialized technical and scientific competencies on the domain-specific knowledge of the proposal under evaluation. The reliance on expertise is rooted in the norms of peer review (Chubin and Hackett 1990; Polanyi 1962)

and ubiquitous in funding evaluations.[1] When appointing a panel or study section in charge of evaluating applications, a criterion for selection is that the experts cover the spectrum of domains and areas that the program targets (Chubin and Hackett 1990; Cole and Cole 1981; Lamont 2009; Li 2017).[2] The practice is enabled with specialized software that screens individual publishing profiles to recognize suitable evaluators.[3] Experts themselves place great reliance on specialized knowledge, respecting the sovereignty of disciplines and the areas of competence of their colleagues. In difficult evaluations, they often turn to the most specialized expert among them for the last word (Lamont and Huutoniemi 2011; Ottaviani and Sørensen 2001).

The seminal studies of expertise are rooted in 1970s information processing theory (Cyert and March 1963; Simon 1978, 1979). Since these early works, expertise has been understood as a superior endowment of information, stored in a structured way in the long-term memory that experts can retrieve at need (Chase and Simon 1973; Simon 1979). The proficiency of experts would thus depend on both on their memory and the cognitive structure of their mental processes (Lord and Maher 1990; Simon 1978). Experts are seen as having the ability to recall knowledge in a reliable way and identify known patterns within large problems, without being nudged to do so (Fiske, Kinder, and Larter 1983; McKeithen et al. 1981). They can then "chunk" problems into smaller and more tractable subunits that can be solved with specific knowledge, heuristics or known solution paths (Chi, Feltovich, and Glaser 1981a; Ericsson and Kintsch 1995). By contrast, non-experts are often influenced by more salient or superficial information that may not be or fruitful for identifying effective solutions, while they overlook less noticeable, yet potentially more useful and valid cues (Chi, Feltovich, and Glaser 1981b; Fiske et al. 1983; McKeithen et al. 1981). A second advantage of experts relates to their metacognitive skills, that is to their information regarding the broader knowledge domain

[1] By way of example, the Center for Scientific Review of the NIH in 2023 engaged approximately 19,000 distinct reviewers every year to evaluate the NIH grant proposals. https://public.csr.nih.gov/AboutCSR/Evaluations#overview. Accessed January 16, 2023.

[2] The NIH stresses the following: "Expertise is the paramount consideration when developing/updating a roster. Each scientific area reviewed by the scientific review group needs expert representation." https://public.csr.nih.gov/ForReviewers/BecomeAReviewer/CharteredReviewers. Accessed January 16, 2023. The reliance on expertise is also mirrored in the realm of innovation funding. Companies give decision power to internal committees of experts selected among technical leaders, senior managers, or heads of the R&D labs (Criscuolo et al. 2017; Lerner and Wulf 2007). Business angels and startup mentors are often executives with successful entrepreneurial backgrounds (Huang and Pearce 2015; Scott et al. 2020).

[3] E.g., Web of Science Reviewer Locator: (https://clarivate.com/products/scientific-and-academic-research/research-publishing-solutions/web-of-science-reviewer-locator/), Springer Nature Reviewer Finder (https://www.springernature.com/gp/editors/resources-tools/reviewer-finder). Accessed Jan 10, 2024.

(Flavell 1979). Metacognition involves a greater aptitude to assess the trustworthiness of information presented as factual, and the capacity to conduct further scrutiny to ascertain reliability, based on direct knowledge about the source or about the context of the information (Alter and Oppenheimer 2007; Pintrich 2002). For example, consider research that proposes to adopt a certain method to solve a challenging problem, claiming this is the best possible strategy. An evaluator who does not know the technicalities of this research method would not be able to tell if this assertion is valid and should take the proposed solution at its face value. One who knows the technicalities of the strategy but is not acquainted with the specific field of research could assess the feasibility of the solution in the abstract but could not tell if the strategy is good in relative terms, compared to alternative strategies that others are attempting or have attempted, if these are not discussed in the proposal. Only an evaluator who knows both the technicalities and the field of research could do so. Consequently, the meta-knowledge of the field enables domain experts to question the significance of what is proposed or identify possible caveats by looking beyond the information provided (Brand-Gruwel et al. 2017; Lucassen and Schraagen 2011).

2.2 The fallacy of expertise

Around the same time that information processing studies were focusing on the problem-solving abilities of experts, social psychologists were focused on the cognitive biases of the human mind. One question that arose naturally was if expertise was an antidote to the defects of human cognition (Kahneman 2011). Contrary to the initial expectations, the studies indicated that experts are subject to general human biases (Camerer and Johnson 1991; Cooke 1991; Kahneman and Klein 2009). Empirical studies have shown that expert judgments are influenced by extraneous factors, such as moods and emotions, especially when they are based on intuition, instead of slow logical thinking (Danziger, Levav, and Avnaim-Pesso 2011; Dushnitsky and Sarkar 2022; Hirshleifer and Shumway 2003; Kahneman and Klein 2009). Other studies have also shown that experts often disagree with each other or provide inconsistent opinions over time (Litvinova et al. 2022; Mumpower and Stewart 1996). In addition, experts are prone to overconfidence, understood as the tendency to maintain high confidence in erroneous beliefs or to provide confidence intervals that are overly narrow (Ben-David et al. 2013; Bradley 1981).

A separate stream of studies has focused on forecasting, i.e., the ability to make predictions under conditions of uncertainty (Armstrong 1991; Clemen and Winkler 1985; Morgan 2014; Tetlock and Gardner 2015). Here too, the results indicate that humans are

generally poor forecasters and that experts do not generally do better than basic forecasting algorithms, such as those assuming the continuation of existing trends (Armstrong 1991; Tetlock 2017). The small groups of people who appear to systematically predict better than chance, called "superforcasters", are not easy to identify by externally observable characteristics, including those who had domain-specific knowledge (Mellers et al. 2015; Tetlock 2017; Tetlock and Gardner 2015). A common problem with these studies is that they adopt variable and inconsistent definitions of expertise, conflating domain expertise with experience (e.g., length of exposure), training (e.g., formal education in a subject), decision-making authority (e.g., holding specialized or top positions in organizations), proficiency (high levels of performance), and so on (Budescu and Chen 2015). Moreover, many of these studies focus on contexts such as consumer behavior, geo-political events, sports, or popular culture, where expertise is difficult to define and to measure. In contrast, research funding decisions are an area where some level of specialized domain knowledge seems inherently necessary to understand even the basic content of the sophisticated technical details provided in proposals. In this area, evaluators' domain knowledge is also well documented in the body of publications that the person has done in the past.

2.3 Predictive validity of research proposal evaluations

To understand the degree to which reviewers are capable of scrutinizing proposals and identify those more likely to be productive, it is necessary to compare the scores provided before the research is funded to the outcomes eventually occurred after funding. Because of the limited availability of data, few empirical studies have done so. A summary of these is provided in Table 1. Two things are worth noting. First, none of the studies has individual reviewers' scores. Instead, they have evaluations that express the collective opinion of a panel of experts, in rank format. Second, four of the five studies consider only funded proposals and do not model the selection into funding, which creates sample selection issues. Looking at the findings, three works report no or very weak correlation of panel evaluations to outcomes (Danthi et al. 2014; Doyle et al. 2015; Reinhart 2009). The two remaining studies are based on NIH data, and are connected, in the sense that they are a study and its re-analysis. Collectively, they indicate that panels' ranking predicts the outcomes of top-ranked proposals, while they are a weak predictor of proposals ranked closer to the payline (Fang et al. 2016; Li and Agha 2015).

[Table 1]

None of the studies mentioned examines the role of expertise as a moderator of predictive power. Likewise, no study to our knowledge has analyzed prediction errors in science funding decisions. Worth mentioning is a study by Li (2017) that has information on the panel members involved in making the collective judgment of the review group. The study looks at the performance of articles whose titles were similar to those of the applications for both funded and unfunded proposals. It finds that panels including permanent committee members who were familiar with the research of PIs, measured by citations to the PI's work, were generally more capable of identifying higher quality proposals, but not more capable to identify lower quality proposals. This, along with simulations calibrated on real data that highlight the importance of evaluators' expertise (Feliciani et al. 2022), underscores the need for comprehensive empirical studies to delve deeper into the role that evaluators' expertise plays in evaluation accuracy. This is the scope of the present paper.

3. DATA AND METHODS

3.1 Research context: Evaluation of research proposals at NNF

We use data from the Novo Nordisk Foundation (NNF), an independent non-profit funder of scientific research, based in Denmark. Between 2016 and 2023, the NNF provided the equivalent of over $6.8 billion in grants predominantly in the medical and health sciences and is currently the world largest private funder of scientific research.[4] The NNF approach to evaluating proposals is a multi-step process, similar to the process of NIH and other funding institutions.[5] First, applications undergo an initial check of eligibility for administrative requirements, conducted by the program officers of NNF. Those clearing the check are forwarded to a review committee, typically composed of 5-12 internationally recognized scholars. The review committees are officially appointed by NNF and its members, who are hired, normally stay in the committee between 2 and 7 years. The members of the committee serve as the evaluators (or *reviewers*) of the proposals. The evaluation process takes place in two steps. In the first step, each proposal is assigned by the program officer to a minimum of two committee members (mean=3.09; SD= 2.31), who have reasonably-close expertise in the

[4] See https://novonordiskfonden.dk/en/facts-and-figures/. Accessed 2024-06-18.
[5] The one outlined is the main peer-review process followed by NNF. Some calls may include variations. See *https://novonordiskfonden.dk/en/news/novo-nordisk-and-novozymes-prizes-awarded/*. On NIH funding model see e.g., https://grants.nih.gov/grants/peer-review.htm#Initial. Accessed 2024-01-08.

proposal and have no conflicts of interest.[6] Important to our methodology, is that reviewers do not have much control over the number of proposals assigned to them in a given round. We will return to this when we discuss reviewers' workload as the exclusion restriction. Each reviewer initially evaluates the proposals assigned independently from the other reviewers.[7] The individual evaluations (or *reviews*, or *opinions*) are provided in the form of a single numerical score, with a brief comment substantiating the score.[8] Each evaluation is then independently submitted to the program officer, who uses all the scores received to prepare for the second and final evaluation step, in which the entire committee meets. In preparation for the meeting, the officer sets aside the proposals that received the poorest scores, which will not be discussed again, unless a committee member explicitly asks to rescue the proposal. The panel discussion leads to a final collegial recommendation to fund or not, which is then upheld by the board.[9] This comprehensive review process typically spans 6 months and ends with the PIs being notified of decisions.

To conduct our study, we had access to full proposals (funded and unfunded), the identity of the PIs and of the reviewers, the scores that the reviewers provided independently in the first step, prior to the committee meeting and the final decision. Until 2017 NNF adopted a score scale from 1 (excellent) to 5 (poor). In 2018, the scale was changed from 1 (excellent) to 6 (poor). Despite the change, the score of 6 was virtually never used. Scores are integer, although there are a few exceptions of .5 scores. To account for the change in the scale and the different uses of these, we first normalize the scores at the call level, by dividing each score by the average score of all applications in the call and then take the quintile score rank.

3.2 Data sample

We use all research proposals for research grants (or *applications*) submitted to NNF between 2012 and 2018, involving 759 funding calls related to 10 different disciplines, with a strong focus on life sciences. We excluded all funding calls for scholarships, non-research proposals, and startup incubation. We also exclude 561 proposals-scores pairs (56 related to

[6] Committee members have an obligation to report any potential conflict of interest. Program officers resolve conflicts by reassigning the evaluation to a committee member that is in no position of conflict. There is a strict rule that people who are in a potential conflict of interest must not participate in any part of the decision.
[7] Criteria are described in the guidelines but are not scored separately. These normally include quality, novelty, potential impact, the budget, and the qualifications of the PI. Additional criteria may be added in specific calls.
[8] Proposals are not ranked in relative terms and ties are admitted.
[9] Unlike the NIH study sections, the panelists do not review or update individual scores after the meeting and a final score is not published. As a result, it can be reasonably assumed that the scores provided in the first step are independent of each other.

funded proposals and 505 related to unfunded proposals) from reviewers who served in a single call and could not be estimated using reviewers fixed effects. The resulting dataset comprises 6,636 evaluations of 5,769 unique applications and 72 unique reviewers.[10] The data comprises both funded and unfunded applications. The average funding rate in our sample was 17.4%.

3.3 Outcome variables

In our dataset, 2,495 evaluations relate to applications that were subsequently accepted for funding. For these applications, we collect data on funding outcomes (until the end of 2023) from the API version of the database Dimensions.ai (*https://www.dimensions.ai*.) at the beginning of 2024, allowing a minimum of 5 years from the award of the grant. We collect three measures of academic performance: publications, citations, and cumulative Field Citation Ratio (FCR). Cumulative FCR is the sum of citations received by all publications acknowledging the grant, divided by the average citations received by publications in the same field and year. Consequently, FCR is normalized to consider the advantage that older grants and larger fields have in accumulating citations and is particularly suitable to compare grants from different fields and years. We complement the three measures of academic performance with a metric of public influence, the Altmetric score, which measures the attention received by research outputs in public outreach, such as social media, coverage in the mainstream press and in blogs, citations in policy reports, Wikipedia, and other documents accessible online. This outcome measure enables us to provide a more holistic view of an application's success by considering both traditional academic impact and broader non-academic influence.

3.4 Measure of domain expertise

We identify reviewers based on their full names, surnames, and institutional affiliations in the database Dimensions.ai. We retrieve the complete publication history for each reviewer, including titles and abstracts of their publications. We use these data to build a measure of the expertise of the reviewer about the specific application that she or he is scoring. As said, the concept of expertise is multifaceted and various operationalizations have been proposed to capture its different dimensions. In this paper we want to measure *domain* expertise. We then take the approach of Boudreau et al. (2014), who aligns expertise with intellectual proximity, a view echoed by earlier researchers (Libby and Frederick 1990), and measured the cosine-similarity between the research proposal and the publications of the evaluators. While

[10] Adding reviewers fixed effects in our estimates (see Section 4) implies a loss of 561 proposals-scores observations (56 related to funded proposals and 505 related to unfunded proposals).

Boudreau and colleagues computed the cosine similarity between the MeSH terms extrapolated by the applications and reviewers' publications, we instead calculate the cosine similarity between the complete titles and summary of the proposals and the titles and abstracts of the reviewer's publications. To this aim, we leverage BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art large language model to generate high-dimensional vector representations that capture the semantic essence of the documents and the underlying concepts with considerable precision. The resulting metric is specific to each pair of application-reviewer at the time of the review.

Cosine similarity can potentially range between 1 (perfect congruence) and -1 (complete dissimilarity). In our sample, the measure ranges between 0.285 and 0.957 (mean 0.741; SD 0.119). We use both the continuous measure of cosine similarity (*expertise*) and a binary variable (*expert evaluator*), in which 1 denotes high expertise, equivalent to the top quartile of cosine similarity (0.829 or more). The threshold was carefully chosen with supplemental analyses and controlled with manual reading (see robustness checks and Appendix). To illustrate the metric with an example, consider an evaluator who is an endocrinologist with extensive experience in clinical research on endocrinology and metabolism. This reviewer would have a cosine similarity=0.41, equivalent to the 1st percentile (virtually no domain expertise) in evaluating a proposal on neuroscience research regarding the neuronal activity of patients with brain disorders. She would have a cosine similarity=0.78, equivalent to the 50th percentile (low domain expertise) in evaluating a proposal on the clinical treatment of chemotherapy-resistant metastatic colorectal cancer (possibly due to the contiguity of the endocrine glands on the intestine) and a cosine similarity of 0.93, equivalent to the 99th percentile (very high expertise) in evaluating a proposal on the clinical treatment of pre-diabetes in obese patients. The same person would have a cosine similarity of 0.829 (equivalent 75th percentile) and equal to our threshold of high expertise in evaluating a proposal regarding the beneficial effects of gut microbiome on host physiology and metabolism.

3.5 Variables description

Table A1 of the Appendix provides a detailed description of all the variables of observation and the controls used in the analysis. Summary descriptive statistics are provided in Table 2 and the Pearson's pairwise correlation matrix is provided in Table 3.

[Table 2 and Table 3]

4. RESULTS

4.1 Descriptive statistics

Table 4 shows the distribution of the quintile scores that expert and non-expert evaluators have given to all applications scores (left) and to the subset of applications that received funding (right). Expert and non-expert evaluators demonstrate a distinct pattern in their evaluations, as indicated by a different distribution across all applications (χ^2(df=4) =85.780; p-value =0.000). In particular, the difference in the mean quintile score between experts (2.846) and non-experts (3.013) computed across the entire sample is statistically significant (p-value = 0.000), indicating that experts generally give more positive evaluations (i.e., lower quintile scores) than non-experts. A similar pattern applies to the subsample of funded applications, which also exhibit non-equal distributions (χ^2(df=4) =86.644; p-value =0.000). Not surprisingly, applications that were funded have a higher incidence of positive evaluations (quintile scores 1 and 2) from both experts and non-experts.

Table 5 shows the mean values of the four post-funding outcomes of funded proposals, arrayed by the evaluations provided before funding and by experts and non-experts. We notice that the mean outcome of the applications evaluated by experts is consistently lower across all measures of post-funding outcome. Moreover, and quite interestingly, the difference in the mean outcomes between experts and non-experts tends to increase from excellent to poor scores. This indicates that the opinions of experts and non-experts tend to diverge more towards the bottom range of the evaluation scale, i.e., for worst scores. Although these statistics do not consider selection, they suggest that non-experts tend to underestimate (give more negative scores) applications that perform quite well after funding, hinting a correlation between expertise and predictive power. We will turn to this in the next sub-section.

[Table 4 and Table 5]

4.2 Predictiveness of evaluations

To understand the predictive power of evaluation scores provided by reviewers with various degrees of expertise, we want to model the relationship between the evaluation scores and post-funding outcomes and study if the evaluator's expertise moderates this relationship. The approach poses important methodological challenges. A first problem is that, although we observe the scores for all funded and unfunded proposals, actual outcomes are observable only for funded proposals. Consequently, estimates are exposed to sample-induced endogeneity that

occurs when an omitted variable influences both the probability of entering the sample (i.e., being funded) and the dependent variable (i.e., the post-funding outcome), causing a correlation of the error terms of the selection and outcome equations. In our case, it can be argued that application quality, which is unobservable, is positively correlated with the error term of the selection for funding (i.e., better applications are more likely to be funded), and with the error term of the outcome equation (i.e., better applications have better post-funding outcomes). To address this issue, we use a two-stage estimation approach (Heckman 1979) where the first stage (selection equation) models the influence of the scores on the selection for funding with a probit model, and the second stage (outcome equation) models the capacity of the scores to predict the post-funding outcomes with an OLS. The Inverse Mills Ratio obtained from the selection equation is included in the outcome equation (lambda), to correct for sample selection bias. The Heckman model should include in the first stage a strictly exogenous variable (i.e., uncorrelated to the omitted variable generating the sample selection bias), known as exclusion restriction, which should affect the probability of entering the sample, but not the dependent variable in the outcome equation (Certo et al. 2016). We use as the exclusion restriction the *workload* of the evaluator, measured by the total number of applications (from all calls) scored by the reviewer in the same month. The rationale of this choice is that a higher workload influences the selection decision, by reducing the amount of time the evaluator can spend reviewing each application, making the selection noisier, but it does not affect the research or the outcomes that this will produce after funding. Essential to our argument is that reviewer workload is a source of exogenous variation affecting selection and it is not correlated to the unobserved quality of proposals. Based on our interviews with NNF program officers, reviewers, who are all hired as committee members, do not have much control on the number of proposals they are asked to score in each round. Their workload varies primarily based on the total number of submissions received for that call, and the number of calls in which the committee is involved at the same time. Workload varies considerably by call and over time. Moreover, the calendar of calls has varied from one year to the other and some committees were involved in multiple overlapping calls. Consequently, workload has a high absolute variation, ranging from a low of 1 to a high of 189 applications to review (mean 72.7; SD 41.8). We provide additional information and discussion on workload in Section A1 of the Appendix.

A second methodological challenge concerns the possibility that one of our main variables of interest, domain expertise, is directly or indirectly correlated to heterogeneity in prediction difficulty, causing identification problems. A specific concern is that some research

outcomes may be easier to predict than others and an unobserved mechanism causes easier prediction tasks to be assigned more often to experts then to non-experts. For example, mainstream proposals may be easier to predict because reviewers have more background material and information at their disposal. Mainstream proposals may also be more likely to be assigned to a domain expert, compared to proposals that are not mainstream, because committees are composed to represent all major research areas. It is also possible that reviewers with better forecasting skills are themselves more involved in research areas that are easier to predict.

Our estimation approach is correlational, and we cannot address identification entirely. However, we mitigate the problem by controlling for reviewer fixed effects in all estimates. The rationale is that reviewers fixed effects should capture the heterogeneity that relates to research domains, as well as individual forecasting skills.

[Table 6]

Table 6 reports the results of the estimates of the Heckman two-stage models with reviewers fixed effects. Panel A reports the baseline models correlating evaluation quintile scores to outcomes, prior to adding our measure of evaluator's expertise. Column 5 of Panel A shows the estimates of the first-stage equation, modelling the probability of being funded as depending on the exclusion restriction, the score, and a large set of controls. The coefficient of the workload variable is negative and significant ($p<0.01$), indicating that the funding rate is higher for lower values of the exclusion restriction. We note that, net of control variables, the evaluation scores predict funding. Recalling that the score scale is reverted (top scores are low numbers), more negative evaluations (i.e., higher scores) are associated with a lower probability of being funded ($p<0.01$). Columns 1 to 4 of Panel A report the baseline estimates of the outcome equation. We also note that the evaluation scores have no or weak predictive power. In particular, reviewers' opinions are weakly associated with citations ($p<0.10$), while they do not significantly predict other post-funding outcome measures.

Panel B of Table 6 reports our main analysis, where we add a continuous variable of evaluators' *expertise*, standardized to ease interpretation (mean zero, unit SD), plus its interaction term to the quantile score. Prior to commenting on the results, we note that the lambda coefficient in the outcome equations (Columns 1-4) is always positive and significant. This, together with the statistical significance of the evaluation score in the selection equation

(Column 5), confirms the existence of a sample selection bias and the need for the Heckman correction.[11]

Column 5 of Panel B reports the estimated coefficients of the selection equation. As in the prior model, quintile scores predict funding. Moreover, the interaction term between the score and the evaluator's expertise is negative and statistically significant ($p<0.05$), indicating that better (i.e., lower) scores provided by experts are increasing the probability of funding. This is consistent with the view that experts' evaluations carry more weight than non-experts' evaluations in deciding funding. Columns 1 to 4 of Panel B show that quintile scores and the interaction term between these and *expertise* are significant ($p<0.01$), indicating that expertise moderates the evaluator's capacity to predict outcomes. Specifically, a one-unit worst quintile score given by an evaluator with average expertise (i.e., cosine similarity = 0.74) is associated with 11% fewer publications, 31% fewer citations, 21% lower FCR and 20% lower public influence. A one-unit worst quintile score given by an expert evaluator (i.e., with a level of expertise that is 1 standard deviation above the mean (cosine similarity = 0.86), is associated with 21% fewer publications, 52% fewer citations, 38% lower FCR and 33% lower public influence.

In Panel C of Table 6, we provide an alternative estimation in which we replace our continuous measure of expertise with the binary variable *expert evaluator*, taking a value of 1 for the 75^ percentile of domain expertise (cosine similarity ≥ 0.83). As before, the estimates confirm the need for the Heckman correction. The coefficient of the interaction term *quintile score X expert evaluator* in the selection equation (column 5, Panel C) is negative ($p<0.01$), confirming the previous evidence that experts' scores are more influential in the decision to fund the proposal. Quite interestingly, the coefficient of *expert evaluator* is positive ($p<0.01$). This suggests that applications evaluated by experts have a greater chance of being funded, reflecting the evidence of higher average scores by expert reviewers observed in the descriptive statistics. The estimates of the outcome equations (columns 1-4) indicate that the interaction term *scores x expert evaluator* is negative ($p<0.01$), confirming that experts' scores predict post-funding outcomes better than non-experts', net of selection into funding. The magnitude of the estimated effects is sizable. A one-unit worst quintile score given by an expert is associated with 20% fewer publications, 63% fewer citations, 44% lower FCR and 44% lower

[11] The existence of a sample selection bias requires both a significant coefficient for the main variable of interest (in our case, the evaluation score) in the selection equation and a significant lambda in the outcome equation (Certo et al., 2016).

public influence. The quintile score is not statistically significant at conventional levels, indicating that the evaluations of non-experts do not predict post-funding outcomes.

[Table 6]

In supplemental analyses included in section A2 of the Appendix (Table A2), we explore alternative variable specifications that capture alternative/more granular levels of expertise. The results confirm the effects found in the main analyses.

4.2 Accuracy of evaluations

The estimates just seen modelled the *predictiveness* of evaluations and indicated that the predictive power of evaluations is contingent on domain expertise. We next want to understand more about the *prediction errors* (*accuracy*) made by experts and non-experts and how these vary along the score scale. Following the approach of the forecasting literature (Broomell and Budescu 2009; Larrick and Soll 2006; Lehmann and Casella 1998; Tetlock and Gardner 2015), we measure the *prediction error* as the difference between the *actual* and *predicted* value of proposals, considering the quintile scores provided before funding as the evaluator's *prediction* regarding the potential value that the proposal would generate for science and society if funded and the quintile rank of post-funding outcomes (on various indicators) as the *actual* value of the proposal observed after funding. The legitimacy of these assumptions is corroborated by a positive rank correlation of distributions observed before and after funding.[12] We then calculate the prediction *error* of the evaluator *i* as the absolute difference between the actual quintile rank of proposal *j* based on its actual post-funding outcome *k* and the quintile score rank of proposal *j* predicted by evaluator *i* before funding, as in the following formula:

$$Prediction\ Error_{ij} = |Actual\ Rank_{jk} - Pre\text{-}funding\ Rank_{ji}|$$

An evaluation is fully correct (or accurate) when the quintile score predicted matches the actual quintile rank of real-world outcomes so that the prediction error is zero. It is incorrect as the two values diverge. For instance, if an evaluator assigns a score placing an application

[12] Spearman rank correlation (r=0.22, p-value=0.00) significant at conventional statistical levels.

in the fourth quintile (score=4), but its cumulative FCR positions it in the third quintile, the resulting prediction error would be calculated as |3 - 4| = 1.[13]

Table 7 shows the mean values of the prediction errors of the four post-funding outcomes, divided by the quintile scores of experts and non-experts. We notice that the mean prediction error of experts is consistently higher across all post-funding outcome measures, suggesting that domain experts are less accurate than non-experts. Quite interestingly, the differences in the mean prediction errors between experts and non-experts are positive for lower quintile scores and negative for higher scores, suggesting that non-experts are more accurate than experts when they give positive scores, while experts are more accurate than non-experts when they give negative scores.

[Table 7]

Table 8 reports the results of the multivariate analysis of prediction errors associated with the four post-funding outcomes (Columns 1 to 4). The selection-equation of the Heckman model is identical to the one reported in Column 5 of Table 6-Panel C. The coefficients of the *expert evaluator's* variable are positive ($p<0.01$) and confirm that, on average, domain experts exhibit larger average errors (i.e., are less accurate) than non-experts.

However, the negative coefficients of the interaction terms ($p<0.01$) between *expert evaluator* and the quintile score are indicative of important differences across the spectrum of the scores. To shed further light on this, we show in Figure 1 the estimated difference in prediction errors (vertical axes) made by domain experts V non-experts, at the five values of the score (horizontal axes) with 95% confidence bars. We note that the experts are more correct than non-experts when they assign the worst scores (i.e., 4 and 5). The difference in the prediction errors (experts – non-experts) is indeed negative and significant ($p<0.05$) for all funding outcome measures when the quintile score is higher than 2. Experts are instead less correct (make larger prediction errors) than non-experts when they assign top scores (i.e., 1). The latter difference is positive and significant ($p<0.05$) for all outcome measures.

We conclude that, while only domain experts' scores have predictive power, experts are not always equally accurate. Specifically, they demonstrate higher accuracy (make smaller

[13] Other approaches use quadratic, instead of absolute error. Results in our case do not change (see robustness checks).

prediction errors) than non-experts when giving negative evaluations. However, they demonstrate lower accuracy (make larger prediction errors) compared to non-experts, when they give positive evaluations.

[Table 8 and Figure 1]

4.3 Supplemental analyses

We run a supplemental analysis to understand whether the ability of expert evaluators to predict the post-funding outcome is also exhibited by evaluators with a strong reputation of scientific excellence in research. To do so, we construct a binary variable (*prestige*) that equals 1 if the evaluator has published articles in Science or Nature prior to the time of the evaluation.[14] In Table 9 we include both the binary variable *expert evaluator* (75^ percentile of cosine similarity) and the binary variable *prestige*, and their interaction terms in the same econometric model. The coefficient of *expert evaluator* and its interaction remain stable and consistent with the previous estimates. The variable denoting evaluators with prestigious publications (*prestige*) is weakly significant ($p<0.1$) in the selection equation (column 5), indicating that prestigious reviewers too are influential in funding decisions. However, the interaction terms of the outcome equations (columns 1-4) indicate that their scores are not more predictive than those of average reviewers. Indeed, the coefficients of *quintile score X prestige* are either not statistically significant or weakly significant ($p<0.10$), but with signs opposite to those expected (Panel A and B of Table 9). In conclusion, the results indicate that, although evaluations from scholars with prestigious publications are also more influential in deciding funding, they do not carry more value than those of evaluators with average or low expertise, suggesting that the reviewer's prestige is not a substitute for domain expertise in evaluations of research proposals.

[Table 9]

4.4 Robustness checks

In unreported estimates, we run four additional robustness checks, available from the authors upon request. First, we replicate our main analyses by using OLS instead of the

[14] The variable is not correlated to *expertise* (Pearson's pairwise correlation coefficient is 0.037).

Heckman two-stage models. Results from OLS confirm that experts' scores predict post-funding outcomes better than non-experts' scores. Results on prediction errors models also confirm that the prediction errors of experts increase (decrease) when they assign more positive (negative) scores. Second, results are robust when employing Poisson or Negative Binomial to model outcome variables that are counts (publications and citations), rather than continuous variables. Third, we tested different thresholds of cosine similarity to codify *expert evaluators*. All results are robust to using threshold at the 70th, 80th and 90th percentiles of cosine similarity. Fourth, results remain stable when introducing supplementary control variables, such as the number of evaluators per application, or evaluators' research productivity, measured by their publication counts before the application date. Fifth, the models of accuracy give comparable results if we use the quadratic error (Davis-Stober et al. 2014), instead of the absolute error, in the related estimates.

5. DISCUSSION AND CONCLUSION

Funding decisions for research and innovation rely on evaluations from reviewers who have domain expertise on what is evaluated. The predictive power and accuracy of these evaluations are critical to organizational success and societal advancement. However, empirical studies are constrained by data availability and important methodological challenges, such as sample selection and measurement of research outcomes. Domain expertise is also difficult to conceptualize and measure outside of academia, casting doubts about the generalizability of results obtained by studies of forecasting conducted in different domains. Understanding the link between expertise and predictive ability is also of paramount importance in view of the growing availability of AI-based applications that can, support, integrate or even replace human judgment.

We analyzed a new dataset containing 16,636 independent evaluations of research proposals that sought funding from NNF, the world largest private funding institution. The findings indicate that the predictive power of peer review opinions is contingent on domain expertise. Experts provide opinions that are on average predictive of the actual research performance, as measured by publications, citations, field-citation ratios, and influence (altmetrics) of funded applications. In supplemental analysis, we find that the research prestige of the evaluators, as measured by having publications in Science & Nature, is not a substitute for domain-specific expertise. Indeed, while of both domain experts and excellent/prestigious scholars have stronger influence on funding decisions, only the opinions of domain experts and

not those of excellent scholars, are predictive of subsequent research outcomes. This suggests that, although both excellence and expertise carry weight in the decision process, it is expertise that makes a real difference.

We further examine the accuracy of the evaluators, measured as the granular prediction error, or difference between the proposal rank predicted and the actual rank observed post-funding. We notice that, although experts have predictive power, they make larger errors than non-experts on average. Specifically, experts are more correct than non-experts when giving negative evaluations, but they make larger errors than non-experts when giving positive evaluations. In simple words, experts are accurate when they say "no", but are prone to large overestimations when they say "yes".

Overall, our findings indicate that domain-specific expertise is critical to making good funding decisions, but experts are also likely to promote several proposals in their areas of expertise that will later underperform.

Multiple mechanisms could explain and be compatible with our findings. We can only speculate on some of these, ruling-off some possible interpretations. In particular, the evidence provided does not seem to be compatible with bias arising from affect or emotional attachment to one's own area of expertise, since this would likely inflate both positive and negative evaluations equally. A high incidence of error for positive evaluations could also be caused by a preference of experts for more volatile or risky proposals, which they would score as very good, but would subsequently lead to high incidence of failures. To explore this possibility, we computed a metric of combinatorial novelty (Shibayama, Yin, and Matsumoto 2021) for a subsample of research proposals in our data submitted after 2015, of which we have references that were cited in the text. However, the post-hoc analysis of proposal novelty (available upon request to the authors) indicates that experts do not generally favor novel proposals. This is also in line by previous evidence (Boudreau et al. 2016), suggesting that this is probably not the mechanism behind our results. Additional explanations are possible. The evidence of higher average errors in positive opinions may be caused by experts overestimating the importance of the research outcomes produced in their domain, relative to other domains. Experts may also be more exposed to confirmation biases, inflating the scores of those proposals that resonate with their expectations, some of which turnout to be incorrect (Nickerson 1998). Finally, experts may behave strategically and try to favor several underperforming proposals in their

areas of expertise (Krishna and Morgan 2001; Ottaviani and Sørensen 2006). Future research may investigate these and other possible mechanisms further.

Our results have limitations. Research output is inherently difficult to conceptualize and measure. Although we have collected four different measures, each measure is prone to its own biases. It is possible, for instance, that experts and non-experts weigh the perspective value of research differently. Future work with alternative measures may shed more light on this important caveat. A second limitation relates to sample selection bias, which we tackled with two-stage estimates. Future research may use alternative approaches. Finally, researchers fixed effect may not entirely capture potential factors that may confound the relationship of expertise and the predictive validity or accuracy of the scores. Furthermore, our data are based on only one funding agency. It would be interesting to see if the results are generalizable to other sets of data.

Despite the limitations, our results are based on original data with unprecedented level of detail and look not only at the predictiveness but also at the accuracy of evaluations. Moreover, the results obtained are very robust to alternative models, specifications, and outcome variables. In terms of the implications, our findings suggest that the availability of expert reviewers is critical to have good evaluations and that good reviewers are those with technical and scientific competences that align closely to the content of the proposal, not necessarily those with high academic prestige. Many funding agencies use panelists as individual evaluators, rather than appointing ad-hoc reviewers. As a result, unusual or off-the-beaten-path evaluations may fall outside the area of expertise of panel members; these evaluations would not only be more problematic but also less favored, as non-expert reviewers tend to undervalue several worthy proposals. The additional use of ad-hoc reviewers for proposals outside the panel members' area of expertise may be a potential solution to this problem. Finally, our analyses focus on individual predictions. However, each proposal is normally evaluated by multiple experts, so greater accuracy of evaluations can be achieved not only by more accurate individual evaluations but also by averaging. Averaging eliminates random errors, but not necessarily biases (Broomell and Budescu 2009; Larrick and Soll 2006). Our results show that experts have a positive bias when giving top scores, indicating that grouping more experts may not solve this problem, and may even exacerbate it. Future research should examine the *average* accuracy that can be achieved by pooling evaluations of reviewers with heterogeneous biases.

6. BIBLIOGRAPHY

Alter, Adam L., and Daniel M. Oppenheimer. 2007. "Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning." *Journal of Experimental Psychology: General* 136(4):569–76.

Armstrong, J. Scott. 1991. "Prediction of Consumer Behavior by Experts and Novices." *Journal of Consumer Research* 18(2):251–56. doi: 10.1086/209257.

Åstebro, Thomas, and Samir Elhedhli. 2006. "The Effectiveness of Simple Decision Heuristics: Forecasting Commercial Success for Early-Stage Ventures." *Management Science* 52(3):395–409.

Ben-David, Itzhak, John R. Graham, and Campbell R. Harvey. 2013. "Managerial Miscalibration." *The Quarterly Journal of Economics* 128(4):1547–84. doi: 10.2307/26372532.

Boudreau, K. J., E. C. Guinan, K. R. Lakhani, and C. Riedl. 2016. "Looking across and Looking beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science." *Management Science* 62(10):2765–83. doi: 10.1287/mnsc.2015.2285.

Boudreau, Kevin, Eva Guinan, Karim R. Lakhani, and Christoph Riedl. 2014. "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance and Resource Allocation in Science." *Management Science* 62(10):2765–2783.

Bradley, James V. 1981. "Overconfidence in Ignorant Experts." *Bulletin of the Psychonomic Society* 17(2):82–84.

Brand-Gruwel, S., Y. Kammerer, L. van Meeuwen, and T. van Gog. 2017. "Source Evaluation of Domain Experts and Novices during Web Search." *Journal of Computer Assisted Learning* 33(3):234–51. doi: 10.1111/jcal.12162.

Broomell, Stephen B., and David V. Budescu. 2009. "Why Are Experts Correlated? Decomposing Correlations between Judges." *Psychometrika* 74(3):531–53. doi: 10.1007/S11336-009-9118-Z.

Budescu, David V., and Eva Chen. 2015. "Identifying Expertise to Extract the Wisdom of Crowds." *Management Science* 61(2):267–80. doi: 10.1287/mnsc.2014.1909.

Camerer, Colin F., and Eric J. Johnson. 1991. “The Process-Performance Paradox in Expert Judgment: How Can Experts Know so Much and Predict so Badly?” Pp. 195–217 in *Toward a General Theory of Expertise: Prospects and Limits*, edited by K. A. Ericsson and J. Smith. Cambridge: Cambridge University Press.

Certo, S. Trevis, John R. Busenbark, Hyun Soo Woo, and Matthew Semadeni. 2016. “Sample Selection Bias and Heckman Models in Strategic Management Research.” *Strategic Management Journal* 37(13):2639–57. doi: 10.1002/SMJ.2475.

Chase, William G., and Herbert A. Simon. 1973. “Perception in Chess.” *Cognitive Psychology* 4(1):55–81. doi: 10.1016/0010-0285(73)90004-2.

Chi, Michelene T. H., Paul J. Feltovich, and Robert Glaser. 1981a. “Categorization and Representation of Physics Problems by Experts and Novices.” *Cognitive Science* 5(2):121–52. doi: 10.1207/s15516709cog0502_2.

Chi, Michelene T. H., Paul J. Feltovich, and Robert Glaser. 1981b. “Categorization and Representation of Physics Problems by Experts and Novices.” *Cognitive Science* 5(2):121–52. doi: 10.1207/s15516709cog0502_2.

Chubin, Daryl E., and Edward J. Hackett. 1990. *Peerless Science: Peer Review and U.S. Science Policy*. Albany, NY: State University of New York Press.

Clemen, Robert T., and Robert L. Winkler. 1985. “Limits for the Precision and Value of Information from Dependent Sources.” *Operations Research* 33(2):427–42.

Cole, J. R., and S. Cole. 1981. *Peer Review in the National Science Foundation*. Washington, DC: National Academic Press.

Cole, Stephen, Jonathan R. Cole, and Gary A. Simon. 1981. “Chance and Consensus in Peer Review.” *Science* 214(4523):881–86. doi: 10.1126/science.7302566.

Cooke, Roger M. 1991. *Experts in Uncertainty. Opinions and Subjective Probabilities in Science*. New York Oxford: Oxford University Press.

Criscuolo, Paola, Linus Dahlander, Thorsten Grohsjean, and Ammon Salter. 2017. “Evaluating Novelty: The Role of Panels in the Selection of R&D Projects.” 60(2):433–60.

Cyert, Richard M., and James G. March. 1963. “A Behavioral Theory of the Firm.” Englewood Cliffs, NJ: Prentice-Hall.

Danthi, Narasimhan, Colin O. Wu, Peibei Shi, and Michael Lauer. 2014. “Percentile Ranking and Citation Impact of a Large Cohort of National Heart, Lung, and Blood Institute–Funded Cardiovascular R01 Grants | Enhanced Reader.” *Circulation Research* 114(4):600–606.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. “Extraneous Factors in Judicial Decisions.” *Proceedings of the National Academy of Sciences of the United States of America* 108(17):6889–92. doi: 10.1073/pnas.1018033108.

Davis-Stober, Clintin P., David V. Budescu, Jason Dana, and Stephen B. Broomell. 2014. “When Is a Crowd Wise?” *Decision* 1(2):79–101. doi: 10.1037/dec0000004.

DellaVigna, Stefano, and Devin Pope. 2018. “Predicting Experimental Results: Who Knows What?” *Journal of Political Economy* 126(6):2410–56.

Doyle, J. M., K. Quinn, Y. A. Bodenstein, C. O. Wu, N. Danthi, and M. S. Lauer. 2015. “Association of Percentile Ranking with Citation Impact and Productivity in a Large Cohort of de Novo NIMH-Funded R01 Grants.” *Molecular Psychiatry* 20(9):1030–36. doi: 10.1038/mp.2015.71.

Dushnitsky, Gary, and Sayan Sarkar. 2022. “Here Comes the Sun: The Impact of Incidental Contextual Factors on Entrepreneurial Resource Acquisition.” *Academy of Management Journal* 65(1):66–92. doi: 10.5465/amj.2019.0128.

Ericsson, K. Anders, and Walter Kintsch. 1995. “Long-Term Working Memory.” *Psychological Review* 102(2):211–45. doi: 10.1037/0033-295X.102.2.211.

Fang, Ferric C., Anthony Bowen, and Arturo Casadevall. 2016. “NIH Peer Review Percentile Scores Are Poorly Predictive of Grant Productivity.” *ELife* 5. doi: 10.7554/eLife.13323.

Feliciani, Thomas, Michael Morreau, Junwen Luo, Pablo Lucas, and Kalpana Shankar. 2022. “Designing Grant-Review Panels for Better Funding Decisions: Lessons from an Empirically Calibrated Simulation Model.” *Research Policy* 51(4):104467. doi: 10.1016/j.respol.2021.104467.

Fiske, Susan T., Donald R. Kinder, and W. Michael Larter. 1983. “The Novice and the Expert: Knowledge-Based Strategies in Political Cognition.” *Journal of Experimental Social Psychology* 19(4):381–400. doi: 10.1016/0022-1031(83)90029-X.

Flavell, John H. 1979. “Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry.” *American Psychologist* 34(10):906–11. doi: 10.1037/0003-066X.34.10.906.

Heckman, James J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47(1):153–61. doi: 10.2307/1912352.

Henderson, Andrew D., and James W. Fredrickson. 1996. “Information Processing Demands as a Determinant of CEO Compensation.” *Academy of Management Journal* 39(3):575–606. doi: 10.2307/256656.

Hirshleifer, David, and Tyler Shumway. 2003. “Good Day Sunshine: Stock Returns and the Weather.” *The Journal of Finance* 58(3):1009–32.

Huang, Laura, and Jone L. Pearce. 2015. *Managing the Unknowable: The Effectiveness of Early-Stage Investor Gut Feel in Entrepreneurial Investment Decisions*. Vol. 60.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, Daniel, and Gary Klein. 2009. “Conditions for Intuitive Expertise: A Failure to Disagree.” *American Psychologist* 64(6):515–26.

Krishna, Vijay, and John Morgan. 2001. “A Model of Expertise.” *The Quarterly Journal of Economics* 116(2):747–75.

Lamont, Michele. 2009. *How Professors Think. Inside the Curious World of Academic Judgement*. Cambridge, MA: Harvard University Press.

Lamont, Michele, and Katri Huutoniemi. 2011. “Comparing Customary Rules of Fairness Evaluative Practices in Various Types of Peer Review Panels.” Pp. 209–32 in *Social Knowledge in the Making*, edited by C. Camic, N. Gross, and M. Lamont. Chicago and London: University of Chicago Press.

Larrick, Richard P., and Jack B. Soll. 2006. “Intuitions About Combining Opinions: Misappreciation of the Averaging Principle.” *Management Science* 52(1):111–27.

Lehmann, E. L., and George Casella. 1998. *Theory of Point Estimation. Second Edition*. Springer.

Lerner, Josh, and Julie Wulf. 2007. "Innovation and Incentives: Evidence from Corporate R&D." *The Review of Economics and Statistics* 29(4):634–44.

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH." *American Economic Journal: Applied Economics* 9(2):60–92. doi: 10.1257/app.20150421.

Li, Danielle, and Leila Agha. 2015. "Big Names or Big Ideas: Do Peer-Review Panels Select the Best Science Proposals?" *Science* 348(6233):434–38.

Libby, Robert, and David M. Frederick. 1990. "Experience and the Ability to Explain Audit Findings." *Journal of Accounting Research* 28(2):348–67. doi: 10.2307/2491154.

Litvinova, Aleksandra, Ralf H. J. M. Kurvers, Ralph Hertwig, and Stefan M. Herzog. 2022. "How Experts' Own Inconsistency Relates to Their Confidence and between-Expert Disagreement." *Scientific Reports* 12(1):1–12. doi: 10.1038/s41598-022-12847-5.

Lord, Robert G., and Karen J. Maher. 1990. "Alternative Information-Processing Models and Their Implications for Theory, Research, and Practice." *Academy of Management Review* 15(1):9–28.

Lucassen, Teun, and Jan Maarten Schraagen. 2011. "Factual Accuracy and Trust in Information: The Role of Expertise." *Journal of the American Society for Information Science and Technology* 62(7):1232–42. doi: 10.1002/asi.

McKeithen, Katherine B., Judith S. Reitman, Henry H. Rueter, and Stephen C. Hirtle. 1981. "Knowledge Organization and Skill Differences in Computer Programmers." *Cognitive Psychology* 13(3):307–25. doi: https://doi.org/10.1016/0010-0285(81)90012-8.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10(3):267–81.

Morgan, M. Granger. 2014. "Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy." *Proceedings of the National Academy of Sciences* 111(20):7176–84.

Mumpower, Jeryl L., and Thomas R. Stewart. 1996. “Expert Judgement and Expert Disagreement.” *Thinking and Reasoning* 2(2/3):191–212.

Nickerson, Raymond S. 1998. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology* 2(2):175–220.

Ottaviani, Marco, and Peter Sørensen. 2001. “Information Aggregation in Debate: Who Should Speak First?” *Journal of Public Economics* 81(3):393–421. doi: 10.1016/S0047-2727(00)00119-5.

Ottaviani, Marco, and Peter Norman Sørensen. 2006. “The Strategy of Professional Forecasting.” *Journal of Financial Economics* 81(2):441–66.

Pintrich, Paul R. 2002. “The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing.” *Theory Into Practice* 41(4):219–25.

Polanyi, Michael. 1962. “The Republic of Science: Its Political and Economic Theory 54-73.” *Minerva* I(1):54–73.

Reinhart, Martin. 2009. “Peer Review of Grant Applications in Biology and Medicine. Reliability, Fairness, and Validity.” *Scientometrics* 81(3):789–809.

Scott, Erin L., Pian Shu, and Roman M. Lubynsky. 2020. “Entrepreneurial Uncertainty and Expert Evaluation: An Empirical Analysis.” *Management Science* 66(3):1278–99. doi: 10.1287/mnsc.2018.3244.

Shibayama, Sotaro, Deyun Yin, and Kuniko Matsumoto. 2021. “Measuring Novelty in Science with Word Embedding.” *PLoS ONE* 16(7 July):1–16. doi: 10.1371/journal.pone.0254034.

Simon, Herbert A. 1978. “Rationality as Process and as Product of Thought.” *The American Economic Review* 68(2):1–16.

Simon, Herbert A. 1979. “Information Processing Models of Cognition.” *Annual Review of Psychology* 30(1):363–96. doi: 10.1146/annurev.ps.30.020179.002051.

Tetlock, Philip E. 2017. *Expert Political Judgement. How Good Is It? How Can We Know? - New Edition*. Princeton. New Jersey: Princeton University Press.

Tetlock, Philippe, and Dan Gardner. 2015. *Super-Forecasting. The Art and Science of Prediction*. London: Penguin Random House.

TABLES AND APPENDIX

Table 1. Summary of articles on the predictive validity of peer-review

	Reinhart 2009	Danthi et al. 2014	Li and Agha 2015	Fang et al. 2016	Doyle et al. 2015
Sample of applications	Funded and unfunded	Only funded	Only funded	Only funded	Only funded
Data source	ESRC	NHLBI	NIH	NIH	NIMH
Period	1998	2001-2008	1980-2008	1980-2008	2000-2009
Country	Switzerland	US	US	US	US
Field	Biology and Medicine	Medicine	Medicine	Medicine	Medicine
Observations	4,000	1,492	137,215	102,740	1,755
Dependent variable (evaluation)	Mean Score (panel level)	Percentile ranking (panel level)	Percentile Score (panel level)	Percentile score (panel level)	Percentile ranking (panel level)
Outcome variables	Publications Citations	Publications Citations H-index	Publications Citations Patents	Publications Citations	Publications Citations
Correlation	Yes	No	Yes	Yes, but weak	No

Table 2. Summary statistics of the main variables

Variable	Obs	Mean	SD	Min	Max
Publications (log)	2,495	1.535	1.184	0	5.220
Citations (log)	2,495	3.642	2.642	0	9.770
FCR (log)	2,495	2.598	2.070	0	8.146
Altmetric (log)	2,495	3.038	2.441	0	9.644
Quintile score	16,636	2.971	1.386	1	5
Expertise	16,636	0.741	0.119	0.285	0.957
Expert evaluator	16,636	0.253	0.435	0	1
Female PI	16,636	0.334	0.472	0	1
Young PI	16,636	0.237	0.425	0	1
Prior grant	16,636	0.252	0.434	0	1
Size (log)	16,636	14.971	1.200	5.485	17.910
Competition	16,636	0.822	0.181	0	1
Funded	16,636	0.150	0.357	0	1
Workload	16,636	72.686	41.757	1	189

Table 3. Pairwise correlations

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	Publications (log)	1.00												
(2)	Citations (log)	0.92	1.00											
(3)	FCR (log)	0.93	0.97	1.00										
(4)	Altmetric (log)	0.88	0.89	0.90	1.00									
(5)	Quintile score	0.09	0.01	0.04	0.04	1.00								
(6)	Expertise	-0.23	-0.16	-0.19	-0.15	-0.07	1.00							
(7)	Expert evaluator	-0.14	-0.13	-0.14	-0.11	-0.05	0.61	1.00						
(8)	Female PI	-0.12	-0.13	-0.12	-0.16	0.06	0.05	0.05	1.00					
(9)	Young PI	0.05	0.06	0.06	0.06	0.07	-0.02	-0.01	0.05	1.00				
(10)	Prior grant	0.16	0.17	0.16	0.17	-0.17	0.03	0.02	-0.08	-0.15	1.00			
(11)	Size (log)	0.33	0.22	0.27	0.31	0.07	-0.13	-0.16	-0.14	0.00	0.12	1.00		
(12)	Competition	-0.20	-0.07	-0.14	-0.16	0.05	0.25	0.06	-0.04	0.04	-0.07	0.12	1.00	
(13)	Funded	.	.	.	.	-0.35	-0.05	0.02	-0.01	-0.06	0.21	-0.07	-0.49	1.00
(14)	Workload	0.00	0.09	0.06	0.02	0.07	0.03	-0.06	-0.05	0.04	0.07	0.23	0.38	-0.20

Pearson's pairwise correlations. Correlations with post-funding outcomes are calculated on the scores of the 2,495 applications that were subsequently accepted for funding.

Table 4. Distribution of scores by experts and non-experts

	All applications				Funded applications			
	Expert		Non-expert		Expert		Non-expert	
Quintile score	N	%	N	%	N	%	N	%
1	928	22.0	2,073	16.7	413	60.6	823	45.4
2	946	22.5	3,082	24.8	195	28.6	639	35.3
3	874	20.7	2,376	19.1	60	8.8	132	7.3
4	778	18.5	2,391	19.3	8	1.2	97	5.4
5	688	16.3	2,500	20.1	6	0.9	122	6.7
Total	4,214	100.0	12,422	100.0	682	100.0	1,813	100.0

Person's χ^2 statistic (df =4), comparing the distribution of scores between experts and non-experts across all applications is 85.780 (p-value =0.000). Person's χ^2 (df =4) comparing the distribution of scores between experts and non-experts across funded applications is 84.644 (p-value =0.000).

Table 5. Mean post-funding outcomes by score distribution for experts and non-experts

	Publications (log) - Mean			Citations (log) - Mean		
Quintile score	[1]	[2]	[3]	[1]	[2]	[3]
	Expert	Non-expert	Diff (SE)	Expert	Non-expert	Diff (SE)
1	1.358	1.537	-0.179 (0.073)**	3.372	3.760	-0.388 (0.165)**
2	1.075	1.639	-0.565 (0.088)***	2.694	3.938	-1.244 (0.203)***
3	1.301	1.872	-0.571 (0.172)***	2.897	4.022	-1.125 (0.368)***
4	0.657	1.812	-1.155 (0.458)**	1.957	3.551	-1.594 (0.924)*
5	1.079	1.923	-0.844 (0.556)	1.409	3.969	-2.560 (1.150)**
	1.261	1.638	-0.377 (0.053)***	3.103	3.845	-0.742 (0.118)***
	FCR (log) - Mean			Altmetric (log) - Mean		
Quantile score	[1]	[2]	[3]	[1]	[2]	[3]
	Expert	Non-expert	Diff (SE)	Expert	Non-expert	Diff (SE)
1	2.353	2.650	-0.298 (0.127)**	2.891	3.052	-0.161 (0.150)
2	1.772	2.836	-1.064 (0.160)***	2.142	3.239	-1.097 (0.190)***
3	2.107	2.922	-0.815 (0.283)***	2.350	3.381	-1.031 (0.340)***
4	1.252	2.746	-1.494 (0.804)*	1.933	3.497	-1.564 (0.959)
5	1.009	3.090	-2.081 (0.948)**	1.421	3.569	-2.149 (1.034)**
	2.140	2.770	-0.630 (0.092)***	2.605	3.200	-0.595 (0.109)***

Note. Standard errors are in parentheses. Column 3 reports T-tests. H0: [1]-[2]=0. * p<0.10; ** p<0.05; *** p<0.01.

Table 6. Post-funding outcome
Panel A – Baseline

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.029	-0.116*	-0.063	-0.083	-0.609***
	(0.030)	(0.068)	(0.054)	(0.064)	(0.017)
Female PI	-0.095**	-0.263**	-0.189**	-0.410***	-0.011
	(0.046)	(0.104)	(0.082)	(0.099)	(0.037)
Young PI	0.115**	0.379***	0.301***	0.354***	0.013
	(0.057)	(0.128)	(0.101)	(0.122)	(0.045)
Prior grant	0.187***	0.441***	0.310***	0.383***	0.548***
	(0.049)	(0.110)	(0.087)	(0.105)	(0.037)
Size (log)	1.528***	3.183***	2.521***	2.506***	0.045
	(0.278)	(0.621)	(0.492)	(0.592)	(0.209)
Size $(\log)^2$	-0.048***	-0.098***	-0.077***	-0.073***	0.001
	(0.010)	(0.022)	(0.017)	(0.021)	(0.008)
Competition	-0.819***	-1.790***	-1.183***	-1.299***	-6.115***
	(0.206)	(0.461)	(0.365)	(0.440)	(0.247)
Workload					-0.009***
					(0.001)
Constant	-10.582***	-22.716***	-18.074***	-18.845***	-0.079
	(2.122)	(4.749)	(3.761)	(4.528)	(117.148)
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.131*	0.262	0.125	0.205	
	(0.077)	(0.172)	(0.136)	(0.164)	

Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Panel B – Expertise predictiveness

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.113***	-0.309***	-0.214***	-0.202**	-0.606***
	(0.037)	(0.083)	(0.066)	(0.079)	(0.017)
Expertise (std)	-0.011	0.056	0.044	-0.025	0.030
	(0.056)	(0.125)	(0.099)	(0.120)	(0.047)
Score x Expertise (std)	-0.090***	-0.208***	-0.162***	-0.131***	-0.032**
	(0.021)	(0.047)	(0.037)	(0.045)	(0.016)
Female PI	-0.111**	-0.295***	-0.213***	-0.432***	-0.013
	(0.046)	(0.103)	(0.082)	(0.099)	(0.037)
Young PI	0.098*	0.347***	0.276***	0.328***	0.011
	(0.057)	(0.127)	(0.101)	(0.121)	(0.045)
Prior grant	0.210***	0.500***	0.357***	0.415***	0.548***
	(0.050)	(0.112)	(0.089)	(0.107)	(0.037)
Size (log)	1.472***	3.098***	2.456***	2.417***	0.017
	(0.276)	(0.619)	(0.490)	(0.592)	(0.209)
Size $(\log)^2$	-0.046***	-0.095***	-0.075***	-0.070***	0.002
	(0.010)	(0.022)	(0.017)	(0.021)	(0.008)
Competition	-0.890***	-2.062***	-1.407***	-1.375***	-6.031***
	(0.231)	(0.517)	(0.410)	(0.495)	(0.255)
Workload					-0.009***
					(0.001)
Constant	-10.079***	-21.828***	-17.389***	-18.080***	0.068
	(2.108)	(4.723)	(3.741)	(4.519)	(191.050)
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.292***	0.637***	0.422***	0.432**	
	(0.089)	(0.200)	(0.159)	(0.192)	

Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Panel C – Expert evaluators' predictiveness

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.019	-0.083	-0.039	-0.061	-0.574***
	(0.030)	(0.067)	(0.053)	(0.064)	(0.019)
Expert evaluator	0.191*	0.456*	0.317	0.333	0.238***
	(0.114)	(0.256)	(0.202)	(0.244)	(0.086)
Score x Expert	-0.183***	-0.542***	-0.396***	-0.374***	-0.128***
	(0.061)	(0.136)	(0.108)	(0.130)	(0.037)
Female PI	-0.099**	-0.276***	-0.198**	-0.418***	-0.013
	(0.046)	(0.103)	(0.082)	(0.099)	(0.037)
Young PI	0.115**	0.381***	0.302***	0.354***	0.012
	(0.057)	(0.127)	(0.101)	(0.121)	(0.045)
Prior grant	0.205***	0.487***	0.344***	0.417***	0.546***
	(0.050)	(0.111)	(0.088)	(0.106)	(0.037)
Size (log)	1.531***	3.180***	2.518***	2.506***	0.046
	(0.277)	(0.618)	(0.490)	(0.591)	(0.209)
Size $(\log)^2$	-0.048***	-0.098***	-0.078***	-0.074***	0.001
	(0.010)	(0.022)	(0.017)	(0.021)	(0.008)
Competition	-0.935***	-2.070***	-1.392***	-1.507***	-6.051***
	(0.215)	(0.480)	(0.380)	(0.458)	(0.250)
Workload					-0.009***
					(0.001)
Constant	-10.545***	-22.516***	-17.919***	-18.724***	-0.104
	(2.119)	(4.729)	(3.747)	(4.520)	(114.407)
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.197**	0.437**	0.256*	0.331*	
	(0.083)	(0.184)	(0.146)	(0.176)	

Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Table 7. Mean accuracy (prediction errors) by score and by expertise

	Publications (log) - Mean			Citations (log) - Mean		
Quintile score	[1]	[2]	[3]	[1]	[2]	[3]
	Expert	Non-expert	Diff (SE)	Expert	Non-expert	Diff (SE)
1	2.293	2.077	0.216 (0.088)**	2.196	1.971	0.225 (0.090)**
2	1.631	1.319	0.312 (0.086)***	1.626	1.327	0.299 (0.086)***
3	1.167	1.197	-0.030 (0.119)	1.217	1.061	0.156 (0.120)
4	0.875	1.649	-0.774 (0.430)*	1.125	1.268	-0.143 (0.387)
5	1.333	2.369	-1.036 (0.632)	0.833	2.098	-1.265 (0.618)**
	1.979	1.742	0.237 (0.060)***	1.922	1.649	0.274 (0.060)***
	FCR (log) - Mean			Altmetric (log) - Mean		
Quantile score	[1]	[2]	[3]	[1]	[2]	[3]
	Expert	Non-expert	Diff (SE)	Expert	Non-expert	Diff (SE)
1	2.211	2.039	0.172 (0.090)*	2.097	2.063	0.034 (0.090)
2	1.687	1.366	0.321 (0.086)***	1.605	1.380	0.225 (0.087)**
3	1.267	1.045	0.221 (0.120)*	1.183	1.114	0.070 (0.120)
4	0.875	1.526	-0.651 (0.392)	1.250	1.505	-0.255 (0.403)
5	0.667	2.221	-1.555 (0.610)**	1.333	2.230	-0.896 (0.609)
	1.949	1.714	0.234 (0.061)***	1.859	1.735	0.125 (0.060)**

Note. Standard errors are in parentheses. Column 3 reports t-test on the mean differences. H0: [1]-[2]=0. * p<0.10; ** p<0.05; *** p<0.01.

Table 8. Expert evaluators' accuracy (prediction errors)

	(1) Error Publications (log)	(2) Error Citations (log)	(3) Error FCR (log)	(4) Error Altmetric (log)
Quintile score	0.146***	-0.029	0.022	0.029
	(0.038)	(0.037)	(0.038)	(0.038)
Expert evaluator	0.490***	0.532***	0.478***	0.401***
	(0.143)	(0.141)	(0.144)	(0.145)
Score x Expert	-0.270***	-0.264***	-0.236***	-0.242***
	(0.076)	(0.075)	(0.077)	(0.077)
Female PI	0.045	0.071	0.059	0.163***
	(0.058)	(0.057)	(0.058)	(0.059)
Young PI	-0.001	-0.023	-0.057	-0.097
	(0.071)	(0.070)	(0.072)	(0.072)
Prior grant	-0.150**	-0.071	-0.091	-0.125**
	(0.062)	(0.062)	(0.063)	(0.063)
Size (log)	-1.457***	-1.190***	-1.262***	-1.112***
	(0.347)	(0.342)	(0.348)	(0.351)
Size $(\log)^2$	0.047***	0.037***	0.040***	0.034***
	(0.012)	(0.012)	(0.012)	(0.012)
Competition	1.785***	1.182***	1.308***	1.219***
	(0.269)	(0.265)	(0.270)	(0.272)
Constant	11.412***	10.446***	10.865***	10.097***
	(2.654)	(2.614)	(2.663)	(2.684)
Year FE	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495
Lambda	-0.680***	-0.277***	-0.423***	-0.399***
	(0.103)	(0.102)	(0.104)	(0.105)

Note. Heckman two-stage estimation models on prediction errors (columns 1-4). Selection equation in column 5 of Table 6. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Table 9. Evaluator's scientific prestige
Panel A – Predictiveness

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	-0.044	-0.137*	-0.086	-0.103	-0.550***
	(0.032)	(0.072)	(0.057)	(0.069)	(0.021)
Expert evaluator	0.186	0.444*	0.307	0.321	0.237***
	(0.114)	(0.255)	(0.202)	(0.244)	(0.086)
Score x Expert	-0.182***	-0.536***	-0.394***	-0.369***	-0.128***
	(0.061)	(0.136)	(0.108)	(0.130)	(0.037)
Prestige	0.030	-0.072	0.054	-0.023	0.255*
	(0.185)	(0.412)	(0.327)	(0.394)	(0.149)
Score x Prestige	0.081*	0.182*	0.155*	0.144	-0.098**
	(0.045)	(0.100)	(0.079)	(0.095)	(0.038)
Workload					-0.009***
					(0.001)
Constant	-10.710***	-22.745***	-18.221***	-18.933***	-0.491
	(2.125)	(4.743)	(3.757)	(4.535)	(111.851)
Controls	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.202**	0.448**	0.266*	0.334*	
	(0.082)	(0.184)	(0.146)	(0.176)	

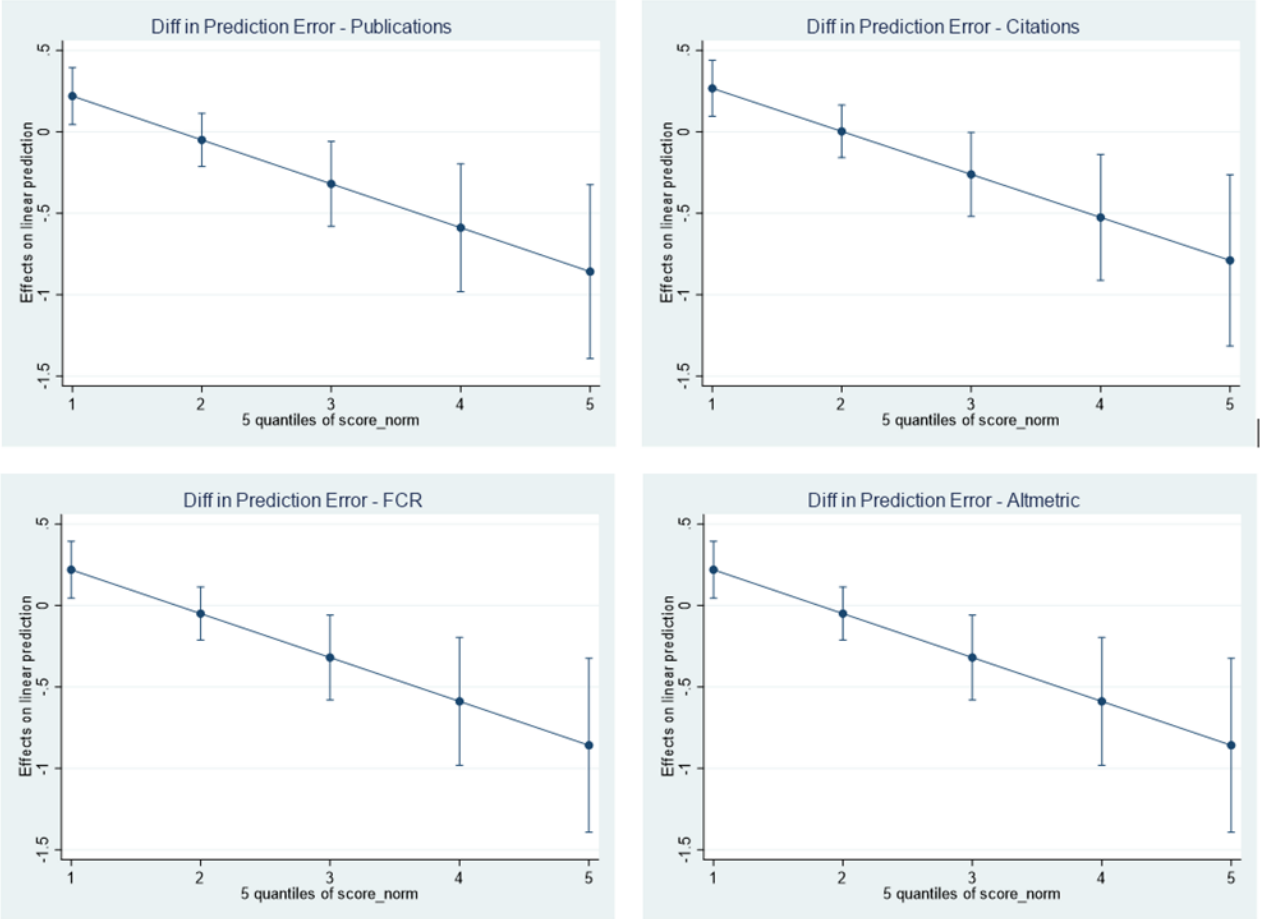
Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Panel B – Accuracy (prediction errors)

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)
Quintile score	0.154***	-0.032	0.024	0.024
	(0.041)	(0.040)	(0.041)	(0.041)
Expert evaluator	0.488***	0.527***	0.476***	0.397***
	(0.143)	(0.141)	(0.144)	(0.145)
Score x Expert	-0.270***	-0.263***	-0.236***	-0.241***
	(0.077)	(0.075)	(0.077)	(0.077)
Prestige	0.037	0.020	0.021	0.021
	(0.231)	(0.228)	(0.232)	(0.234)
Score x Prestige	-0.017	0.020	0.001	0.022
	(0.056)	(0.055)	(0.056)	(0.057)
Constant	11.457***	10.423***	10.877***	10.078***
	(2.662)	(2.623)	(2.672)	(2.693)
Controls	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495
Lambda	-0.692***	-0.286***	-0.429***	-0.406***
	(0.103)	(0.102)	(0.104)	(0.104)

Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5 of Panel A. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Figure 1. Estimated differences in prediction between experts and non-experts for different score quintiles



Appendix

Table A1. Variable construction

Variable	Description
Publications (log)	Log-transformed sum of the number of publications (until the end of 2023) that acknowledge funding from a given NNF grant.
Citations (log)	Log-transformed sum of the number of citations (until the end of 2023) to publications that acknowledge funding from a given NNF grant.
FCR (log)	Log-transformed sum of the FCR, i.e., the citations (until the end of 2023) to publications that acknowledge funding, divided by the average number of citations received by documents published in the same field and year.
Altmetric (log)	Log-transformed sum of Altmetric scores (until the end of 2023) of publications that acknowledge funding from a given NNF grant.
Quintile score	Quintile of the evaluator's assessment score. Scores have been normalized at the call level, by dividing each score by the average score given by all evaluators in a funding call.
Expertise	Cosine similarity between the evaluator's publications and the application content, represented as vectors, ranging from -1 to 1.
Expert evaluator	Dummy set to 1 if the cosine similarity falls within the top quartile.
Female PI	Dummy set to 1 if the PI is female, 0 otherwise.
Young PI	Dummy set to 1 if the PI is under 35 years old at the time of applications, 0 otherwise.
Prior grant	Dummy set to 1 if the PI has been awarded a NNF grant before the application date, 0 otherwise.
Size (log)	Log-transformed requested funding amount of the applications.
Competition	Number of unfunded applications divided by the total number of applications submitted within a single funding call.
Funded	Dummy set to 1 for funded applications, 0 otherwise.
Workload	Number of applications reviewed by the evaluator during the month of the application date.

A1. Workload

We compute workload as the total number of proposals that a panel member has to review in the same month of the application.

Proposals are assigned by the program officers shortly after the closure of the submission deadline. In assigning the proposals, the program officers primarily assign proposals to competent members. Our interviews with panel officers suggested that, while they try to give some load to all panel members, they also try to avoid excess overloading of any given member. Once the proposals are assigned, there is no or minimal reassignment, due almost entirely to conflicts of interest, signaled immediately after the first assignment. Consequently, there is no reason to believe that reviewers may accept or reject workload depending on the quality of the proposals that they are assigned (our unobserved omitted variable).

The level of workload varies primarily in dependence of the number of proposals received, which varies considerably from one call to another. Moreover, some committees are involved in multiple calls that occasionally overlap, creating additional workload variation. As a consequence, workload has a high absolute variation, ranging from a low of 1 to a high of 189 applications to review (mean 72.7; SD 41.8). Looking at the distribution of workload by month, we also notice that the patterns vary considerably from year to year, picking in three periods of the year, approximately coordinated with the Danish holiday calendar: in the spring (March to May), late summer (August -September) and late fall (November - December). While 58% of reviewers are based in Denmark, a share of reviewers is based abroad (17 countries). Variations of holidays and working days across countries create an additional source of exogenous variation.

A2. Alternative levels of expertise

In Table A2 we report further evidence regarding the predictiveness of evaluations at the different levels of expertise. In particular we run a model similar to the one reported in Table 6, with 4 levels of expertise, using the first quartile (i.e., cosine similarity < 0.29) as the baseline. In Panel A of Table A2, the coefficients of the quintile scores (columns 1-4) are not statistically significant at conventional levels, suggesting that the scores of the less expert evaluators are not predictive of post-funding outcomes, net of the selection into funding. Conversely, the negative and significant coefficients of the interaction terms of the scores with the quartile 4 of expertise confirm the findings reported in the Panel C of Table 6.

When looking at the prediction error (Panel B of Table A2), the positive and statistically significant coefficients ($p<0.01$) of the expert evaluators' variables (quartiles 2, 3 and 4) indicate that the average error is lower for evaluators in the quartile 1. However, the negative and significant coefficients ($p<0.01$) of the interaction terms of the scores associated to the highest quartiles of expertise are in line with the findings that experts are more correct when they assign worst scores.

Table A2. Expertise quartiles
Panel A – Predictiveness

	(1) Outcome Publications (log)	(2) Outcome Citations (log)	(3) Outcome FCR (log)	(4) Outcome Altmetric (log)	(5) Selection Funded
Quintile score	0.026	0.006	0.030	0.017	-0.608***
	(0.033)	(0.074)	(0.059)	(0.071)	(0.036)
Expertise (2^ qtl)	0.053	0.132	0.135	0.123	-0.193*
	(0.123)	(0.274)	(0.217)	(0.262)	(0.113)
Expertise (3^ qtl)	-0.238	-0.205	-0.183	-0.570*	-0.080
	(0.152)	(0.340)	(0.269)	(0.324)	(0.126)
Expertise (4^ qtl)	-0.023	0.216	0.121	-0.132	0.129
	(0.159)	(0.356)	(0.282)	(0.340)	(0.127)
Score x Exp (2^ qtl)	-0.163***	-0.311***	-0.239***	-0.280**	0.077
	(0.052)	(0.117)	(0.092)	(0.111)	(0.047)
Score x Exp (3^ qtl)	-0.098	-0.233*	-0.182	-0.146	0.019
	(0.063)	(0.141)	(0.112)	(0.135)	(0.047)
Score x Exp (4^ qtl)	-0.266***	-0.719***	-0.533***	-0.505***	-0.094*
	(0.069)	(0.154)	(0.122)	(0.147)	(0.048)
Workload					-0.009***
					(0.001)
Constant	-9.510***	-20.843***	-16.680***	-16.865***	-0.390
	(2.121)	(4.746)	(3.761)	(4.532)	(187.745)
Controls	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495	16,636
Lambda	0.270***	0.612***	0.389**	0.435**	
	(0.091)	(0.204)	(0.162)	(0.195)	

Note. Heckman two-stage estimation models on post-funding outcome (columns 1-4). Selection equation in column 5. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.

Panel B – Accuracy (prediction errors)

	(1) Error Publications (log)	(2) Error Citations (log)	(3) Error FCR (log)	(4) Error Altmetric (log)
Quintile score	0.259***	0.038	0.097**	0.104**
	(0.041)	(0.041)	(0.042)	(0.042)
Expertise (2^ qtl)	0.919***	0.676***	0.720***	0.697***
	(0.153)	(0.151)	(0.154)	(0.155)
Expertise (3^ qtl)	1.315***	0.827***	0.859***	0.885***
	(0.189)	(0.187)	(0.191)	(0.192)
Expertise (4^ qtl)	1.485***	1.221***	1.190***	1.124***
	(0.198)	(0.197)	(0.200)	(0.202)
Score x Exp (2^ qtl)	-0.372***	-0.245***	-0.272***	-0.274***
	(0.065)	(0.064)	(0.066)	(0.066)
Score x Exp (3^ qtl)	-0.466***	-0.240***	-0.263***	-0.266***
	(0.078)	(0.078)	(0.079)	(0.080)
Score x Exp (4^ qtl)	-0.571***	-0.441***	-0.430***	-0.438***
	(0.086)	(0.085)	(0.087)	(0.087)
Constant	10.240***	9.531***	9.975***	9.244***
	(2.642)	(2.620)	(2.668)	(2.688)
Controls	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Grant type FE	Y	Y	Y	Y
Reviewer FE	Y	Y	Y	Y
N	2,495	2,495	2,495	2,495
Lambda	-0.300***	-0.057	-0.184	-0.157
	(0.114)	(0.113)	(0.115)	(0.116)

Note. Heckman two-stage estimation models on prediction errors (columns 1-4). Selection equation in column 5 of Panel A. Standard errors are in parentheses. * p<0.10; ** p<0.05; *** p<0.01.