The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations

HARVARD | BUSINESS | SCHOOL

The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations

Kevin J. Boudreau
Eva C. Guinan
Karim R. Lakhani
Christoph Riedl

Working Paper

13-053

December 4, 2012

The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations

Kevin J. Boudreau, Eva C. Guinan, Karim R. Lakhani, Christoph Riedl*

Abstract

Central to any innovation process is the evaluation of proposed projects and allocation of resources. We investigate whether novel research projects, those deviating from existing research paradigms, are treated with a negative bias in expert evaluations. We analyze the results of a peer review process for medical research grant proposals at a leading medical research university, in which we recruited 142 expert university faculty members to evaluate 150 submissions, resulting in 2,130 randomly-assigned proposal-evaluator pair observations. Our results confirm a systematic penalty for novel proposals; a standard deviation increase in novelty drops the expected rank of a proposal by 4.5 percentile points. This discounting is robust to various controls for unobserved proposal quality and alternative explanations. Additional tests suggest information effects rather than strategic effects account for the novelty penalty. Only a minority of the novelty penalty could be related to perceptions of lesser feasibility of novel proposals.

Keywords: Project evaluation and resource allocation, expert review, open science, scientific paradigms, field experiment.

*Kevin J. Boudreau, London Business School & Harvard Institute for Quantitative Social Science (kboudreau@london.edu); Eva Guinan, Dana Farber Cancer Institute and Harvard Medical School (Eva_Guinan@dfci.harvard.edu); Karim R. Lakhani, Harvard Business School & Harvard Institute for Quantitative Social Science (k@hbs.edu); Christoph Riedl, Harvard Institute for Quantitative Social Science (criedl@iq.harvard.edu). KJB, KRL & CR are members of the Harvard-NASA Tournament Laboratory. Eric Lonstein provided excellent support with the execution of the review process and data collection. Griffin Weber provided valuable assistance in selecting evaluators based on their field of expertise. Eric Lin provided access to the PubMed citation data. Executives at InnoCentive greatly assisted us with the platform to collect the new research hypotheses. This paper has benefited from the feedback of participants at the Open and User Innovation Conference at Harvard University, NBER Workshop on Open Science, Washington University at St. Louis strategy seminar, REER conference at Georgia Tech, and the Bocconi University technology management seminar. Alfonso Gambardella, Shane Greenstein, Danielle Li, Jackson Nickerson, Lamar Pierce, Gary Pisano, Eric von Hippel & Todd Zenger provided feedback on earlier versions of the paper. CR acknowledges support from the German Research Foundation under grant code RI 2185/1-1. Supported by Harvard Catalyst, NCRR and NCATS Award 8UL1TR000170-05, -03S and -04S, Harvard Business School Division of Research and Faculty Development and the Harvard-NASA Tournament Laboratory. All mistakes remain our own.

1 Introduction

Is there a bias against novel ideas and research hypotheses? As resources for scientific and technical advance are scarce, society and organizations have developed a range of institutions to help select the most meritorious ideas and to allocate resources to them for further development. Often project evaluation cannot benefit from objective measures, given the inherently uncertain nature of research and innovation, and thus we depend on the subjective evaluation of experts (Stephan, 2012) who use their existing knowledge base and experience to assess the merits of a research proposal. In the academic science sphere, the approach most relied upon for enabling research agendas and providing research funds is expert peer review. In this paper we investigate how nascent scientific hypotheses are evaluated by designing a randomized expert peer review process. Specifically, we investigate a longstanding hypothesis that novel research ideas outside currently accepted scientific paradigms are susceptible to being discounted, rejected, or ignored (Kuhn, 1962; Fleck, 1935).

Expert peer review of new research proposals in academic science is now a large organized practice in its own right. In 2011 the combined budgets of the US federal civilian agencies dedicated towards academic research in the sciences (National Institutes for Health (NIH) and National Science Foundation (NSF)) exceeded $37.9 billion and was apportioned to roughly 40,000 research applications through peer review. The cost of a faculty member reviewing grant applications for the NIH was estimated at $1,700 per proposal (Stephan, 2012). Apart from the scale and considerable effort and investment of volunteer time, the process generates considerable overhead. For example, the NSF in 2010 arranged for over 19,000 scientists to come to the Washington DC area to participate in proposal evaluation. Moreover, expert evaluation of grants is highly consequential to the direction and production of scientific ideas, also likely impacting promotion and labor market outcomes (Chubin and Hackett, 1990). Chubin and Hackett (1990) find that between 33% to 48% of scientists who face an initial rejection from funding sources in natural sciences in fact stop working on the rejected line of inquiry subsequent to the rejection. The importance of expert peer review in science and expert judgments on innovative projects more generally has accorded this area considerable research attention, particularly in natural and life sciences (for reviews and surveys of this work see: Marsh et al., 2008; Langfeldt, 2006; for recent studies, see: Li, 2012; Luukkonen, 2012).

In this paper we study expert peer review in the academic sciences. We focus on a hypothesis articulated by Thomas Kuhn, in his now 50-year old book on "The Structure of Scientific Revolutions" (1962). He conjectured that there is an inherent bias against novel ideas in science. The prevailing scientific paradigm and its practitioners establish which questions are deemed interesting, how they can legitimately be addressed, and what a solution should look like. This should tend research towards incrementalism and "normal science"–and a corresponding disinclination to novel, non-paradigmatic research paths. It follows there might plausibly be less novel research proposed in the first place. However, conditional on novel research indeed being proposed, it might also be

the case that this research is penalized by gatekeepers conferring resources. This could be rooted in any number of mechanisms. For example, there could be inherent informational challenges in evaluating new ideas on the basis of old ones, or possible evaluator conflicts of interest or other kinds of strategic effects. Inasmuch, as the existing paradigm was hard-won and now constitutes a working path to continued progress, it is also plausible that deviations from the path could reasonably be expected to be of lower productivity or at least higher risk, on average.

To empirically study whether expert evaluators indeed treat novel proposals differently–and to provide insight as to why–we worked closely with the leadership team of a very large medical school at a top tier research university to manipulate features of an internal grant proposal process focused on an endocrine-related disease. Effectively we "layered" a field experiment within a normal grant process, to allow us to derive relevant inferences beyond what observing a naturally-occurring process would make possible. Our design began with recruiting a large number (142) of accomplished medical researchers to act as evaluators, while at the same time taking steps to attract a relatively large number of proposals (150) of varying novelty. We then randomly assigned multiple research proposals to each reviewer and multiple evaluators to each research proposal, resulting in 2,130 proposal-evaluator pairs. Researchers' and evaluators' identities were blinded from one another, and evaluators' identities or involvement were also kept confidential, resulting in a "triple-blinded" evaluation process.

In our analysis we find that evaluators uniformly and systematically give lower scores to proposals with increasing novelty; i.e. there is an economically significant novelty discount. We measured novelty in terms of the extent to which research proposals reflected unique combinations of descriptive knowledge keywords (Medical Subject Headings or "MeSH terms") that had not previously appeared in the published medical sciences literature (consisting of over 22 million records). A number of alternative formulations of the novelty measure produce similar results. Our main challenge in the analysis is to estimate the effect of novelty, while assuring this estimate is not somehow biased by unobserved proposal characteristics and quality. We find, on average, a standard deviation increase in our measure of novelty resulted in a 4.5 percentile point drop in ranking or a 7-position drop in rank order within our particular sample of 150 proposals. A combination of features of the research design and diagnostics suggest the effect cannot be explained by lower quality of novel proposals; composition of evaluators; imperfect "blinding" of proposals and responses to the identities of researchers rather than their content. Our analysis goes on to discern among several possible underlying mechanisms causing the novelty penalty.

Therefore, we provide evidence in support of a longtime conjecture concerning incrementalism in science. In particular, we present evidence that initial funding assessments can prematurely shut down novel work by rewarding work within the existing map of science. Moreover, our findings provide evidence that built into the institutions of scientific innovation is a novelty paradox, based on what may be related to the fundamental limits of expert evaluation and peer review. We would expect that exploring some degree of novel research paths should be productive in maintaining the

advance of science and technology (Weitzman, 1998; Simonton, 1999; Fleming, 2001). Whatever that optimal balance might be, it seems unlikely that the application of a relatively indiscriminate and uniform novelty penalty should be an effective means of attaining this balance.

2 Expert Evaluation in Science

One of the fundamental distinguishing features of academic research is the extensive and almost exclusive use of peer evaluations in all stages of scientific work; from allocating resources to initiate new lines of inquiry (grant reviews); to judging the quality and significance of output (journal reviews); and ultimately determining promotion outcomes (tenure proceedings) and community and societal accolade (prizes for recognition of important work). Academic science as a self-governing system uses peer evaluators as important gatekeepers that continually ensure that high standards and quality are met; that good work is recognized and rewarded and poor quality is dismissed (Zuckerman and Merton, 1971; Langfeldt, 2006). The degree of uncertainty related to the assessing the quality of scientific work is, however, a function of the stage that is being assessed; with considerably more ambiguity when embryonic scientific ideas are being judged for potential impact as compared to making a tenure decision or giving an award.[1] Chubin and Hackett (1990) find that between 33% to 48% of scientists who face an initial rejection from funding sources in natural sciences in fact stop working on the rejected line of inquiry subsequent to the rejection. Hence, peer-based grant evaluation is highly consequential to the direction and production of scientific ideas, also likely impacting promotion and labor market outcomes (Chubin and Hackett, 1990).

2.1 Noise and Bias in Peer Evaluation

Peer evaluators assessing research proposals are faced with the considerable challenge of not being able to directly observe the true quality of the research applications they are tasked with evaluating. At best, expert peer evaluation generates an approximation of the true unobserved quality. Even ex-post, the true ex-ante quality of an idea is vexingly difficult to discern. For example, even if accurate measurement were possible, the judgements of the evaluation process might themselves influence later outcomes. Therefore, under the very best of circumstances, we might interpret evaluator scoring as the true quality and some inevitable subjective error. In this best case, the error is zero-mean and devoid of any unnecessary noise. Thus evaluation can be judged to be a problem of simply aggregating or averaging the signals generated by evaluators in a way that minimizes error. Indeed, the majority of NIH funding decisions are based on taking a simple average of the score given by evaluators in study section committees that range in size from 25 to 50 members (Li,

[1] We limit our analysis of the existing literature to studies focused on grant evaluations. There is quite a substantial literature on peer evaluation of journal manuscripts, however, this relates to scientific work already completely as opposed to at a nascent proposal stage. See Ware (2011) for a comprehensive review of issues related to manuscript review.

2012). Evidence from assessing national grant review systems has shown that inter-rater reliability can be improved if more evaluators are added to the review team (Marsh et al., 2008). Extensions to this principle of aggregation include weighting signals through review boards and through the composition of those boards (Langfeldt, 2006; Mutz et al., 2012), and creating panel and group processes that create opportunities for disclosing otherwise unshared information among evaluators (Olbrecht and Bornmann, 2010).

The literature studying the efficacy of grant peer review, largely begun in policy and natural science journals in the 1970s, consists of a considerable number of testimonials, surveys and several field studies that introduce hypotheses concerning possible distortions in the peer review process and data consistent with some of these hypotheses (see Langfeldt 2006 and Marsh et al. 2008 for reviews). Evidence suggests that noise and bias generated in the expert peer review process may well exceed the above mentioned "best case" imperfect benchmark. For example, in their seminal study of NSF grants, Cole et al. (1981) found that the variance among evaluators for a given proposal was higher than the variance between proposals. Jaysinghe et al.'s (2003) analysis of the Australian Research Council's grant-making history of more than 2000 proposals and over 6000 evaluators showed that very few proposals were significantly different from the cutoff value of funding and thus the decision to fund or not was essentially driven by chance. Glaser et al. (2002) argue this unnecessary noise and bias generates considerable Type I and Type II error, impacting both poor quality proposals (positively) and truly meritorious proposals (negatively).

A first set of issues documented in past research relates to how evaluators respond to the characteristics of the submitting researcher. Most funding agencies employ a single-blind review process where the identities of the proposal submitters are known to the evaluators but not vice versa; many review protocols also ask for an assessment of the ability of the investigator to carry out the proposed research; leading to a so called Mertononian (1968) "Matthew Effect".[2] Characteristics of the researcher (and particularly the relatedness of these characteristics to those of evaluators) have also been found to correlate with scoring and evaluation outcomes, in dimensions such as race, gender, education and affiliation (e.g., Bornmann et al. 2007; Marsh et al 2008; Reinhart 2009). In the context of single-blinded NIH committees, Li (2012) finds causal evidence showing that relatedness between evaluators and researchers (in terms of cites made by reviewer to submitting researchers) led to more positive evaluations. However, she also found that these more related evaluators also provided appraisals that were more highly correlated to independent measures of the research quality–indicating some informational benefits, apart from any possible bias.

Beyond the characteristics of the submitter, past work has also pointed to how evaluators may respond to characteristics of research proposals, themselves. For example, Kotchen et al (2004) study of NIH grants between 1997 and 2002 shows that clinical research proposals are under-rated

[2]While not tested formally for grants, Ceci and Peters' (1982) experiment in the area of journal peer review showed that more than 90% of already accepted papers, having their attribution changed to lower prestige authors, were rejected when sent back into the review process.

and less likely to be funded as compared to laboratory research via peer evaluations, potentially due to concerns about safety and privacy of human subjects. Peer evaluations may also be influenced by "strategic effects" such that the content of a proposal might either be too close to the work conducted by the reviewer or the new work may diminish the importance of the evaluators' own work resulting in lower scores for the proposal (Horrobin 1990). On the other hand, experts may be inclined to more positively evaluate research in their own field to promote the importance of that field (Fang, 2011; Travis & Collins, 1991). More generally there may be a "tower of babel" issue (Laudel 2006) amongst various schools of thought in scientific disciplines such that those belonging to one stream of literature or another may fail to fully recognize or comprehend the precepts, methods, or conclusions of those of other literatures (Carter, 1982; Wade, 1973; Travis and Collins, 1991).

2.2 The Treatment of Novelty

A range of scholars have converged on identifying that novel and unique combination of ideas and technologies, more generally knowledge, in an evolutionarily recombinant search process is one of the basic drivers of creativity, innovation, scientific discoveries and technical change (Basalla, 1988; Weitzman, 1998; Simonton, 1999; Fleming, 2001; Sorenson and Fleming, 2004). Indeed Singh and Fleming (2010, 43) note that "the idea that novelty is a unique combination is at least as old as Adam Smith (1766)." Fleming (2001), makes important progress in using patent data to show that novel recombinations of technological components tend to lead to lesser success, on average, yet to greater variability and breakthrough innovations. However, rather than a simple question of rationally "optimizing the innovation portfolio" of projects along novel and non-novel research paths, it is a long held presumption that the institutional and educational structure of science leads novelty to be systematically discounted, rejected, or ignored (Kuhn, 1962; Fleck, 1935).[3] Evidence on the discounting of interdisciplinary work is at least consistent with this hypothesis (*e.g.*, Porter and Rossini, 1985).

According to Kuhn (1962, 175), science proceeds incrementally within paradigms: "the entire constellations of beliefs, values, techniques, and so on, shared by the members of a given community." By Kuhn's account, science has a tendency towards incremental advance within the established paradigm, as it is this prevailing paradigm that establishes which scientific questions are deemed interesting, how they can legitimately be addressed, and what a solution might look like. It is further conjectured this tendency to incrementalism is sufficiently strong that only through "scientific revolutions" can established paradigms be overthrown–and only when contrary evidence

[3]Leafing through research on evaluation from the natural science, one finds abundant presumptions of bias. Peer evaluators working within the context of normal science are described by Carter (1982) as "so committed to their paradigm that they cannot (or will not) see the value of unorthodox ideas" (Carter, 1982, 11). Hacking (p. xxvi), in the preface to the 2012 edition of Kuhn's book, asserts: "normal science does not aim at novelty but at clearing up the status quo. It tends to discover what it expects to discover." Horrobin (1990) provides a litany of anecdotes that relate to the suppression of innovation in the medical sciences due to peer review.

has accumulated or shows an inability to solve problems based on the prior theoretical apparatus (Luukkonen, 2012).

We might simply expect that knowledge must necessarily accumulate as an accumulative stockpile of new ideas, resting on past ideas, experience and findings. It might therefore simply be an inevitable feature of the accumulation of knowledge that advances are made easier, more predictable and better understood when they continue from an existing body of well-established knowledge. Where there is not a "surrounding" or "adjacent" body of theoretical and experimental results to contextualize a proposal, we might expect there simply to be greater uncertainty. This might be akin to a particularly pernicious form of the earlier mentioned "Tower of Babel" problem (Laudel, 2006); however, rather than a disconnect between two existing research approaches, it is a disconnection between established research and yet to be pursued research. The education, incentives and attention of new generations of scientists, within the established paradigm, might exacerbate such information effects. For example, apart from focusing attention and effort towards the existing paradigm and normal science, this context of bounded cognition could also lead research to be regarded as a potentially less legitimate "school of thought", as above. Inasmuch as these different proposed research paths compete for resources and attention with existing research traditions, conflicts of interest or strategic effects could also plausibly shape evaluator decisions. More prosaically than the above information or incentive effects, novel research might also be prejudiced by mundane features of proposal administration — such as a need to present past data or results and established body of research—which might systematically prejudice evaluators against more novel work (Langfeldt 2001, 2006; Stephan, 2012). Given likely very different payoffs from pursuing novel research, we might also expect that those that select themselves into novel, non-paradigmatic research possess rather different characteristics from those in the mainstream, perhaps creating additional biases related to (non)relatedness of gatekeeping evaluators and those proposing the research.

3 Research Design & Context

The central goal of our empirical analysis is to test whether novel proposals are treated differently by expert evaluators–and particularly whether the prediction of a systematic bias against novel ideas is borne out in the data. We are particularly interested in how evaluators respond to research proposals as such, rather than the identity of researchers (a separate issue beyond the scope of this paper and given more attention in past research). In this section we describe our research design to "evaluate the evaluators" in a grant allocation process. The design, summarized here, involves several manipulations and alterations to a grant solicitation and evaluation process. Broadly, the design is intended to allow us to draw relevant statistical inferences, while preserving key features of a typical grant proposal setting.

The research setting involved our working closely with a research grant allocating body within a leading medical research university to study results from the first stage of a $1M grant process

related to a major endocrine system disease. This disease category is a major economic and health burden on society and it is the focus of significant research efforts at the host medical school and related teaching hospitals. The university gives out internal grants to allow investigators to bootstrap their research efforts to generate preliminary data for NIH grant applications.

3.1 A Call for Research Proposals from the "First Phase" of a Grant Process

The ideal sample of research proposals should allow us to observe meaningful variation in levels of novelty, but at the same time be sufficiently narrow and constrained so as to allow us to precisely define and measure what it means to deviate from existing research pathways. Further, sampling should be defined in relation to something other than the literature itself, to avoid circular and tautological definitions of what constitutes paradigmatic work. Thus, the design of the grant problem was defined in terms of a disease area (i.e., "nature" rather than existing human knowledge), and avoided mention of existing literature or other constraints of the existing body of knowledge and established research pathways. The problem was simply to make significant progress in research and treatment of the endocrine-system related disease.

While problem definition allowed for the possibility of observing a range of novel and paradigmatic research proposals, it did not itself guarantee large numbers, or variation in the nature of submissions. We implemented several alterations of the process with this in mind. For example, to encourage submissions, the university president was enlisted to communicate via email an open call to all members of the medical school and the broader university community. Rather than take actions to attempt to promote variation in (high and low) novelty, as such, we found it practically simpler to encourage new projects and "out of the box" proposals. We did so on the expectation that within-paradigm normal science proposals would be attracted in any case, and the greater challenge would be in attracting novelty.[4]

A more profound research design choice taken to encourage participation and variation was to partition the grant proposal process into two distinct phases. A first phase involved a solicitation of proposals for approaches and ideas related to a wide range of possible avenues and research pathways for making progress on the disease–from diagnosis, to treatment, to prophylaxis. This first phase—essentially a call for research hypotheses—is most relevant to the goals of this paper. Separating the two phases accomplished several goals, including allowing us to focus on the most relevant unit of analysis in the first phase: the idea and hypothesis and proposed research pathway. Partitioning the first phase also reduced "entry costs" for prospective submitters, as documenting

[4]An implication of taking this approach of emphasizing novelty is that the evaluators observing these messages will have effectively been primed to understand that a goal of this exercise is to seek novel research ideas. However, our design is not sensitive to oversampling on any one type of proposal–only that there be meaningful variation. Further, it is not clear that the colloquial usage of "novelty" should necessarily be interpreted as our precise usage of *non-paradigmatic* research.

the basic research idea required a much shorter proposal than a full-length proposal. The average proposal in this exercise was just several pages. The shorter research proposal also avoided the problem of having the submission process have requirements that might on their own prejudice against novel submissions (Langfeldt, 2001, 2006).

Important to note, partitioning the process should have also altered the precise structure of incentives for participating in the grant process and, particularly, the first phase. The first phase involved several explicit and implicit payoffs. Rather than a direct share of the $1M research funds to be disbursed in the second phase, 12 winners were each awarded $2,500 cash prizes. Perhaps more important, the invitation to the first phase noted that the winning an award would also mean that the second stage call for proposals would directly include calls for research in areas of the winning research hypotheses, thus increasing the odds of winning a substantial grant in the second phase. (Indeed 4 of the 12 first stage winners went on to win second stage research grants.) Importantly, the first phase of the process also served as a platform for high-profile exposure among peers and university leaders. Awards were conferred by the dean of the medical school in a public formal ceremony, attended by colleagues and members of the media and press. Therefore, while there is necessarily some tradeoff in gaining the benefits of partitioning while altering incentives, at least the categorical structure of incentives continues to appeal to typical incentives to attain "gold, ribbons, and puzzles" (i.e., financial rewards, reputation and acclaim and intrinsically motivating work), the usual incentives of scientific researchers (Stephan, 2012). In total, the process generated 150 research proposals. Of the research proposals, 72 of them came from researchers within the host university. The remaining 78 came from external researchers outside of the host university.

3.2 Recruiting Expert Evaluators

We recruited 142 evaluators to participate in this exercise. The typical practice of selecting expert evaluators regularly used by major funding agencies is to choose those with specialized knowledge, closely-related to that contained in the research proposal (Langfeldt, 2006). For example, in the case of an *ad hoc* referee team, this might typically include a small handful of specialized researchers whose phenomenological interest, research methods and/or questions relate to the research proposal (Jayasinghe et al., 2003)–perhaps five or seven evaluators (Langfeldt, 2006). In larger evaluation processes, as in the work of the NIH and NSF, meant to attend to larger flow of proposals, standing committees and subcommittees might reasonably form around topic area themes, with proposals directed to most appropriate subcommittees (Li, 2012). Such larger committees can grow as large as 30 or 50 researchers (Li, 2012).

To assure a large number and variety of evaluators while maintaining a representative group of "leading" researchers who might potentially represent typical "gatekeepers", we separately recruited from three distinct groups: i) those with at least one publication in the same endocrine disease; ii) those without publications in the disorder, but with at least one coauthored publication on a

separate topic with someone who had published in the disorder; and iii) those without any direct or first-degree link to disease-specific publications. We recruited evaluators from the host university's roster of faculty. This allowed us to observe relatively detailed characteristics of each evaluator, using the university's systematized data collection on researcher careers. Drawing on faculty from the host university also assured reasonably high-caliber participants, independent of rank. Strong institutional support and commitment also helped minimize drop out of recruited evaluators. Our approach to identifying evaluators was simply to rank order faculty at the medical school in each of these three groups according to publications, inviting the top-ranked individuals. The pool is obviously non-random and therefore our analysis relies on the observable variation of evaluators, rather than on any claims of randomness or representativeness of this group in relation to the typical teams that are formed above. Both the number and variety of evaluators comprised by our group should exceed typical levels. The composition of evaluators is described in Table 1.

<Table 1>

As can be seen in Table 1, evaluators are relatively balanced in numbers across the three groups. The proportions are also relatively evenly balanced across senior and junior faculty, with roughly half being senior (associate or full) professors for each group. Each group also includes considerable diversity in gender, age and training (in terms of M.D. or Ph.D.). Each group uniformly includes (very) highly accomplished researchers, with an average publication count of 101. (While submitters are themselves accomplished, they are clearly much more junior, on average, with roughly a tenth of the number of publications as these most accomplished researcher leaders.) Variation in mean levels of publication across each group is difficult to interpret, given we should expect that numbers of publications should differ across different fields. However, the counts remain high in each case.

3.3 Randomized Assignment Procedure

A total of 15 randomly selected research proposal submissions (among the 150 total) were assigned to each of the 142 evaluators. Equivalently, an average of 14.2 randomly-selected faculty evaluated each proposal. Practical considerations of having medical school personnel prepare and distribute copies of proposals, while reliably corresponding with, supporting and recording evaluations led to an additional simplification/constraint in these random assignments. Rather than perfectly randomize in an unconstrained fashion we instead created 10 "blocks", which contained 15 randomly selected proposals to which evaluators were then randomly assigned. Therefore, each research proposal within a given block was evaluated by the same evaluators. Thus, the data should be considered "block randomized." This should not introduce any systematic error, but creates potential correlated error within each evaluator block.

3.4 The Evaluation Process

Following convention in medical research grant proposal evaluations, the task of evaluators was to score proposals using a 10-point scale to summarize their estimated assessment of the potential impact of the ideas, hypotheses, and research pathways contained in each of the 15 proposals they evaluated.[5] The score was to respond to the question, "On a scale of 1 to 10 (1 Lowest - 10 Highest) please assess the impact on (the disease) care, patients, or science of research." We also devised a secondary measure for validating the measure as indicative of perceived quality. This secondary question asked the evaluator to assign a total of 100 points across proposals in terms of his or her perception of merit in the proposals. Given our interest in novel research proposals in particular, we also chose to include a similar question to evaluate perceptions of the feasibility of proposals with an analogous 10-point scale.

Given our interest in having evaluators respond to the content of proposals (and particularly their novelty) rather than to the submitting researchers' identities, we designed the process with the aim of minimizing the probability that identities would be found out. Submitters names were blinded on proposals. The identity of evaluators was also blinded. Further, each evaluator performed his or her evaluation independently, was not prompted to interact with other evaluators, nor were they given the names of other evaluators, and had access only to the 15 proposals they were assigned. Thus, evaluators were effectively blinded from one another–and the overall evaluation process can be regarded as "triple blinded". The novel format, the call for new research and the wide participation might have also served to increase the likelihood of anonymizing identities.

4 Data & Variables

With 142 evaluators and 15 evaluations per reviewer, our data set contains 2,130 proposal-evaluation pair observations, with proposals and evaluators randomly assigned to one another. Several data sources were brought together for the analysis. These include: each of the evaluator score sheets, the database of prior academic publications and citations of submitting researchers (of the subset coming from the host university) and detailed backgrounds and c.v.'s of all evaluators. Here we review the definition and construction of our main variables.

The main dependent variable in our study is the main score out of ten given by evaluators (Section 3.1) as an overall assessment of the potential impact of a research proposal (*Score*). We also use the secondary dimensions that evaluators judged in their assessments, including the points allocation and feasibility (Section 3.1). These and other main variables and other variables used in the analysis are defined and basic descriptive statistics provided in the following tables. These are divided in terms of whether they vary with the proposal, the researcher (beyond the proposal), the evaluator, or the proposal-evaluator pair.

[5]Funding decisions at the NIH are based on a 9-point "impact/priority" scale (1-exceptional - 9-poor); http://enhancing-peer-review.nih.gov/scoring&reviewchanges.html

<Tables 2 & 3>

4.1 The Key Construct: Novelty of a Research Proposal

In our study, novelty is intended to reflect the extent to which a proposed research pathway departs from established science. In developing a measure of novelty, we exploit the controlled lexicon of keywords customarily used to describe and categorize the content of research and research proposals in the life sciences. These are referred to as "Medical Subject Heading" (MeSH) terms. Contrary to many other academic fields, MeSH keywords are not assigned by authors. Rather, they are assigned by professional science librarians, trained specifically to perform this task. The use of this controlled vocabulary is intended to assure a global and consistent assignment of keywords across the life science research community (Coletti and Bleich, 2001). We hired a professional librarian trained in standardized procedures for evaluating the content of research according to guidelines used by National Library of Medicine (NLM) at the NIH to code each of the proposals. On average, proposals in our sample had 12.42 MeSH terms (std. dev. = 5.42).

Our main measure of novelty compares the MeSH term combination of a proposal with those in the existing published literature. To generate a measure of whether and the extent to which a proposal departs from the existing literature, we examine each of the possible pairs of MeSH terms (i.e., for N terms there would be $N(N-1)/2$ pairs). We determine what fraction of these pairs have not yet appeared in the existing literature. We refer to the entirety of publications listed in the PubMed database (currently about 22 million articles), as representing the published literature. Therefore, our measure *Novelty* can vary from zero to one. In our data it varies from 0 to .47. While a small share of our observations, about a tenth, have zero novelty, the bulk of the observations vary in the degree to which they depart from the existing literature, without being utterly disconnected.

While the use of a standardized and fixed lexicon is what makes this comparison possible, one potential limitation is that we will not observe deviations and departures that go beyond the controlled lexicon. Thus, we may miss additional sources or types of novelty. To the extent there are forms of novelty that involve utter new categories of innovation, rather than novel recombinations, our measure will undercount novelty. More importantly, the effect of novelty we estimate will be downward (upward) biased to the extent that any unobserved novelty is positively (negatively) correlated with recombinant novelty. In its face, we speculate it to be unlikely that utterly novel ideas and recombinant novelty should necessarily be strongly negatively correlated. We also rely here on claims that recombinant novelty should be an important source of truly novel advances (Weitzman, 1998; Fleming, 2001). The hierarchical nature of MeSH terms also provides some assurance that all novel advances are covered by higher-levels of categorization within the MeSH hierarchy, if not lower levels.

Another possible limitation of the *Novelty* measure is that it exploits the entire existing literature, when the state of the field might be better described by just recent years of work. However,

we found almost no difference in our results when using the entire literature as a benchmark versus using just the last 10 years of the PubMed record. This might possibly be explained by there simply being a large weight of papers published in recent years, or perhaps relate to the cumulativeness of science in medical research whereby more dated findings and research pathways remain relevant, or at least informative.

More generally, it should also be noted that there may be any number of ways of algebraically constructing a statistic that captures how a proposal's MeSH term vector maps onto a vector describing the content of the existing literature. For example, we have focused on new-to-the-world two-way (pair) combinations of MeSH terms; we might have considered different "dimensionality" of the combinations we compared. However, one-way (singlet) combinations are ruled out because they are each present in the published literature and three-way (triplet) and four-way (quadruplet) combinations do not result in qualitative differences in results. Another plausible dimension in which we might reconsider the construction of *Novelty* is that our measurement of the existing literature discriminates between whether a MeSH pair has previously occurred or not, rather than its frequency or popularity. But on conceptual grounds, it is more salient here to consider prior existence rather than frequency of use.

A number of other considerations led us to perform additional tests (to follow, in the analysis). For example, we considered whether we might somehow privilege those combinations that are novel to the disease area, rather than novel to the broader literature. This question is potentially somewhat more interesting and less straightforward than earlier considerations. However, rather than artificially constraining our definition of novelty, we maintain the strict definition (of novel to the existing body of human knowledge, as a whole). Instead, within the analysis we more directly deal with the question of how individuals from different parts of the literature behave by accounting for their "intellectual distance" from a proposal. We might also expect there to be qualitative differences between something that has zero novelty (unambiguously within-paradigm) versus something that departs, at all; or the "degree" or level of novelty might also play a role. We test for these discrete and possible non-linear effects in our analysis.

At a more mechanical level, a characteristic of the *Novelty* measure that needs to be acknowledged is that it is constructed on the basis of unique pairwise combinations of MeSH terms–and therefore is closely tied to the count of MeSH terms involved. As the simple count of MeSH terms might itself be related to the scoring and evaluation of proposals, we explicitly control for this count in our analysis.

5 Estimation Approach

5.1 A General Baseline Empirical Framework

The central goal of the empirical analysis is to answer the question of whether novel research proposals are treated differently by evaluators — and following Kuhn's hypothesis, whether novelty is systematically discounted. To illustrate how this might be done, we first summarize potential factors influencing evaluator scores in a general framework, including: characteristics of the evaluator (indexed by e); characteristics of the proposal itself (indexed by p), including Novelty; and characteristics of the researcher submitting the proposal (indexed by r). Scores might also be influenced by relatedness of evaluator attributes with either those of the proposal or researcher. (See Section 2). We denote variables describing the link between evaluators and these factors with subscripts of $e-p$ and $e-r$. Beyond these categories of factors, we also allow for some zero-mean error term, . Thus, our general framework for comprehending problems of statistical inference is as follows:

$$Score_{ep} = f(X_e, Novelty_p, X_p, X_r, X_{e-p}, X_{e-r}) + _{ep} \tag{1}$$

The advantage of this general framework is that it does not (nor does it need to) specifically identify and discriminate among the many complex factors that could influence actual and perceived merits of research proposals. Rather, the framework accounts for them all, while allowing us to recognize the challenges of isolating and measuring the effect of $Novelty$.

5.2 Linear Regression Model Implementation

Evaluator Characteristics. To account for evaluator characteristics in our model's implementation, we exploit the assignment of multiple proposals to each evaluator by including individual evaluator dummy variables ($_e$), as in expression (2). In principle, introducing these controls should not alter estimates of the effect of novelty, as the random assignment procedure should have eliminated any correlation between these characteristics and $Novelty$. They should nonetheless increase precision of the model estimates. These controls are also appealing in that they allow us to refer to an intuitive interpretation of the novelty effects, in terms of how individual evaluators differ in their evaluation across proposals (rather than broader comparisons across evaluators and proposals, at once). With this unambiguous and intuitive interpretation, we can then turn our analysis to assessing why this is so.

Researcher Characteristics. The question of researcher characteristics is an altogether different kind of question. If our design has been successful, the combination of a "triple-blind" process (Section 3.1), the call for new and potentially less-known research, wide participation, and unorthodox proposal format should have obscured researcher identities. If successful, then $cov\left(Score_{ep}, X_r\right) = 0$ and $cov\left(Score_{ep}, X_{e-r}\right) = 0$, and all researcher-related terms should drop out of the model, as in

expression (2). (Note, we later re-introduce researcher-related terms as predictors of proposal quality, rather than as direct determinants of the score–an important distinction.) Apart from relying on the design to assure these factors do not play a role, we perform explicit diagnostic tests as part of robustness checks.

Research Proposal Characteristics. Accounting for research proposal characteristics–other than novelty–is the single greatest estimation challenge in our analysis. To appreciate this point, consider that the ideal experiment is to compare research proposals that are identical in every way, except for their novelty. In such a setup, simply comparing matched pairs of this kind would allow us to infer any novelty effect. Of course, the characteristic of novelty should not be entirely separable or uncorrelated with other characteristics of a research proposal. Our basic econometric challenge is then to assure that our estimated effect of novelty does not somehow reflect characteristics and quality of proposals that i) affect score; ii) are correlated with $Novelty$; and iii) are not themselves a direct cause or consequence of novelty.[6] Failing to account for such characteristics and quality has the potential to generate omitted variable bias. Adding complication, differences in characteristics and quality across proposals are not readily observable.

Unfortunately, it is inherently not possible to exploit randomized assignment or other features of experimental design to overcome this fundamental estimation challenge. For example, including proposal dummy variables would simply lead $Novelty$ to drop out of the model altogether. Another approach might be to simply introduce a selection of research proposal covariates as controls; however, no number of "control" variables can wholly and assuredly account for unobserved characteristics and remove any doubt of omitted variable bias.

Our main empirical strategy for dealing with these issues is practically similar to that of introducing control variables, but with important differences in its aims and in which variables we select to add to the model. Here, our approach is to address possible omitted variable bias by introducing regressors that should serve as powerful independent predictors of (unobserved) proposal characteristics and quality. Rather than have the goal of wholly controlling for these unobserved characteristics, our goal is to observe how or whether accounting for unobserved characteristics and quality lead the estimated coefficient on $Novelty$ to change, if at all. As predictors of proposal characteristics and quality, we mainly rely on measures of the quality of the researchers themselves. In particular we exploit researcher's publication records (with data drawn from outside the experiment). The intuition here is that a "high quality" researcher should be systematically more likely to produce a "high quality" proposal, deserving of a high score.[7] If unobserved proposal characteristics and quality are correlated with both the scoring and with $Novelty$, introducing these predictors

[6]Controlling for direct causes or consequences of novelty would only serve to "soak up" any estimated novelty effect, rather than help us properly estimate and calibrate the effect. Analysis of possible causes and consequences of novelty effect is left for subsequent analysis where we attempt to interpret the novelty effect.

[7]As a reminder, our ability to exploit this approach depends crucially on successful blinding of researchers' identities.

should alter the estimated coefficient on *Novelty*.

To provide still greater assurance of this approach, we include still more predictors including those related to the proposal itself. The challenge here is to select predictors that might somehow be reasonably expected to predict unobserved proposal characteristics and quality, but which should not relate to novelty, nor its causes or consequences. Here, we elect to use the word count from each proposal. These might roughly capture, for example, the level of effort that was put into a proposal.

Based on these above considerations, our linear implementation of the generalized framework can be summarized as follows:

$$Score_{ep} = {}_{e} + Novelty_{p}, +_{p}Quality_{p} + {}_{e-p}X_{e-p} + {}_{ep} \tag{2}$$

where *Quality* should be understood as a series of proxies for the true and unobserved characteristics and quality of a proposal with bearing on the score, independent of its novelty. The term ${}_{ep}$ is a zero-mean error term, redefined from that in expression (1) to reflect the linear approximation in this model. We estimate the model as ordinary least squares (OLS). Given possible idiosyncratic errors across evaluator groups, evaluators and research proposals, we use bootstrapped standard errors.

In our main estimates, geared to estimating the direct effect of novelty on evaluations, we do not investigate relationships between evaluator characteristics and proposal characteristics (*i.e.*, ${}_{e-p}X_{e-p}$). (In not explicitly controlling for these characteristics in the initial calibration of effects we are exploiting the randomized assignment of evaluators to assure this term is not correlated with our error term.) The explicit use of these terms becomes more relevant in our follow-on investigation into the causes and explanations of novelty effects.

6 Analysis & Results

6.1 Main Results

This section implements the model described in expression (2). Our first OLS results are reported in Table 4. Model (1) regresses *Score* on *Novelty* and a constant term. We also control for the simple count of MeSH terms, to assure our estimated coefficient on *Novelty* is not simply indirectly picking up the effects of this term (Section 4.1). We find the coefficient on *Novelty* is statistically significant and negative (-2.38), consistent with a novelty discount. Subsequent estimates will not substantially differ from this first and simplest estimate. The coefficient on the MeSH term count is itself positive and significant and alters the coefficient on *Novelty* from what it would have otherwise been (-1.53 significant at $p < 1\%$). Introducing a more flexible series of dummies corresponding to different counts of MeSH terms does not alter our coefficient on *Novelty*, as in model (3); nor does introducing dummies for individual types of MeSH terms alter the results. Therefore, we simply

maintain the linear control for MeSH term count in subsequent models.

<Table 4>

We proceed to control for differences across evaluators. We begin by introducing dummy variables for each of the 10 blocks of evaluators (recall random assignment was implemented through block randomization, as described in Section 3.4). As reported in model (4), this does not affect the estimated coefficient on $Novelty$, and contributes little explanatory power (comparing the adjusted-R^2 statistic in model (4) with that of model (1)). Therefore, our use of block randomization rather than simple randomization has no impact on results. Also consistent with effective randomization, neither does adding individual evaluator dummies affect model coefficients, as reported in model (5). Evaluator differences do, however, account for a large share of variation; the adjusted-R^2 increases from .03 in model (1) to .22 in model (5). We retain evaluator dummy variables as part of our preferred specification given this considerable explanatory power, and in order to allow us to then interpret the coefficient on $Novelty$, as literally representing how individual evaluators treat more novel proposals differently (*i.e.*, comparing responses across proposals and within evaluators, rather than across evaluators).

To fully implement the model described in expression (2), we turn to accounting for unobserved research proposal characteristics and quality (Section 5.2). Here, we exploit data which are only available for the 689 proposal-evaluator pairs related to submissions coming from the host university. We report results on this subsample in Table 5. We begin by re-regressing the earlier model (5) from Table 4 on this subsample to confirm similar results, as reported in model (1) of Table 5. The estimated coefficient on *Novelty* for this subsample, -3.37, is similar to the earlier estimate. (The coefficient estimated on just the excluded data is itself significant and negative.) In regressions performed here, our goal is to introduce variables that serve as predictors of the unobserved characteristics and quality of research proposals, apart from novelty (Section 5.2). Model (2), for example, introduces the number of publications the submitting researcher has in the endocrine-related disease area. The positive and significant coefficient clearly indicates this count serves as a meaningful predictor of proposal quality and scoring, despite coming from outside the experiment. More crucial to our analysis, introducing this predictor has no significant impact on the coefficient estimated on *Novelty*. Model (3) alternatively introduces the total citation count in these publications and finds similar results. There is no such positive relationship between score and either overall publications or overall citations, as in models (4) and (5), and just the sign on total publications is significant. These coefficients are also difficult to interpret, as they may reflect not only the general success of a researcher but also differences across fields. Nonetheless the significance, at least on publications in model (4), suggests it is somehow predictive of scores. Most importantly, the coefficient on $Novelty$ is again unchanged.

<Table 5>

In model (6), we go further by including each of these four measures of publications and citations at once, along with 23 other statistics making a "long list" of 27 predictors.[8] Remarkably, the point estimate and significant of the estimated coefficient on *Novelty* still does not change. This is especially of note given that including this long list of predictors of unobserved proposal characteristics leads the adjusted-R^2 statistic to jump from .3 to .4. Therefore, this bundle of predictors does seem to be offering substantial predictive power for unobserved proposal characteristics. In model (7) we add still another variable, the word count of proposals. This, too, does not significantly change the estimate. Therefore, we see overwhelmingly stable estimates of the coefficient on *Novelty* despite introducing these meaningful predictors of unobserved proposal characteristics and quality. Therefore, subject to further robustness checks, it appears we may return to the complete data set without these additional controls.

6.2 Robustness

Unobserved Proposal Characteristics, *i.e.*, $cov\left(Novelty_{p,}, _{ep}\right) = 0$. Our most important concern overall in estimating the coefficient on $Novelty$, the point just dealt with above, is that novelty may be correlated with unobserved proposal quality and characteristics that affect scoring. Perhaps a first and most important assurance is that including a barrage of predictors for unobserved proposal characteristics, as in the preceding section and in Table 5, did not affect the estimated coefficient on $Novelty$. This is an immensely convenient feature of the structure of the data and underlying mechanisms that we could not have predicted in the initial design of this study.

A second source of assurance is the extent to which the variables intended to capture unobserved proposal characteristic appear to explain a large share of the variation that is potentially attributable to unobserved proposal differences. Whereas regressing a model with only evaluator dummies produces an R^2 statistic of .40, adding $Novelty$ and the "long list" of predictors of unobserved proposal quality leads the R^2 statistic to increase to .55. (adjusted-R^2 statistic = .40). Importantly, we can compare this to the total amount of variation that might potentially be explained in the best of cases by comparing this R^2 to what we get when regressing the score on both evaluator dummies and research proposal dummies at once. In this case, the R^2 statistic is .61. Therefore, our "long list" of predictors of unobservable proposal characteristics and quality can be understood to explain a remarkable 71% of all possible variation that might be explained by cross-sectional proposal

[8]Controls include: counts of endocrine publications; counts of endocrine publication citations; counts of total publications; counts of total publication citations; re-counts of the earlier four measures, based on just the past 3, 5, and 7 years; counts of publications divided by author position (first author, second author, last author, second last author, other position); counts of publications with MeSH terms as in the research proposal; counts of publication citations with MeSH terms as in the research proposal; counts of publications with MeSH terms as in the research proposal; counts of publication citations with MeSH terms as in the research proposal; counts "hit" publications based on achieving top citation percentiles within year of publication (including separate measures for 99th, 95th, 90th and 75th percentiles); counts of citations for most cited paper in endocrine publications; counts of citations for most cited paper in total publications; counts of citations for most cited paper in publications with at least one MeSH term as the research proposal.

differences (or $(.55 - .40)/(.61 - .40)$). Therefore, we have explained a large share, a majority of unobserved variation with our predictor strategy and have found this variation is orthogonal to variation in novelty.

A third source of assurance relates to implementing research proposal dummies. Of course, using these dummies leads *Novelty* or any other proposal covariate to drop out of the model. However, we can at least test to see whether the model remains stable for characteristics of proposal-evaluator pairs (that do not drop out) when adding both proposal and evaluator dummies, as shown in Table 6. Here we use a measure of the intellectual *Distance* between evaluators and research proposal. (Relationships with this measure are studied substantively in Section 6.4; here we simply use the measure instrumentally to test model robustness.) Model (1) first re-regresses our baseline model with the entire data set (i.e., model (5) of Table 4) while adding our measure of *Distance* between evaluators and proposals (Table 2). Model (2) regresses the same model, but this time includes research proposal dummies so as to wholly control for cross-proposal differences. Of course the *Novelty* variable drops out. Crucially, the coefficient on *Distance* is statistically unchanged. This is notable given the radical re-specification of the model with proposal dummies (as indicated by the jump in adjusted-R^2 statistic from .22 to .48). Models (3) and (4) repeats such a comparison, but this time also including the interaction between $Novelty \times Distance$. While the estimated coefficients on the interaction in both cases are insignificant, the point estimates are almost identical with and without proposal dummies. And again, the estimated coefficients on *Distance* remain almost identical across these models. The close similarity of coefficients with and without proposal dummies provides added assurance our estimation approach in dealing with unobserved proposal characteristics has been successful. This is fortunate, given this estimation challenge is an inherent limitation that should afflict any study attempting to measure a novelty effect.

<Table 6>

"Blind" Evaluations, *i.e.* $cov\,(X_{e-r}, _{ep}) = 0$**.** Important to our analysis is that evaluators responded to the contents of research proposals, rather than to the identities of researchers behind the proposals. The combination of a "triple-blind" process (Section 3.1), the emphasis in the call on proposing new research (rather than potentially well-known existing work), the wide call for participation and the wide set of evaluators, and the unorthodox format of the submissions should have all served to obscure identities of researchers to a considerable degree. Fortunately, we are able to explicitly test for effects of relatedness between evaluators and researchers in these data. We have 39 instances in which submitting researchers and the evaluators examining their proposals were from the same hospital or organization within the broader university. We also have 11 cases in which submitting researchers and evaluators were coauthors in previous work. Models (5) and (6) introduce indicators for same organization and coauthors, while model (7) introduces them together. Consistent with the successful blinding process, coefficients on these variables are statistically indistinguishable from zero, while the coefficient on *Novelty* is unchanged.

The Novelty Measure. Another feature of the analysis deserving greater scrutiny is our main explanatory variable, *Novelty*. The bulk of our discussion on this point appears in Section 4.1. Beyond this, we might simply ask the question "*How novel is novel*?". For example, should we expect to see a discrete change as a research proposal moves from zero *Novelty* to non-zero levels of *Novelty*? Might there be non-linear effects? To assess these questions, we re-estimated our model. Models (8) and (9) each feature a dummy to distinguish zero from non-zero levels of *Novelty*. Model (8) further allows for a possible convex or concave relationship with *Novelty* in allowing for an affine relationship, including quadratic and linear terms. Model (9) is similar, except specifying the relationship with *Novelty* as a flexible, non-parametric relationship. The non-parametric partial relationship between *Score* and *Novelty* of the model is reported in Figure 1. The results in both models suggest no stark or qualitative departures from the linear model. Both models (8) and (9) find no discrete difference between zero and non-zero levels of *Novelty* and indicate the relationship between *Score* and *Novelty* is slightly concave, becoming slightly more negative for higher levels of *Novelty*.

<Figure 1>

Alternative Dependent Variables. Our basic assumption in our analysis is that our dependent variable, the overall impact score (customary in evaluations), is the most relevant dimension of the evaluation to focus on. At the same time, we should expect that patterns in relation to our other measures of evaluators' judgments, *Feasibility* and *Allocation*, should not yield fundamentally different patterns. To confirm this supposition, models (10) and (11) in Table 6 replace the dependent variable in our baseline model with these variables. We find that replacing the dependent variable in this way yields similar signs and significant results; although, as should be expected, with lower statistical significance.

6.3 Magnitude of Effects

We can use the model coefficients to gain a better appreciation for how the novelty discount should affect outcomes. We begin with a simple illustration. The score drops -2.21 points (rated on the 10-point scale) for every unit change of our *Novelty* measure. Therefore, for every standard deviation of *Novelty* (standard deviation = .11), evaluators reduce the score by .24 points. If we ask how such a systematic discounting would affect the outcome of a given research proposal, consider that a .24-point discount implies a .17 standard deviation drop in terms of the average scores of proposals (standard deviation = 1.44). If we approximate the distribution of mean proposal scores as Normal, this drop of .17 standard deviations relative to the mean implies a 7% drop in the percentile. In the case of 150 proposals, this means a drop of 10.5 rank positions.[9]

[9]Another appreciation of magnitude of effects can be achieved by re-regressing this model, but replacing the dependent variable and *Novelty* with the log-transforms to allow the coefficient on *Novelty* (-.58) to be interpreted as an elasticity measure. Therefore, on average, for every 1% increase in novelty, the score drops .6%.

A limitation of the above estimate of the impact on rank order is it is in relation to the mean of the distribution–the "fattest" part of the distribution where a drop in score has the greater impact on one's rank. Estimating the average effect would effectively involve conducting the earlier comparison, but doing so across the entire distribution to estimate the average effect. At the same time, we can improve our ability to assess the impact for individual proposals by moving from the parameterized Normal estimate of the distribution, used above, to using the actual distribution of scores. To simulate these outcomes, we effectively recalculate the ranks of individual proposals if their scores were to drop by .24 points. Results are summarized in Figure 2. The average drop in rank over the entire sample would be -6.8 places or a 4.5 percentiles. These values are sensibly slightly smaller than the estimates in the preceding paragraph where we calculated the effect in relation to the mean proposal. Apart from the population average effects, we should also be especially interested in the marginal effect of novelty in the right tail–or the very best proposals. For example, eight of the top ten proposals would drop at least one rank if they were one standard deviation more novel. It should also be noted in regard to Figure 2 that the number of simulated incidences with zero changes in rank is a minority of cases.

<Figures 2,3>

Apart from these marginal effects, it is also informative to consider the infra-marginal effects–or what would have happened if novelty simply did not play a role at all in shaping evaluations. To simulate this effect, we adjust each of the received scores by their total effect of novelty. Here we subtract -2.21, multiplied by the particular value of $Novelty$ for each proposal. Recall, we showed earlier the effect was linear over the entire domain of $Novelty$.) We then re-rank the proposals and compare those ranks with the original, actually received ranks. As can be seen in Figure 3, the effects are large, with a standard deviation in change of rank of 6.4–implying that roughly a third of the sample would experience a change in rank larger than this. (By definition, the average change of rank over a list of fixed size must be zero.)

6.4 Interpretation of Novelty Discount

Having shown a novelty discount, above, and established the robustness of the result, we turn to the question of what mechanisms might be the reason for the novelty discount. Several explanations can be set aside prior to proceeding to any analysis. For example, the measured effect is not caused by selection and sorting of novel proposals to certain types of evaluators; the randomization procedure addressed this possibility (Section 3.4). The measured novelty discount is also not the result of prejudicial administration of grants (ex: requiring pre-existing data, etc.). The format of proposals was designed to avoid this possibility (Section 3.2). Nor was the discount caused by some sort of prejudice against the authors of these research proposals or their relationships with evaluators. Researcher identities were blinded (Sections 3.1 and 6.2). It is also not the case that

novel proposals were simply of lower quality. This was the key point in our estimation strategy and robustness checks (Section 6.1). In the following analysis, we consider how a range of possible incentive and information effects and perceptions of novel proposals might play a role.

Departures from Existing "Map of Science" or Departures from Evaluators' Own Specialized Knowledge? To better triage possible explanations for the novelty discount, we first test whether novelty, a "departure" from the body of existing research, is any different from a departure from an evaluator's own specialized knowledge. We are able to exploit varying proposal-evaluator distance to explicitly estimate the effect of distance from the evaluator, as distinct from novelty. As a measure of *Distance*, we calculate the angular separation between vectors representing the MeSH terms in a given proposal and counts of MeSH terms in the publication history of the evaluator, generating a scale-free measure of the proximity that varies between zero and one. In order for this value grow larger with distance, we define *Distance*, as one minus this angular separation. Model (1) of Table 7 begins by reporting the earlier preferred baseline model (model 5 of Table 4) to ease comparison. Model (2) introduces this *Distance* measure and model (3) does so while also including dummies for individual research proposals. The coefficient on *Distance* is positive and significant and statistically the same in these two models.[10] This indicates that intellectual distance is indeed an important factor shaping evaluation, and that evaluators tend to give more stringent evaluations to research proposals that are closer to their own areas of expertise.[11] Evaluators might have some combination of both greater ability and greater incentives to evaluate proposals "close" to them more stringently. More crucial to our interpretation of the novelty discount, introducing this variable has no effect on our estimated coefficient on *Novelty*. This suggests that a research proposal's departure from the existing map of science has a distinct and separate effect from the question of being distant from an evaluator's own area of deep specialized expertise.[12]

<Table 7>

Incentive Effects: Strategic Incentives or Conflicts of Interests? While novel proposals should by definition depart from existing research pathways, they might still serve as alternative or substitute pathways to certain bodies of existing research. This could plausibly create incentives

[10] While the particular question of relatedness of an evaluator and the evaluated proposal is not the thrust of our analysis and has been examined in earlier research, our estimate here is perhaps useful in several regards. It exploits both dummy variables for the evaluator and for the proposal. It focuses on relatedness to the content of the proposal (through blinding), rather than to the identity of the researcher. Our estimate is also based on relatedness to a "first phase" generation of an idea and research hypothesis, rather than for completing an entire proposal. Here we also focus on an individual evaluator's assessment, as the unit of observation, rather than discrete approval outcomes or the overall evaluation of a group of evaluators.

[11] We also considered possible quadratic and indicator-based relationships with *Distance*. None of these alternative specifications changes the estimated coefficient on *Novelty*.

[12] We find similar results when replacing the angular separation measure of distance to one that is algebraically more similar to how we constructed the *Novelty* measure, whereby we measure the fraction of MeSH term pairs that the given evaluator has not previously encountered.

to alter evaluations among those evaluators most proximate and most likely to have a stake in the outcomes of novel research. For example, a novel approach could plausibly represent a challenge to existing research paths and compete for resources and attention. Alternatively, a novel research proposal might somehow build interest in an existing research area and act as a complement to existing proximate areas. To seek evidence that might be consistent with any sort of strategic or incentive effects, we test whether evaluators most proximate to novel research proposals respond to novelty any differently than those who are more distant and therefore potentially more disinterested. To do so, model (4) of Table 7 introduces an interaction term between *Novelty* and *Distance* into the model. Given we are now interested in estimating the coefficients on characteristics of evaluator-proposal pairs, we continue to include both evaluator and proposal dummies in the model to assure most precise estimates. (Consequently, all direct effects in these and following interactions need not be explicitly included.) The estimated coefficient on *Distance* remains unchanged, but that on the interaction term is not statistically distinguishable from zero. To further vet the possibility that more proximate evaluators respond differently to novelty, model (5) introduces an interaction of *Novelty* with an indicator for whether the evaluator conducts research in the endocrine-related disease area; and model (6) interacts *Novelty* with a count of the number of the disease-related publications of the evaluator. Neither interaction is found to be significant. In model (7) we then tested whether a more discrete measure of distance might better capture any strategic effects. Here we interacted *Novelty* with an indicator switched on when *Distance* is more than two standard deviations below the mean of that variable (*i.e*, .72). Again, as reported in model (7), we find no indication of a significant interaction term and therefore no evidence that more proximate evaluators treat novelty any differently.

In finding no evidence that more proximate evaluators treat novelty differently, we fail to find any strong indication of strategic incentives or conflicts of interest playing a role in the discounting of novelty. Further, these results more generally underline a qualitative difference between departing from the "map" of (normal) science, and simply departing from one's own area of expertise. While intellectual "distance" from a proposal clearly shapes evaluations (consistent with earlier research), we find no evidence from these tests that more proximate evaluators treat novelty any differently.

Information Effects: The Limits of Knowledge in Evaluating Novel Ideas? Having failed to find evidence consistent with the strategic incentives of evaluators playing a role in the novelty discount, here we seek evidence of whether it might be inherently and unavoidably challenging to assess the quality of novel ideas on the basis of existing knowledge–an information effect of sorts affecting the evaluation of novelty. Certainly a relatively uniform discounting of novelty across the population, as exhibited in earlier tests, provides evidence consistent with this possibility. However, here we intend to implement more discriminating tests. We begin with the conjecture that if it is difficult to evaluate novel ideas, that there might still be certain individuals with at least incrementally greater capabilities to perform these evaluations. In particular, it may be those who are both

i) most proximate to the proposal and ii) with highest experience and knowledge.[13] Therefore, we go beyond the earlier two-way interactions between novelty and proximity and introduce an added three-way interaction with measures of knowledge and experience. Thus, we effectively ask whether proximate evaluators of relatively high quality and experience treat novelty any differently. (Introducing three-way interactions, of course, we also introduce all constituent two-way interactions and direct effects not already dealt with by proposal dummies and evaluator dummies in the model.)

We investigate two different measures of knowledge and experience. These are, first, an indicator for whether the faculty is senior and, second, the number of publications. We find no evidence that senior faculty, as a group, respond differently to novelty when they are close to the novel proposal, as in model (8). However, we do find significant interactions where we measure knowledge and experience in terms of numbers of publications, as in model (9).[14] In this model (9), the negative interaction between *Distance* and the number of publications indicate that more accomplished evaluators tend to respond more negatively, in general, to more distant proposals than do less accomplished evaluators.

As regards the interactions with *Novelty*, the negative interaction between *Novelty* and the number of publications, it appears that more widely published evaluators are even more critical in general of novel proposals than lesser published evaluators already are. It is difficult, however, to suggest that this comes from strategic incentives in the sense of a conflict of interest: the three-way interaction in model (9) indicates it is precisely where proposals are proximate to most published evaluators where they give a slight boost in the evaluation. The positive three-way interaction is by no means a wholly discriminating test; however, it is consistent with more accomplished evaluators having an incrementally greater ability to interpret novel proposals, and therefore being less inclined to reflexively discount these particular proposals.

If novel proposals are indeed intrinsically difficult to evaluate on account of their novelty and on account of the bounded ability of researchers to draw on existing knowledge to evaluate new ideas, one possibility is we might simply see greater variance and disagreement across evaluations in the case of novel proposals. Greater variance or uncertainty might itself account for discounting, despite no systematic differences in expected quality *per se*. To assess this possibility, we re-estimate our model, but allow model variance to vary as a parametric function of *Novelty*, as we simultaneously estimate the conditional mean. As reported in model (1) of Table 8, we find no systematic relationship between model variance and *Novelty*; the coefficient indicating the relationship between model variance and *Novelty* is insignificant, small, and even nearly zero (while coefficients in the conditional mean part of the model remain roughly unchanged). To further probe this possibility,

[13] It is possible to develop counter hypotheses that those who are closest and most experienced are least able to recognize useful novel solutions that depart from the orthodoxy. Most important here, we seek evidence of any significant patterns whatsoever.

[14] Given we are now introducing multiple interactions into the model, we assure the signs and significance are indeed meaningful, by dropping the three-way interaction to assure the two-way interactions retain their sign and significance.

we re-specify our analysis altogether, to focus on (all proposals for) each proposal as the unit of analysis (*i.e.*, 150 proposal observations, rather than 2,130 evaluator-proposal observations). Model (3) has as its dependent variable the range (*i.e.*, max - min) of evaluator scores for each proposal, Model (4) has as its dependent variable the standard deviation of all scores for each proposal. We find no significant relationships with $Novelty$. Therefore, we find no evidence of greater variance or disagreement in scores with increased novelty. Therefore, we find no evidence that novelty is leading to a wider range of subjective assessments; only a uniform discounting of novel proposals.

<Table 8>

Another potential interpretation of discounting that could potentially account for the novelty discount is a general expectation of greater risk, despite no inherent quality differences. To gain greater insight into this possibility, we exploit evaluators' own self-reported subjective $Feasibility$ scores. Model (1) of Table 9 begins by reporting our preferred baseline model to ease interpretation. Model (2) adds this new variable $Feasibility$ in attempts to "explain away" at least part of the novelty discount. We find that the overall evaluation is indeed related to $Feasibility$ significantly and positively. More importantly, it explains 23% of the novelty discount (i.e., $\|(-2.21 - 1.70)/-2.21)\|$). Therefore, while the formulation of these beliefs around feasibility might themselves be subjective, the response to these beliefs–a discount–would seem entirely appropriate. Given it appears we have found an important effect, we wish to better assure we have properly specified and calibrated it. We begin in model (3) with a most flexible specification of $Feasibility$ as a series of dummies for each integer value of the variable. We do not find any difference in fit or the impact on the $Novelty$ coefficient (*i.e.*, how much of the novelty discount is explained). Therefore, this test suggests we have explained as much of the novelty discount that can be explained with the earlier linear model. In model (4), we attempt to assess whether it is the perception of feasibility or some true underlying feasibility that individuals are responding to. We are limited in what we can measure in this respect, however as a rough indication model (4) replaces each evaluator's perceived $Feasibility$ measure with an average of this measure across all evaluators on a proposal, in case this might approach something closer to an "objective" measure of feasibility. However, the fit of this model drops appreciably, as per the adjusted-R^2 falling to .41 in model (4) from .56 in model (2). Therefore, subjective feasibility scoring appears to play an important role above any absolute or objective feasibility.

<Table 9>

7 Conclusions

In the paper we presented results and analysis from a medical research grant proposal process in which we investigated how expert evaluators–elite researchers from a leading medical school–treated

novel research proposals. We designed and implemented this study to "evaluate the evaluators" by manipulating and altering features of the grant proposal process in order to derive relevant inferences. The experimental design assured a blind process and that evaluators responded to the content of the proposals rather than to the identity of submitting researchers. We also exploited random assignment to assure any measured effects were not related to selection and sorting of research proposals viz. evaluators. We found a large, robust and relatively stable and uniform novelty discount across a wide range of tests. A standard deviation increase in novelty, all else being equal, led evaluators to provide a score that resulted in a roughly 4.5 percentile drop in ranking, on average. The key estimation challenge in the analysis was to assure that unobserved characteristics and quality of research proposals were not themselves the cause of the measured discount. We confirmed the relationship and underlying assumptions of the research design across a wide range of robustness tests and diagnostics.

In additional testing, we found that novelty–a departure from existing science–shaped evaluator scoring in qualitatively different ways than did mere "distance" from the specialized expertise of an evaluator. Whereas novel research proposals were relatively generally discounted, evaluators tended to be more critical of proposals that were closer to their area of expertise–and these two effects of novelty and intellectual distance appeared to work largely independently of one another. Therefore being "off the map of science" appears to have generally qualitatively different effects on evaluators' assessments than does simply being out of one's depth.

We found evidence that the most accomplished (*i.e.*, most published) of researchers tended to discount novelty more heavily than the rest of the population, but discounted novelty slightly less so when proposals were intellectually proximate to their own area of expertise. The results are more easily interpreted as consistent with inherent information challenges in evaluating novel ideas on the basis of existing knowledge than they are in terms of, say, a conflict of interest or strategic incentives in the evaluation process. We could not find evidence that novel proposals created greater disagreement or subjective evaluations of their merit. Rather greater novelty simply led to relatively uniform discounting. We see a novelty discount of similar magnitude across a range of tests, subsamples, different types of evaluators.

While we did not seek to (and are unable to) assess the productivity aspects of novelty, its treatment may be consequential to the organization of peer evaluation. Fleming (2001) showed that on average increasing novelty in the patent database yields lower success but higher variability leading to a disproportionate share of breakthroughs. One interpretation of our finding is that peer evaluators have internalized, within the normal science paradigm, the average effects of novelty and thus discount it uniformly, for potential concerns about the lower success rates. This appears to be true, at least to some extent, as roughly a quarter of the discount related to subjective evaluations of lower feasibility of novel proposals. However, this censoring of novel projects means that the experiments never get a chance to be run and thus the benefits of generating variance and greater diversity of experiments are curtailed. This should be of concern to policy makers and

society in general as precious research funds are being allocated more towards incremental research as compared to high variability and potentially breakthrough efforts. In the natural sciences, the capital intensive nature of most laboratories means that the lack of funding for the initial grant application almost certainly ensures that the scientists will all together drop that line of research and steer themselves towards more normal science.

Bias in evaluations due to evaluators' responses to the identity of researchers can be fixed by altering the process (insisting on double and triple blind evaluation) (Goldin and Rouse, 2000) or creating incentives and systems for accurate information sharing (Li, 2012) or perhaps changing the composition of evaluators. By contrast, it is more difficult to imagine how an evaluator's score would not be impacted by the limits of their own knowledge and, more particularly, the bounds of the established map of science. Indeed, the paradox exists because innovation requires novelty–but novelty, as we have shown, is not appreciated and is in fact penalized. We speculate there might possibly be remedies such as the introduction of new scoring metrics, the education of evaluators, the use of algorithm based complementary scoring and possibly other approaches, which each deserve deeper study. Further, and perhaps more profoundly, the question of what the optimal level of novelty in the research portfolio remains an open question.

References

Basalla, G. 1988. *The Evolution of Technology*. Cambridge University Press, Cambridge.

Carter, G. M. 1982. What We Know and Do Not Know About the NIH Peer Review System. Tech. rep., RAND Corporation.

Chubin, D. E., E. J. Hackett. 1990. *Peerless Science: Peer Review and US Science Policy*. SUNY Press, Albany.

Cole, S., J. R. Cole, G. A. Simon. 1981. Chance and consensus in peer review. *Science* **214**(4523) 881–886.

Coletti, M. H., H. L. Bleich. 2001. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association* **8**(4) 317–324.

Fleck, L. 1935. *The Genesis and Development of a Scientific Fact*. 1979th ed. University of Chicago Press, Chicago.

Fleming, L. 2001. Recombinant uncertainty in technological search. *Management Science* **47**(1) 117–132.

Gläser, J., G. Laudel, S. Hinze, L. Butler. 2002. Impact of evaluation-based funding on the production of scientific knowledge: What to worry about, and how to find out. *Expertise for the German Ministry for Education and Research* .

Goldin, Cl., C. Rouse. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review* **90**(4) 715–741.

Horrobin, D F. 1990. The philosophical basis of peer review and the suppression of innovation. *Journal of the American Medical Association* **263**(10) 1438–41.

Jayasinghe, U. W., H. W. Marsh, N. Bond. 2003. A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **166**(3) 279–300.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press, Chicago.

Langfeldt, L. 2001. The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science* **31**(6) 820–841.

Langfeldt, L. 2006. The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. *Research Evaluation* **15**(1) 31–41.

Laudel, G. 2006. Conclave in the Tower of Babel: how peers review interdisciplinary research proposals. *Research Evaluation* **15**(1) 57–68.

Li, D. 2012. Information, Bias, and Effciency in Expert Evaluation: Evidence from the NIH. *MIT Working Paper* 1–57.

Luukkonen, T. 2012. Conservatism and risk-taking in peer review: Emerging ERC practices. *Research Evaluation* **21**(1) 48–60.

Marsh, H. W., U. W. Jayasinghe, N. W. Bond. 2008. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *American Psychologist* **63**(3) 160.

Merton, R. K. 1968. The Matthew Effect in science. *Science* **159**(3810) 56–63.

Mutz, R., L. Bornmann, H.-D. Daniel. 2012. Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS One* **7**(10).

Olbrecht, M., L. Bornmann. 2010. Panel peer review of grant applications: what do we know from research in social psychology on judgment and decision-making in groups? *Research Evaluation* **19**(4) 293–304.

Peters, D. P., S. J. Ceci. 1982. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences* **5**(2) 187–255.

Simonton, D. K. 1999. *Origins of Genius: Darwinian Perspectives on Creativity*. Oxford University Press, New York.

Singh, J., L. Fleming. 2010. Lone inventors as sources of breakthroughs: Myth or reality? *Management Science* **56**(1) 41–56.

Sorenson, O., L. Fleming. 2004. Science and the diffusion of knowledge. *Research Policy* **33**(10) 1615–1634.

Stephan, P. E. 2012. *How Economics Shapes Science*. Harvard University Press, Cambridge, MA.

Ware, M. 2011. Peer review: recent experience and future directions. *New Review of Information Networking* **16**(1) 23–53.

Weitzman, M. L. 1998. Recombinant growth. *Quarterly Journal of Economics* **113**(2) 331–360.

Zuckerman, H., R. K. Merton. 1971. Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva* **9**(1) 66–100.

FIGURES

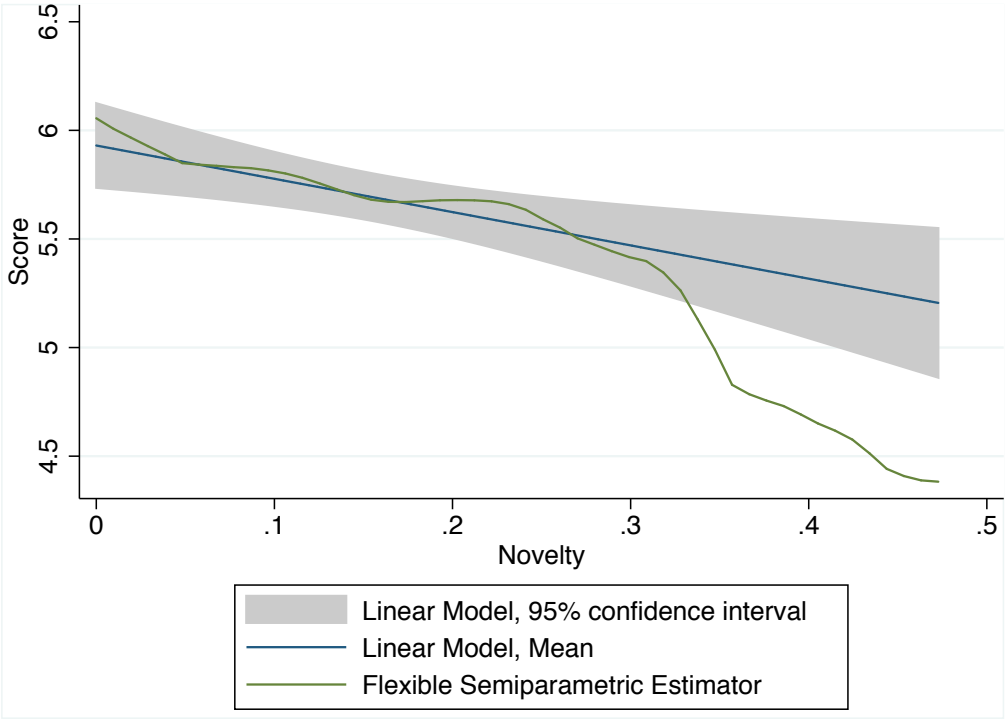

Figure 1 Linear versus Semi-Parametric Specifications of the Relationship Between *Score* and *Novelty*
Notes. The graph presents the estimated partial relationship between *Score* and *Novelty* from our preferred baseline linear model, model (5) of Table (4), and an alternative flexible, semi-parametric estimate of this relationship. The semi-parametric model is estimated on a two-stage estimate, first estimating parametric coefficients and then separately estimating the non-parametric relationship with *Novelty* using locally-weighted regression methods (Yatchew 1998).

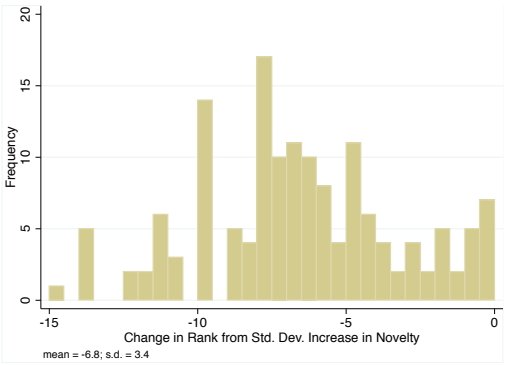

Figure 2 Simulation of the Effect of a Standard Deviation Increase in Novelty for Each Research Proposal on Rank Outcomes

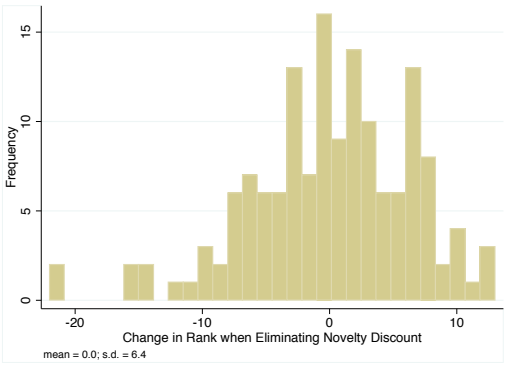

Figure 3 Simulation of the Effect of Eliminating the Novelty Discount on Rank Outcomes

TABLES

Table 1 Evaluator Characteristics

	Disease-Domain Publications		Coauthor with Disease-Domain Publications		No Disease-Domain Publications	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
From Inside University	1.00	.00	1.00	.00	1.00	.00
Assistant Professor or Instructor	.50	.50	.50	.50	.48	.50
Full Professor	.50	.50	.50	.50	.52	.50
Female	.40	.49	.30	.46	.48	.50
Year Birth	1962	11	1962	10	1958	13
Year Final Degree Awarded	1989	13	1990	10	1986	15
Publication Count	118	164	77	74	107	150
Ph.D.	.62	.49	.52	.50	.37	.48
M.D.	.74	.44	.65	.48	.74	.44

Notes. Number of observations = 142 evaluators.

Table 2 Means, Standard Deviations and Correlations

Variable	Mean	S.D.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Proposal Characteristics																
(1) *Score*	5.7	2.6														
(2) *Novelty*	.16	.11	-.06													
(3) *MeSH Term Count*	12.43	5.42	.15	.20												
(4) *Feasibility*	5.23	2.56	.69	-.01	.15											
(5) *Points Allocation*	6.67	9.17	.51	.00	.15	.49										
(6) *Word Count*	1366	2489	.02	.10	.14	.06	.03									
Researcher Characteristics																
(7) *Total Publications*	9.13	24.01	.03	-.09	.23	.06	.07	.00								
(8) *Total Citations*	99	521	.07	.04	.31	.10	.09	.05	.90							
(9) *Endocrine Publications*	.22	1.07	.07	-.07	.12	.12	.19	.05	.05	-.02						
(10) *Endocrine Citations*	1.40	6.79	.06	-.02	.14	.11	.15	.03	.05	-.02	.86					
(11) *University Affiliation*	.47	.50	.13	-.02	-.09	.16	.14	-.12	.	.	.	.				
Evaluator Covariates																
(12) *Total Publications*	101	138	-.12	-.01	-.01	-.05	.00	.01	.00	-.01	.01	.00	-.01			
Proposal-Evaluator Characteristics																
(13) *Distance*	.87	.07	.16	.14	.33	.15	.08	.03	.08	.14	.24	.18	-.03	-.28		
(14) *Coauthor*	.01	.07	.01	.02	.00	.00	.02	-.02	.17	.11	.11	.12	.08	.03	-.01	
(15) *Same Hospital*	.02	.13	.01	-.06	.00	.01	.02	-.01	.06	.02	-.03	-.01	.14	.01	-.04	.14

Notes. Number of observations = 2,130 research proposal-evaluator pairs, except for (7), (8), (9), (10) for which there are 689 research proposal-evaluator pairs.

Table 3 Variable Definitions

Variable	Description
Proposal Characteristics	
Score	Integer score between 1 and 10 response to the following question: "*On a scale of 1 to 10 (1 Lowest - 10 Highest) please assess the impact on T1D care, patients, or science of research that successfully addressed the questions or successfully followed the approaches proposed in the submissions*."
Novelty	Fraction of MeSH term dyads (from all possible combinations of terms associated with a given proposal) that are not observed in prior published research
MeSH Term Count	Number of MeSH terms coded by expert to reflect the contents of the research propsal
Feasibility	Integer score between 1 and 10 response to the following question: "*On a scale of 1 to 10 (1 Lowest – 10 Highest), please assess the likelihood that a research proposal could be executed based on this submission*"
Points Allocation	Integer score between 1 and 10 response to the following question: *Imagine you had limited resources to fund the proposals. Please allocate 100 points amongst all the proposals. Higher point allocations mean that you give a higher priority to that proposal being funded and developed.*"
Word Count	The character count of the submitted research proposal, divided by 6
Individual Mesh Term Dummies	Among the 737 different MeSH terms in the proposals, this is a series of dummy variables for the 75 that appear in at least 5 research proposals; all other are captured with an "other" dummy
Researcher Characteristics	
Total Publications	Count of all prior publications listed in PubMed
Total Citations	Count of the citations of all prior publications listed in PubMed
Endocrine Publications	Count of all prior publications listed in PubMed related to the endocrine system disease area
Endocrine Citations	Count of the citations of all prior publications listed in PubMed related to the endocrine system disease area
"Long List" of Proposal Quality Predictors	The four preceding variables, in addition to 23 other measures. Controls include: counts of endocrine publications; counts of endocrine publication citations; counts of total publications; counts of total publication citations; re-counts of the earlier four measures, based on just the past 3, 5, and 7 years; counts of publications divided by author position (first author, second author, last author, second last author, other position); counts of publications with MeSH terms as in the research proposal; counts of publication citations with MeSH terms as in the research proposal; counts of publications with MeSH terms as in the research proposal; counts of publication citations with MeSH terms as in the research proposal; counts "hit" publications based on achieving top citation percentiles within year of publication (including separate measures for 99th, 95th, 90th and 75th percentiles); counts of citations for most cited paper in endocrine publications; counts of citations for most cited paper in total publications; counts of citations for most cited paper in publications with at least one MeSH term as the research proposal.
University Affiliation	Indicator variable switched to one for researchers from the host university
Evaluator Covariates	
Total Publications	Count of all prior publications listed in PubMed
Evaluator Block Dummies	Evaluators were assigned to assigned randomly to 10 separate blocks of 15 randomly chosen research proposals; this is a series of 10 dummy variables corresponding to those randomized blocks
Proposal-Evaluator Characteristics	
Distance	Angular separation between the vectore representing the MeSH term coverage of the research proposal and the MeSH term coverage of the evaluator's publication history
Coauthors	Indicator switched on if evaluator and submitting researcher have been coauthors in the past
Same Hospital	Indicator switched on in cases in which the evaluator and submitting researcher are employed with the the same hospital or division within the university

Table 4 Baseline Estimated Novelty Effect

Dependent Variable:	*Score*				
Model:	1	2	3	4	5
	Linear MeSH Count	Flexible MeSH Count	MeSH Term Dummies	Evaluator Block Dummies	Individual Evaluator Dummies
Research Proposal Characteristics					
Novelty	-2.38***	-2.35***	-2.3015***	-2.21***	-2.21***
	(.53)	(.48)	(.50)	(.61)	(.51)
MeSH Term Count	.083***			.081***	.081***
	(.011)			(.012)	(.008)
MeSH Count Dummies		Y	Y		
Individual MeSH Term Dummies			Y		
Evaluator Characteristics					
Evaluator Block Dummies				Y	
Evaluator Dummies					Y
Constant	5.04***				
	(.10)				
Adj-R^2	.03	.07	.15	.03	.22

Notes. OLS estimates; *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively; bootstrapped standard errors are reported; number of observations = 2,130 research proposal-evaluator pairs.

Table 5 Estimated Novelty Effect, Accounting for Unobserved Proposal Characteristics and Quality

Dependent Variable:	*Score*						
Model:	1	2	3	4	5	6	7
	Baseline Model	Related Pubs	Related Cites	Total Pubs	Total Cites	"Long" List of Controls	Proposal Word Count
Research Proposal Characteristics							
Novelty	-3.37***	-3.13***	-3.28***	-3.74***	-3.42***	-2.97**	-3.35**
	(.92)	(.93)	(.92)	(.94)	(.93)	(1.49)	(1.45)
MeSH Term Count	.090***	.072***	.074***	.087***	.082***	.103***	.074*
	(.016)	(.017)	(.017)	(.017)	(.018)	(.037)	(.043)
Number of Words							.000
							(.000)
Researcher Characteristics							
Endocrine Publications		.209***					
		(.062)					
Endocrine Citations			.025**				
			(.012)				
Total Publications				-.0063*			
				(.0030)			
Total Citations					-.0001		
					(.0000)		
"Long" List of Proposal Quality Predictors[(1)]						Y	Y
Evaluator Characteristics							
Evaluator Dummies	Y	Y	Y	Y	Y	Y	Y
Adj-R^2	.30	.30	.30	.30	.30	.40	.40

Notes. OLS estimates; *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively; heteroskedasticity-autocorrelation robust standard errors are reported; Number of observations = 689 proposal-evaluator pairs and pertain only to submitting researchers from within the host university.

Table 6 Robustness

Dependent Variable:	*Score*									*Feasibility*	*Allocation*
	With and Without Proposal Dummies				Response to Researcher Identity			Affine	Semi-Param.	Alternative Dependent Variables	
	1	2	3	4	5	6	7	8	9	10	11
Research Proposal Characteristics											
Novelty	-2.35***		-5.88		-2.21***	-2.18***	-2.18***	3.63**	*See Figure 2*	-.780*	-3.19*
	(.41)		(5.64)		(.49)	(.49)	(.49)	(1.72)		(.41)	(1.66)
Novelty^2								-14.9***			
								(3.6)			
I{Novelty > 0}								.187	.340		
								(.27)	(1.28)		
MeSH Term Count	.068***		.067***		.081***	.081***	.081***	.066***	.076***	.078***	.277***
	(.01)		(.01)		(.010)	(.010)	(.010)	(.010)	(.043)	(.01)	(.05)
Research Proposal Dummies		Y		Y							
Evaluator Characteristics											
Evaluator Dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Proposal-Evaluator Characteristics											
Distance	3.44***	3.71***	3.00**	3.14**							
	(1.22)	(1.20)	(1.35)	(1.35)							
Novelty x Distance			4.06	5.02							
			(6.43)	(5.94)							
Coauthor					.17		.05				
					(.738)		(.759)				
Same Hospital						.45	.45				
						(.388)	(.390)				
Adj-R^2	.22	.48	.22	.48	.30	.30	.40	.30	.20	.17	.04

Notes. OLS estimates; *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively; bootstrapped standard errors are reported; number of observations = 2,130 research proposal-evaluator pairs.

Table 7 Interpretation of the Novelty Discount

Dependent Variable:	*Score*								
	1	2	3	4	5	6	7	8	9
	Baseline Novelty Model	*Novelty vs. Distance*	*Distance*	*How "Close" Evaluators Treat Novelty*				*How "Close" Proposals are Treated by Most Accomplished Evaluators*	
Research Proposal Characteristics									
Novelty[(1)]	-2.21***	-2.35***							
	(.49)	(.49)							
MeSH Term Count	.081***	.068***							
	(.01)	(.01)							
Proposal Dummies			Y	Y	Y	Y	Y	Y	Y
Evaluator Characteristics									
Evaluator Dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y
Proposal-Evaluator Characteristics									
Distance		3.44***	3.71***	3.14**	3.62***	3.71***	3.76**	4.22*	3.83**
		(1.31)	(1.21)	(1.40)	(1.22)	(1.21)	(1.48)	(2.33)	(1.57)
Novelty x Distance				5.02				-2.36	-1.44
				(5.72)				(8.93)	(6.31)
Novelty x T1D Researcher					-1.13				
					(.82)				
Novelty x No. Endocrine Pubs						-.02			
						(.05)			
I{Distance>.72}							-.05		
							(.46)		
I{Distance>.72} x Novelty							.30		
							(3.06)		
Senior x Distance								-2.28	
								(2.43)	
Senior x Novelty x Distance								(17.0)	
								(10.4)	
Pubs x Distance									-.007*
									(.004)
Pubs x Novelty									-.057**
									(.024)
Pubs x Novelty x Distance									.067**
									(.027)
Adj-R^2	.59	.30	.36	.23	.22	.24	.26	.36	.36

Notes. OLS estimates; *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively; bootstrapped standard errors are reported; number of observations = 2,130 research proposal-evaluator pairs.

Table 8 Novelty and Disagreement or Variance of Evaluations

Dependent Variable:	*Score*		*Range(Scores)*	*Std.Dev.(Scores)*
	1	2	3	4
Unit of Analysis:	Each Evaluation of Each Proposal		Each Proposal	Each Proposal
Parametric Mean				
Novelty	-2.19***		-.51	-.51
	(.47)		(1.14)	(1.14)
MeSH Term Count	.08***		.00	.00
	(.01)		(.02)	(.02)
Research Proposal FE		Y		
Evaluator FE	Y	Y		
Parametric Variance				
Novelty	.06	-.11	n/a	n/a
	(.32)	(.36)		
Constant	1.57***	1.13	n/a	n/a
	(.06)	(.07)		
Number of Observations	2130	2130	150	150

Notes. Models (1) and (2) estimated with maximum likelihood; model (3) estimated with OLS; *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively; bootstrapped standard errors are reported.

Table 9 Subjective Perceived Feasibility of Novel Proposals

Dependent Variable:	*Score*			
	1	2	3	4
	Baseline Preferred	Feasibility	Feasibility Dummies	Average Feasibility Score
Research Proposal Characteristics				
Novelty[(1)]	-2.21***	-1.70***	-1.69***	-1.60***
	(.49)	(.37)	(.37)	(.43)
MeSH Term Count	.081***	.031***	.031***	.020**
	(.010)	(.007)	(.007)	(.009)
Feasibility Score		.65***		
		(.02)		
Feasibilility Score Dummies			Y	
Average Feasibility Score				.78***
				(.03)
Evaluator Characteristics				
Evaluator FE	Y	Y	Y	Y
Adj-R^2	.22	.56	.57	.41

Notes. OLS estimates; *, **, and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively; bootstrapped standard errors are reported; number of observations = 2,130 research proposal-evaluator pairs.