Expertise vs. Bias in Evaluation: Evidence from the NIH *

Danielle Li
Northwestern University†

First draft: August 1, 2011
This draft: May 20, 2013

Abstract

Evaluators with expertise in a particular field may have an informational advantage in separating good projects from bad. At the same time, they may also have personal preferences that impact their objectivity. This paper develops a framework for separately identifying the effects of expertise and bias on decision making and applies it in the context of peer review at the National Institutes of Health (NIH). I find evidence that evaluators are both biased in favor of and better informed about projects in their own area. On net, the benefits of expertise tend to dominate the costs of bias; limiting the influence of personal preferences may also reduce the quality of funding decisions.

*I am very grateful to Pierre Azoulay, Michael Greenstone, and especially David Autor for detailed feedback on this project. I also thank Jason Abaluck, Leila Agha, Josh Angrist, Manuel Bagues, David Berger, David Chan, Esther Duflo, Brigham Frandsen, Alex Frankel, Richard Freeman, Bob Gibbons, Nathan Hendren, Ben Jones, Niko Matouschek, Xiao Yu May Wang, Ziad Obermeyer, Amanda Pallais, Chris Palmer, Michael Powell, Heidi Williams, and numerous seminar participants for helpful comments and suggestions. I am grateful to George Chacko, Raviv Murciano-Goroff, Joshua Reyes, and James Vines for assistance with data. All errors are my own.

†danielle-li@kellogg.northwestern.edu

1 Introduction

When decisions are complex and technical, it is natural to turn to experts for advice. Lawmakers, corporate boards, venture capital groups, and regulatory bodies, for instance, all seek input from industry insiders. But how much should one trust expert advice? While experts may have valuable insights about a project's potential, they may also have personal preferences that compromise their objectivity. As a result, attempts to limit bias by seeking only impartial evaluators may come at the direct cost of reducing expertise.

Understanding how experts shape investment decisions is particularly crucial in the innovative sector, where the payoffs to specific investments are notoriously uncertain (Arrow, 1962). Because ideas are so difficult to assess and because their value may take years or even decades to be realized, there is both greater value placed on expertise and greater scope for obfuscation.[1]

This paper develops a framework for separately identifying the effects of expertise and bias on decision making and provides, to my knowledge, the first empirical estimate of the efficiency trade-off between the two. I do so in a context important for medical innovation: grant funding at the National Institutes of Health (NIH). With an annual budget of $30 billion, the NIH is the world's largest funder of biomedical research, spending half as much on R&D as the entire US pharmaceutical industry and playing a role in the development of most FDA-approved drugs.[2]

To receive funding for a project, individual scientists submit grant applications to the NIH, which are then evaluated by committees of other active scientists. Because the majority of NIH funds are allocated in this way, peer review is the key institution responsible for consolidating thousands of investigator-initiated submissions into a concrete, publicly funded research agenda.

The success of this system, then, depends on the ability of reviewers to identify and fund the most promising ideas in their areas of speciality. Yet advice in this setting may be distorted by the fact that reviewers have self-selected into their preferred research areas and have made substantial investments—from pursuing graduate studies to establishing labs—toward acquiring their expertise. Reviewers may, for example, favor lower-quality applicants whose work is related to their own over higher-quality applicants whose work is unrelated. In a guide aimed at scientists describing the NIH grant review process, one reviewer writes: "*If I'm sitting in an NIH study section, and I believe the real area of current interest in the field is neurotoxicology [the reviewer's own speciality], I'm thinking if you're not doing neurotoxicology, you're not doing interesting science.*"[3] Alternatively, reviewers may be biased against applicants whose work is related if they perceive them as competitors.

[1] See, for example, Aghion and Tirole (1994) and David, Mowery, and Steinmueller (1992).

[2] In 2006, pharmaceutical companies spent close to 50 billion dollars on R&D. CBO "Research and Development in the Pharmaceuticals Industry" (2006). Over two-thirds of FDA priority review drugs cite NIH-funded research. See Sampat and Lichtenberg (2011).

[3] See http://www.clemson.edu/caah/research/images/What_Do_Grant_Reviewers_Really_Want_Anyway.pdf.

I formalize this intuition with a model of misaligned incentives with strategic communication derived from Crawford and Sobel (1982) and apply it to peer review. In this model, a reviewer evaluates a grant application which may or may not be related to his own work. The reviewer has better information about the quality of related applications—ones in his own area—but is also biased in that he derives a personal payoff from funding that application, independent of its quality. I then show that the reviewer's expertise and bias will both affect equilibrium funding decisions and that these effects can be empirically distinguished. In particular, expertise means that high-quality applicants should benefit from being evaluated by related reviewers who can more accurately observe their quality, while low-quality applicants should be hurt for the same reason; bias means that reviewers should be systematically more (or less) likely to fund applicants in their own area regardless of quality.

Peer review at the NIH presents a rare opportunity to obtain empirical traction on these issues. To do so, this paper assembles a new, comprehensive dataset linking almost 100,000 NIH grant applications to the committees in which they were evaluated. For each application, I observe its score, funding status, the identity of the applicant, and the names of all reviewers who are present. Using publication databases such as Web of Science, I can also build detailed publication and grant histories for each applicant. These data are used to construct two key variables: application quality and applicant-reviewer relatedness. Together, these ingredients allow me to 1) estimate the causal effect of being evaluated by related reviewers on an applicant's likelihood of receiving funding; 2) decompose this effect into a portion that comes from expertise and one that comes from bias; and 3) assess the efficiency consequences of potentially biased reviewers in terms of the quality of research that the NIH supports, as measured by citations and publications.

The first of the variables I construct is a measure of the quality for aIl grant applications. The quality of funded grants can be measured by tracking the stream of citations that accrue to the publications that they subsequently support. A key challenge in this setting, however, is finding a way to assess the quality of unfunded applications as well. Linking unfunded applications to publications is possible because, for the large NIH grants this paper focuses on, standards for preliminary results are so high that researchers often submit applications based on nearly completed research. As a result, it is common to publish the work proposed in an application even if the application itself goes unfunded. To find these related publications, I use a text-matching approach that links grant application titles with the titles and abstracts of semantically related publications by the same applicant. I further restrict my analysis of application quality to articles published soon enough after grant review to not be directly affected by any grant funds.

Next, I define an applicant and a reviewer to be related if the reviewer has cited the applicant in the five years prior to the committee meeting. This measure of relatedness is, however, likely to be correlated with quality. This poses a challenge for identification because any measure of quality will necessarily contain some error; if my measure of relatedness is correlated with quality, then

this measurement error in quality will bias estimates of the effect of relatedness.

In order to identify the causal effect of relatedness on funding decisions, I exploit the organizational structure of NIH review committees to generate variation in relatedness that is uncorrelated with quality. Specifically, the committees I study consist of two types of members, "permanent" and "temporary."[4] I show that while permanent and temporary reviewers have similar qualifications as scientists, permanent members have more influence within the committee. My analysis then compares outcomes for different applicants reviewed in the same meeting who are cited by the same *total* number of committee members but by different numbers of *permanent* reviewers. This identifies the effect of being related to a more *influential* set of reviewers under the assumption that the quality of an applicant is not correlated with the composition of reviewers who cite her, conditional on the total number.

I also show that my results do not rely on the distinction between permanent and temporary reviewers by using applicant fixed effects to compare outcomes for the same applicant across meetings in which she is related to different numbers of reviewers. This alternative specification identifies the effect of being related to an *additional* reviewer under the assumption that the time-variant unobserved quality of an application is not correlated with relatedness.

My paper has three primary findings. First, I show that reviewers are more likely to fund applicants in their own area: holding total relatedness constant, every additional permanent member whose work is related to an applicant increases her chances of being funded by 3.1 percent, the equivalent of a one-quarter standard deviation increase in application quality. Second, I decompose the effect of expertise and bias. I find that while reviewers are biased in favor of applicants whose work is related to their own, they are also better able to identify high-quality research among these applicants; the correlation between scores and funding outcomes is over 50 percent higher for applicants who work in the same area as at least one permanent member (conditional on the same total number of reviewers with whom an applicant shares a research interest). Finally, on net, I show that the gains associated reviewer expertise tend to dominate the losses associated with bias. Treating applicants related to influential members as if they were unrelated—thereby reducing both bias and expertise—would reduce the quality of the NIH-supported research portfolio by two to three percent, as measured by future citations and publications. In addition to quantifying the role that bias and information play on average, I also document substantial and persistent variation in how well grant review committees perform; eliminating the effects of relatedness appears to hurt performance across the entire distribution of committee performance.

My empirical work is relevant for innovation policy. A key debate in this literature focuses on what mechanisms are most effective for encouraging innovation: while patents may distort subsequent access to and use of new knowledge, a concern with research grants and other R&D

[4] "Permanent" members are not actually permanent; they serve four-year terms. See Section 5.3 for a discussion of permanent versus temporary reviewers.

subsidies, however, is that the public sector may make poor decisions about which projects to fund. Currently, there is little empirical evidence on how—and how successfully—governments make these research investments.[5] Different organizations, moreover, allocate funds in different ways; NIH's reliance on peer review of individual grants, for example, stands in contrast to major European funding agencies, which often support large groups of scientists working in predetermined priority areas. Understanding the strengths and weaknesses of these models is important because, by making investments in specific people, labs, and ideas, funding not only affects near-term scientific output but also shapes the allocation of future research attention and resources.

The trade-off between expertise and bias is also important in settings outside of innovation: regulators and lobbyists routinely move between the public and private sectors, prompting concerns over conflicts of interest; the past experiences of central bankers may impact both their expertise and preferences; social ties may lead to better job referrals but may also encourage nepotism; academics are well informed about the quality of their colleagues but may show bias when making promotion and editorial decisions.[6] In these settings, it is often difficult to attribute differences in the treatment of connected individuals to either better information or bias because it is difficult to observe the quality of decisions that are not made or people who are not promoted. This paper contributes by studying these issues in a context where this challenge can be more readily overcome.

The remainder of this paper proceeds as follows. In the next section, I discuss the details of NIH grant review. I discuss my conceptual and statistical frameworks in Sections 3 and 4, respectively. Section 5 explains and provides support for my empirical strategy. Section 6 presents my main results and Section 7 discusses efficiency. The final section concludes.

2 Institutional Context

Each year, thousands of scientists travel to Bethesda, Maryland where they read approximately 20,000 grant applications and allocate over 20 billion dollars in federal grant funding. During this process, more than 80 percent of applicants are rejected even though, for the vast majority of biomedical researchers, winning and renewing NIH grants is crucial for becoming an independent investigator, maintaining a lab, earning tenure, and paying salaries (Stephan, 2012; Jones, 2010).

The largest and most established of these grant mechanisms is the R01, a project-based, renewable research grant that constitutes half of all NIH grant spending and is the primary funding source for most academic biomedical labs in the United States. There are currently 27,000 out-

[5] See Acemoglu, 2008; Kremer and Williams, 2010; Grilliches, 1992; and Cockburn and Henderson, 2000 for surveys. One recent exception is Hegde (2009), which considers the political economy of NIH congressional appropriations.

[6] Blanes i Vidal, Draca, and Fons-Rosen (2012) document evidence that lobbyist wages are tied to access; Hansen, McMahon, and Velasco Rivera (2012) study preferences and private information at the Bank of Englands Monetary Policy Committee; Topa (2012) surveys the role that networks play in the labor market; Bagues and Zinovyeva (2012) and Brogaard, Engleberg, and Parsons (2012) study the role of connections in academic promotion and publishing, respectively.

standing awards, with 4,000 new projects approved each year. The average size of each award is 1.7 million dollars spread over three to five years.

To apply for an R01, the primary investigator submits an application, which is then assigned to a review committee (called a "study section") for scoring and to an Institute or Center (IC) for funding. The bulk of these applications are reviewed in one of about 180 "chartered" study sections, which are standing review committees organized around a particular theme, for instance "Cellular Signaling and Regulatory Systems" or "Clinical Neuroplasticity and Neurotransmitters."[7] These committees meet three times a year in accordance with NIH's funding cycles and, during each meeting, review between 40 to 80 applications. My analysis focuses on these committees.

Study sections are typically composed of 15 to 30 "permanent" members who serve four-year terms and 10 to 20 "temporary" reviewers who are called in as needed. The division of committees into permanent and temporary members plays an important role in my identification strategy, which I discuss in greater detail in Section 5.3. Within a study section, an application is typically assigned up to three reviewers, mostly permanent, who provide an initial assessment of its merit.

The process of assigning applications to study sections and reviewers is nonrandom. In practice, applicants are usually aware of the identities of most permanent study-section members, suggest a preferred study section, and usually get their first choice (subject to the constraint that, for most applicants, there are only one or two study sections that are scientifically appropriate). Study-section officers, meanwhile, assign applications to initial reviewers on the basis of intellectual fit. I will discuss the implications of this nonrandom selection on my identification strategy in Section 5.3.

Once an application has been assigned, initial reviewers read and score the application on the basis of five review criteria: *Significance* (does the proposed research address an important problem and would it constitute an advance over current knowledge?), *Innovation* (are either the concepts, aims, or methods novel?), *Approach* (is the research feasible and well thought out?), *Investigator* (is the applicant well-qualified?), and *Environment* (can the applicant's institution support the proposed work?). Based on these scores, weak applications (about one-third to one-half) are "triaged" or "unscored," meaning that they are rejected without further discussion. The remaining applications are then discussed in the full study-section meeting. During these deliberations, an application's initial reviewers first present their opinions, and then all reviewers discuss the application according to the same five review criteria listed above.

Following these discussions, all study-section members anonymously vote on the application, assigning it a "priority score," which, during my sample period, ranged from 1.0 for the best application to 5.0 for the worst, in increments of 0.1. The final score is the average of all member

[7]The NIH restructured chartered study sections during my sample period and my data include observations from 250 distinct chartered study sections. These changes do not affect my estimation because I use within-meeting variation only.

scores. This priority score is then converted into a percentile from 1 to 99, where a percentile reflects the percentage of applications from the same study section and reviewed in the same year that received a better priority score. According to this system, a lower score is better, but, for ease of exposition and intuition, I report inverted percentiles (100 minus the official NIH percentile, e.g., the percent of applications that are *worse*), so that higher percentiles are better. In my data, I observe an application's final score (records of scores by individual reviewers and initial scores are destroyed after the meeting).

Once a study section has scored an application, the Institute to which it was assigned determines funding. Given the score, this determination is largely mechanical: an IC lines up all applications it is assigned and funds them in order of score until its budget has been exhausted. When doing this, the IC only considers the score: NIH will choose to fund one large grant instead of two or three smaller grants as long as the larger grant has a better score, even if it is only marginally better. The worst percentile score that is funded is known as that IC's payline for the year. In very few cases (less than four percent), applications are not funded in order of score; this typically happens if new results emerge to strengthen the application. Scores are never made public.[8]

Funded applications may be renewed every three to five years, in which case they go through the same process described above. Unfunded applications may be resubmitted, during the period of my data, up to two more times. My analysis includes all applications that are reviewed in each of my observed study-section meetings, including first-time applications, resubmitted applications, and renewal applications.

3 How Do Relationships Impact Funding Decisions? Conceptual Framework

The following model of decision making demonstrates how bias and expertise can affect grant funding through strategic communication. In this model, a grant application is evaluated by a committee that either approves or rejects the application based on advice from a reviewer who has personal preferences that are potentially different from those of the committee. I characterize the equilibrium of the model and use it to motivate my empirical strategy, discussed in Section 4.

A grant application has some true quality Q^* and, if approved, the committee receives a payoff of Q^*. If the application is rejected, the committee receives its outside option U, where $U > E(Q^*)$. Applications are either related ($R = 1$) to the reviewer or not ($R = 0$). Neither the committee nor the reviewer observes Q^*, but the reviewer observes a signal Q_R about Q^* where I think of Q_1 as giving a more precise signal than Q_0.[9] After observing the signal, the reviewer

[8] For more details on this process, see Gerin (2006).

[9] For simplicity, I assume that the signals Q_R are real numbers with continuous unconditional distributions such that $E(Q^*|Q_R)$ is increasing in Q_R.

sends a message to the committee about the application's quality and the committee then decides whether to approve the grant. When determining what message to send, a reviewer considers his payoffs: for an unrelated application, this is identical to that of the committee, but for a related application, the reviewer now receives $Q^* + B$ if the application is funded and U otherwise. The term B represents his bias. The timing is as follows:

1. The application's true quality Q^* is realized.
2. The application's type (related or unrelated) is determined and is observed by the committee.
3. The reviewer observes the signal Q_R.
4. The reviewer sends a costless and unverifiable message M to the committee from some message space $\mathbf{M}$.
5. The committee, observing M, makes a decision $D \in \{0, 1\}$ of whether to fund the grant.
6. True quality is revealed and the reviewer and committee both receive their payoffs.

Proposition 1 describes the perfect Bayesian equilibria of this game. There are always uninformative equilibria in which messages are meaningless and the grant is never funded. This proposition therefore focuses on informative equilibria, i.e. those in which the committee's decision depends on the reviewer's message. An informative equilibrium is unique if all other informative equilibria are payoff-equivalent for the parties.

Proposition 1 *The equilibria of the game is summarized by the following two cases:*
CASE 1*: $R = 0$. There exists a unique informative equilibrium in which*

1. *The reviewer reports a message Y if $E(Q^*|Q_0) > U$ and N otherwise.*[10]
2. *The committee funds the grant if and only if the message is Y.*

CASE 2*: $R = 1$. There exists a level of bias $B^* > 0$ such that for bias $B \leq B^*$ there is a unique informative equilibrium such that*

1. *The reviewer reports a message Y if $E(Q^*|Q_1) > U - B$ and N otherwise.*
2. *The committee funds the grant if and only if the message is Y.*

When $B > B^*$, only uninformative equilibria exist and the grant is never funded.

Proof See Appendix 3.

Here, funding decisions can be distorted because the committee is unable to distinguish situations when an application should be funded (e.g., when $E(Q^*|Q_1) > U$) from ones in which it

[10] I assume there are at least two elements in the message space $\mathbf{M}$ which, without loss, I call Y and N.

should not (e.g., when $U > E(Q^*|Q_1) > U - B$). When deciding whether to trust a reviewer's assessment of a related application, the committee must decide whether enough information about its true quality is communicated in spite of the reviewer's bias. When bias is small, committees listen to biased reviewers because they value expertise. As bias increases, however, more undeserving applications are funded until, finally, when bias becomes too large, committees stop taking advice about related applications and the informative equilibrium breaks down.

This model makes the simplifying assumption that committees can observe whether an application is related to a reviewer. In Appendix 3, I allow the application's relatedness to be unknown to the committee and show that all the same qualitative features of this model continue to hold.

The committee's decision rule in the informative equilibria of this model is given by

$$\begin{aligned} D &= \underbrace{\mathbb{I}(E(Q^*|Q_0) > U)}_{\text{baseline for unrelated}} + \underbrace{[\mathbb{I}(U > E(Q^*|Q_1) > U - B)]}_{\text{bias for related (+)}} R \\ &\quad + \underbrace{[\mathbb{I}(E(Q^*|Q_1) > U) - \mathbb{I}(E(Q^*|Q_0) > U)]}_{\text{additional information for related (+/-)}} R. \end{aligned} \tag{1}$$

The first term of Equation (1) indicates that committees listen to advice about unrelated applications. The remaining terms show that a reviewer can influence a committee's decision about related applications in two ways. Bias increases the probability that a low-quality related application is funded. Bias, then, lowers the quality of peer review, as measured by the expected payoff to the committee. Meanwhile, expertise increases the likelihood that a high-quality related application is funded, while decreasing this probability for low-quality related applications. This information effect improves the quality of grant review. The net effect of relatedness on the quality of decisions is thus ambiguous.

Many popular critiques of NIH peer review assume that differences in funding likelihood among applicants with the same quality must be due to bias (see Ginther et al., (2011)). Equation (1) shows, however, that this need not be the case. In particular, the difference in the expected likelihood of funding between related and unrelated applications of the same quality is given by

$$\begin{aligned} E[D|Q^*, R = 1] - E[D|Q^*, R = 0] &= \Pr(U > E(Q^*|Q_1) > U - B) \\ &\quad + \Pr(E(Q^*|Q_1) > U) - \Pr(E(Q^*|Q_0) > U). \end{aligned}$$

This expression will be nonzero if reviewers are biased ($B \neq 0$). Funding differentials, however, can arise even in the absence of bias; because reviewers can more confidently attest to the quality of related applications, committees may have a stronger posterior belief about the quality of related relative to unrelated applications because they update more following a favorable review. Distinguishing between bias and information driven explanations is important because they have different implications for whether relatedness enhances the quality of peer review.

4 How Do Relationships Impact Funding Decisions? Statistical Framework

This section describes the empirically testable predictions of my model. In particular, the committee decision rule described by Equation (1) can be thought of as a data generating process for the funding decisions I observe in my data. To identify the effects of expertise and bias, I make the following simplifying assumptions: for $R = 0, 1$, the reviewer's signal Q_R can be written as $Q_R = Q^* + \varepsilon_R$ where $\varepsilon_R \sim U[-a_R, a_R]$ and $E(Q^*|Q_R)$ can be approximated by λQ_R for some constant λ_R. Given this, an application's conditional likelihood of funding can be expressed as[11]

$$\begin{aligned}
E[D|Q^*, R] &= \Pr(\lambda_0(Q^* + \varepsilon_0) > U) + \Pr(U > \lambda_1(Q^* + \varepsilon_1) > U - B)R \\
&\quad + \left[\Pr(\lambda_1(Q^* + \varepsilon_1) > U) - \Pr(\lambda_0(Q^* + \varepsilon_0) > U)\right] R \\
&= \frac{a_0 - U/\lambda_0 + Q^*}{2a_0} + \frac{B}{2a_1\lambda_1}R + \left[\frac{a_1 - U/\lambda_1 + Q^*}{2a_1} - \frac{a_0 - U/\lambda_0 + Q^*}{2a_0}\right] R \\
&= \frac{1}{2} + \underbrace{\frac{1}{2a_0}}_{\text{Quality corr.}} Q^* + \underbrace{\frac{B}{2a_1\lambda_1}}_{\text{Bias term}} R + \underbrace{\left[\frac{1}{2a_1} - \frac{1}{2a_0}\right]}_{\text{Add. corr. for related}} RQ^* \\
&\quad - \frac{U}{2a_0\lambda_0} + \left[\frac{1}{2a_0\lambda_0} - \frac{1}{2a_1\lambda_1}\right] RU. \qquad (2)
\end{aligned}$$

This expression shows how bias and expertise can be separately identified in my data. In particular, consider the regression analogue of Equation (2):

$$D = \alpha_0 + \alpha_1 Q^* + \alpha_2 R + \alpha_3 RQ^* + \alpha_4 U + \alpha_5 RU + X\beta + \epsilon, \qquad (3)$$

where X includes other observable I can condition on. Here, the coefficient on relatedness R tests for bias: it is nonzero if and only if $B \neq 0$. Second, the coefficient on RQ^* tests for expertise. To see this, notice that α_1 captures, for unrelated applicants, how responsive funding decisions are to increases in quality. In the model, this is determined by the precision of the reviewer's signal of quality for unrelated applications. The coefficient on RQ^*, meanwhile, captures the additional correlation between quality and funding for related applicants. A high coefficient on RQ means that a committee is more sensitive to increases in the quality of related applicants than to increases in the quality of unrelated applicants. In the model, this is determined by the difference in the precision of signals for related and unrelated applications.

The intuition for how I separately identify bias and expertise is simple: if I find that related applications are more (or less) likely to be funded regardless of their quality, then this is a level

[11] The limited support of the error distribution means that if an application is extremely high (low) quality, the committee will choose to approve (reject) it regardless of what the reviewer says. As such, Equation (2) is valid for candidates with quality such that $Q^* + \varepsilon_R$ cannot be greater than U or less than U for all possible ε_R.

effect of relatedness that I attribute to bias. If I find that quality is more predictive of funding for related applications than for unrelated applications, then this is an interaction effect of relatedness and quality that I attribute to expertise. In practice, both these effects may exist.[12]

Finally, the terms U and RU control for funding selectivity; for high cutoffs U, the correlation between funding and quality will be low even in the absence of bias or differential information because the marginal unfunded application is already very high-quality. The RU term, meanwhile, ensures that relationships are not credited for changing the correlation between funding and quality simply by lowering the threshold at which grants are funded. The exact form of Equation (2) depends on distributional assumptions, but my results are robust to allowing for nonlinear effects of relatedness and quality. These results are discussed in Appendix C and Appendix Table E.

Equation (2) says that, as long as Q^* is perfectly observed, exogenous variation in relatedness is not needed to identify the presence of bias. This is because exogenous variation in relatedness is necessarily only when aspects of an application's quality are potentially omitted; if quality were observed, one could directly control for any correlation between relatedness and quality.

In practice, however, I do not observe an application's true quality Q^*. Instead, I observe a noisy signal $Q = Q^* + v$. Thus, instead of estimating Equation (3), I estimate

$$D = a_0 + a_1 Q + a_2 R + a_3 RQ + a_4 U + a_5 RU + Xb + e. \tag{4}$$

Measurement error in quality can potentially pose problems for identification. Proposition 2 describes the conditions that must be met in order to consistently estimate bias from observed data.

Proposition 2 *Given observed quality $Q = Q^* + v$, the bias parameter α_2 in Equation (3) is consistently estimated by a_2 in Equation (4) when the following conditions are met:*

1. $Cov(R, Q^*|U, RU, X) = 0$ *and* $Cov(R^2, Q^*|U, RU, X) = 0$,
2. $E(v|U, RU, X) = 0$,
3. $Cov(v, R|U, RU, X) = 0$.

Proof : See Appendix B.

Condition 1 requires that my measure of relatedness, R, be uncorrelated, conditional on observables, with true application quality. If this were not the case, any mismeasurement in true quality Q^* would bias estimates of α_2 through the correlation between Q^* and R. Thus, in my study, exogenous variation in relatedness is required only to deal with measurement error.

[12] These predictions hold when reviewers and committees are in an informative equilibrium. If the equilibrium were not informative, then advice from related reviewers would not be taken; I would find no effect of bias and a lower correlation between funding and quality for related applications. My results are not consistent with a non-informative equilibrium.

Condition 2 requires that measurement error be conditionally mean zero. This means that, after controlling for observable traits of the application or applicant, my quality measure cannot be systematically different from what committees themselves are trying to maximize. Otherwise, I may mistakenly conclude that committees are biased when they are actually prioritizing something I do not observe but which is not mean zero different from my quality measure.

Finally, Condition 3 requires that the extent of measurement error not depend, conditional on observables, on whether an applicant is related to a reviewer. This may not be satisfied if related applicants are more likely to be funded and funding itself affects my measure of quality. Suppose, for instance, that two scientists apply for a grant using proposals that are of the same quality. One scientist is related to a reviewer and is funded because of bias. The funding, however, allows her to publish more articles, meaning that my measure of quality—future citations—may mistakenly conclude that her proposal was better than the other scientist's to begin with. Mismeasurement of ex ante grant quality makes it *less* likely that I would find an effect of bias.

Another reason why Condition 3 may not be satisfied is given by the Matthew Effect, a sociological idea wherein credit and citations accrue to established investigators simply because they are established (see Merton, 1986 and Azoulay, Stuart, and Wang, 2011). Were this the case, more related applicants would receive more citations regardless of the true quality of their work, meaning that measurement error v would be correlated with relatedness. This would lead me to underestimate bias; related applicants may receive higher scores simply for being established, but this bias would look justified by a citation-based measure of quality (which reflects bias in the scientific community at large).

Together, Conditions 1-3 are weaker than assuming classical measurement error, but they place restrictions on how I can measure quality and relatedness. In particular, to satisfy these conditions, I construct quality and relatedness measures to meet the following standards:

1. Quality $Q = Q^* + v$ must be consistently measured for funded and unfunded grants and not be directly affected affected by whether an applicant receives funding. As described above, failure to measure quality properly may lead to violations of Conditions 2 and 3.
2. My measure of relatedness must be independent of quality, conditional on variables I can observe. This is simply Condition 1.

In the next section, I describe my data, explain how I construct my quality and relatedness measures, and provide evidence that these measures satisfy the identifying conditions in this section.

5 Data and Empirical Strategy

5.1 Data

I have constructed a new dataset describing grant applications, review-committee members, and their relationships for almost 100,000 applications evaluated in more than 2,000 meetings of 250 chartered study sections. My analytic file combines data from three sources: NIH administrative data for the universe of R01 grant applications, attendance rosters for NIH peer-review meetings, and publication databases for life-sciences research. Figure 1 summarizes how these data sources fit together and how my variables are constructed from them.

I begin with two primary sources: the NIH IMPAC II database, which contains administrative data on grant applications, and a series of study section attendance rosters obtained from NIH's main peer-review body, the Center for Scientific Review. The application file contains information on an applicant's full name and degrees, the title of the grant project, the study-section meeting to which it was assigned for evaluation, the score given by the study section, and the funding status of the application. The attendance roster lists the full names of all reviewers who were present at a study-section meeting and whether a reviewer served as a temporary member or a permanent member. These two files can be linked using meeting-level identifiers available for each grant application. Thus, for my sample grant applicants, I observe the identity of the grant applicant, the identity of all committee members, and the action undertaken by the committee.

Next, I construct detailed measures of applicant demographics, grant history, and prior publications. Using an applicant's first and last name, I construct probabilistic measures of gender and ethnicity (Hispanic, East Asian, or South Asian).[13] I also search my database of grant applications to build a record of an applicant's grant history as measured by the number of new and renewal grants an applicant has applied for in the past and the number he has received. This includes data on all NIH grant mechanisms, including non-R01 grants, such as post-doctoral fellowships and career training grants. To obtain measures of an applicant's publication history, I use data from Thomson-Reuters Web of Science (WoS) and the National Library of Medicine's PubMed database. From these, I construct information on the number of research articles an applicant has published in the five years prior to submitting her application, her role in those publications (in the life sciences, this is discernable from the author position), and the impact of those publications as measured by citations. In addition to observing total citations, I can also identify a publication as "high impact" by comparing the number of citations it receives with the number of citations received by other life-science articles published in the same year.

My final sample consists of 93,558 R01 applications from 36,785 distinct investigators over the period 1992-2005. Of these applications, approximately 25 percent are funded and 20 percent are

[13] For more details on this approach, see Kerr (2008). Because Black or African American names are typically more difficult to distinguish, I do not include a separate control for this group.

from new investigators, those who have not received an R01 in the past. This sample is derived from the set of grant applications that I can successfully match to meetings of study sections for which I have attendance records, which is about half of all R01 grants reviewed in chartered study sections. Table 1 shows that my sample appears to be comparable to the universe of R01 applications that are evaluated in chartered study sections.

So far, I have discussed how I measure the prior qualifications of an applicant. As Conditions 1-3 of Section 4 indicate, however, I also need a direct measure of grant quality that is not directly affected by funding and a measure of relatedness that is conditionally independent of my quality measure. I discuss each of these requirements in turn.

5.2 Measuring Quality

A strength of this project lies in my ability to go beyond using past applicant characteristics to assess application quality. Instead, I construct a direct measure of application quality by examining the publications and citations it produces in the future. Condition 2 of my identifying conditions requires that this measure of grant quality not be systematically (e.g. non-mean-zero) different from what committee members are looking for, after conditioning on observables. While there is no way to formally test this assumption (as the objective function of the committee is unobserved), I address this concern in two ways.

First, I attempt to construct quality measures that are informative and flexible. While imperfect, future citations are nonetheless a standard and useful measure of quality that can capture different aspects of grant quality. Total citations reflect how well regarded a project is on average. Policymakers, however, may care mostly about whether a publication is truly pathbreaking. In this case, I can measure whether a publication is a "hit" based on where it falls in the distribution of citations for all other publications in its cohort (same field, same year); that is, a hit publication can be defined as one which is cited at the 99th, 95th, etc., percentiles of similar publications. Further, because my sample begins in 1992 and my citation data go through 2008, I can capture a fairly long run view of quality for almost all publications associated with my sample grants (citations for life-sciences articles typically peak one to two years after publication). This allows me to observe whether a project becomes important in the long run, even if it is not initially highly cited. If reviewers are using their expertise to maximize a welfare function based on long-run impact or the number of hit publications, then my quality measure would capture this.

Second, my analysis will include very detailed controls for many applicant or application characteristics—probabilistic gender and ethnicity, education, institutional affiliation, past publication characteristics—including some that reviewers themselves cannot observe like grant-application history and the number of citations that an applicant's past publications eventually accrue after the date of grant review. This allows my framework to identify bias even if, for instance, committees

take diversity preferences into account when assessing quality.

Finally, even if Condition 2 is violated, my estimate of the welfare implications of seeking advice from related reviewers will still be consistent *with respect to* the number of citations and hit publications produced by the NIH (see Section 7). This in itself is a metric of decision-making quality that is relevant for policy.

The primary challenge of constructing a measure of application quality is finding a way to do this for unfunded applications. I address this challenge by finding publications that are associated with the research described in the preliminary results section of an application. This is possible because the grants I study, the R01, are intended for projects that have demonstrated a substantial likelihood of success, meaning that applicants describe publishable research in their preliminary results. In fact, the bar for preliminary results is so high that the NIH provides a separate grant mechanism, the R21, for pursuing exploratory research leading to an R01 application.

To find these related publications, I look for articles published by a grant's primary investigator around the time of grant review. Because my method of constructing quality needs to be consistent for funded and unfunded grants (see Section 4), I cannot use data on grant acknowledgements because they are, naturally, only available for funded grants. Instead, I compare the titles and abstracts of an applicant's publications with the title of her grant proposal to determine which publications are related. For instance, if I see a grant application titled "Traumatic Brain Injury and Marrow Stromal Cells" reviewed in 2001 and an article by the same investigator entitled "Treatment of Traumatic Brain Injury in Female Rats with Intravenous Administration of Bone Marrow Stromal Cells," published within one year of this grant application, I conclude that this publication and its future citations can be used as a measure of the quality of the application. Text-matching ensures that I can measure quality using the same procedure for all grant applications. These publications (and the citations that accrue to them) form the basis of my quality measure.

The identifying conditions in Section 4 also require that my quality measure not be directly affected by funding.[14] This may occur in two ways: funding can be used to subsidize research on a different topic from the original proposal or it can be used to extend research on the same topic.

Text matching limits the set of publications I use to infer application quality to those on the same topic as the grant. This reduces the possibility of my measure of application quality being contaminated by unrelated research that the grant is used to subsidize. Funding, however, may also increase the number of publications on the same topic as the grant. To address this concern, I

[14]Grant funding, for instance, can be used to start new experiments related to the proposed project or to subsidize research on unrelated projects. Existing evidence on the effect of grant funding on research outcomes suggests that this effect is likely to be small; using a regression-discontinuity approach, Jacob and Lefgren (2011) find that receiving an R01 increases the number of articles a PI publishes in the next five years by 0.85, from a mean of 14.5. This figure includes all publications by a PI, including ones that may be on a different topic from the original application. Jacob and Lefgren's analysis, however, only documents the effect of grant receipt for marginal applicants. The effect of funding on future publications and citations could be larger elsewhere in the distribution, and I take additional precautions to create a measure of quality not affected by funding.

also restrict my quality calculations to articles published in a short time window surrounding grant review. These articles are likely to be based on research that was already completed or underway at the time the grant application was written. To compute the appropriate window, I consider funding, publication, and research lags. A grant application is typically reviewed four months after it is formally submitted, and, on average, another four to six months elapse before it is officially funded.[15] In addition to this funding lag, publication lags in the life sciences typically range from three months to over a year. It is thus highly unlikely that articles published up to one year after grant review would have been directly supported by that grant. Instead, the research underlying these articles are likely what would have been proposed and discussed in the preliminary results section of the grant application. Thus, my measure of an application's quality examines the total number of citations that accrue to publications 1) on the same topic as the grant proposal and which are 2) published within one year of grant review. I also demonstrate that my results are robust to the choice of window.

5.2.1 Assessing Validity of Quality Measures

Figure 2 demonstrates that my matching strategy can identify publications related to unfunded grant applications. In fact, using the measure of quality described above, I find that funded and unfunded grants are almost equally represented among the subset of grant applications that generate many citations. Figure 2 also shows, however, that unfunded grants are more likely to produce few citations. There are two possible explanations for this finding: 1) unfunded applications are of lower quality and should thus be expected to produce fewer citations, or 2) funding directly improves research output, meaning that I fail to measure quality consistently for funded and unfunded grants.

I distinguish between these explanations by using variation in whether grants with the same score are funded. Because budgets vary across ICs, applications from the same meeting with the same score are sometimes funded and sometimes not. If funding has a direct impact on my measure of quality, then I should mistakenly attribute higher quality to funded applications than to unfunded ones with the same score. Figure 3 shows this is not the case. Each dot represents the mean number of citations associated with *funded* applications that receive a particular score, regression-adjusted to account for differences across meetings; the crosses represent the same for *unfunded* applications. The dots do not lie systematically above the crosses, meaning that measured quality for funded grants does not appear to be systematically higher than for unfunded grants with the same score.

The accompanying statistical test is reported in Table 2. I compare measured quality for funded and unfunded grant applications with similar scores from the same meeting. Funding status can vary because pay lines at different ICs differ within the same year. Columns 1 and 2 show that, among the set of scored applications, funded grants tend to be of higher quality, but this effect

[15] See http://grants.nih.gov/grants/grants_process.htm.

disappears once I control for a smooth function of scores. Columns 3 and 4 expand the sample to the full set of applications, with low scores imputed for applications that were considered too low-quality to be scored, and find the same results. Together with Figure 3, this finding mitigates concerns that my measure of quality is directly affected by funding.

Appendix C discusses several robustness tests for my measure of quality. It is possible, for instance, that not receiving a grant may slow down a scientist's research (if, for example, they need to spend time applying for more grants). If this is the case, then a grant can directly impact the research quality of funded vs. non-funded applicants even before any funding dollars are disbursed. To address this concern, I estimate an alternative specification focusing on publications on the same topic that were published one year *prior* to the grant-funding decision; these articles are likely to inform the grant proposal, but their quality cannot be affected by the actual funding decision. In general, my results are robust to other windows; this is unsurprising because I will show in the next section that relatedness to permanent reviewers (conditional on relatedness to total reviewers) is uncorrelated with applicant quality.

Another potential concern with my quality measure is that it does not include later publications, potentially on different topics, that a review committee could anticipate would be supported by the grant. It is common for grant funding to subsidize research on future projects that may not be closely related to the original grant proposal; even though reviewers are instructed to restrict their judgements to the merits of the research proposed in the grant application, it is possible that they may attempt to infer the quality of an applicant's future research pipeline and that related reviewers might have more information about this. Appendix C addresses this concern. To test whether my results are robust to this possibility, I use data on grant acknowledgements to match grants to *all* subsequent publications, not just to the research that is on the same topic or which is published within a year of grant review. Because grant acknowledgment data exist only for funded grants, this specification can only examine whether relatedness impacts the scores that funded applicants receive. I show that results using data on grant acknowledgments are largely similar.

Appendix C also reports another test of the validity of my quality measure. If my results were driven by changes in measured grant quality near the payline, I would find no effect of relatedness on scores for the subset of applications that are either well above or well below the payline. However, in both of these subsamples, I find evidence that being related to a permanent member increases scores and increases the correlation between scores and quality. Because relatedness cannot affect actual funding status in these subsamples, the effect I find cannot be driven by differences in how well quality is measured.

Finally, it is also worth emphasizing that, as discussed in Section 4, overcrediting funded applications relative to unfunded applications would lead me to *underestimate* the extent of bias.

5.3 Measuring Relatedness

Next, I construct a measure of applicant-committee relatedness that is uncorrelated with the quality of an application. This is done by first using citation histories available in Web of Science to determine whether an applicant's work is related to every individual reviewer present at the study-section meeting. Specifically, I define an applicant's work to be related to a reviewer if the reviewer has cited the applicant in the five years prior to the review meeting.

Using citations to measure relatedness has several benefits. First, citations capture a form of relatedness that, as demonstrated by the quote in the Introduction, may strongly influence a reviewer's personal preferences: reviewers may prefer work that they find useful for their own research. Second, citations capture this form of intellectual connection more finely than other measures, such as departmental affiliation, allowing for more-informative variation in relatedness. Third, using data on whether the reviewer cites the applicant (as opposed to the applicant citing the reviewer) reduces concerns that my measures of relatedness can be strategically manipulated by applicants.

One may also consider more-social measures of relatedness, such as coauthorship or being from the same institution. These relationships, however, are often subject to NIH's conflict-of-interest rules; reviewers who are coauthors, advisors, advisees, or colleagues, etc. are prohibited from participating in either deliberations or voting. Intellectual relatedness is a form of relatedness that likely matters for grant review but which is not governed by conflict-of-interest rules.

Table 3 describes the characteristics of the sample study sections. In total, I observe 18,916 unique reviewers. On average, each meeting is attended by 30 reviewers, 17 of whom are permanent and 13 temporary. The average applicant has been cited by two reviewers, one temporary and one permanent. The average permanent and average temporary reviewer both cite four applicants.

The number of reviewers who have cited an applicant is likely to be correlated with applicant quality; better applicants may be more likely to be cited by reviewers and may, independently, submit higher-quality proposals. Using this as a measure of relatedness, then, would violate Condition 1 of Section 4. I instead exploit the structure of chartered NIH study sections to find exogenous variation in reviewer-applicant relatedness. As discussed in Section 2, review committees consist of "permanent" and "temporary" members. My identification strategy examines how the number of permanent members who cite an applicant, call this R^P, affects the committee decision, conditional on the *total* number of reviewers who cite the applicant, R. That is, I compare the outcomes of scientists whose applications are reviewed in the same meeting, who have similar past performance, and who, while related to the same total number of reviewers, differ in the number of influential members they are related to. In order for this strategy to be valid, I need to show that 1) permanent reviewers are indeed more influential within a study section but that 2) permanent and temporary reviewers are otherwise identical, meaning that being related to a permanent or temporary reviewer

is uncorrelated with an applicant's quality.

5.3.1 Assessing Validity of Relatedness Measures

There are many reasons why permanent reviewers have more influence over an applicant's score. Most basically, these reviewers do more work. As discussed in Section 2, reviewers are responsible for providing initial assessments of a grant application before that application is discussed by the full committee. These initial assessments are extremely important for determining a grant application's final score because they 1) determine whether a grant application even merits discussion by the full group and 2) serve as the starting point for discussion. In many study sections, there is also a rule that no one can vote for scores outside of the boundaries set by the initial scores without providing a reason. While I do not have data on who serves as one of an application's three initial reviewers, permanent reviewers are much more likely to serve as an initial reviewer; they are typically assigned eight to ten applications, compared with only one or two for temporary reviewers. In addition, permanent members are required to be in attendance for discussions of all applications; in contrast, temporary members are only expected to be present when their assigned grants are discussed, meaning that they often miss voting on other applications. Finally, permanent members work together in many meetings over the course of their four-year terms; they may thus be more likely to trust, or at least clearly assess, one another's advice, relative to the advice of temporary reviewers with whom they are less familiar.

To test whether permanent members seem to have more influence, I use the fact that I observe almost 5,000 unique reviewers in meetings in which they are permanent and in meetings in which they are temporary. For each of these reviewers, I find the set of applicants with whom they are related and show that a larger proportion of those applicants are funded when the reviewer is permanent rather than temporary. These regressions include controls for applicant characteristics and reviewer fixed effects, meaning that similarly qualified applicants related to the same reviewer are more likely to be funded when that reviewer is permanent than when the reviewer is temporary. These results are presented in Appendix C as well.

In addition to providing evidence that permanent members are more influential, I also need to demonstrate that permanent members and temporary members are comparable as scientists: if this were the case, then being related to a permanent member instead of a temporary member (conditional on total relatedness) should not be indicative of quality. Figure 4 and Table 4 compare the observable characteristics of permanent and temporary members and show that they have similar publication histories and demographics. Figure 4, in particular, shows that the distribution of quality, as measured by previous publications and citations, is essentially identical for permanent and temporary reviewers. The bottom panel of Table 4 suggests why this may not be surprising: permanent and temporary reviewers are often the same people; 35 percent of permanent reviewers

in a given meeting will be temporary reviewers in a future meeting and 40 percent of temporary reviewers in a given meeting will be permanent in the future.

Even if permanent and temporary members are identical as scientists, though, there may still be concerns arising from the fact that reviewers are not randomly assigned to applications. This selection is nonrandom in two ways. First, membership rosters listing the permanent (but not temporary) members associated with a study section are publicly available, meaning that grant applicants know who some of their potential reviewers may be at the time they submit their application. The scope for strategic submissions in the life sciences, however, is small: for most grant applicants, there are only one or two appropriate study sections and, because winning grants is crucial for maintaining one's lab and salary, applicants do not have the luxury of waiting for a more receptive set of reviewers. Second, once an application has been received, the study-section administrator assigns it to initial reviewers on the basis of 1) intellectual match and 2) reviewer availability. If, for instance, not enough permanent reviewers are qualified to evaluate a grant application, then the study section administrator may call in a temporary reviewer. Temporary reviewers may also be called if the permanent members qualified to review the application have already been assigned too many other applications to review.

This process may raise concerns for my identification. For example, suppose that two applicants, one better known and higher quality, submit their applications to a study section that initially consists of one permanent reviewer. The permanent reviewer is more likely to be aware of the work of the better-known applicant and thus there would be no need to call on a related temporary member. To find reviewers for the lesser-known applicant, however, the administrator calls on a temporary reviewer. Both applicants would then be related to one reviewer in total but, in this example, the fact that one applicant is related to a temporary member is actually correlated with potentially unobserved aspects of his quality. This would be a violation of Condition 1 in Section 5.3, which says that relatedness to permanent members, conditional on total relatedness, should not be correlated with quality.

I deal with this and other similar concerns in two ways. First, I provide direct evidence that the characteristics of the applicants and the quality of the applications do not appear to be systematically related to whether an applicant is related to a permanent or temporary member, conditional on total relatedness. Table 5 describes the demographic and past performance characteristics of grant applicants, divided by the total number of reviewers they are related to and by the composition of those reviewers. Most notably, applicants who are related to more reviewers in total tend to be more established: they have more past publications and citations and are less likely to be new investigators. Conditional on total related reviewers, however, there appear to be few differences among applicants: applicants related to one permanent reviewer are virtually identical to those related to one temporary reviewer. Among applicants related to two reviewers, those related to two permanent or one of each look identical. Those related to two temporary

reviewers appear to have slightly fewer past publications, consistent with the concern raised above, but this difference is less than five percent of a standard deviation. Approximately 75 percent of my sample fall into the categories reported in Table 5, but these figures are similar for applicants related to three or more reviewers.

Figure 5 provides more evidence for identifying Condition 1, that the number of permanent members an applicant is related to is not correlated with her quality, conditional on total relatedness. Instead of comparing applicant-level characteristics, however, Figure 5 directly examines the quality of the submitted application itself. The upper-left-hand panel shows the distribution of my measure of application quality for applicants related to exactly one reviewer. The solid line shows the distribution of quality among applicants related to one permanent member, and the dotted line, among those related to one temporary member. These distributions are essentially identical. Similarly, the upper-right-hand panel shows the same, but with quality measured using the number of publications associated with a grant. The bottom two panels of Figure 5 repeat this exercise for applicants who are related to a total of two reviewers. In this case, there are now three possibilities: the applicant is related to two temporary reviewers, two permanent, or one of each. In all of these cases, the distribution of applicant quality is again essentially identical.

Because applicant characteristics are not correlated with the composition of related reviewers, examining the effect of relatedness to permanent members addresses concerns about the Matthew Effect. Because my identification holds scientific esteem as measured by total relationships constant, there is no reason to believe that applicants related to permanent members would be more or less likely to be cited than applicants related to temporary members.

Second, another way to address concerns about the assignment of temporary and permanent members is to show that my results are robust to an alternative specification that does not rely on this distinction. In my main specifications, I control for the *total* number of related reviewers in order to restrict my comparisons to applicants of similar quality. This approach controls for both time-varying and time-invariant unobserved quality of applicants under the assumption that the unobserved quality of an application is not correlated with the composition of permanent and temporary reviewers an applicant is related to.

Another approach is to simply control for applicant fixed effects. In this specification, I compare the funding outcomes of applications from the *same* applicant across meetings in which the applicant is related to different total numbers of reviewers. The downside of this approach is that applicant fixed effects only control for time-invariant unobserved quality. If there are aspects of the quality of an applicant's proposal that are not controlled for with information on past publications and grant histories, then this may bias my results.

This second approach also captures a slightly different causal effect: the effect of being related to an additional reviewer, as opposed to being related to a more influential reviewer. The relative magnitudes of these effects are theoretically ambiguous: if only permanent reviewers have influence,

then the effect of being related to a permanent reviewer (conditional on total relatedness) will be larger than the effect of being related to an additional member (because that additional member may be temporary and thus, in this example, inconsequential). If, on the other hand, temporary members have as much influence as permanent ones, then the composition of related reviewers would not matter, but the number would. I show that both identification strategies yield similar results.

5.4 Estimating Equations

Having defined my relatedness and quality measures, the causal effect of being related to a more influential member can be estimated from the following regression:

$$\text{Decision}_{icmt} = a_0 + a_1 \#\text{ Related Permanent}_{icmt} + a_2 \#\text{ Related}_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt}. \quad (5)$$

Decision_{icmt} is a variable describing the decision (either the score or the funding status) given to applicant i whose proposal is evaluated by committee c in meeting m of year t. $\#\text{ Related Permanent}_{icmt}$ is the number of permanent, and $\#\text{ Related}_{icmt}$ the number of total, reviewers to whom the applicant is related. The covariates X_{icmt} include indicators for sex; whether an applicant's name is Hispanic, East Asian, or South Asian; quartics in an applicant's total number of citations and publications over the past five years; indicators for whether an applicant has an M.D. and/or a Ph.D.; and indicators for the number of past R01 and other NIH grants an applicant has won, as well as indicators for the number to which she has applied. The δ_{cmt} are fixed effects for each committee meeting so that my analysis compares outcomes for grants that are reviewed by the same reviewers in the same meeting. Standard errors are clustered at the committee-fiscal-year level. Given these controls, a_1 captures the effect of being related to an additional permanent reviewer on the likelihood of an applicant being funded.

The overall impact of relatedness on an applicant's likelihood of funding, a_1, however, does not distinguish between bias and expertise for two reasons. First, expertise can have effects on the funding process that are not captured by a_1. This could happen if, for example, reviewers with expertise in a subject area are more likely to fund high-quality research in that area and less likely to fund low-quality research in that area; this may change the identities of the related applicants who are funded without affecting the overall likelihood that a related applicant is funded. Second, relatedness can increase the overall likelihood that an applicant is funded because of expertise as well as bias. This may happen if committees fund applicants whose quality they believe is above a threshold; if the precision of signals for related applicants is higher, then more of these related applicants would be expected to fall above that threshold.

To separately identify the roles of expertise and bias, I introduce information on quality. My theoretical model makes two predictions: first, that applicants related either to more or more-

influential reviewers will be more likely to be funded and, second, that funding decisions for this group will be more responsive to information about quality. To test these predictions, I estimate

$$\begin{aligned} D_{icmt} &= a_0 + a_1 \#\text{ Related Permanent}_{icmt} + a_2 \text{Quality}_{icmt} \times \text{Related to Permanent}_{icmt} \\ &+ a_3 \text{Quality}_{icmt} + a_4 \#\text{ Related}_{icmt} + \mu X_{icmt} + \delta_{cmt} + \varepsilon_{icmt}. \end{aligned} \tag{6}$$

In Equation (6), the coefficient a_1 captures the effect of relatedness on funding that is attributable to bias: Does being related to one additional permanent reviewer, conditional on total relatedness, affect an applicant's likelihood of being funded for reasons unrelated to quality? The coefficient a_2, meanwhile, measures the effect of relatedness that is attributable to expertise: Are higher quality applications more likely to be funded when they are evaluated by related permanent members, conditional on total relatedness? The model in Section 3 also includes terms RU and U to control for the degree of selectivity in a committee (recall that U was the outside option of the unbiased reviewer). In my empirical implementation, I proxy for selectivity using the percentile pay line of the committee. I thus include a level control for pay line (this is absorbed in the meeting fixed effect) as well as the pay line interacted with relatedness. My results are not affected by either the inclusion or exclusion of these variables.[16]

6 Main Results

Table 6 considers the effect of being related to a committee member on funding and scores. The first column reports the raw within-meeting association between the number of permanent related reviewers and an applicant's likelihood of being funded. Without controls, each additional related permanent member is associated with a 3.3 percentage point increase in the probability of funding, off an average of 21.4 percent. This translates into a 15.3 percent increase. Most of this correlation, however, reflects differences in the quality of applications; applicants may be more highly cited by reviewers simply because they are better scientists. Column 2 adds controls for applicant characteristics such as past publication and grant history. This reduces the effect of an additional permanent related reviewer on funding probability to 1.5 percentage points, or 7.1 percent. Even with these controls, relatedness may still be proxying for some unobserved aspect of application quality. Finally, I control for the total number of reviewers by whom each applicant has been cited. Given this, my identification comes from variation in the composition of an applicant's related reviewers; I am comparing outcomes for two scientists with similar observables, who are

[16] In my alternative specification using applicant fixed effects, the analogous regression equation is given by:

$$\begin{aligned} D_{icmt} &= a_0 + a_1 \#\text{ Related}_{icmt} + a_2 \text{Quality}_{icmt} \times \text{Related to a reviewer}_{icmt} \\ &+ a_3 \text{Quality}_{icmt} + \mu X_{icmt} + \delta_i + \varepsilon_{icmt}. \end{aligned}$$

cited by the same total number of reviewers but by different numbers of influential reviewers. In Column 3, I find that an additional permanent related reviewer increases an applicant's chances of being funded by 0.7 percentage points, or 3.1 percent. This is my preferred specification because it isolates variation in relatedness that is plausibly independent of an application's quality.

Columns 4–6 and 7–8 report the effect of relatedness on an applicant's percentile score and likelihood of being scored at all (e.g., rejected early in the process due to low initial evaluations), respectively. In both cases, I find a similar pattern, though an economically smaller effect. Being related to a more influential set of reviewers increases an applicant's score by a quarter of a percentile and her likelihood of being scored by just over half a percent.

Table 7 reports my main regressions, decomposing the effects of bias and expertise. Columns 1, 3, and 5 reproduce the estimates of the level effect of relatedness on funding and scores from Table 6. Column 2 reports estimates of the coefficients from Equation (6) for funding status. The coefficient of 0.0049 on the number of related permanent reviewers says that, due to bias, each additional related permanent reviewer increases the likelihood that an application is funded by 0.5 percentage points, or 2.3 percent.[17] The magnitude of this effect appears to be sizable. To see this, notice that Column 2 also reports the increase in funding likelihood associated with an increase in application quality. The figure of 0.0067 means that a one standard deviation (302 future citations) increase in quality is associated with a 3.02*0.0067=2.02 percentage point increase in an applicant's likelihood of funding. The sensitivity of committees to changes in application quality highlights the magnitude of the bias effects that I find: being related to an additional permanent reviewer, conditional on total relatedness, increases an applicant's chances of being funded by 0.5 percentage points or as much as a one-quarter standard deviation increase in quality.

Column 2 of Table 7 also shows that review committees do a better job of discerning quality when an applicant is related to a permanent member, conditional on the total number of related reviewers. To see this, consider an applicant who is related to one permanent member versus an applicant who is related to one temporary member. A one standard deviation increase in quality for the former applicant increases her likelihood of funding by 0.42+0.67 = 1.09 percentage points compared to 0.67 percentage points for the latter applicant. Thus, despite overall positive bias in favor of related applicants, being related to a permanent member may not be beneficial for all applicants. Because reviewers have more information about the quality of related applicants, related applicants with lower-quality proposals end up receiving lower scores. These results are consistent with the predictions of my model: relationships increase distortion arising from bias but also decrease the variance of the committee's signal of quality.

My results also alleviate a potential concern about this empirical approach, which is that I

[17]This effect is slightly smaller in magnitude than the overall effect of relatedness estimated in Table 6; this is consistent with the explanation that expertise can also increase the overall likelihood of a related applicant being funded, by making committee members more confident about the quality of related applicants relative to unrelated applicants.

may label reviewers biased if they are maximizing some unobserved aspect of application quality that is systematically different from my citation-based measure (this would violate Condition 2, that measurement error in quality is conditionally mean zero). If, for example, reviewers are better at identifying "undervalued" research in their own area, then they may be more likely to fund low-citation related research over higher-citation unrelated research—not because of bias, but because of better information about the true quality of related projects. This behavior, however, would tend to *decrease* the correlation between citations and funding likelihood for related applicants, relative to unrelated applicants. The fact that reviewers appear to be more sensitive to citation-based counts of quality for applicants in their own area, as indicated by Column 2, provides some evidence that citation counts do convey information about quality that reviewers care about.

Columns 4 and 6 consider the effect of relatedness on other outcome measures. Being related to an additional permanent reviewer increases an applicant's score by one-fifth of a percentile or about as much as would be predicted by a one-quarter standard deviation increase in quality.[18] I find a positive, but not statistically significant, effect of relatedness on the correlation between quality and scores and no such information effect on the likelihood of being scored. The magnitudes of these estimates suggest that relatedness and quality have a greater impact on an application's funding status than on its score or likelihood of being scored. This suggests that reviewers both pay more attention to quality for applications at the margin of being funded and are more likely to exercise their bias when this bias might be pivotal for funding.

Table 8 reports results under an alternative identification strategy of applicant fixed effects to control for unobserved application quality. This specification identifies the effect of being related to an additional reviewer, as opposed to the effect of being related to a greater proportion of permanent reviewers. My results, however, are similar: due to bias, an additional related reviewer increases an applicant's chances of being funded by 0.71 percentage points. Interestingly, conditional on applicant fixed effects, the quality of an application is essentially not predictive of funding likelihood except for applicants who are related to reviewers.

7 How Do Relationships Affect the Efficiency of Grant Provision?

My main results show that 1) applicants who are related to study-section members are more likely to be funded, independent of quality, as measured by the number of citations that their research eventually produces; and 2) the correlation between eventual citations and funding likelihood is higher for related applicants, meaning that study-section members are better at discerning the quality of applicants in their own area.

Next, I embed my analysis of the effect of relationships on decisions into a broader analysis

[18] To see this, note that 3.02 (one standard deviation of quality) times 0.24 (the coefficient on quality) is equal to 0.71 percentage points. My estimate of bias is 0.19 percentiles or about one-quarter of the effect of quality.

of their effect on overall efficiency. Assessing the efficiency consequences of related experts requires taking a stand on the social welfare function that the NIH cares about; without one, it would be impossible to assess whether distortions arising from the presence of related experts brings the the grant review process closer to or further from the social optimum.

In this section, I assume that policymakers care about maximizing either the number or impact of publications and citations associated with NIH-funded research. An important disclaimer to note is that an efficiency calculation based on this measure of welfare may not always be appropriate. If, for instance, the NIH cares about promoting investigators from disadvantaged demographic or institutional backgrounds, then a policy that increases total citations may actually move the NIH further from the goal of encouraging diversity. Yet, while citations need not be the only welfare measure that the NIH cares about, there are compelling reasons why policy-makers should take citation-based measures of quality in account when assessing the efficacy of grant review. In addition to being a standard measure of quality used by both economists when studying science and by scientists themselves, citations can also be used to construct, as discussed in Section 5, flexible metrics that capture both high-quality normal science and high-impact work. My citation data, moreover, extend beyond my sample period, allowing me to observe the quality of a publication as judged in the long run. This alleviates concerns that citations may underestimate the importance of groundbreaking projects that may not be well cited in the short run.

Given these caveats, I begin by comparing the actual funding decision for an application to the counterfactual funding decision that would have been obtained in the absence of relationships. Specifically, I define

$$\begin{aligned} \text{Decision}_{icmt}^{\text{Benchmark}} &= \text{Decision}_{icmt} \text{ (actual funding)} \\ \text{Decision}_{icmt}^{\text{No Relationship}} &= \text{Decision}_{icmt} - \widehat{a_1} \#\text{ Related}_{icmt} \\ &- \widehat{a_2} \text{Quality}_{icmt} \times \text{Related to permanent}_{icmt}, \end{aligned}$$

where $\widehat{a_1}$ and $\widehat{a_2}$ are estimated from Equation (6) of Section 5.4.[19] The counterfactual funding decision represents what the committee would have chosen had applicants related to permanent members been treated as if they were unrelated.

I summarize the effect of relationships by comparing the quality of the proposals that would have been funded had relationships not been taken into account with the quality of those that actually are funded. Specifically, I consider all applications that are funded and sum up the number of publications and citations that accrue to this portfolio. This is my benchmark measure of the quality of NIH peer review. I then simulate what applications would have been funded had relationships not been taken into account. To do this, I fix the total number of proposals that

[19] Even though $\text{Decision}_{icmt}^{\text{No Relationship}}$ is constructed using estimates from Equation (6), it does not rely on the model to interpret those coefficients.

are funded in each committee meeting but reorder applications by their counterfactual funding probabilities. I sum up the number of publications and citations that accrue to this new portfolio of funded grants. The difference in the quality of the benchmark and counterfactual portfolio provides a concrete, summary measure of the effect of relationships on the quality of research that the NIH supports.

For a fuller sense of how committees affect decisionmaking, I create a measure of committee-specific performance and examine how relationships affect the distribution of performance among NIH peer-review committees. First, I define a committee's *value-added*. Suppose two scientists submit applications to the same committee meeting. A good committee is one that systematically funds the application that is of higher quality. Good committees, moreover, should bring insights beyond what can simply be predicted by objective measures of an applicant's past performance. In particular, suppose now that two scientists with identical objective qualifications submit applications to the same committee meeting. A committee with high value-added is one that systematically funds the application that subsequently generates more citations, even though the applications initially look similar. My measure of committee value-added formalizes this intuition:

$$\text{Decision}_{icmt} = a + b_{cmt}\text{Quality}_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt}. \tag{7}$$

Here, the dependent variable is either an application's actual funding status $\text{Decision}_{icmt} = D_{icmt}^{\text{Benchmark}}$ or its counterfactual funding status $\text{Decision}_{icmt} = \text{Decision}_{icmt}^{\text{No Relationship}}$. The committee fixed effects δ_{cmt} restrict comparisons of applications to those evaluated in a single meeting and the X_{icmt} control for past applicant qualifications. The coefficients of interest are the b_{cmt}. These are meeting-specific slopes that capture the relationship between an application's quality Quality_{icmt} and its likelihood of being funded Decision_{icmt}. Each b_{cmt} is interpreted as the percentage-point change in the likelihood that an application will be funded for a one-unit increase in quality. This forms the basis of my committee value-added measure.

This concept of committee value-added differs from the classical notion of value-added commonly used in the teacher or manager performance literature (see Kane, Rockoff, and Staiger, 2007, and Bertrand and Schoar, 2003). Teacher value-added, for instance, is typically estimated by regressing student test scores on lags of test scores, school fixed effects, and teacher fixed effects. A teacher's fixed effect, the average performance of her students purged of individual, parental, and school-wide inputs, is taken to be the basic measure of quality.

This traditional measure, however, does not capture value-added in my setting. Good committees are not ones in which all applications are high-performing; after all, committees have no control over which applications are submitted. Rather, good committees are ones in which funded grants perform better than unfunded grants. I measure a committee's performance by the relationship between an application's quality and its likelihood of being funded because, unlike a teacher,

a committee's job is not to *improve* the quality of grant applications but to *distinguish* between them. One concern with the estimated $\hat{b}_{cmt}$ is that idiosyncratic variation in grant performance may lead me to conclude that some committee meetings do an excellent job of identifying high-quality applications when in fact they are simply lucky. I correct for this using a standard Bayesian shrinkage approach, discussed in Appendix D.

7.1 Results

Table 9 estimates the effect of relationships on the quality of research that the NIH supports. In effect, I ask what the NIH portfolio of funded grants would have been had committees treated applicants who are related to permanent members as if they were not, holding all else fixed. In my sample, I observe 93,558 applications, 24,404 of which are funded. Using this strategy, I find that 2,500, or 2.7 percent, of these applications change funding status under the counterfactual.

On average, being related to a greater composition of influential reviewers helps an applicant obtain funding; ignoring them would decrease the number of related applicants who are funded by 4.5 percent. These applications from related reviewers, however, are on average better than the applications that would have been funded had relationships not mattered. The overall portfolio of funded grants under the counterfactual produces two to three percent fewer citations, publications, and high-impact publications.

This pattern is underscored by Figure 6, which graphs the distribution of value-added under the benchmark and counterfactual cases. First, Figure 6 shows that there is substantial variation in the ability of committees to identify grant applications that subsequently produce high-impact research. For a study section with median value-added, a one standard deviation increase in the quality of an application evaluated by the median committee would increase its likelihood of being funded by approximately 6.3 percent. For 75th percentile committees, this figure is 13.3 percent. A striking feature of this distribution is that the bottom quarter to third of committees actively subtract value, meaning that increases in quality are correlated with *decreases* in the likelihood that an application will be funded. As explained in Section 7, these figures account for sampling variation so that a committee is deemed to have negative value-added only if it systematically does so from meeting to meeting. This could happen if committees systematically favor other factors that are negatively correlated with future citations, for example, the funding of new investigators.

Ignoring the role of intellectually related reviewers tends to worsen committee value-added throughout the middle of the value-added distribution. A study section with the median value-added, as calculated using an application's counterfactual funding status, falls from 6.3 percent to 4.7 percent, the 75th percentile falls to 11.8 percent, and the 25th percentile falls from -1.0 percent to -2.9 percent. The magnitudes of these declines are small compared to the overall distribution; understanding other reasons for this dispersion is an important area for future research.

8 Conclusion

This paper develops a conceptual and statistical framework for understanding and separately identifying the effects of bias and expertise in grant evaluation. My results show that, as a result of bias, being related to a more influential member of a review committee increases an application's chances of being funded by 3.1 percent. Viewed in terms of how committees respond to increases in application quality, this bias increases the chances that an application will be funded by the same amount as would be predicted by a one-quarter standard deviation increase in application quality. The expertise that reviewers have about research in their own area, however, also improves the quality of review: working in the same area as a permanent committee member increases the responsiveness of the committee to proposal quality by over 50 percent. On net, ignoring relationships reduces the quality of the NIH-funded portfolio as measured by numbers of citations and publications by two to three percent.

My results suggest that there may be scope for improving the quality of peer review. I document significant and persistent dispersion in the ability of committees to fund high-quality research. Finding ways to eliminate the lower tail of committees, for which increases in quality are actually associated with *decreases* in funding likelihood, could lead to large improvements in the quality of NIH-funded research as measured by citations. The magnitude of these potential benefits is not small when viewed in dollar terms. NIH spending for my sample of approximately 25,000 funded grants totaled more than 34 billion dollars (2010 dollars). These grants generated approximately 170,000 publications and 6.8 million citations.[20] This means that, in my sample, the NIH spent about 250,000 dollars per publication, or about 5,000 dollars per single citation. Even if these numbers do not represent the social value of NIH-funded research, they suggest that the value generated by high-quality peer review can be substantial. This paper shows that reviewers can improve peer review even if they are biased. Understanding and quantifying other factors affecting committee performance is an important area for future work. Here, the uniformity of NIH's many chartered study sections is helpful because it allows for the possibility of targeted randomized experiments, holding other institutional features constant. For instance, to understand the impact of committee composition on peer-review quality, applicants could be assigned to intellectually broad or narrow committees. Answers to these questions can provide insights on how to improve project evaluation at the NIH and elsewhere.

[20] I have 170,000 publications linked to grants via formal grant acknowledgments computed from the PubMed database. PubMed, however, undercounts citations because it only counts citations from a subset of articles archived in PubMed Central. To arrive at the 6.8 million citations figure, I use total publications calculated via text matching (about 100,000 publications) and the total citations accruing to those publications (4.3 million) to compute the average number of citations per publication. I then scale this by the 170,000 publications found in PubMed.

References

[1] Acemoglu, Daron. (2008) *Introduction to Modern Economic Growth*, Princeton University Press.

[2] Aghion, Philippe and Jean Tirole. (1994) "The Management of Innovation." *Quarterly Journal of Economics*, 109(4), 1185-1209.

[3] Arrow, Kenneth. (1962). "Economic welfare and the allocation of resources for invention." in Richard Nelson, ed., *The Rate and Direction of Inventive Activity*, Princeton University Press.

[4] Azoulay, Pierre, Toby Stuart, and Yanbo Wang. (2011) "Matthew: Fact or Fable?" Working paper. Available online: http://pazoulay.scripts.mit.edu/docs/shmatus.pdf

[5] Bagues, Manuel and Natalia Zinovyeva. (2012) "The Role of Connections in Academic Promotions." IZA Discussion Paper #6821.

[6] Blanes i Vidal, Jordi, Mirko Draca, and Christian Fons-Rosen. (2012) "Revolving Door Lobbyists." *American Economic Review*, 102(7): 3731-48.

[7] Brogaard, Jonathan, Joseph Engelberg and Christopher Parsons. (2012) "Network Position and Productivity: Evidence from Journal Editor Rotations." mimeo.

[8] Bertrand, Marianne and Antoinette Schoar. (2002) "Managing With Style: The Effect Of Managers On Firm Policies." *The Quarterly Journal of Economics* 118(4), 1169-1208.

[9] Cockburn, Iain, and Rebecca Henderson. (2000) "Publicly Funded Science and the Productivity of the Pharmaceutical Industry." *Innovation Policy and the Economy*, 1: 1-34.

[10] Congressional Budget Office. (2006) "Research and Development in the Pharmaceuticals Industry." Available online at: http://www.cbo.gov/ftpdocs/76xx/doc7615/10-02-DrugR-D.pdf

[11] Crawford, Vincent and Joel Sobel. (1982) "Strategic Information Transmission." *Econometrica*, 50(6):1431-1451.

[12] David, Paul, Mowery David, and W.E. Steinmueller. (1992) "Assessing the Economic Payoff from Basic Research." *Economics of Innovation and New Technology* 73-90.

[13] Gerin, William. (2006) *Writing the NIH grant proposal: a step-by-step guide.* Thousand Oaks, CA: Sage Publications.

[14] Ginther, Donna, Walter Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel Haak, and Raynard Kington. (2011) "Race, Ethnicity, and NIH Research Awards." *Science*, 333(6045): 1015-1019.

[15] Griliches, Zvi. (1992) "The search for R&D spillovers." *Scandinavian Journal of Economics*, 94 (supplement), S29-S47.

Paper #4240

[16] Hansen, Stephen, Michael McMahon, and Carlos Velasco Rivera. (2012) "How Experts Decide: Preferences or Private Assessments on a Monetary Policy Committee?" mimeo.

[17] Hegde, Deepak. (2009) "Political Influence behind the Veil of Peer Review: An Analysis of Public Biomedical Research Funding in the United States" *Journal of Law and Economics*, 52(4): 665-690.

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

[19] Jones, Benjamin. (2010) "Age and Great Invention." *Review of Economics and Statistics* 92(1): 1-14.

[20] Kane, Thomas, Jonah Rockoff, and Doug Staiger. (2007) "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6), 615-631.

[21] Kerr, William. (2008) "Ethnic Scientific Communities and International Technology Diffusion." *The Review of Economics and Statistics*, 90(3): 518-537.

[22] Kremer, Michael and Heidi Williams. (2010) "Incentivizing innovation: Adding to the toolkit," in Josh Lerner and Scott Stern, eds., *Innovation Policy and the Economy*, Vol. 10. Chicago, IL: University of Chicago Press.

[23] Merton, Robert. (1968) "The Matthew Effect in Science" *Science* 159(3810): 5663.

[24] Sampat, Bhaven and Frank Lichtenberg. (2011) "What are the Respective Roles of the Public and Private Sectors in Pharmaceutical Innovation?" *Health Affairs*, 30(2): 332-339.

[25] Stephan, Paula. (2012) *How Economics Shapes Science.* Cambridge, MA: Harvard University Press.

[26] Topa, Giorgio. (2012) "Labor Markets and Referrals" *Handbook of Social Economics*

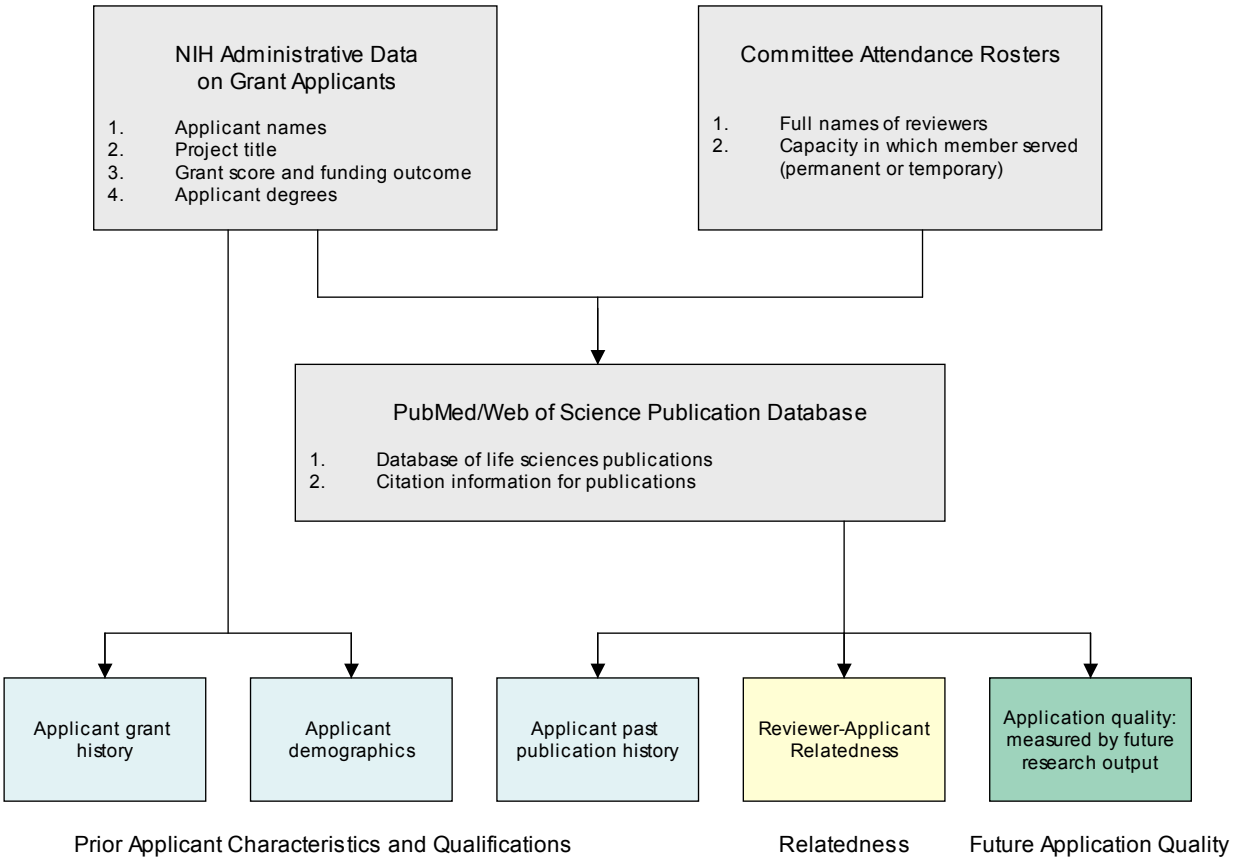

FIGURE 1: DATA SOURCES AND VARIABLE CONSTRUCTION

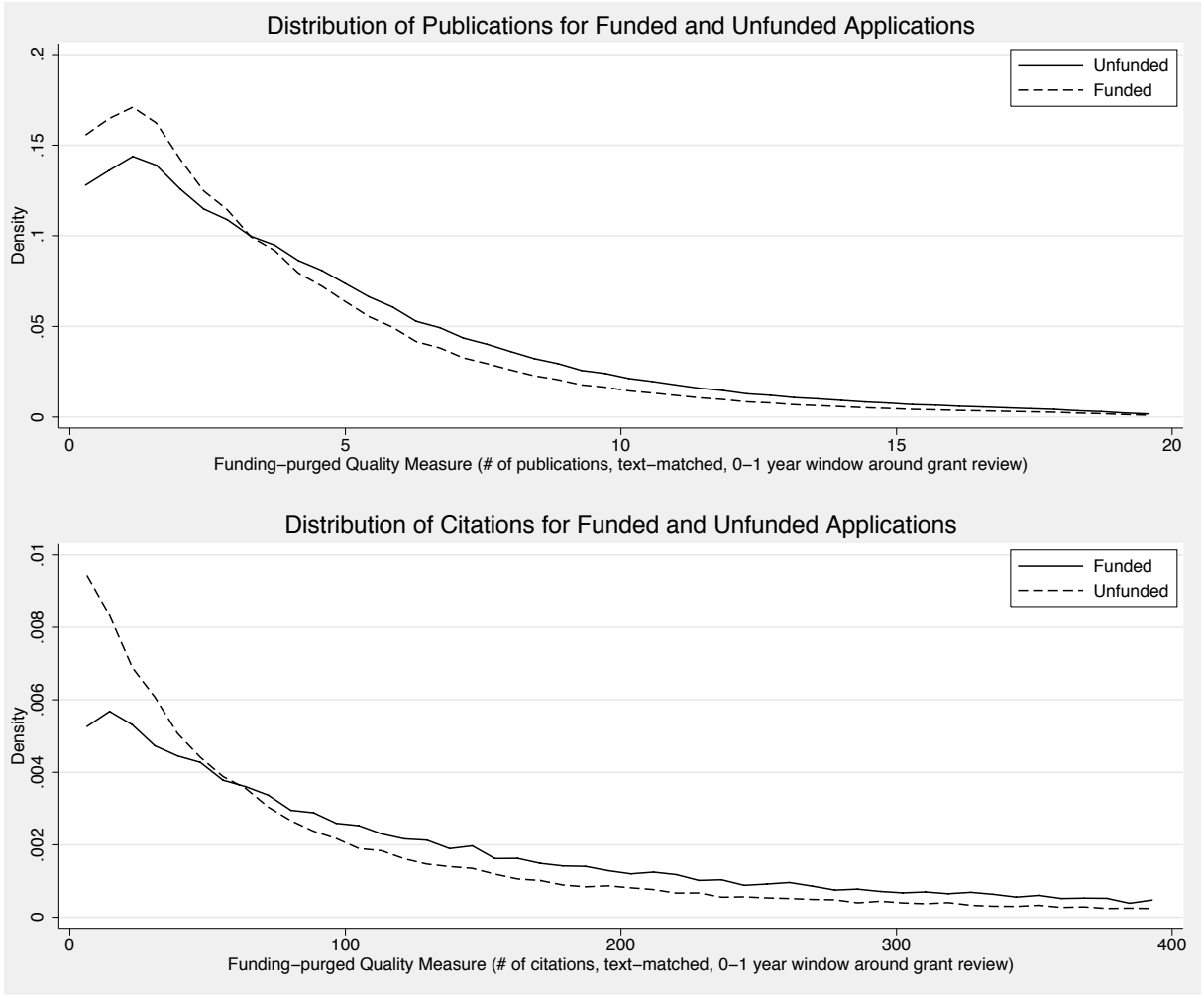

FIGURE 2: DISTRIBUTION OF APPLICATION QUALITY: FUNDED AND UNFUNDED GRANTS

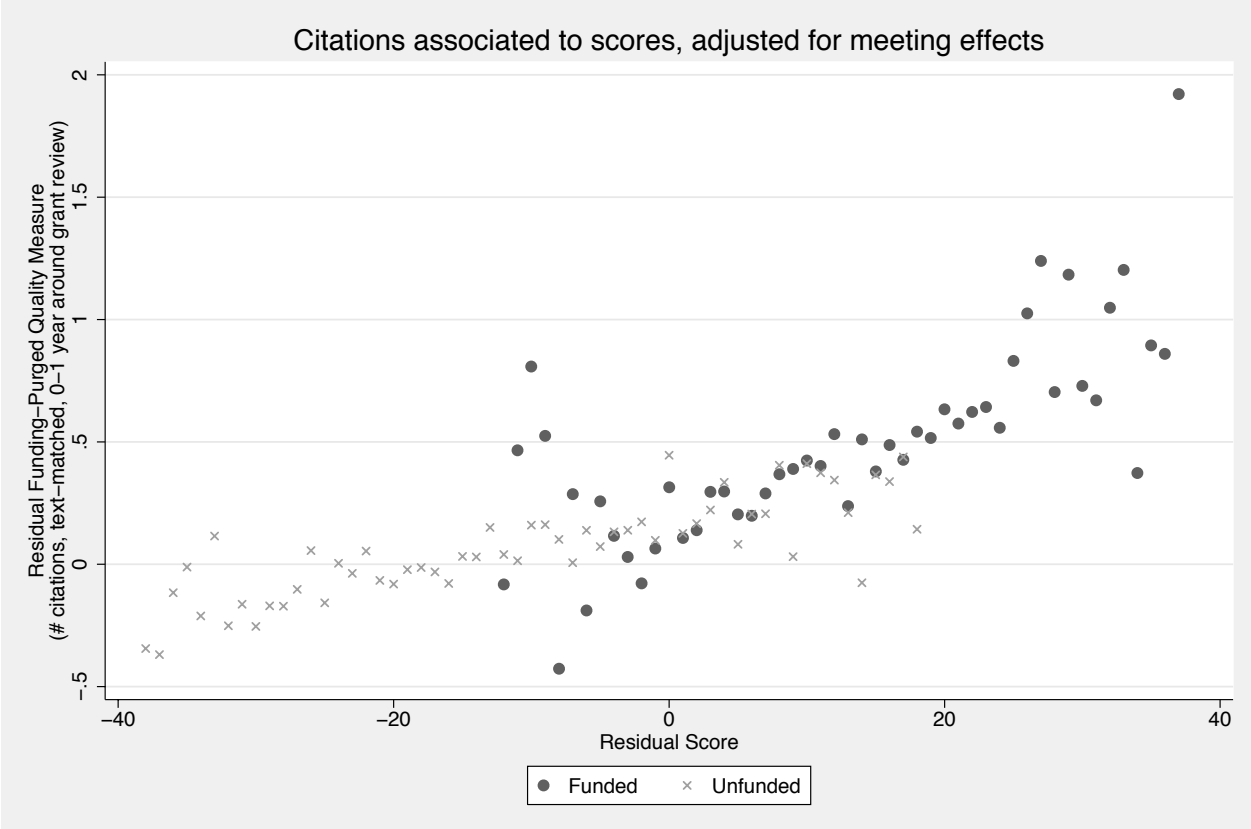

FIGURE 3: MEAN APPLICATION QUALITY BY SCORE: FUNDED AND UNFUNDED GRANTS

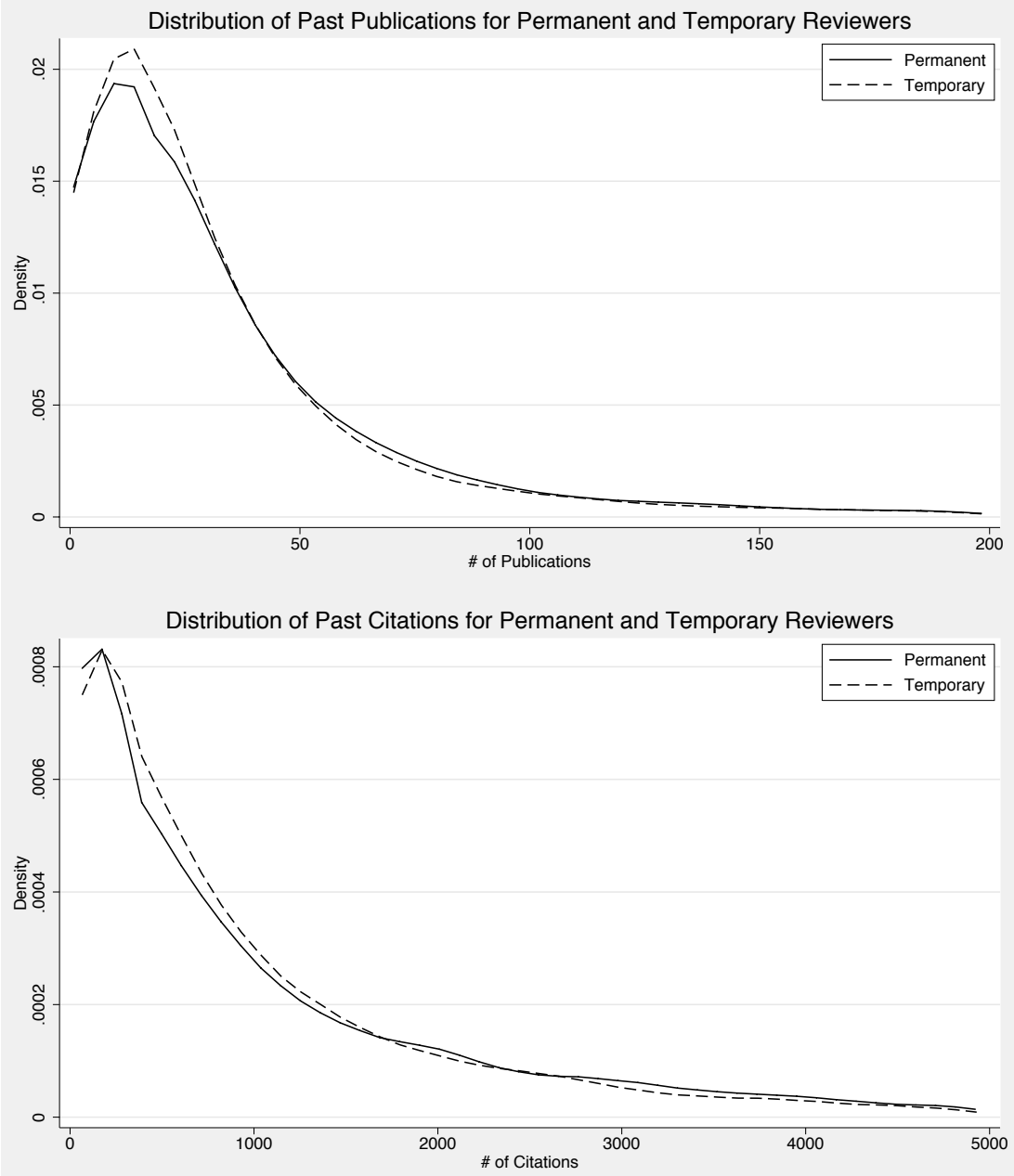

FIGURE 4: DISTRIBUTION OF PAST CITATIONS: PERMANENT AND TEMPORARY REVIEWERS

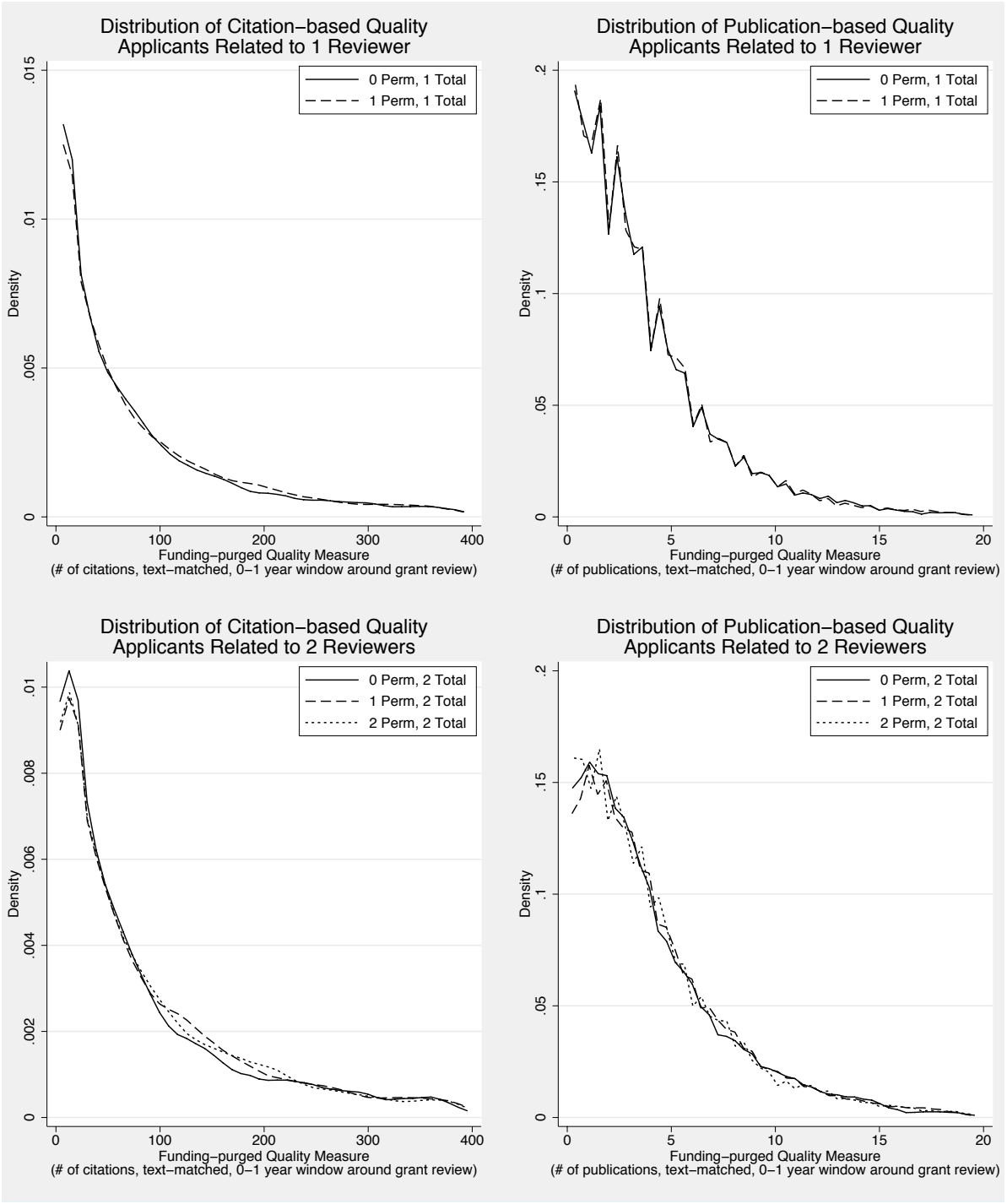

FIGURE 5: APPLICATION QUALITY CONDITIONAL ON TOTAL RELATED REVIEWERS

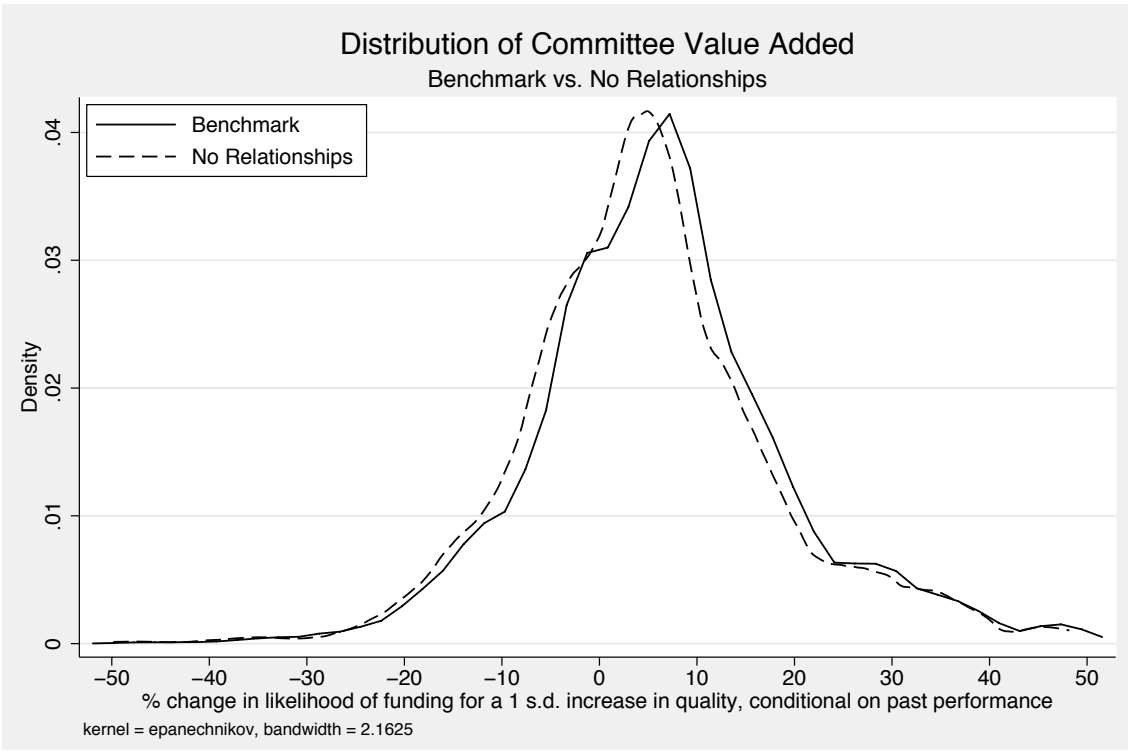

FIGURE 6: DISTRIBUTION OF MEETING-LEVEL VALUE-ADDED

Table 1: Applicant Characteristics

	Roster-Matched Sample		Full Sample	
Sample Coverage		Std. Dev.		Std. Dev.
# Grants	93,558		156,686	
# Applicants	36,785		46,546	
Years	1992-2005		1992-2005	
# Study Sections	250		380	
# Study Section Meetings	2,083		4,722	
Grant Characteristics				
% Awarded	26.08		30.48	
% Scored	61.58		64.04	
% New	70.31		71.21	
Percentile Score	70.05	18.42	71.18	18.75
# Publications, grant-publication matched (median)	2	5	2	5
# Citations, grant-publication matched (median)	36	265	38	302
PI Characteristics				
% Female	23.21		22.58	
% Asian	13.96		13.27	
% Hispanic	5.94		5.79	
% M.D.	28.72		29.26	
% Ph.D.	80.46		79.69	
% New investigators	19.70		20.02	
# Publications, past 5 years	15	60	15	55
# Citations, past 5 years	416	1431	423	1474

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the year following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. The full sample includes data from any new or competing R01 grant evaluated in chartered study sections from 1992 to 2005. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped.

Table 2: Does being funded directly affect my measure of quality?

	(1)	(2)	(3)	(4)
	Subsample of Scored Applications		**Full Sample (Score Imputed for Unscored Applications)**	
Dep var: Grant Quality	No score controls	Controls for smooth function of score	No score controls	Controls for smooth function of score
1(Grant is funded)	0.4692***	-0.0286	0.7492***	0.0683
	(0.032)	*(0.059)*	*(0.035)*	*(0.055)*
Observations	57,613	57,613	93,558	93,558
R-squared	0.1169	0.1221	0.1021	0.1119
Meeting Fixed Effects	X	X	X	X

Notes: Coefficients are reported from a regression of grant quality on an indicator for whether the grant was funded and meeting fixed effects. Columns (2) and (4) include controls for quartics in the applicant score. Column (2) compares grant applications with the same score and evaluated in the same meeting, but which differ in funding status because they are assigned to different Institutes with different paylines. Scores are available only for applications that were not triaged; Columns (3) and (4) assign scores of zero to triaged applications to test with the full sample.

Table 3: Committee Descriptives

	Roster Matched Sample	
Reviewer Characteristics		Std. Dev.
# Reviewers	18,916	
# Permanent reviewers per meeting	17.23	4.52
# Temporary reviewers per meeting	12.35	7.44
# Meetings per permanent reviewer	3.69	3.03
# Meetings per temporary reviewer	1.78	1.30
# Applications	53.73	17.31
Relationship Characteristics		
# Reviewers who cite applicant	1.94	2.81
# Permanent reviewers who cite applicant	1.11	1.73
# Applicants cited by permanent reviewers	4.12	5.32
# Applicants cited by temporary reviewers	4.12	5.09

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the 2 years following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped.

Table 4: Characteristics permanent and temporary members

	Permanent	Temporary
Number of reviewers	9371	14067
Reviewer Characteristics		
% Female	31.68	24.28
% Asian	14.99	13.08
% Hispanic	6.40	5.05
% M.D.	27.42	25.85
% Ph.D.	79.45	80.99
# Publications, past 5 years (median)	22	21
# Citations, past 5 years (median)	606	590

Reviewer Transitions

	% Permanent in the Past	% Permanent in the Future	% Temporary in the Past	% Temporary in the Future
Current Permanent Members	61.87	63.71	38.11	35.45
Current Temporary Members	16.25	41.30	32.73	50.13

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the 2 years following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped. Transitions are calculated based on whether a reviewer is present in the roster database during the full sample years from 1992 to 2005. Means are taken for the years 1997 to 2002 in order to allow time to observe members in the past and future within the sample.

Table 5: Applicant characteristics, by number and composition of related reviewers

Related to 0 Reviewers			
% Female	27.50		
% Asian	15.35		
% Hispanic	6.88		
% M.D.	25.40		
% Ph.D.	82.73		
% New investigators	27.22		
# Publications, past 5 years (median)	9		
	(31)		
# Citations, past 5 years (median)	172		
	(713)		
N	37757		
Related to 1 Reviewer Total	**1 Permanent**	**1 Temporary**	
% Female	22.24	23.97	
% Asian	13.51	15.09	
% Hispanic	5.79	5.57	
% M.D.	27.11	26.71	
% Ph.D.	81.63	82.24	
% New investigators	19.34	19.76	
# Publications, past 5 years (median)	15	15	
	(49)	*(52)*	
# Citations, past 5 years (median)	442	443	
	(1102)	*(1080)*	
N	10980	7049	
Related to 2 Reviewers Total	**2 Permanent**	**1 Each**	**2 Temporary**
% Female	20.26	20.89	22.93
% Asian	12.54	13.17	13.69
% Hispanic	5.14	5.02	5.82
% M.D.	28.64	29.28	28.53
% Ph.D.	79.88	80.02	81.04
% New investigators	15.88	16.25	17.06
# Publications, past 5 years (median)	18	18	17
	(31)	*(50)*	*(45)*
# Citations, past 5 years (median)	563	556	510
	(1336)	*(1233)*	*(1050)*
N	4841	5094	2403

Notes: The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. Future publications refers to the number of research articles that the grant winner publishes in the year following the grant which share at least one salient word overlap between the grant project title and the publication title. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. Investigators with common names are dropped as are any for which the covariates are missing. Social science study sections are dropped. Transitions are calculated based on whether a reviewer is present in the roster database during the full sample years from 1992 to 2005. Means are taken for the years 1997 to 2002 in order to allow time to observe members in the past and future within the sample.

Table 6: What is the effect of being related to a reviewer on an applicant's likelihood of funding?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	1(Score is above the payline)			**Score**			**1(Scored at all)**		
	Mean = 0.214, SD = 0.410			Mean = 71.18, SD = 18.75			Mean = .640, SD = .480		
Related Permanent Reviewers	0.0328***	0.0155***	0.0066***	1.1067***	0.5192***	0.2371**	0.0500***	0.0248***	0.0042**
	(0.001)	*(0.001)*	*(0.002)*	*(0.054)*	*(0.052)*	*(0.093)*	*(0.002)*	*(0.001)*	*(0.002)*
Total Related Reviewers			0.0066***			0.2105***			0.0153***
			(0.001)			*(0.060)*			*(0.001)*
Observations	93,558	93,558	93,558	57,613	57,613	57,613	93,558	93,558	93,558
R-squared	0.0630	0.0906	0.0909	0.1186	0.1390	0.1392	0.0775	0.1230	0.1243
Committee × Year × Cycle FE	X	X	X	X	X	X	X	X	X
Past Performance, Past Grants, and		X	X		X	X		X	X

Notes: Coefficients are reported from a regression of committee decisions (above payline, score, or scored at all) on the number of permanent members related to an applicant, controlling for meeting level fixed effects. Columns (2) (5) and (8) include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to. Columns (3) (6) and (9) include an additional control for the total number of related reviewers. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review.

Table 7: What is the contribution of expertise vs. bias?

	(1)	(2)	(3)	(4)	(5)	(6)
	1(Score is above the payline)		**Score**		**1(Scored at all)**	
	Mean = 0.214, SD = 0.410		Mean = 71.18, SD = 18.75		Mean = .640, SD = .480	
Related Permanent Reviewers	0.0066***	0.0049**	0.2371**	0.1946**	0.0042**	0.0032
	(0.002)	*(0.002)*	*(0.093)*	*(0.094)*	*(0.002)*	*(0.002)*
Related to Permanent Reviewers × Future		0.0042***		0.0741		0.0007
		(0.001)		*(0.059)*		*(0.001)*
Future Citations		0.0067***		0.2356***		0.0109***
		(0.001)		*(0.062)*		*(0.001)*
Total Related Reviewers	0.0066***	0.0072***	0.2105***	0.2282***	0.0153***	0.0158***
	(0.001)	*(0.001)*	*(0.060)*	*(0.060)*	*(0.001)*	*(0.001)*
Observations	93,558	93,558	57,613	57,613	93,558	93,558
R-squared	0.0909	0.0937	0.1392	0.1405	0.1243	0.1265
Committee × Year × Cycle FE	X	X	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. Columns (1) (3) and (5) are reproduced from Table 6. Columns (2), (4) and (6) add controls for application quality and application quality interacted with relatedness to permanent reviewers. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant′s previous research in the 5 years prior to grant review. Future citations are calculated using all publications by an applicant in the year after grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant′s name is Hispanic, East Asian, or South Asian, quartics in an applicant′s total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Table 8: What is the contribution of expertise vs. bias? Applicant fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)
	1(Score is above the payline)		**Score**		**1(Scored at all)**	
	Mean = 0.214, SD = 0.410		Mean = 71.18, SD = 18.75		Mean = .640, SD = .480	
Total Related Reviewers	0.0065***	0.0071***	0.2907***	0.2914***	0.0112***	0.0108***
	(0.001)	*(0.001)*	*(0.060)*	*(0.061)*	*(0.001)*	*(0.001)*
Related to Reviewers × Future Citations		0.0042*		0.0969		0.0077***
		(0.002)		*(0.095)*		*(0.002)*
Future Citations		0.0016		0.0529		-0.0022
		(0.002)		*(0.097)*		*(0.002)*
Observations	93,558	93,558	57,613	57,613	93,558	93,558
R-squared	0.4524	0.4648	0.5448	0.5450	0.5629	0.5632
Applicant FE	X	X	X	X	X	X
Past Performance and Past Grants	X	X	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. All regressions include controls for applicant effects. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant′s previous research in the 5 years prior to grant review. Future citations are calculated using all publications by an applicant in the -1 to 2 years after grant review, with text matching. Applicant characteristics include quartics in an applicant′s total number of citations and publications over the past 5 years and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Table 9: What is the effect of relationships on the quality of research that the NIH supports?

	Benchmark	**No Relationships**
Number of Funded Grants	24,404	24,404
Number of Grants that Change Funding Status	2,500	2,500
Total # Citations	6,751,209	6,593,302
(% change relative to benchmark)		*-2.34*
Total # Publications	151,662	146,674
(% change relative to benchmark)		*-3.29*
Total # in Top 99% of Citations	6,642	6,475
(% change relative to benchmark)		*-2.51*
Total # in Top 90% of Citations	13,010	12,645
(% change relative to benchmark)		*-2.81*
Total # Related Applicants Funded	18,615	17,782
(% change relative to benchmark)		*-4.47*

Notes: Benchmark refers to characteristics of grants ordered according to their predicted probability of funding, using the main regression in Table 6 of funding status on relationships and other characteristics. No relationships refers to ordering of grants under the assumption that relatedness to permanent members and relatedness to permanent members interacted with quality do not matter (their coefficients are set to zero). Expected citations are calculated as fitted values from a regression of citations on relationships, past performance, demographics, and meeting fixed effects. The number of projects that are funded is kept constant within meeting. See text for details.

APPENDIX MATERIALS

A Proof of Proposition 1

A perfect Bayesian equilibrium for this game is characterized by a message strategy for the reviewer, a set of beliefs about Q^* by the committee for each message, and a decision strategy for the committee. Having defined the equilibrium concept, I proceed with the proof of Proposition 1.

Case 1. Suppose that the reviewer reports her exact posterior and the committee to believes it. In this case, the committee maximizes its utility by funding the proposal if and only if $Q_0 > U$. The reviewer has no incentive to deviate from this strategy because she is receiving her highest payoff as well.

Suppose, now, that there were another informative equilibrium. Each message $M \in \mathbf{M}$ induces a probability of funding $D(M)$. Let the messages be ordered such that $D(\mathbf{M}_1) \leq \cdots \leq D(\mathbf{M}_K)$ where $\mathbf{M}_i$ are the set of messages M_i that induce the same probability of funding $D(M_i)$. For reviewers of type $E(Q^*|Q_0) > U$, the reviewer strictly prefers that the grant be funded. She thus finds it optimal to send the message $\mathbf{M}_K$ that maximizes the probability that the grant is funded. Call this set Y. For $E(Q^*|Q^* + \varepsilon_0) < U$ the reviewer strictly prefer $E(Q^*|Q_0) = U$. Because the distribution of Q_R is assumed to be continuous on $\mathbb{R}$ and such that $E(Q^*|Q_R)$ is increasing in Q_R, this occurs with probability zero. Thus, with probability one, the space of possible messages is equivalent to $\mathbf{M} = \{Y, N\}$. For this equilibrium to be informative, it must be that $D(N) < D(Y)$. Given this, the committee's optimal reaction is to fund when $M = Y$ and to reject otherwise.

If the we allow uninformative equilibria, $D(\mathbf{M}_1) = \cdots = D(\mathbf{M}_K)$ and any reviewer message is permissible. It must be that $D(M_i) = 0$ for all M_i because the outside option U is assumed to be greater than the committee's prior on quality.

Case 2. Now consider the case of a reviewer evaluating a related application. As in Case 1, the set of messages is equivalent, with probability one, to $\mathbf{M} = \{Y, N\}$. In this case, however, reviewers of type $E(Q^*|Q_1) > U - B$ send $M = Y$ and reviewers of type $E(Q^*|Q_1) < U - B$ send $M = N$. The only reviewer who sends any other message is one for which $E(Q^*|Q_1) = U - B$.

Given this messaging strategy, a committee's expectation of Q^* given $M = N$ is $E(Q^*|E(Q^*|Q_1) < U - B)$. Since this is less than U, the grant goes unfunded. The committee's expectation of Q^* given $M = Y$ is $E(Q^*|E(Q^*|Q_1) < U - B)$. When this is larger than U, the committee listens to the reviewer's recommendation and we can verify that $D(Y) > D(N)$. When $E(Q^*|E(Q^*|Q^* + \varepsilon_1) < U - B) < U$, the grant is never funded: $D(Y) = D(N) = 0$. In this case, only babbling equilibria exist.

If the we allow uninformative equilibria, $D(\mathbf{M}_1) = \cdots = D(\mathbf{M}_K)$ and any reviewer message

is permissible. It must be that $D(M_i) = 0$ for all M_i because the outside option U is assumed to be greater than the committee's prior on quality.

Unobserved relatedness: Next, I consider a modification of Proposition 1 where the committee cannot observe whether the application is related to the reviewer.

Proposition A.2 *Assume that p is the probability that an application is related to a reviewer. Then, for every p, there exists a level of bias, B', such that for $B < B'$ there is a unique informative equilibrium:*

The reviewer reports a message Y if his posterior, $E(Q^*|Q_1)$, is greater than $U-B$ and N otherwise.

1. *An unrelated reviewer reports a message Y if his posterior, $E(Q^*|Q_0)$, is greater than U and N otherwise.*
2. *A related reviewer reports a message Y if his posterior, $E(Q^*|Q_1)$, is greater than $U - B$ and N otherwise.*
3. *The committee funds the grant if and only if the message is Y.*

For $B \geq B'$, only uninformative equilibria exist and the grant is never funded.[21]

Proof In this case, the reviewer's messaging strategy remains the same as in Proposition 1: because reviewers themselves know whether they are related, they form, with probability one, strict preferences about whether an application should be funded. Related reviewers for which $E(Q^*|Q_1) > U - B$ send $M = Y$ and those for which $E(Q^*|Q_1) < U - B$ send $M = N$. Similarly, unrelated reviewers of type $E(Q^*|Q_0) > U$ send $M = Y$ and unrelated reviewers of type $E(Q^*|Q_0) < U$ send $M = N$.

The committee, however, does not observe the relatedness and, as such, forms the following expectation of quality conditional on observing $M = Y$:

$$K\left[E(Q^*|E(Q^*|Q_0) > U)\right] + (1 - K)\left[E(Q^*|E(Q^*|Q_1) > U - B)\right]$$

The first term $E(Q^*|E(Q^*|Q_0) > U)$ is the committee's expectation of quality if it knows that the $M = Y$ message is sent by an unrelated reviewer. Similarly, the second term $E(Q^*|E(Q^*|Q_1) > U - B)$ is the committee's expectation of quality if it knows that the message is sent by a related reviewer. The term K is the probability that the committee believes a Y message comes from an unrelated reviewer, that is, $K = E(R = 0|M = Y)$. By Bayes' Rule, this is given by $K = E(R =$

[21] Again, in all cases where an informative equilibrium exists, there also exist uninformative equilibria where the grant is never funded.

$0|M = Y) = \frac{E(R=0,M=Y)}{E(M=Y)}$. The overall probability of a Y message is thus given by

$$E(M = Y) = (1 - p)(E(Q^*|Q_0) > U) + p(E(Q^*|Q_1) > U - B)$$

Similarly, the probability that the message is Y and the reviewer is unrelated is given by $(1 - p)(E(Q^*|Q_0) > U)$. As such, we have

$$K = \frac{(1 - p)(E(Q^*|Q_0) > U)}{(1 - p)(E(Q^*|Q_0) > U) + p(E(Q^*|Q_1) > U - B)}.$$

and for

$$K\left[E(Q^*|E(Q^*|Q^* + \varepsilon_0) > U)\right] + (1 - K)\left[E(Q^*|E(Q^*|Q^* + \varepsilon_1) > U - B)\right] > U$$

the committee funds the application. Again, we can verify that $D(Y) > D(N)$. For any fixed p, the threshold B' can be defined to set this expression equality. There also exist uninformative equilibria where all grants are rejected. This term is less than U, then the grant is never funded: $D(Y) = D(N) = 0$. In this case, only babbling equilibria exist.

B Proof of Proposition 2

Measurement error in Q^* can potentially affect the estimation of α_2 in Equation (3). The presence of U, RU, and X, however, will not affect consistency; for simplicity, I rewrite both the regression suggested by the model and the actual estimating equation with these variables partialed out. The remaining variables should then be thought of as conditional on U, RU, and X

$$D = \alpha_0 + \alpha_1 Q^* + \alpha_2 R + \alpha_3 R Q^* + \epsilon$$

$$\begin{aligned} D &= a_0 + a_1 Q + a_2 R + a_3 RQ + e \\ &= a_0 + W + a_2 R + e, W = a_1 Q + a_3 RQ \end{aligned}$$

The coefficient a_2 is given by:

$$a_2 = \frac{\text{Var}(W)\text{Cov}(D,R) - \text{Cov}(W,R)\text{Cov}(D,W)}{\text{Var}(W)\text{Var}(R) - \text{Cov}(W,R)^2}$$

Consider $\text{Cov}(W,R)$:

$$\begin{aligned} \text{Cov}(W,R) &= \text{Cov}(a_1(Q^* + v) + a_3 R(Q^* + v), R) \\ &= a_1\text{Cov}(Q^*,R) + a_1\text{Cov}(v,R) + a_3\text{Cov}(RQ^*,R) + a_3\text{Cov}(Rv,R) \end{aligned}$$

Under the assumption that R and Q^* are conditionally independent, this yields:

$$\begin{aligned} \text{Cov}(W,R) &= a_3\text{Cov}(RQ^*,R) + a_3\text{Cov}(Rv,R) \\ &= a_3\left[E(R^2Q^*) - E(RQ^*)E(R)\right] + a_3\left[E(R^2v) - E(Rv)E(R)\right] \\ &= a_3\left[E(R^2)E(Q^*) - E(R)^2E(Q^*)\right] + a_3\left[E(R^2)E(v) - E(R)^2E(v)\right] \\ &= a_3\left[E(R^2)0 - E(R)^20\right] + a_3\left[E(R^2)0 - E(R)^20\right] \\ &= 0 \end{aligned}$$

With this simplification, the expression for the estimated coefficient on a_2 becomes:

$$\begin{aligned}
a_2 &= \frac{\text{Var}(W)\text{Cov}(D,R) - \text{Cov}(W,R)\text{Cov}(D,W)}{\text{Var}(W)\text{Var}(R) - \text{Cov}(W,R)^2} \\
&= \frac{\text{Var}(W)\text{Cov}(D,R)}{\text{Var}(W)\text{Var}(R)} \\
&= \frac{\text{Cov}(D,R)}{\text{Var}(R)} \\
&= \frac{\text{Cov}(\alpha_0 + \alpha_1 Q^* + \alpha_2 R + \alpha_3 RQ^* + \varepsilon, R)}{\text{Var}(R)} \\
&= \frac{\alpha_2 \text{Var}(R) + \alpha_3 \text{Cov}(RQ^*, R)}{\text{Var}(R)} \\
&= \frac{\alpha_2 \text{Var}(R) + \alpha_3 \left[E(R^2)E(Q^*) - E(R)^2 E(Q^*)\right]}{\text{Var}(R)} \\
&= \alpha_2
\end{aligned}$$

C Robustness Checks

Appendix Table A provides evidence that permanent members do indeed have more influence. In my sample, I observe almost 5,000 reviewers serving both as permanent and as temporary members. For this subset of reviewers, I show that a larger proportion of the applicants whom they have cited are funded when the reviewer is permanent than when the reviewer is temporary, conditional on applicant qualifications. I also show that mean scores for applicants related to a reviewer are higher when that reviewer is permanent. These regressions include reviewer fixed effects, meaning that an applicant related to the same reviewer is more likely to be funded when that reviewer is permanent as opposed to temporary.

The next set of results in this section support the assertion that quality is measured consistently. Appendix Table B addresses concerns that funding may directly influence the number of citations produced by a grant by, for example, freeing up an investigator from future grant writing so that he can concentrate on research. Instead of including articles published after the grant is reviewed, Appendix Table B restricts my analysis to articles published one year before a grant is reviewed. These publications are highly likely to be based off research that existed before the grant was reviewed, but cannot have been influenced by the grant funds. Using this metric, I find nearly identical measures of bias and information.

Another potential concern is that, because I restrict my main quality measure to be based on articles that are closely related to the grant proposal topic, I am potentially missing other research that reviewers might be anticipating when they evaluate a grant proposal. To test whether this is the case, I use grant acknowledgement data recorded in the National Library of Medicine's PubMed database to match funded grants to all the articles that it produces, regardless of topic or date

of publication. For the set of funded grants, Appendix Table C reruns my core regressions using citations to publications that explicitly acknowledge a grant as my measure of quality. This analysis differs slightly from my main results using citations because general citations cannot be computed for publications in PubMed. A limited set of citations can, however, be computed using publications in PubMed Central (PMC). PMC contains a subset of life sciences publications made available for free. While this is not as comprehensive a universe as that of Web of Science, it contains, for recent years, all publications supported by NIH dollars. Undercounting of publications would, further, not bias my result as long as it does not vary systematically by whether an applicant is related to a permanent or to a temporary member. I find results that are consistent with my primary findings, though of a slightly smaller magnitude.

Another test of my assumption that citations are not directly affected by funding is to ask whether I find bias in the review of inframarginal grants, that is grants that are well above or well below the funding margin. All grants in either group have the same funding status so any bias I find cannot be attributed to differences in funding. Because I hold funding status constant, I can only assess the impact that related permanent members have on an applicant's score not on an applicant's funding status. Appendix Table D reports these results. In Columns 3–4 and 5–6, I report estimates of the effect of bias and information in the sample of funded and unfunded grants, respectively. In both cases, I still find evidence that bias exists. The magnitudes are somewhat smaller than in my main regression; because these are subsamples, there is no reason to expect that the magnitude of the effect of relationships should be the same for high- and low-quality grants as it is for the entire sample.

Appendix Table E adds nonlinearity to Equation (6) in order to show that my results are robust to the assumption in Section 3 that $Q_R = Q^* + \varepsilon_R$ for ε_R uniform and $E(Q^*|Q_R) \approx \lambda_R Q_R$. Without these assumptions, the association between relatedness and quality would, in general, be nonlinear. To show that this does not make a material difference for my results, I allow for the effects of quality and relatedness to vary flexibly by including controls for cubics in Q, as well as cubics of Q interacted with whether an applicant is related to a permanent member. I find similar results, both qualitatively and quantitatively.

My results are robust to non-parametric controls for the total number of related applicants (meeting by number of related reviewers fixed effects) and using alternative definitions of relatedness, including using applicant-reviewer mutual citations and citations defined only on publications for which applicants and reviewers are primary authors (first, second, and last position). These and other detailed tables are available from the author.

D Estimating Committee Value-Added

I estimate committee value-added using the following regression:

$$\text{Decision}_{icmt} = a + b_{cmt}\text{Quality}_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt} \tag{8}$$

D_{icmt} is either the actual or counterfactual funding decision for applicant i reviewed during meeting m of committee c in year t. Q_{icmt} is a measure of application quality such as the number of citations it produces in the future and X_{icmt} are detailed controls for the past performance of the applicant, including flexible controls for number of past publications and citations, number and type of prior awarded grants and prior applications, and flexible controls for degrees, gender, and ethnicity. Finally, δ_{cmt} are committee meeting level fixed effects. The coefficients b_{cmt} capture, for each meeting, the correlation between decisions and quality, conditional on X_{icmt}.

Variation in b_{cmt} include sampling error so that $\hat{b}_{cmt}$ is a combination of true value-added plus a noise term. I assume this luck term to be independent and normal:

$$\hat{b}_{cmt} = b^*_{cmt} + \nu_{cmt} \tag{9}$$

Under this assumption, $\text{Var}(\hat{b}_{cmt}) = \text{Var}(b^*_{cmt}) + \text{Var}(\nu_{cmt})$ so that the estimate of true variance is upwardly biased from the additional variance arising from estimation error. To correct for this, I note that the best estimate for b^*_{cmt} is given by $\text{E}(b^*_{cmt}|\hat{b}_{cmt}) = \lambda_{ct}\hat{b}_{cmt} + (1-\lambda_{ct})\overline{\hat{b}}_{ct}$ where $\overline{\hat{b}}_{ct}$ is the mean of meeting quality for that committee-year and $\lambda_{ct} = \frac{\sigma^2_{b^*_{cmt}}}{\sigma^2_{b^*_{cmt}} + \sigma^2_{\nu_{cmt}}}$ is a Bayesian shrinkage term constructed as the ratio of the estimated variance of true committee effects, $\sigma^2_{b^*_{cmt}}$, to the sum of estimated true variance $\sigma^2_{b^*_{cmt}}$ and estimated noise variance $\sigma^2_{\nu_{cmt}}$.

To derive this shrinkage term, I use the correlation in meeting quality across the three different funding cycles of a committee fiscal year. In particular, if meeting-specific errors are independent, then $\text{Cov}(\hat{b}_{cmt}, \hat{b}_{cm't}) = \text{Var}(b^*_{cmt}) = \hat{\sigma}^2_{b^*_{cmt}}$. This can be estimated at the committee-year level because a committee meets three times during the year. I construct

$$\hat{\lambda}_{ct} = \frac{\hat{\sigma}^2_{b^*_{cmt}}}{\hat{\sigma}^2_{b^*_{cmt}} + \hat{\sigma}^2_{\nu_{cmt}}} \tag{10}$$

so that the adjusted committee value-added is given by:

$$VA_{cmt} = \hat{\lambda}_{ct}\hat{b}_{cmt} \tag{11}$$

Because committee membership is not fixed across funding cycles within the same fiscal year (temporary members rotate, permanent members do not), variation in VA_{cmt} represents a conservative lower bound on the variance of committee quality.

Appendix Table A: Do permanent reviewers have more influence?

	(1)	(2)
	Proportion of Related Applicants who are Funded	**Average Score of Related Applicants**
Related Reviewer is Permanent	0.003***	0.336**
	(0.001)	*(0.144)*
Observations	15871	15870
R-squared	0.954	0.571
Reviewer FE	X	X
Past Performance, Past Grants, and Demographics	X	X

Notes: This examines how outcomes for related applicants vary by whether the related reviewer is serving in a permanent or temporary capacity. The sample is restricted to 4909 reviewers who are observed both in temporary and permanent positions. An applicant is said to be related by citations if a reviewer has cited that applicant in the 5 years prior to the meeting. Applicant characteristics include indicators for sex and whether an applicant′s name is Hispanic, East Asian, or South Asian, quartics in an applicant′s total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table B: What is the contribution of expertise vs. bias? Grant quality measured from articles published before grant review

	(1)	(2)	(3)	(4)	(5)	(6)
	1(Score is above the payline)		**Score**		**1(Scored at all)**	
	Mean = 0.214, SD = 0.410		Mean = 71.18, SD = 18.75		Mean = .640, SD = .480	
Related Permanent Reviewers	0.0066***	0.0054**	0.2371**	0.2009**	0.0042**	0.0036*
	(0.002)	*(0.002)*	*(0.093)*	*(0.094)*	*(0.002)*	*(0.002)*
Related to Permanent Reviewers × Future		0.0051**		0.0800		0.0006
		(0.002)		*(0.095)*		*(0.002)*
Future Citations		0.0117***		0.4330***		0.0144***
		(0.002)		*(0.093)*		*(0.002)*
Total Related Reviewers	0.0066***	0.0071***	0.2105***	0.2283***	0.0153***	0.0157***
	(0.001)	*(0.001)*	*(0.060)*	*(0.060)*	*(0.001)*	*(0.001)*
Observations	93,558	93,558	57,613	57,613	93,558	93,558
R-squared	0.0909	0.0945	0.1392	0.1412	0.1243	0.1263
Committee × Year × Cycle FE	X	X	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. Columns (1) (3) and (5) 1 are reproduced from Table 6. Columns 2 and 4 add controls for application quality and application quality interacted with relatedness to permanent reviewers. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant′s previous research in the 5 years prior to grant review. Future citations are calculated using all publications by an applicant in the year before grant review to the year of grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant′s name is Hispanic, East Asian, or South Asian, quartics in an applicant′s total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table C: What is the contribution of bias and information? Explicit grant acknowledgements for the sample of funded grants

	(1)	(2)
Dep var: Score Mean = 71.18, SD = 18.75	**Explict Grant Acknowledgements**	
Related Permanent Reviewers	0.1384* *(0.0724)*	0.1285* *(0.0734)*
Related to Permanent Reviewers × Future Citations		0.0749 *(0.1004)*
Future Citations		0.4806*** *(0.0770)*
Total Related Reviewers	-0.0074 *(0.0456)*	0.0086 *(0.0472)*
Observations	24395	24395
R-squared	0.1743	0.1793
Committee × Year × Cycle FE	X	X
Past Performance, Past Grants, and Demographics	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. The analytic sample includes all awarded R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant′s previous research in the 5 years prior to grant review. Future citations are calculated explicit grant acknowlegments. Applicant characteristics include indicators for sex and whether an applicant′s name is Hispanic, East Asian, or South Asian, quartics in an applicant′s total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table D: What is the contribution of expertise vs. bias? Inframarginal grant applications

	(1)	(2)	(3)	(4)	(5)	(6)
Dep var: Score Mean = 71.18, SD = 18.75	**All**		**Funded**		**Not Funded**	
Related Permanent Reviewers	0.2371**	0.1946**	0.1348*	0.0942	0.1694*	0.1383
	(0.093)	*(0.094)*	*(0.073)*	*(0.073)*	*(0.089)*	*(0.091)*
Related to Permanent Reviewers × Future		0.0741		0.1064**		0.0520
		(0.059)		*(0.051)*		*(0.068)*
Future Citations		0.2356***		-0.0590		0.1523**
		(0.062)		*(0.045)*		*(0.063)*
Total Related Reviewers	0.2105***	0.2282***	-0.0040	0.0043	0.1387**	0.1435**
	(0.060)	*(0.060)*	*(0.046)*	*(0.046)*	*(0.058)*	*(0.058)*
Observations	57,613	57,613	24,395	24,395	33,218	33,218
R-squared	0.1392	0.1405	0.1728	0.1731	0.1857	0.1866
Committee × Year × Cycle FE	X	X	X	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X	X	X	X

Notes: Coefficients are reported from a regression of score on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant's previous research in the 5 years prior to grant review. Future citations are standardized to be mean zero, standard deviation 1 within each committee-year. Future citations are calculated using all publications by an applicant in the year after grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, quartics in an applicant's total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.

Appendix Table E: What is the contribution of bias and information?
Nonlinear controls for quality and relatedness

	(1)	(2)	(3)
	Awarded	**Score**	**Scored**
Related Permanent Reviewers	0.0036*	0.1460	0.0026
	(0.002)	*(0.095)*	*(0.002)*
Related to Permanent Reviewers × Future Citations	0.0094***	0.2350**	-0.0023
	(0.002)	*(0.109)*	*(0.003)*
Related to Permanent Reviewers × Future Citations^2	-0.0002	-0.0043	0.0009***
	(0.000)	*(0.008)*	*(0.000)*
Related to Permanent Reviewers × Future Citations^3	0.0000	0.0000	-0.0000***
	(0.000)	*(0.000)*	*(0.000)*
Future Citations	0.0136***	0.5356***	0.0281***
	(0.002)	*(0.111)*	*(0.003)*
Future Citations^2	-0.0003	-0.0138	-0.0015***
	(0.000)	*(0.009)*	*(0.000)*
Future Citations^3	0.0000	0.0001	0.0000***
	(0.000)	*(0.000)*	*(0.000)*
Total Related Reviewers	0.0071***	0.2285***	0.0157***
	(0.001)	*(0.059)*	*(0.001)*
Observations	93,558	57,613	93,558
R-squared	0.0953	0.1418	0.1282
Committee × Year × Cycle FE	X	X	X
Past Performance, Past Grants, and Demographics	X	X	X

Notes: Coefficients are reported from a regression of committee decisions (score or funding status) on the variables reported, controlling for meeting level fixed effects and detailed applicant characteristics. The analytic sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2005, for which I have study section attendance data. A reviewer is related to an applicant if the reviewer has cited any of the applicant′s previous research in the 5 years prior to grant review. Future citations are calculated using all publications by an applicant in the year after grant review, with text matching. Applicant characteristics include indicators for sex and whether an applicant′s name is Hispanic, East Asian, or South Asian, quartics in an applicant′s total number of citations and publications over the past 5 years, indicators for whether an applicant has an M.D. and/or a Ph.D., and indicators for the number of past R01 and other NIH grants an applicant has won and indicators for how many she has applied to.