American Economic Journal: Applied Economics 2017, 9(2): 60–92
https://doi.org/10.1257/app.20150421

Expertise versus Bias in Evaluation: Evidence from the NIH[†]

By Danielle Li*

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 examines these issues in the context of peer review at the US National Institutes of Health. I show that evaluators are both better informed and more biased about the quality of projects in their own area. On net, the benefits of expertise weakly dominate the costs of bias. As such, policies designed to limit bias by seeking impartial evaluators may reduce the quality of funding decisions. (*JEL* D82, H51, I10, I23, O38)

A key debate in the economics of innovation focuses on what mechanisms are most effective for encouraging the development of new ideas and products: while patents may distort access to new knowledge ex post, a concern with research grants and other R&D subsidies is that the public sector may make poor decisions about which projects to fund ex ante.

In the United States, the vast majority of public funding for biomedical research is allocated by the National Institutes of Health (NIH), through a system of peer review in which applications are evaluated by committees of scientists working on similar issues. The collective opinion of these scientists is responsible for consolidating thousands of investigator-initiated submissions into a publicly funded research agenda.

But how much should we trust their advice? While reviewers may have valuable information about the potential of projects in their research areas, advice in this setting may also be distorted precisely because reviewers have made so many investments in acquiring their domain expertise. For example, in a guide aimed at scientists describing the NIH grant review process, one reviewer highlights his preference for work related to his own: "*If I'm sitting in an NIH study section, and I*

* Harvard Business School, 211 Rock Center, Boston, MA 02163 (e-mail: dli@hbs.edu). 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, Brigham Frandsen, Alex Frankel, Richard Freeman, Bob Gibbons, Nathan Hendren, Ben Jones, Bill Kerr, Jiro Kondo, Josh Lerner, Niko Matouschek, Xiao Yu May Wang, Ziad Obermeyer, Marco Ottaviani, Dimitris Papanikolaou, Amanda Pallais, Chris Palmer, Michael Powell, Amit Seru, 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.

[†] Go to https://doi.org/10.1257/app.20150421 to visit the article page for additional materials and author disclosure statement or to comment in the online discussion forum.

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*."[1] Alternatively, reviewers may be biased against applicants in their own area if they perceive them to be competitors.

This paper examines the impact of intellectual proximity between reviewers and applicants (hereafter "proximity" or "relatedness") on the quality of funding decisions. I provide evidence that reviewers are better informed about the quality of related candidates but also biased in their favor. Overall, the benefits of expertise appear to outweigh the costs of bias. To show this, I assemble a new, comprehensive dataset linking almost 100,000 NIH grant applications to the committees in which they were evaluated.

My analysis requires two key ingredients: (i) a source of exogenous variation in the intellectual proximity between grant applicants and the more influential members of their review committees and (ii) a measure of quality for grant applications, including that of unfunded applications. Given these, the intuition underlying my empirical work is as follows: if reviewers are only biased and not more informed, then related applicants should receive better (or worse) evaluations regardless of their quality. If related reviewers also have better information, then the effect of working in the same area as a more influential reviewer should differ for high– and low–quality applicants. Strong applicants should benefit from being evaluated by influential reviewers who can more accurately assess their quality, but weak applicants should be hurt for the same reason. In this case, the impact of proximity should be increasing in the quality of applications. I now provide more detail about my proximity and quality measures in turn.

I begin with a baseline measure of the intellectual proximity of individual applicants and reviewers: whether a reviewer has cited an applicant's work in the five years prior to the committee meeting. This captures whether the applicant's work has been of use to the reviewer, but is likely to be correlated with quality because better applicants are more likely to be cited. To identify exogenous variation proximity between candidates and review committees, I take advantage of the distinction between "permanent" and "temporary" members in NIH review committees.[2] Permanent members play a greater role in the grant evaluation process and have more influence on committee scores. I define intellectual proximity to the review committee as the number of *permanent* reviewers that have cited an applicant's work—controlling for the total number of such reviewers, both permanent and temporary. This strategy identifies the causal impact 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 work.

The next part of my analysis considers the role of bias and expertise. To do so, I require information on application quality. The primarily challenge in measuring application quality is doing so for unfunded applications: it is natural, after all, to think that the research described in unfunded applications does not get produced

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

[2] "Permanent" members are not actually permanent; they serve four-year terms. See Sections I and IIIA for a discussion of permanent versus temporary reviewers.

and thus its quality cannot be observed. At the NIH, however, this is not the case. Standards for preliminary results for large research grants 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 compares grant application titles with the titles and abstracts of publications to find research by the same applicant on the same topic as the grant. 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. For consistency, I use this same approach to measure the quality of funded applications as well.

I present three key findings. First, related applicants receive higher scores and are more likely to be funded. Each additional permanent reviewer in an applicant's area, holding constant total related reviewers, increases that applicant's chances of being funded by 2.2 percent. While this may seem like a small effect, it is substantial when viewed relative to reviewers' sensitivity to application quality: it is the same increase in funding probability that we would expect from a one-quarter standard deviation increase in the quality of the application itself, as measured by citations to research that it produces. This large effect suggests that when quality is difficult to assess, reviewer opinions play a comparably large role in funding decisions.

Second, I show that these findings are most consistent with reviewers both having better information about the quality of related applications and being biased in their favor. If reviewers were biased but not better informed, related applications should be more likely to be funded and receive higher scores regardless of their quality. Instead, I find that higher quality applicants benefit more from being evaluated by related reviewers: the impact of proximity on funding likelihood and scores are both increasing in quality. At the same time, if reviewers were more informed but unbiased, I would expect the impact of proximity to be negative for low quality applications. For all but possibly the lowest decile of applications, this is not what I find. Rather, I find that low-quality applications receive no benefit from being evaluated by related reviewers. This pattern strongly suggests that worse information for this group cancels out the positive overall effects of bias.

A potential concern with my empirical strategy is the fact that permanent and temporary reviewers differ significantly in terms of their past publications and citations, and applicants cited by more permanent members, even conditional on total relatedness, have significantly more past citations. If stronger applicants are more likely to be cited by permanent reviewers, then this may be an alternative explanation for my finding that related applicants are more likely to be funded. I provide several different pieces of evidence that this does not drive my results.

First, it is not the case that permanent reviewers are more qualified than temporary reviewers: they tend to have more citations but fewer publications. Second, I show that, conditional on the total number of reviewers an applicant has been cited by, there is no correlation between the number of permanent reviewers that cite an applicant and my text-matched measure of application quality. Further, the impact of relatedness that I estimate in my main tables does not change when I include detailed controls for applicant characteristics and publication histories. Finally, I provide two additional complementary sets of analysis. The first uses reviewer fixed

effects to show that applicants are more likely to be funded when the reviewer that has cited them is serving as a permanent reviewer, compared to when that reviewer is serving as a temporary reviewer. 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 cited by 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 proximity.

Finally, I provide suggestive evidence that the gains associated with reviewer expertise dominate the losses associated with bias: the average quality of applications funded by committees in which a greater share of applicants are related to reviewers tends to be higher than meetings of the same committee in which fewer applicants are related. This suggests that enacting a policy that restricts close intellectual ties may reduce the quality of the NIH-supported research portfolio, as measured by future citations.

Of course, such a conclusion regarding intellectual ties in science may not hold in other settings. For instance, biases associated with financial conflicts of interest may well outweigh any informational advantages that those decision makers may have. Nonetheless, the results in this paper have implications for how organizations treat conflicts of interest. In many settings, personal preferences develop alongside expertise, as a result of individuals self-selecting and making investments into a particular domain. These biases are particularly challenging to address: in contrast with race or gender discrimination, eliminating bias stemming from intellectual ties can directly degrade the quality of information that decision makers have access to. This paper demonstrates that conflict of interest policies necessarily entail efficiency trade-offs.

The question of how organizations should use information from potentially conflicted experts has also been of long-standing theoretical interest (Crawford and Sobel 1982; Li, Rosen, and Suen 2001; Garfagnini, Ottaviani, and Sørenson 2014), but has remained relatively understudied empirically. Emerging work shows that these issues are relevant in many empirical settings ranging from financial regulation to judicial discretion to academic promotion and publication.[3] In these and other settings, it is often challenging to attribute differences in the treatment of connected individuals to either better information or bias because it is difficult to observe the counterfactual quality of decisions that are not made. This paper contributes by studying these issues in the context of public investments in R&D, a setting that is both independently important, and in which various empirical challenges can be more readily overcome.

Finally, there is currently little empirical evidence on how—and how successfully—governments make research investments, and existing studies in this area find mixed results.[4] This paper demonstrates the value of expert advice in this setting.

[3] See, for instance, Agorwal et al. (2014); Hansen, McMahon, and Rivera (2014); Kondo (2006); Fisman, Paravisini, and Vig (2012); Zinovyeva and Bagues (2015); Blanes i Vidal, Draca, and Fons-Rosen (2012); Brogaard, Engleberg, and Parsons (2011); and Laband and Piette (1994).

[4] See Acemoglu (2009), Kremer and Williams (2010), Grilliches (1991), and Cockburn and Henderson (2000) for surveys. Li and Agha (2015) document a positive correlation between scores and outcomes, but Boudreau et. al (2016) and Azoulay, Graff-Zivin, and Manso (2011) raise concerns about the ability to support recognize and foster

I. Context

A. *Grant Funding at the NIH*

The NIH plays an outsized role in supporting biomedical research. Over 80 percent of basic life science laboratories in the United States receive NIH funding, and half of all FDA approved drugs, and over two-thirds of FDA priority review drugs, explicitly cite NIH-funded research (Sampat and Lichtenberg 2011). The decision of what grants to support is made by thousands of scientists who act as peer reviewers for the NIH. Each year, they collectively read approximately 20,000 grant applications and allocate over $20 billion 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 outstanding awards, with 4,000 new projects approved each year. The average size of each award is $1.7 million spread over three to five years.

Because R01s entail such large investments, the NIH favors projects that have already demonstrated a substantial likelihood of success. As evidence of how high this bar is, the NIH provides a separate grant mechanism, the R21, for establishing the preliminary results needed for a successful R01 application. The fact that R01 applications are typically based on research that is already very advanced makes it possible to measure the quality of unfunded grants, which is a key part of my empirical strategy.[5] See Section IIB for a detailed discussion.

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."[6] 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.

novel research. Hegde (2009) considers congressional appropriations for NIH funding. Jacobs and Lefgren (2011) find few effects of individual NIH grants output related to marginally unfunded applicants.

[5] This emphasis on preliminary results was one point of critique that the NIH peer review reform of 2006 was designed to address; under the new system, the preliminary results section has been eliminated to discourage this practice. My data come from before the reform but, anecdotally, it is still the norm to apply for R01s. For a satirical take from 2011, see http://www.phdcomics.com/comics/archive.php?comicid=1431.

[6] 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.

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. Within a study section, an application is typically assigned up to three reviewers who provide an initial assessment of its merit. Permanent members are responsible for performing initial assessments on eight to ten applications per meeting, compared to only one to three for temporary members. The division of committees into permanent and temporary members plays an important role in my identification strategy: permanent reviewers have more influence over the scoring process, but are otherwise similar to temporary members in terms of their scientific credentials. In Section IIIA, I discuss why this might be the case and provide empirical evidence.

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 IIIA.

Once an application has been assigned to a study section, it is assigned to three initial reviewers who 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. 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 scores. This priority score is then converted into a percentile from 1 to 99.[7] 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

[7] At the NIH, a grant's percentile score represents 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, this paper reports inverted percentiles (100 minus the official NIH percentile, e.g., the percent of applications that are *worse*), so that higher percentiles are better.

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 4 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.

B. *Expertise and Bias among Reviewers*

How likely is it that reviewers have better information about the quality of applications in their own area? In informal interviews with scientists serving on peer review committees, I found that the majority of scientists have more confidence in their assessments of related proposals; for many, this translates into speaking with greater authority during deliberations. Reviewers are also more likely to be assigned as initial reviewers for applications in their area, forcing them to evaluate the proposal in more detail. Even when they are not assigned as initial reviewers, many reviewers said they were more likely to carefully read applications in their own area. These mechanisms suggest that reviewers may have greater "expertise" about related applications, either because they know more to begin with or because they pay more attention.

How likely is it that reviewers in my setting are biased? NIH reviewers have little to no financial stake in the funding decisions they preside over, and conflict of interest rules bar an applicant's coauthors, advisers or advisees, or colleagues from participating in the evaluation process.[9] Yet, there is often significant scope for reviewers to have preferences based on their intellectual connections with applicants. Because NIH support is crucial to maintaining a lab, reviewers are well aware that funding a project in one research area necessarily means halting progress in others. Many of the reviewers I spoke with reported being more enthusiastic about proposals in their own area; several went further to say that one of the main benefits of serving as a reviewer is having the opportunity to advocate for more resources for one's area of research. These preferences are consistent with the idea that reviewers have a taste for research that is similar to theirs, or that they perceive this research to be complementary to their own. On the other hand, some study section members also mentioned that other reviewers—not they—were strategic in terms of evaluating proposals from competing labs.[10] This concern is also supported by research indicating that labs regularly compete over scarce resources, such as journal space, funding, and scientific priority (Pearson 2003).

[8] For more details on the NIH review process, see Gerin et al. (2010).

[9] For this reason, I cannot study the impact of these more social connections on funding outcomes.

[10] I conducted 16 informal interviewers with current and past members of NIH study sections. These interviews were off the record but subjects agreed that interested readers could contact the author for more details of these conversations as well as for a full list of the interviewees.

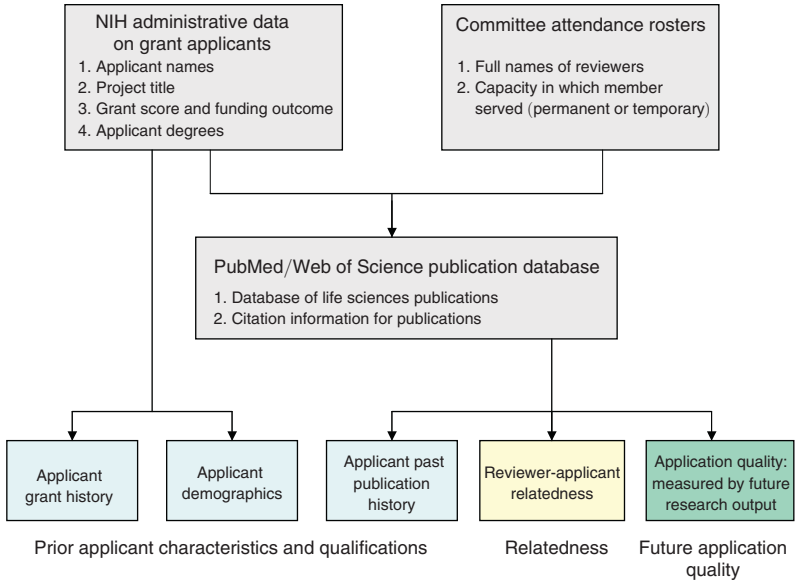

Figure 1. Data Sources and Variable Construction

II. Data

The goal of this paper is to (i) identify how intellectual proximity to influential reviewers affects an applicant's chances of being funded and (ii) to assess the role of bias versus expertise in these funding decisions.

In order to accomplish this, I construct 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.

Table 1—Grant Application Descriptives

	Roster-matched sample		Full sample	
		SD		SD
Sample coverage				
Number of grants	93,558		156,686	
Number of applicants	36,785		46,546	
Years	1992–2005		1992–2005	
Number study sections	250		380	
Number study section meetings	2,083		4,722	
Grant application characteristics				
Percent awarded	26.08		30.48	
Percent scored	61.58		64.04	
Percent new	70.31		71.21	
Percentile score	70.05	18.42	71.18	18.75
Number of publications (text-matched, in first year after grant review)	0.3	0.8	0.3	0.8
Number of citations (up to 2008, to text-matched publications in first year after grant review)	10	51	11	55
Applicant (PI) characteristics				
Percent female	23.21		22.58	
Percent Asian	13.96		13.27	
Percent Hispanic	5.94		5.79	
Percent MD	28.72		29.26	
Percent PhD	80.46		79.69	
Percent new investigators	19.70		20.02	
Number of publications, past five years	15	60	15	55
Number of citations, past five years	416	1,431	423	1,474

Notes: The analytic sample includes new or competing R01 grants evaluated in chartered study sections from 1992 to 2005, for which I have study section attendance data, with social science study sections dropped. The quality of grant applications is measured as follows: number of publications refers to the number of research articles that the grant winner publishes in the year following the grant that share at least one salient word overlap between the grant project title and the publication title. Number of citations refers to the total number of citations that accrue to this restricted set of publications from the time of publication to the end of my citation data in 2008. Applicant characteristics are measured as follows: female, Asian, and Hispanic are all defined probabilistically based on full name. A new investigator is one who has never previously been a PI on an NIH grant. Past publications include any first, second, and last authored articles published in the five years prior to applying for the grant. Number of citations include all citations to those publications, to 2008. Investigators with common names are dropped as are any for which the covariates are missing.

My final sample consists of 93,558 R01 applications from 36,785 distinct investigators over the period 1992–2005. 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. Of these applications, approximately 25 percent are funded and 20 percent are from new investigators, those who have not received an R01 in the past. Seventy percent of applications are for new projects, and the remainder are applications to renewal existing projects. All of these types of applications are typically evaluated in the same study section meeting. Table 1 shows that my sample appears to be comparable to the universe of R01 applications that are evaluated in chartered study sections.

There are three components to these data: (i) a measure of intellectual proximity between applicants and review committees; (ii) a measure of application quality;

(iii) various measures of other applicant characteristics. Sections IIB and IIC first describe how I measure proximity, application quality, and applicant characteristics, respectively. I describe how my empirical strategy uses these measures later in the text, in Sections III and IV.

A. *Measuring Proximity*

I measure the intellectual proximity between an applicant and his or her review committee as the number of permanent reviewers who have cited an applicant's work in the five years prior to the meeting, conditional on the total number of such reviewers. This is a measure of how intellectually connected applicants are to the more influential members of their review committees.

I construct proximity in this way for two reasons. First, using citations to measure proximity has several benefits. Citations capture a form of proximity 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. Citations also capture this form of intellectual connection more finely than other measures, such as departmental affiliation, allowing for more informative variation in proximity. Further, using data on whether the reviewer cites the applicant (as opposed to the applicant citing the reviewer) reduces concerns that my measures of proximity can be strategically manipulated by applicants. Finally, one may also consider more social measures of proximity, such as coauthorship or being affiliated with the same institution. These ties, 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 proximity is a connection that likely matters for grant review but which is not governed by conflict-of-interest rules.

Second, I focus on being cited by permanent reviewers in order to generate variation in proximity that I will argue is unrelated to an applicant's quality. This is because the total number of reviewers who cite an applicant is likely to be correlated with quality: better applicants may be more likely to be cited and may, independently, submit higher quality proposals. By controlling for the total number of reviewers who cite an applicant, I compare applicants that differ in their proximity to more influential reviewers, but not in the quality of their work. I discuss this strategy and provide evidence for its validity in Section IIIA.

Table 2 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 of whom are 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.

B. *Measuring Quality*

I measure application quality using the number of publications and citations that the research it proposes produces in the future. The key challenge to constructing this measure is finding a way to use ex post publication data to assess the ex ante

Table 2—Committee Descriptives

	Roster matched sample	
		SD
Number of reviewers	18,916	
Number of applications	53.73	17.31
Composition		
Number of permanent reviewers per meeting	17.23	4.52
Number of temporary reviewers per meeting	12.35	7.44
Number of meetings per permanent reviewer	3.69	3.03
Number of meetings per temporary reviewer	1.78	1.30
Relatedness		
Total number reviewers who cite applicant	1.94	2.81
Number of permanent reviewers who cite applicant	1.11	1.73
Number of permanent reviewers cited by applicants	4.12	5.32
Number of temporary reviewers cited by applicants	4.12	5.09

Notes: See notes to Table 1 for details on the sample. A reviewer is defined as citing an applicant if the reviewer has published a paper in the past five years that has cited any of the applicant's papers. An applicant is defined as citing a reviewer if the applicant has published a paper in the past five years that cites the reviewer's work.

quality of applications. For example, how does one measure the quality of applications that are unfunded if publications cannot acknowledge grants that do not exist?

To overcome this challenge, I develop a way to identify publications associated with research described in the preliminary results section of each grant application. As discussed in Section I, this is possible because it is extremely common for scientists to submit grant proposals based on nearly completed research, especially for the large R01 grants that I study. To find these publications, I first identify all research articles published by a grant's primary investigator. I then use a text matching technique to identify articles on the same topic as the grant application. This is done by comparing each publication's title and abstract with the title of the applicant's grant proposal. For instance, if I see a grant application entitled "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," I label these publications as related. In my baseline specifications, I require that publications share at least four substantive (e.g., with articles and other common words excluded) overlapping words with the grant project title or its abstract. On average project titles have 10 substantive words, and abstracts have 50. I describe the text matching process I use in more detail in online Appendix B, and show robustness to alternative matching thresholds.

Text matching makes it possible to identify publications associated with unfunded grants. Funding itself, however, may enable scientists to produce more or better research, making it difficult to disentangle the ex ante quality of an application from its ex post publications and citations. This is a particularly important concern for this paper because it affects my ability to distinguish bias from expertise. To see this, suppose that two scientists submit proposals that are of the same ex ante quality, but that one scientist is related to a more influential reviewer, who funds him out of bias. The funding, however, allows this scientist to publish more articles, meaning that an

econometrician that examines ex post outcomes may mistakenly conclude that his proposal was ex ante better. This would lead me to mistake bias for expertise.

Funding may improve a scientist's output both by subsidizing research on topics unrelated to the original application or by supporting work in the same area. I deal with both of these concerns. First, text matching restricts the set of publications I use to assess an application's quality to those that are on the same topic. Second, I also restrict the set of publications I use to assess quality to those published within one year of grant review. This short time window identifies articles based on research that was already completed or underway at the time the application was written. These unlikely to be directly supported by the grant.[11]

This procedure is designed to isolate the set of publications based on the ideas outlined within a grant application. I then use citation information to assess the quality of these ideas. Specifically, for each application, I count the total number of publications, the total number of citations these publications receive through 2012, and the number of "hit" publications, where a hit is defined as being in the ninetieth, ninety-fifth, or ninety-ninth percentiles of the citation distribution relative to all other publications in its cohort (same field, same year). 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.

Figure 2 plots the relationship between ex ante quality and an application's likelihood of funding, measured using future citations to text-matched publications, and shows that, on average, better applications are more likely to be funded. This provides evidence that, on average, peer reviewers are able to identify high quality applications (Li and Agha 2015). Despite this, online Appendix Figure B shows that many unfunded applications go on to generate more citations and publications than funded applications. This can also be seen in online Appendix Table A, which reports detailed comparisons of the distribution of citations and publications associated funded and unfunded applications: the average funded grant produces 10.3 citations and 0.33 publications, compared with 8.7 citations and 0.26 publications for unfunded applications.

One concern with these figures is that it is possible that funding itself impacts my measure of quality, making it appear as though funded applicants were higher quality ex ante when in fact they were not. To provide evidence that this is not the case, I examine a fuzzy regression discontinuity in funding outcomes around the applicant score payline. If my measure of quality is capturing a grant application's ex ante quality, then it should vary smoothly at this discontinuity. If, instead, funding impacts my measure of quality, then I would see a discontinuous jump in quality at

[11] 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. See http://grants.nih.gov/grants/grants_process.htm. 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. My results are robust to other windows. See online Appendix Tables F and G.

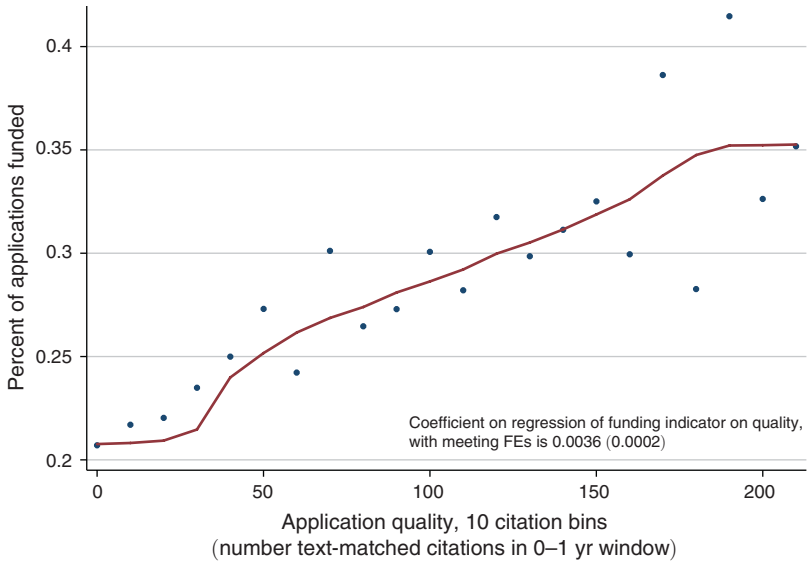

FIGURE 2: RELATIONSHIP BETWEEN APPLICATION QUALITY AND GRANT FUNDING

Notes: Application quality is measured using text-matched publications published within one year of grant review, and then computing all citations to this subset of publications, to 2008. See Section IIB and online Appendix B for additional details about how quality is constructed. The *x*-axis is the number of such citations, divided into ten-citation bins. Each dot represents the percent of applications that are funded, among applications in the same ten-citation bin. Citations are top coded at the ninety-ninth percentile. This is done for legibility only; analyses use the full distribution of both variables. The plotted line presents a locally smoothed polynomial estimated using a Epanechnikov kernel.

the funding threshold. Figure 3 demonstrates that my quality measure is smooth at this funding threshold, even though the likelihood of being funded changes discontinuously. In both panels of Figure 3, the applicant's percentile score, the running variable, is plotted along the *x*-axis. For this figure, the score has been rounded to the nearest integer and re-centered so that 0 represents the funding threshold relevant for that particular application (funding thresholds can differ based on how much funding a particular Institute has been allocated for particular research areas). In panel A of Figure 3, I plot the proportion of applications with that centered score that are ultimately funded, and there is a clear discontinuity at the funding threshold. This is a not a sharp discontinuity because grants can be funded out of order.[12] Panel B of Figure 3 plots centered scores against the average measured quality for each score group. In general, there is a positive slope, indicating that better scoring applications tend to have higher quality, but I find no evidence of a discontinuity at the funding threshold.

The accompanying statistical test is reported in online Appendix Table C. I show that there is no effect of being over the funding threshold, conditional on scores, on measured quality, and further find no effect of funding on measured outcomes,

[12] For example, new investigators may be funded ahead of established investigators with the same score if the funding Institute wants to encourage submissions from younger scientists.

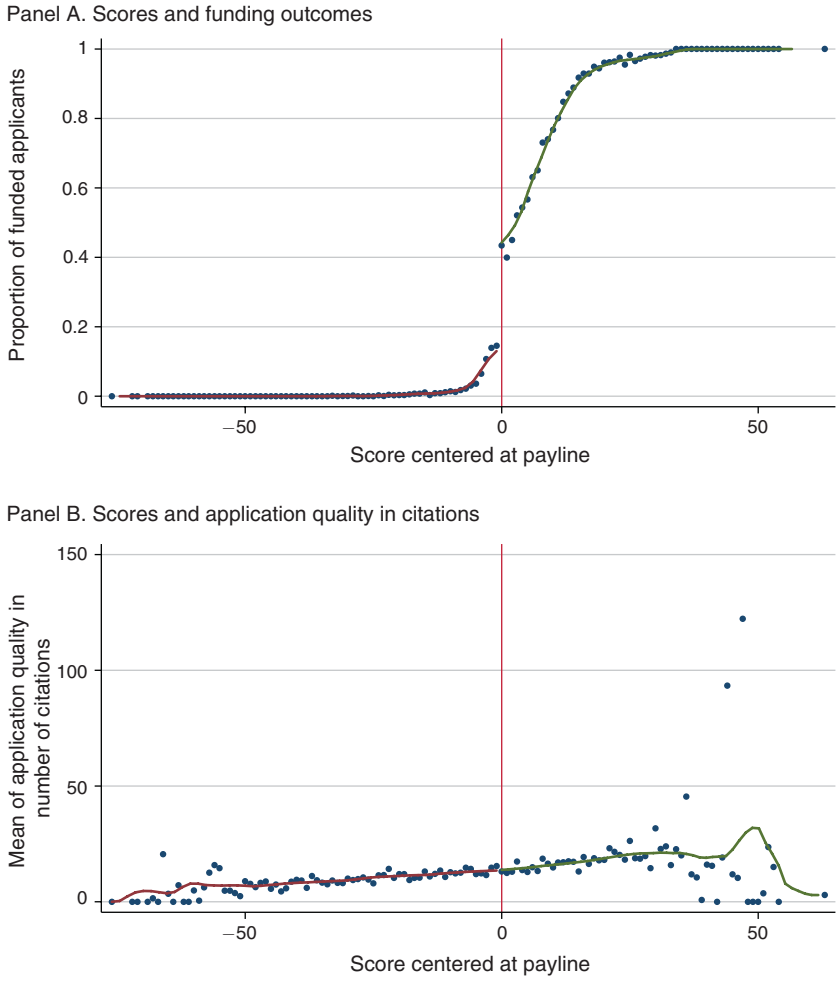

FIGURE 3. REGRESSION DISCONTINUITY IN APPLICATION QUALITY, BY SCORE

Notes: The top panel graphs the relationship between scores and the proportion of applicants in that score range who are funded. Scores normally in percentiles from 0 to 100 are centered at their meeting specific payline, which is set to be 0. The data are then collapsed by round values of this centered score. The *y*-axis plots the proportion of applicants that are funded for applications that share the same centered score, rounded to the neared integer value. The line is calculated using Epanechnikov kernel weighted local polynomial smoothing. The bottom panel is constructed analogously but with mean application quality using text-matched citations as the dependent variable.

using the funding threshold as an instrument. Together with Figure 3, this finding mitigates concerns that my measure of quality is directly affected by funding.

C. *Measuring Applicant Characteristics*

Finally, 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 discernible 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. Sample descriptives for these variables are also provided in Table 1.

III. Identifying the Causal Impact of Proximity

The first part of my empirical analysis estimates the effect of intellectual proximity to more influential committee members:

$$(1) \quad Assessment_{icmt} = a_0 + a_1 Proximity\_to\_Permanent_{icmt} + a_2 Total\_Proximity_{icmt} + \mu X_{icmt} + \delta_{cmt} + e_{icmt}.$$

$Assessment_{icmt}$ is a variable describing the committee's assessment (either the funding status, score, or whether an application was scored at all) of applicant i whose proposal is evaluated by committee c in meeting m of year t. $Proximity\_to\_Permanent_{icmt}$ is the number of permanent reviewers who have cited an applicant's work in the five years prior to the committee meeting, and $Total\_Proximity_{icmt}$ is the total number of such reviewers. 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 MD and/or a PhD; 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.

My coefficient of interest is a_1. a_1 compares the funding outcomes of scientists whose applications are reviewed in the same meeting, who have similar past performance, and who, while cited by the same total number of reviewers, differ in their proximity to permanent reviewers.

[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.

Table 3—Characteristics of Permanent and Temporary Reviewers

Reviewer characteristics		Permanent	Temporary	*p*-value
Demographics				
Percent female		31.68	24.28	0.00
Percent Asian		14.99	13.08	0.00
Percent Hispanic		6.40	5.05	0.00
Education				
Percent MD		27.42	25.85	0.00
Percent PhD		79.00	81.00	0.00
Past citations				
Mean		1,470	1,375	0.00
Median		606	590	0.09
Fifth		0	0	0.00
Ninety-fifth		5,459	5,002	0.00
Past publications				
Mean		53	57	0.05
Median		22	21	0.00
Fifth		0	0	0.00
Ninety-fifth		154	152	0.67
Reviewer transitions (1997 to 2002 subsample)	% perm. in the past	% perm. in the past	% perm. in the past	% perm. in the past
Current permanent	61.87	63.71	38.11	35.45
Current temporary	16.25	41.30	32.73	50.13

Notes: Observations are at the reviewer-study section meeting level. The sample includes all reviewers in chartered study sections from 1992 to 2005, for which I have study section attendance data. Number of reviewer publications include any first, second, and last authored articles published in the five years prior to the study section meeting date for which the reviewer is present. Number of citations refers to all citations accruing to those publications, to 2008. Reviewer characteristics are measured as follows: female, Asian, and Hispanic are all defined probabilistically based on full name. MD and PhD are defined based on rosters in which a reviewer's degree follows his or her name. Reviewer transitions are calculated based on whether a reviewer is present in the roster database during the full sample years from 1992 to 2005. The set of reviewers used in this calculation are those present in meetings from 1997 to 2002 in order to allow time to observe members in the past and future within the sample.

A. *Permanent versus Temporary Reviewers*

In order for a_1 to identify a causal effect, I need to show that proximity to permanent reviewers is not correlated with other characteristics that may also impact funding outcomes, conditional on proximity to all reviewers. Foremost, one may be concerned that being cited by permanent reviewers signals higher quality than being cited by temporary reviewers.

Before providing direct evidence to refute this claim, I first explore the characteristics of permanent versus temporary reviewers. Table 3 compares demographic, educational, and publication characteristics of permanent and temporary reviewers. Permanent reviewers tend to be somewhat more diverse: 32 percent are women, relative to 25 percent of temporary reviewers; 15 percent are Asian; and 6.4 percent Hispanic, compared to 12 percent and 5.1 percent, respectively, among temporary reviewers.[14] This is difference is likely due to the fact that the NIH makes

[14] I do not have information on whether a reviewer is black because my demographic variables come from analyzing names. Black names are more difficult to recognize relative to Asian or Hispanic names.

a conscious effort to ensure diversity on their review panels. Similarly, permanent members are also slightly more likely to be medical doctors with some clinical experience, relative to PhDs. This difference, 27 percent versus 26 percent, is statistically significant but small.

Table 3 also shows that permanent reviewers appear to have slightly stronger publication histories, as measured by the number of publications in the previous five years and the number of forward citations to those publications, measured up to the year 2008. For example, permanent reviewers average 1,470 citations to past publications compared to 1,375 for temporary reviewers. This difference of about 100 citations, while significant, is small relative to the standard deviation of citations, which is approximately 3,000 for both groups. Meanwhile, reviewers appear to be similar in terms of their number of publications. The full distribution is plotted in online Appendix Figure A.

The findings in Table 3 so far highlight several potentially important differences in the qualifications and characteristics of permanent and temporary reviewers. The bottom panel of Table 3 provides some evidence for why permanent and temporary reviewers still nonetheless appear broadly similar: during the course of their career, the same person often serves in both capacities. This difference does not simply reflect a progression from temporary to permanent as reviewers age. Rather, for the set of reviewers observed in the middle of my sample, between 1997 and 2002, 35 percent of permanent reviewers in a given meeting will serve as temporary reviewers in a future meeting while 40 percent of temporary reviewers in a given meeting will serve as permanent reviewers in a future meeting. These common changes in reviewer status across meetings mitigates concerns that permanent reviewers are categorically different from temporary members.

My next set of results explore the matching of applicants to permanent or temporary reviewers, which is nonrandom in two ways. First, rosters listing the permanent (but not temporary) reviewers associated with a study section are publicly available, meaning that 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 intellectually 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, assignment is also nonrandom because study section administrators assign applications to initial reviewers on the basis of (i) intellectual match and (ii) 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

Table 4—Applicant Characteristics, by Number and Composition of Related Reviewers

Dep. var.: Applicant characteristics	Female (1)	Asian (2)	Hispanic (3)	MD (4)	PhD (5)	New investigator (6)	Previous publications (7)	Previous citations (8)	Application quality (9)	Funding propensity (× 100) (10)
Number of proximate permanent reviewers	−0.0003	−0.0013	−0.001	0.0119	0.0036	−0.0013	−0.0057	52.1057	0.0065	0.0433
	(*0.002*)	(*0.002*)	(*0.0010*)	(*0.011*)	(*0.010*)	(*0.002*)	(*0.360*)	(*13.250*)	(*0.0060*)	(*0.0320*)
Observations	93,558	93,558	93,558	93,558	93,558	93,558	93,558	93,558	93,558	93,558
R^2	0.0627	0.0381	0.0252	0.0452	0.0588	0.0679	0.0665	0.2071	0.0312	0.2248
Meeting FEs	X	X	X	X	X	X	X	X	X	X
Number of proximate reviewer FEs	X	X	X	X	X	X	X	X	X	X

Notes: See notes to Table 1 for details about the sample. Coefficients are reported from a regression of applicant characteristics on the number of permanent members related to an applicant, controlling for meeting-level fixed effects, and fixed effects for proximity to all reviewers. Proximity to permanent reviewers is defined as the number of permanent reviewers who have cited any of the applicant's research in the five years prior to grant review. Outcome variables are defined as follows: female, Asian, and Hispanic are all defined probabilistically based on full name. MD and PhD are from administrative grant application records. A new investigator is one who has never previously been a PI on an NIH grant. Previous publications include any authored articles published by the applicant in the five years prior to applying for the grant. Previous citations include all citations to those publications, to 2008. Application Quality is text-matched citations to grant applications, in hundreds, as described in the text. Funding propensity is an aggregate variable constructed from regressing funding outcomes on all demographic variables, education, fixed effects for decile bins for both past publication and citations, and fixed effects for number of past R01 and other grants, and taking the fitted values of funding likelihood from this regression.

reviewer. Both applicants would then be related to one reviewer in total but, in this example, the fact that one applicant works in the same area as a temporary member is actually correlated with potentially unobserved aspects of quality.

Table 4 shows that this may be a potential concern. Table 4 regresses the demographic characteristics and publication records of applicants on the number of permanent reviewers they have been cited by, conditional on meeting fixed effects and fixed effects for the total number of citing reviewers. This compares two applicants to the same meeting, who have been cited by the same total number of reviewers, but who differ in the number of citing permanent and temporary reviewers. I find that applicants cited by more permanent members do not differ on any demographic or educational characteristics, but that there is a significant relationship between being cited by permanent reviewers and previous citations in column 8: each additional permanent reviewer who cites an applicant is associated with 52 more citations, relative to a mean of 1,429 citations, or about 3.6 percent. Column 9, however, shows that, conditional on relatedness to all reviewers, relatedness to permanent reviewers is not predictive of the quality of applications themselves. Next, column 10 shows that proximity to permanent reviewers is not correlated with an application's predicted propensity to be funded. To show this, I regress application funding on applicant gender, race, education, past publications, past citations, and past grant history to construct an index describing his or her propensity to be promoted. I find that there is no relationship between proximity to permanent reviewers and this index.

Despite appearing similar in terms of qualifications, the structure of the NIH is such that permanent reviewers play a larger role in making funding decisions. There are several reasons why this is the case. Most basically, permanent reviewers

do more work. As discussed in Section I, 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 (i) determine whether a grant application even merits discussion by the full group and (ii) serve as the starting point for discussion. Study sections also evaluate 40 to 80 applications per meeting, meaning that it is unlikely that reviewers have had a chance to carefully read proposals to which they have not been officially assigned. In many study sections, moreover, 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. As a result, being evaluated by a close permanent reviewer can impact how an application is treated, even though it is not correlated with a grant's underlying quality.

The results in Tables 3 and 4 raise a potential concern: permanent and temporary reviewers have statistically different publication histories and permanent reviewers are more likely to be assigned to applicants with more previous citations themselves. Before testing whether relatedness impacts funding decisions, I present several pieces of evidence to show that this is unlikely to bias my results.

First, the results in Table 3 should not be interpreted as presenting a case that permanent members have stronger publication histories than temporary reviewers. Rather, permanent reviewers have more past citations while temporary reviewers have more past publications. More generally, differences in the qualifications of permanent and temporary reviewers would only impact my results if it translates into a relationship between proximity to permanent reviewers and an application's quality itself. Table 4, online Appendix Table B, and online Appendix Figure C provide direct evidence that this is not the case. In particular, the upper left-hand panel of online Appendix Figure C shows the distribution application quality (as defined in the previous section) for applicants cited by exactly one reviewer. The solid line shows the distribution of quality among applicants cited by one permanent reviewer and the dotted line does so for those cited by one temporary reviewer. These distributions are statistically indistinguishable: a Kolmogorov–Smirnov test cannot reject the null that these two distributions are equal. 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 online Appendix Figure C repeat this exercise for applicants who have been cited by a total of two reviewers. In this case, there are now three possibilities: the applicant has been cited by two temporary reviewers, two permanent, or one of each. In all of these cases, the distribution

of applicant quality is statistically similar.[15] Online Appendix Table B compares means and other percentiles of these distributions.

Further, if my measure of relatedness were correlated with unobserved factors that also impact an application's likelihood of funding, then I would expect the inclusion of applicant characteristics (publication history, demographics, etc.) to impact my estimates. In Section IIIB, I show that this is not the case: the impact of relatedness that I estimate does not change when I include detailed controls for applicant characteristics and publication histories.

Finally, I provide two additional complementary sets of analysis. The first uses reviewer fixed effects to show that applicants are more likely to be funded when the reviewer that has cited them is serving as a permanent reviewer, compared to when that reviewer is serving as a temporary reviewer. This is presented in Table 6 and discussed in Section IIIB. 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 cited by 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 proximity. This is presented in online Appendix Table K and discussed in Section V, and online Appendix Section D.

B. *Impact of Proximity: Results*

Table 5 estimates the effect of intellectual proximity on funding and scores. The first column reports the raw within-meeting association between proximity to permanent reviewers and an applicant's likelihood of being funded. Without controls, each additional permanent reviewer who has cited an applicant is associated with a 3.3 percentage point increase in the probability of funding, from an overall average of 21.4 percent. This translates into a 15.4 percent increase. Most of this correlation, however, reflects differences in quality—better applicants are more likely to be cited by reviewers. Column 2 adds a full set of fixed effects for the total number of reviewers who have cited an applicant. Once I do this, my identifying variation comes from changes to the *composition* of the reviewers who have cited an applicant—effectively the impact of switching the reviewers an application is related to from temporary to permanent. With these controls, the impact of being cited by a permanent reviewer falls to 0.0050, but remains significant. This says that comparing two scientists reviewed in the same meeting, cited by the same number of applicants, being cited by an additional permanent reviewer increases an applicant's likelihood of funding by $0.0050/0.214$ or 2.3 percent.

To appreciate the magnitude of this effect, it is useful to consider how sensitive funding decisions are to changes in application quality. Recall from Figure 2 that an application's likelihood of funding is increasing in its quality. A regression of funding on application quality, holding constant meeting fixed effects, says that an applicant's likelihood of funding increases by 3.6 percentage points for every 100

[15] Approximately 75 percent of my sample of applications are cited by two or fewer reviewers. This pattern also holds for applicants cited by three or more reviewers.

Table 5—What Is the Impact of Proximity on Committee Assessments?

	1(Score is above the payline) Mean = 0.214, SD = 0.410			Score Mean = 71.18, SD = 18.75			1(Scored at all) Mean = 0.640, SD = 0.480		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Number of proximate permanent reviewers	0.0328 (*0.001*)	0.0050 (*0.002*)	0.0047 (*0.002*)	1.1067 (*0.054*)	0.1641 (*0.094*)	0.1611 (*0.093*)	0.0500 (*0.002*)	0.0012 (*0.002*)	0.0011 (*0.002*)
Observations	93,558	93,558	93,558	57,613	57,613	57,613	93,558	93,558	93,558
R^2	0.0630	0.0688	0.0950	0.1186	0.1224	0.1439	0.0775	0.0899	0.1340
Meeting FEs	X	X	X	X	X	X	X	X	X
Number of proximate reviewer FEs		X	X		X	X		X	X
Past performance, past grants, and demographics			X			X			X

Notes: See notes to Table 1 for details about the sample. 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. Proximity to permanent reviewers is defined as the number of permanent reviewers who have cited any of the applicant's research in the five years prior to grant review. "Past performance, past grants, and demographics" include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, indicator variables for deciles of an applicant's total number of citations and publications over the past five years, indicators for whether an applicant has an MD and/or a PhD, and indicators for the number of past R01 and other NIH grants an applicant has won, as well as indicators for how many she has applied to.

citation increase in quality. A back of the envelope calculation comparing this with the coefficient in column 2 says that proximity helps an applicant get funded by as much as would be expected from a 13.8 citation (or one-fourth standard deviation) increase in the quality of the application itself. This sizable effect suggests that when reviewers cannot easily predict the quality of applications, other factors like relatedness play a comparably larger role.

Finally, column 3 adds controls for applicant characteristics such as demographics, education, past publications, past citations, and grant history. If proximity to permanent members is not correlated with applicant characteristics conditional on proximity to all reviewers, we would not expect the addition of these controls to alter our estimates. We find that this is indeed the case: the addition of this large set of controls changes the estimated coefficient from 0.0050 to 0.0047; the percentage change decreases to 2.2 percent.

Columns 6 and 9 report estimates of the impact of proximity on the score that an application receives and whether an application is scored at all. I find a statistically significant but relatively small effect of proximity in scores: switching to a proximate permanent reviewer increases, holding total proximity constant, an applicant's score by 0.16 points or about 1 percent of a standard deviation. I find no evidence that relatedness increases the overall probability that an applicant is scored at all (recall that about 40 percent of applicants are deemed sufficiently weak that they are not given a score).

Table 6 considers an alternative test: do applicants fare differently when the reviewer they have been cited by serves as a permanent member, compared to when that same reviewer serves as a temporary member? Finding that proximity matters more when reviewers are permanent would strongly suggest that reviewer preferences influence funding decisions, independently of the characteristics of applications themselves.

Table 6—What Is the Impact of Reviewer Status on Funding Outcomes? Reviewer Fixed Effects

	Proportion of proximate applicants who are funded Mean: 0.37, SD: 0.36 (1)	Average score of proximate applicants Mean: 73.3, SD: 14.3 (2)
Reviewer is permanent	0.003	0.336
	(*0.001*)	(*0.144*)
Observations	15,871	15,870
R^2	0.954	0.571
Reviewer FEs	X	X
Past performance, past grants, and demographics	X	X

Notes: This table examines how outcomes for applicants cited by reviewers vary by whether the citing reviewer is serving in a permanent or temporary capacity. The sample is restricted to 4,909 reviewers who are observed both in temporary and permanent positions. An applicant is said to be proximate if a reviewer has cited that applicant in the five years prior to the study section meeting in which the reviewer and applicant are matched. "Past performance, past grants, and demographics" include indicators for sex and whether an applicant's name is Hispanic, East Asian, or South Asian, indicator variables for deciles of an applicant's total number of citations and publications over the past five years, indicators for whether an applicant has an MD and/or a PhD, and indicators for the number of past R01 and other NIH grants an applicant has won, as well as indicators for how many she has applied to.

To test this, 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 reviewer meeting, I identify the grant applicants cited by that reviewer within the previous five years and calculate the proportion of those applications who are funded, and the average score of those applications from cited scientists. I then regress this outcome variable on an indicator for whether or not the reviewer is serving as a permanent member during that meeting, controlling for reviewer fixed effects and average demographic, publication, and grant characteristics about the set of related applicants, weighted by the number of related applicants. In column 1 of Table 6, I show that a larger proportion of related applicants are funded when the reviewer is permanent rather than temporary. In column 2, I also find that the average score of related applicants increases when a reviewer is permanent.

IV. Expertise versus Bias

My results in the previous section show that applicants who work in the same area as more influential committee members are more likely to be funded. Is this a problem for peer review? Not necessarily. Reviewers may advocate for candidates in their area simply because they are more confident in their assessments: receiving more precise signals about related applicants allows reviewers to form higher posterior expectations about their quality. This could lead to a greater proportion of related applicants falling above the funding bar even in the absence of bias. Because this type of behavior improves the quality of peer review, while biases do not, it is important to distinguish between the two explanations.

The results in Tables 5 and 6 provide initial evidence that this effect is not simply driven by better information. To see this, suppose that related reviewers are unbiased but better informed. In this case, the more precise signals they receive should increase the variance of their posteriors over the quality of applications from related applicants, but should not change the mean, as long as scores are linear in beliefs. While this would lead to a greater proportion of related applicants being above the funding threshold, it should not change the average score for these applicants, because reviewers would also have stronger negative posteriors as well. By contrast, Tables 5 and 6 show that average scores are also higher for related applicants.

A more direct evaluation of the role of expertise and bias involves using information on applicant quality. In general, related reviewers can be (i) only biased, (ii) only better informed, or (iii) both.[16] If reviewers are only biased, they will give better assessments to related applicants regardless of their quality. The impact of relatedness should be to increase an applicant's likelihood of funding for any level of quality. By contrast, reviewers can also be unbiased and better informed about the quality of related applicants. In this scenario, high quality applicants benefit from being evaluated by related reviewers who can more accurately observe their quality, but low quality applicants are hurt for the same reason. The estimated impact of relatedness should be increasing in quality and also be negative for particular low quality applications. Finally, related reviewers may be both biased and better informed. Bias means that assessments are shifted up for related applicants regardless of quality, but information means that there is also a stronger relationship between quality and assessments for intellectually related applicants. In this scenario, the impact of relatedness is still increasing in application quality, but it may not necessarily ever be negative. In this section, I examine which of these scenarios best characterizes the committee evaluations I observe.

An important caveat to note about the following analysis is that it is not possible to pin down the exact amount of bias or information without further parametric assumptions about the nature of reviewers' beliefs. In online Appendix F, I present and estimate a model of grant allocation with biased experts in which the separate contribution of bias and expertise can be precisely estimated. In this model, I make a stronger set of distributional assumptions that allow me to attribute differences in the slope of the relationship between quality and funding outcomes (between related and unrelated applicants) to the value of expertise and to attribute level differences in committee assessments to the impact of bias. Because this model relies on strict distributional assumptions, I proceed with the more flexible approach of splitting my sample into quality bins.

A. *Expertise versus Bias: Results*

In this section, I examine how the impact of relatedness differs by application quality. Figure 4 presents the relationship between quality and funding estimated on

[16] For simplicity and because this is suggested by my earlier findings, I show the case in which reviewers are positively biased and better informed, although it is theoretically possible that related reviewers can be negatively biased or less informed.

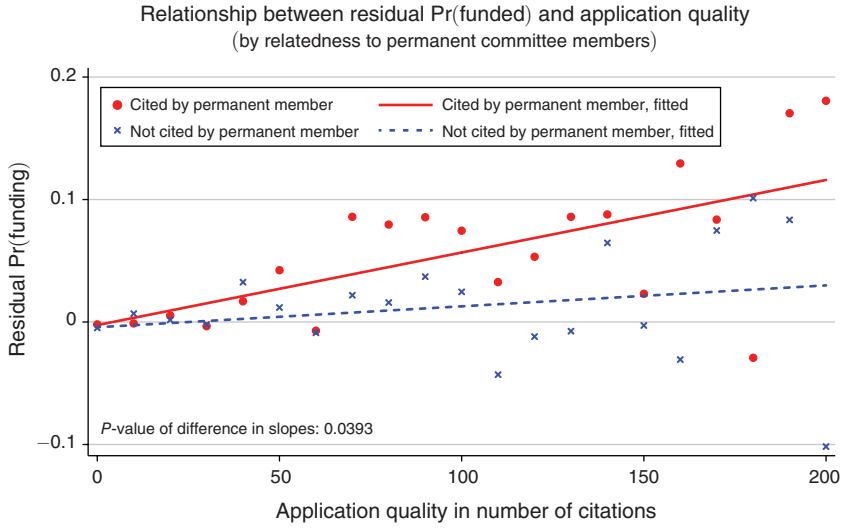

FIGURE 4. REALIZED COMMITTEE ASSESSMENTS BY APPLICATION QUALITY

Notes: This figure examines committee assessments by application quality. The *x*-axis represents ten citation bins of application quality, where citations are measured from text-matched publications within one year of grant review. The *y*-axis represents an application's residual likelihood of funding, after taking into applicant demographics, education, publication history, grant history, and total number of proximate reviewers. For each quality bin, each dot represents the mean residual funding likelihood for applicants with that quality who have been cited by at least one permanent member. The X's represent the same for applicants who have not been cited by a permanent member. See the text in Section IVA for more discussion.

my actual data. The *x*-axis corresponds to application quality, measured in citations to text-matched publications. The *y*-axis plots an application's residual likelihood of being funded, after taking into account demographics, education, publication record, grant history, and total number of related reviewers.

Figure 4 first shows that an applicant's chances of funding, controlling for applicant characteristics, are increasing in quality for both related and unrelated applicants. I find a stronger slope for applicants cited by permanent reviewers, indicating that funding decisions are more responsive to quality when grants are evaluated by more closely related reviewers. This difference between these slopes is significant at the 5 percent level.

The relationship I estimate between quality and funding for related and unrelated candidates suggests that reviewers are both better informed and more biased when it comes to evaluating related applicants. First, the impact of relatedness is increasing in applicant quality: high-quality applicants are more likely to be funded when evaluated by related reviewers, compared to equally high-quality applicants evaluated by unrelated reviewers. This fact is suggestive of expertise. At the same time, if related reviewers were unbiased, we would expect low-quality applications from related candidates to be penalized. The fact that low-quality applications do not benefit but are not hurt suggests that bias can undo some potential benefits of

Table 7—What Is the Impact of Proximity by Application Quality?

		Quartiles of residual application quality			
	All (1)	Bottom (2)	Second (3)	Third (4)	Top (5)
Panel A. Dependent variable: 1(*score above payline*)					
Number of proximate permanent reviewers	0.0047	0.0004	0.0012	0.0038	0.0126
	(*0.002*)	(*0.004*)	(*0.004*)	(*0.005*)	(*0.004*)
Observations	93,558	22,463	23,929	23,360	23,806
R^2	0.0950	0.1680	0.1415	0.1311	0.1613
Panel B. Dependent variable: Score					
Number of proximate permanent reviewers	0.1611	−0.0466	0.0239	0.1927	0.6509
	(*0.093*)	(*0.169*)	(*0.220*)	(*0.293*)	(*0.216*)
Observations	57,613	16,081	14,593	12,056	14,883
R^2	0.1439	0.2139	0.2222	0.2612	0.2223
Panel C. Dependent variable: 1(*scored at all*)					
Number of proximate permanent reviewers	0.0011	0.0007	−0.0064	0.0026	0.0046
	(*0.002*)	(*0.004*)	(*0.004*)	(*0.007*)	(*0.005*)
Observations	93,558	22,463	23,929	23,360	23,806
R^2	0.1340	0.1771	0.1681	0.1667	0.1924
Meeting FEs	X	X	X	X	X
Number of proximate reviewer FEs	X	X	X	X	X
Past performance, past grants, and demographics	X	X	X	X	X

Notes: See notes to Table 1 for details about the sample. Coefficients are reported from a regression of committee decisions (above payline, score, or scored at all) on number of proximate permanent reviewers, controlling for meeting-level fixed effects and fixed effects for total proximity. Panel A regressions use the same specification as column 2 in Table 5; panel B uses the same specification as column 5 in Table 5; panel C uses the same specification as column 8. In particular, column 1 of this table replicates results from Table 5. Columns 2 through 5 split the sample based on quartiles of residual application quality. To calculate residual quality, I regress application quality in citations on dummies for female, Hispanic, east Asian, south Asian, MD, PhD, fixed effects for decile bins for both past publication and citations, and fixed effects for number of past R01 and other grants, and taking the residuals from this regression. Regressions also control for these variables directly.

expertise. The remainder of this section explores this pattern more rigorously and addresses alternative explanations.

Table 7 splits my main sample into quartiles of quality and separately estimates the impact of proximity for each subsample. These quality quartiles are constructed with respect to residual future citations: applications in the highest quartile are those that received more citations than would be predicted given the applicant's initial publications, demographics, and grant history. If closer reviewers had more expertise, we would expect proximity to be particularly beneficial for strong applications that might otherwise not look competitive based on ex ante observables.

Column 1 of Table 7 replicates results from columns 3, 6, and 9 of Table 5—the impact of proximity for the entire sample, controlling for meeting fixed effects, fixed effects for total proximity, and controls for applicant characteristics. Columns 2–5 confirm the pattern laid out by Figure 4. I find no effect of proximity for the first two quartiles, a positive but statistically insignificant effect for third quartile applications, and a large and significant effect for the applications in the top quartile. This means that strong applications benefit much more from being evaluated

by an intellectually related reviewer, relative to applications that are weaker. The coefficient on relatedness for top quartile applications is 0.0126 (column 5). In percentage terms, this says that each additional related permanent reviewer increases the funding likelihood of top quartile applicants by 1.26 percentage points, from top-quartile mean of 22.3 percent—a 5.7 percent increase. By contrast, I estimate a 2.2 percent increase for the whole sample (column 1) and a zero effect for bottom quartile applicants (column 2).

There are, however, several alternative explanations that would generate this pattern even in the absence of expertise. First, my estimates could be misleading if my measure of quality is directly impacted by whether or not an application is funded. To see this, suppose that many low-quality applications from related applicants are funded because of bias. If funding itself makes these applications appear stronger ex post, then they may mistakenly be categorized as high-quality applications. If this were the case, then what should be identified as a large effect of proximity for low-quality applicants becomes identified instead as a large effect for high-quality applicants. This would lead me to mistake bias for expertise. In discussing my measure of quality in Section IIB, I provided evidence that my quality measure was not contaminated by funding status. Specifically, Figure 3 shows that my measure of quality does not discontinuously change as a result of funding, making it very unlikely that my findings in Table 7 are driven by this pattern.

Another possible explanation for my findings is that bias is more pivotal for high-quality applications simply because they are closer to the funding threshold. Were this the case, the impact of proximity would appear to be increasing in quality even if close reviewers were no better informed. I provide several explanations for why this is unlikely to drive my results.

First, it is not empirically the case that only the highest quality grants have a chance of receiving funding. If this were the case, we would expect a grant's probability of funding to be low up to a threshold, then peak and remain high. Instead, Figure 2 shows that a grant's probability of funding is linearly increasing in quality. It is also not the case that low-quality applications have no chance of funding: 21 percent of applications with zero future citations are funded compared to 24 percent of applications with greater than zero future citations.[17] Because the likelihood of funding is similar for these groups, an equally strong push by close reviewers would have similar effects on a grant's likelihood of funding across quartiles, which is not what I find.

Second, if reviewers were simply biased, then we should expect proximity to increase application scores evenly across all quality bins. The middle panel of Table 7, however, shows that the impact of proximity on scores also increases in quality. I find a much larger effect for the top quartile than for any of the other bins. For scores, though not significantly, I in fact estimate a negative coefficient for the bottom half of quality, suggesting that proximity may actually hurt lower quality applicants.

[17] Further, 26.9 percent of grants in the bottom quartile of residual quality are funded, compared with only 22.3 percent in my top residual quality quartile.

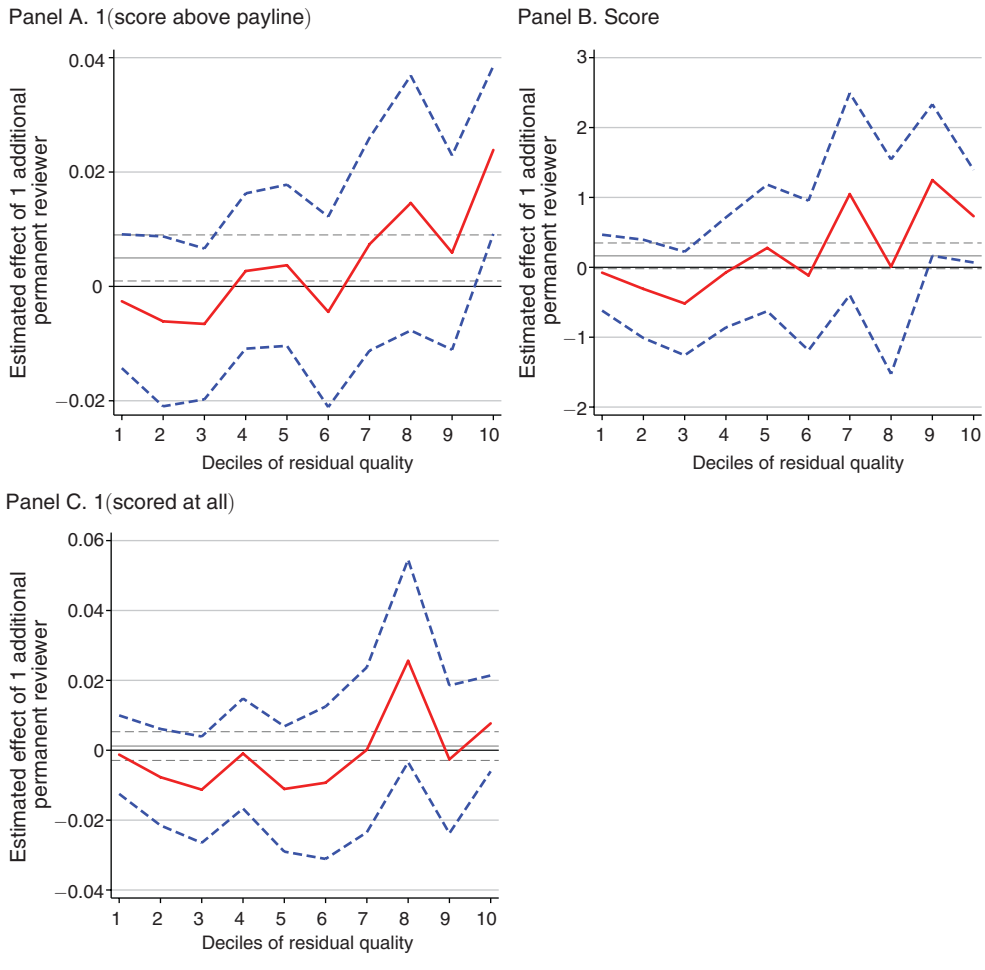

FIGURE 5. IMPACT OF PROXIMITY, BY QUALITY DECILE

Notes: The *x*-axis represents deciles of application quality, where quality is measured as the residual number of citations to text-matched publications that an application receives, after taking into account applicant demographics, education, publication history, and grant history. The *y*-axis represents coefficients from a regression of funding on number of related permanent reviewers, controlling for meeting effects and fixed effects for total related reviewers, for applications in that decile of quality. Dashed lines represent 95 percent confidence intervals. The gray horizontal line represents the effect of proximity (and confidence interval) estimated on the entire sample.

To explore this pattern more thoroughly, Figure 5 plots the estimated effect of proximity by decile of application quality. I find that the estimated impact of proximity generally increases with application quality and that, for particularly low or high deciles, my estimated effect is outside of the 95 percent confidence interval associated with the average effect in the whole sample, shown in gray. When I break quality to this finer level, I find suggestive evidence of a negative impact of proximity on candidates who submit applications that are especially poor relative to their qualifications. The coefficients are not statistically significant but their magnitude is economically meaningful: a bottom decile application evaluated by a close reviewer is 0.3 percentage points less likely to be funded than a similarly poor application

evaluated by an related reviewer, off a baseline funding likelihood of 30 percent for this group, making for a 1 percent decrease. Because each coefficient estimate represents the joint effect of bias and information, it is likely that the coefficient would be even more negative in the absence of potentially offsetting positive biases. By contrast, a proximate applicant in the top decile is 2.4 percentage points more likely to be funded than a similar applicant who is not related, off a baseline for this group of 26 percent, making for a 9.2 percent increase. This pattern strongly suggests that close reviewers have additional expertise about applicants. Figure 5 also shows a similar pattern for application scores.

In online Appendix Table J, I consider how the impact of proximity may differ for new and renewal applications, and for new and experienced investigators. For new grants, I find a sharper negative effect of proximity on funding likelihood for low-quality applications, and similar positive effects for high-quality applicants. This suggests that related reviewers have differential better information about the quality of these new applications. My estimates for new investigators are noisier and I can neither reject my estimates being different from zero or different from my estimates for other samples. These results are discussed in more detail in online Appendix C.

Finally, my results also consider the impact of proximity on an application's likelihood of being scored at all (e.g., rejected early in the process due to low initial evaluations). In general, I do not find significant effects, although my pattern of estimates is broadly consistent with the pattern that I find when examining funding likelihood and scores. See the last panel of Table 7.

B. *Proximity and Overall Portfolio Quality*

My main results show that (i) 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 (ii) better scientists benefit more from proximity, suggesting that study section members are better at discerning the quality of applicants in their own area.

Next, I assess the impact of proximity on the overall quality of funded applications, at the meeting level. For each meeting, I construct variables describing the overall share of applicants who have been cited by permanent reviewers, the overall share who have been cited by any reviewer, and the average ex ante measured quality associated with funded applications, as well as with all applications. I then regress the average quality of funded applicants on the share of related permanent reviewers, controlling for, at the minimum, year fixed effects and study section fixed effects.

Table 8 provides suggestive evidence that, on net, proximity increases the overall quality of funded grants. Meetings in which a greater share of applicants have been evaluated by a closely related permanent reviewer fund grants fund, on average, higher quality applications. Column 1 examines the relationship between share related to permanent including year and study section fixed effects. Year effects are needed to account for the fact that grants funded in later years have less time to accrue applications. Study section effects control for field-level differences in

Table 8—What Is the Effect of Proximity on the Quality of the NIH Portfolio?

	Mean grant application quality, meeting level					
	Awarded grants Mean = 0.144, SD = 0.265			All grants Mean = 0.111, SD = 0.202		
	(1)	(2)	(3)	(4)	(5)	(6)
Share of applicants cited by permanent reviewers	0.0856 (*0.051*)	0.0742 (*0.065*)	0.0714 (*0.067*)	0.0364 (*0.023*)	0.0090 (*0.030*)	0.0119 (*0.028*)
Share of applicants cited by any reviewers		0.0221 (*0.080*)	−0.0153 (*0.084*)		0.0531 (*0.037*)	0.0217 (*0.036*)
Observations	2,056	2,056	2,056	2,063	2,063	2,063
R^2	0.2554	0.2554	0.2805	0.4210	0.4219	0.5070
Year FEs	X	X	X	X	X	X
Study section FEs			X			X
Past performance, past grants, and demographics			X			X

Notes: Regression is at the study section meeting level. The dependent variable in columns 1–3 is the average quality, in citations to text matched publications, for funded grants associated with that meeting. The dependant variable in columns 4–6 is the average quality of all grant applications, including the unfunded ones. During my sample, there were seven meetings with no funded grants, which accounts for the difference in sample sizes between columns 1–3 and 4–6. The share of applicants cited by permanent reviewers is equal to the proportion of applicants in a meeting who have been cited by at least one permanent reviewer. Share cited by any reviewer is defined analogously. Columns 3 and 6 control for meeting level means of applicant demographics, education, past publications, past citations, and past R01 and other grants.

citation rates. The coefficient on share related is 0.0856, which implies that a 1 standard deviation increase in share related (0.182) increases the average quality of funded applications by $0.182 \times 0.0856 = 0.0156$ or $0.0156/0.144$ or 11 percent, significant at the 10 percent level. Column 2 controls for the share of applicants related to any reviewers, so that this specification more closely approximates the variation I use in my main individual-level regressions. Adding this control increases standard errors while decreasing the estimated coefficient slightly so that my results are no longer significant. This reflects the fact that, when aggregated to the meeting level, there is less variation in the share of applicants related to permanent members, conditional on total relatedness. Similarly, column 3 adds a variety of controls for the composition of applicants to a meeting: mean of gender, ethnicity, and education variables, as well as indicator variables for meeting-level means for number of citations (rounded to the nearest 100), and number of publications (rounded to the nearest 10). The coefficient does not change.

Although most of these estimates are not statistically significant, their relative stability across specifications suggests that I am picking up an effect of relatedness to permanent members. The magnitude of the effect I find in column 3 is such that a 1 standard deviation increase in the share of applicants related to permanent members (holding constant overall relatedness) increases the average quality of funded applicants by about 9 percent, which is a plausible magnitude. In online Appendix F.F4, I conduct an additional calculation using stronger distributional assumptions. In that exercise, I also find that proximity on net increases the quality of funded applications.

By contrast, columns 4–6 repeat this exercise with the average quality of all applications as the dependent variable. If it's the case that related reviewers have expertise, then they should increase the quality of funded applicants by choosing the best applicants to fund; there should be no effect of share related on quality of applicants in general. This is what I find. The magnitudes of the coefficients in columns 5 and 6 (having controlled for share related to any reviewers) are almost an order of magnitude smaller.

The analysis in Table 8 assumes that policymakers care about maximizing 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 policymakers should take citation-based measures of quality in account when assessing the efficacy of grant review. My citation data 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.

V. Additional Robustness Checks

The online Appendix discusses a variety of robustness and specification checks.

Online Appendix tables A–C provide supporting details for the data plotted in online Appendix Figures B and C, and Figure 3.

Online Appendix Tables D–I examine the robustness of my results to alternative measures of grant quality: changing the time window I use to measure publications associated with a grant; restricting to authors with very rare names to improve the quality of publication matches; varying my text matching process; and restricting only to publications in which the PI has played a primary role.

For example, not receiving a grant may slow down a scientist's research by requiring her to spend additional time applying for funding. If this is the case, then a grant can directly impact the research quality of funded versus 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. This is described in online Appendix Table F.

My results in Table 7 are based on residualized measures of quality. Residualizing citations allows me to identify whether proximate candidates have better information about an application's quality that cannot easily be gleaned from the primary investigator's CV. Online Appendix Table I shows that my results are robust to splitting my sample based on various non-residualized measures of quality: whether or not an application goes on to produce any citations to text-matched publications within the first year at all; those that produce publications cited at the ninety-fifth percentile of its field-year cohort versus not; and those that produce publications

cited at the ninety-ninth percentile of this distribution versus not. In all these cases, I find a stronger effect of proximity on higher quality applications.[18]

Online Appendix Table J presents my main estimates separately for new versus renewal grants and new versus established investigators.

Online Appendix Table K describes the results of an alternative estimation strategy that does not rely on the distinction between permanent and temporary reviewers. Instead of comparing outcomes for different reviewers who have been cited by different numbers of permanent reviewers, I use applicant fixed effects to examine outcomes for the same applicant over time, across meetings in which he or she is cited by different numbers of total reviewers.[19] I find largely similar results: related applicants are more likely to be funded, and the impact of relatedness is increasing in the quality of applications.

My next set of results describe broader tests of the logic of my empirical strategy. Online Appendix Table L, for instance, reports a different 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 proximity for applications that share the same funding status. To test for this, I examine the impact of proximity on application scores for the subset of applications that are either all funded or all unfunded. In both of these subsamples, I find evidence that being proximate to a permanent member increases scores and increases the correlation between scores and quality. Because proximity cannot affect actual funding status in these subsamples, the effect I find cannot be driven by differences in how well quality is measured.

Another potential concern with my quality measure is that text matching may eliminate publications on topics different from that described in the grant application but which review committees care about. 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. 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 proximity impacts the scores that funded applicants receive. In online Appendix Table M, I show that results using data on grant acknowledgments are largely similar.

Online Appendix Table N takes a different approach to addressing the potential concern that my relatedness measure is capturing unobserved aspects of an application's quality. If this were the case, then we might expect the impact of relatedness to appear similar to the impact of observed measures of quality, insofar as observed

[18] It is not possible to explore the impact of proximity on the funding outcomes of particularly low quality candidates according to unresidualized measures of quality. This is because there is significant bunching of applications at zero publications and citations.

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

$$Assessment_{icmt} = a_0 + a_1 Total\_Proximity_{icmt} + \mu X_{icmt} + \delta_i + \varepsilon_{icmt}.$$

and unobserved quality may be correlated. Online Appendix Table N shows that this is not the case: whereas the benefit of relatedness is found to be increasing in an application's quality, I find that the marginal impact of an applicant's past citations is the same across quality quartiles. This specification is discussed in more detail in online Appendix E.

Finally, online Appendix F estimates a model of committee decision-making in which bias and expertise parameters can be separately identified under a stricter set of distributional assumptions; online Appendix Table O presents estimates of bias and expertise, and online Appendix Table P makes efficiency calculations based on this model.

VI. Conclusion

This paper examines the impacts of bias and expertise on the quality of grant evaluation at the NIH. My results show that the preferences of influential reviewers matters for the funding outcomes of otherwise similar grant applications: being evaluated by an intellectually related reviewer increases an applicant's chances of funding by 2.2 percent, or the equivalent of a one-quarter standard deviation increase in application quality itself. This figure suggests that committees have a hard time predicting quality and, by comparison, reviewer preferences have a relatively large effect on funding outcomes. The fact that I find a positive effect of relatedness shows that although scientists compete for scarce resources such as funding and scientific priority, they nonetheless favor applications in their own area, suggesting that they view the research of others as complements to their own.

At the same time, reviewers do not merely favor all related applicants. Rather, higher quality applicants benefit much more from being evaluated by reviewers in their own area and particularly low quality applicants can even be hurt. This finding strongly suggests that reviewers have better information about the quality of related candidates. In this setting, I find that, on net, the quality of funding grants improves when more potentially biased experts are included in the review process.

My results suggest that there may be scope for improving the quality of peer review. For example, current NIH policy prohibits reviewers from evaluating proposals from their own institution. In the past, the National Bureau of Economic Research was considered a single institution, meaning that economists often recused themselves from evaluating the work of other economists.[20] The findings in this paper demonstrate why such policies may entail efficiency trade-offs.

REFERENCES

Acemoglu, Daron. 2009. *Introduction to Modern Economic Growth.* Princeton: Princeton University Press.

Agarwal, Sumit, David Lucca, Amit Seru, and Francesco Trebbi. 2014."Inconsistent Regulators: Evidence from Banking." *Quarterly Journal of Economics* 129 (2): 889–938.

Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* 42 (3): 527–54.

[20] Current conflict of interest policies apply to members of the same NBER program.

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

Boudreau, Kevin J., Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl. 2016. "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science." *Management Science* 62 (10): 2765–83.

Brogaard, Jonathan, Joseph Engelberg, and Christopher A. Parsons. 2011. "Network Position and Productivity: Evidence from Journal Editor Rotations." http://rady.ucsd.edu/faculty/directory/engelberg/pub/portfolios/editors.pdf.

Cockburn, Iain M., and Rebecca M. Henderson. 2000. "Publicly Funded Science and the Productivity of the Pharmaceutical Industry." In *Innovation Policy and the Economy*, Vol. 1, edited by Adam B. Jaffe, Josh Lerner, and Scott Stern, 1–34. Chicago: University of Chicago Press.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6): 1431–51.

Fisman, Raymond, Daniel Paravisini, and Vikrant Vig. 2017. "Cultural Proximity and Loan Outcomes." *American Economic Review* 107 (2): 457–92.

Garfagnini, Umberto, Marco Ottaviani, and Peter Norman Sørensen. 2014. "Accept or reject? An organizational perspective." *International Journal of Industrial Organization* 34: 66–74.

Gerin, William, Christine H. Kapelewski, Jerome B. Itinger, and Tanya M. Spruill. 2010. *Writing the NIH Grant Proposal: A Step-by-Step Guide.* Thousand Oaks, CA: SAGE Publications.

Ginther, Donna K., Walter T. Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L. Haak, and Raynard Kington. 2011. "Race, Ethnicity, and NIH Research Awards." *Science* 333 (6045): 1015–19.

Griliches, Zvi. 1991. "The Search for R&D Spillovers." National Bureau of Economic Research (NBER) Working Paper 3768.

Hansen, Stephen, Michael McMahon, and Carlos Velasco Rivera. 2014. "Preferences or private assessments on a monetary policy committee?" *Journal of Monetary Economics* 67: 16–32.

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): 665–90.

Jacob, Brian A., and Lars Lefgren. 2011. "The impact of research grant funding on scientific productivity." *Journal of Public Economics* 95 (9–10): 1168–77.

Jones, Benjamin F. 2010. "Age and Great Invention." *Review of Economics and Statistics* 92 (1): 1–14.

Kerr, William R. 2008. "Ethnic Scientific Communities and International Technology Diffusion." *Review of Economics and Statistics* 90 (3): 518–37.

Kondo, Jiro E. 2006. "Self-Regulation and Enforcement in Financial Markets: Evidence from Investor-Broker Disputes at the NASD." https://www.chicagobooth.edu/research/workshops/finance/docs/kondo_jobmkt.pdf.

Kremer, Michael, and Heidi Williams. 2010. "Incentivizing Innovation: Adding to the Tool Kit." In *Innovation Policy and the Economy*, Vol. 10, edited by Josh Lerner and Scott Stern, 1–17. Chicago: University of Chicago Press.

Laband, David N., and Michael J. Piette. 1994. "Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors." *Journal of Political Economy* 102 (1): 194–203.

Li, Danielle. 2017. "Expertise versus Bias in Evaluation: Evidence from the NIH: Dataset." *American Economic Journal: Applied Economics.* https://doi.org/10.1257/app.2015.0421.

Li, Danielle, and Leila Agha. 2015. "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science* 348 (6233): 434–38.

Li, Hao, Sherwin Rosen, and Wing Suen. 2001. "Conflicts and Common Interests in Committees." *American Economic Review* 91 (5): 1478–97.

Pearson, Helen. 2003. "Competition in biology: It's a scoop!" *Nature* 426 (6964): 222–23.

Sampat, Bhaven N., and Frank R. Lichtenberg. 2011. "What Are the Respective Roles of the Public and Private Sectors in Pharmaceutical Innovation?" *Health Affairs* 30 (2): 332–39.

Stephan, Paula E. 2012. *How Economics Shapes Science.* Vol. 1. Cambridge: Harvard University Press.

Zinovyeva, Natalia, and Manuel Bagues. 2015. "The Role of Connections in Academic Promotions." *American Economic Journal: Applied Economics* 7 (2): 264–92.

This article has been cited by:

1. Pierre Azoulay, Christian Fons-Rosen, Joshua S. Graff Zivin. 2019. Does Science Advance One Funeral at a Time?. *American Economic Review* **109**:8, 2889-2920. [Abstract] [View PDF article] [PDF with links]
2. Carola Frydman, Eric Hilt. 2017. Investment Banks as Corporate Monitors in the Early Twentieth Century United States. *American Economic Review* **107**:7, 1938-1970. [Abstract] [View PDF article] [PDF with links]