NBER WORKING PAPER SERIES

ADMINISTRATIVE DISCRETION IN SCIENTIFIC FUNDING: EVIDENCE FROM A PRESTIGIOUS POSTDOCTORAL TRAINING PROGRAM

Donna K. Ginther
Misty L. Heggeness

Working Paper 26841
http://www.nber.org/papers/w26841

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2020

Corresponding Author: Misty L. Heggeness, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20746 (e-mail: misty.heggeness@gmail.com). This paper reports the results of research and analysis undertaken by the National Institutes of Health (NIH) staff, consultants, and contractors. We acknowledge generous support and feedback from Lisa Evans, Henry Khachaturian, Kay Lund, Sherry Mills, Matthew Notowidigdo, Walter T. Schaffer, and Jennifer Sutton. This work would not have been possible without the amazing skills and assistance of Frances D. Carter-Johnson, Maria I. Larenas, Ryan Pfirrmann-Powell, and Ying Zeng within the NIH Division of Biomedical Research Workforce Programs, as well as participants from the 2018 Association for Public Policy Analysis and Management (APPAM) fall research conference and seminar participants at Iowa State University. We appreciate the time and candor given to us by various NIH institute program officers and training directors who so graciously agreed to be interviewed in an effort to help us truly understand the NRSA funding process on the ground. We also give very special thanks to Pierre Azoulay, Hyungjo Hur, Ming Lei, Jon Lorsch, Julie Mason, Michael Spittel, and Jonathan Wiest for taking the time to provide critical feedback. This analysis began while Misty L. Heggeness was employed at the National Institutes of Health. Donna Ginther acknowledges financial support from NIH Division of Biomedical Research Workforce (DBRW) (Contract No. HHSN276201300089U) to NETE Solutions. Any errors in this paper are the sole responsibility of the authors. The views presented here do not necessarily reflect any official position of the National Institutes of Health, U.S. Census Bureau, the Federal Reserve System, the Federal Reserve Bank of Minneapolis, or the University of Kansas. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Administrative Discretion in Scientific Funding: Evidence from a Prestigious Postdoctoral Training Program
Donna K. Ginther and Misty L. Heggeness
NBER Working Paper No. 26841
March 2020
JEL No. J24,O3,O38

ABSTRACT

The scientific community is engaged in an active debate on the value of its peer-review system. Does peer review actually serve the role we envision for it—that of helping government agencies predict what ideas have the best chance of contributing to scientific advancement? Many federal agencies use a two-step review process that includes programmatic discretion in selecting awards. This process allows us to determine whether success in a future independent scientific-research career is more accurately predicted by peer-review recommendations or discretion by program staff and institute leaders. Using data from a prestigious training program at the National Institute of Health (NIH), the Ruth L. Kirschstein National Research Service Award (NRSA), we provide evidence on the efficacy of peer review. We find that, despite all current claims to the contrary, the existing peer-review system works as intended. It more closely predicts high-quality science and future research independence than discretion. We discover also that regression discontinuity, the econometric method typically used to examine the effect of scientific funding, does not fit many scientific-funding models and should only be used with caution when studying federal awards for science.

Donna K. Ginther
Department of Economics
University of Kansas
333 Snow Hall
1460 Jayhawk Boulevard
Lawrence, KS 66045
and NBER
dginther@ku.edu

Misty L. Heggeness
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233
and Federal Reserve Bank of Minneapolis
misty.l.heggeness@census.gov

I. Introduction

For most, peer review has always been a pillar of the scientific enterprise. Recent attacks on the U.S. federal-funding system, however, have questioned whether peer review identifies the best science (Stahel and Moore 2014; Fang and Casadevall 2016; Fang, Bowen, and Casadevall 2016; Gallo, Sullivan, and Glisson 2016; Li and Agha 2015; Li 2017). This is not the first time the peer-review system has come into question (Gustafson 1975; Roy 1985; McNutt et al. 1990; Travis and Collins 1991; Godlee, Gale, and Martyn 1998; Wessely 1998; Smith 2006; Costello 2010). In fact, in 1975 Congress conducted a public debate during the National Science Foundation Peer Review Special Oversight Hearings to evaluate the role of peer review in federal scientific funding. For the most part, it concluded that peer review, compared with all other options (including agency discretion), worked as expected (Baldwin 2018). Gustafson (1975, 1065) argued that "for most types of fundamental research the traditional project grant, selected by peer review, with overall priority among fields and subfields determined at least in part by proposal pressure, appears to provide the best available guarantee of scientific merit and accurate information." In this paper, we study the funding decisions at the National Institutes of Health (NIH) and examine whether discretion or peer review is more effective at identifying future engagement in science. Our results show that peer review identifies future success more often than program officer discretion.

Whether today or in the past, questions continue as to whether the peer review system functions like an old boys' club. On the one hand, we are lucky that in recent decades the number of scientists in the U.S. has been greater than at any other time in history, as more and more get trained and fewer leave the larger ecosystem of science when they reach retirement age (Heggeness et al. 2016, 2017; Blau and Weinberg 2017). More scientists means more intellect being poured into scientific advancements and the betterment of humankind—not only in academia but in industry and government as well. However, more scientists also suggests increased competition—especially for academic science positions (Tilghman and Rockey 2012)—and competition has been heating up for scientific grant funding. Even those with low percentile scores on grant applications are finding themselves out of the loop and unfunded. Frustrated with a system that is not able to fund as many high-quality science projects as are out there, peer review continues to be debated (Scarpa 2006; Marsh, Upali, and Bond 2008; Fang and Casadevall 2016; Fang, Bowen, and Casadevall 2016).

While it may seem like the peer-review system has always been an integral part of the scientific machinery, throughout history journal editors and program managers have exercised varying levels of discretion. In fact, scientific funding and journal review have a much longer history of relying on funding managers and journal editors to make discretionary decisions about what constitutes good science. These discretionary decisions made by leaders within the scientific community drove much of what we knew as scientific advancements well into the 1970s (Smith 2006; Benos et al. 2007; Baldwin 2018). The question remains: Do scientific leaders in control of grant funding (and, more broadly, journal publications) identify quality science and future superstar scientists more precisely than the current peer-review system?

While the issue of peer review in journal publications is equally relevant, its role is different. Journal editors aim to identify whether an already-completed scientific manuscript is novel enough to count as a solid contribution to their journal. In most cases, scientific funding agencies do not know what the results of the proposed project will be and whether it will, in fact, produce results. Funding agencies rely on the scientific track record of the principal investigator and educated opinions about whether the idea is robust and worthy enough to fund. Since scientific funding decisions occur at the first stage of an innovation and are, by nature, a riskier proposition, there is more nuance in determining what will turn out to be novel science. For this reason, we focus on identifying the impact of peer review at an innovative endeavor's first stage of development.

The U.S. federal government is a major contributor to the advancement of the scientific enterprise. Federal agencies that fund scientific research include the National Institutes of Health (NIH), National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), Advanced Research Projects Agency-Energy (ARPA-E), United States Department of Agriculture (USDA), and a handful of other agencies. NIH, the largest scientific agency in the world, is by far the primary federal funder of biomedical research. Each year it provides over $39.2 billion in funding.[1] The NIH budget has increased since 2013 and more than doubled since 2000.[2] While NIH is the primary funder, other agencies also distribute funds for biomedical and scientific research, and while each agency has

[1] For more information, see https://www.nih.gov/about-nih/what-we-do/budget

[2] For more information, see https://officeofbudget.od.nih.gov/pdfs/FY19/Approp%20History%20by%20IC%20FY%202000%20-%20FY%202019%20(V3).pdf

its own system, the systems have similarities. For example, most agencies send a proportion of proposals out for external review, and all have program officers who manage the grant portfolios and help ensure the research funded meets the priorities of the organization. Nevertheless, given the immense size of the NIH funding portfolio, and the fact that the review process is the same across all institutes at the NIH, it is critically important to understand how discretion operates in conjunction with the peer review process.

Peer review is often seen as a singular and independent method to produce high-quality evaluations of scientific projects potentially worth funding. In reality, no peer-review system mechanically follows rules all the time. Owing to the nature of science and the federal managers' responsibility to hold scientists accountable, there is and always will be some element of discretion within the grant-awarding system. All federal agencies have some sort of two-step decision-making process where expert opinion is sought through peer review, then internal deliberations are undertaken with agency leaders and scientific-program staff, and finally budgets are explored. All these steps occur before any final decisions are made. The relevant questions to ask are, (1) How much discretion exists? (2) How is it documented? And (3) what implications does the level of discretion have on the future of science?

We take advantage of this type of two-stage review to examine the effectiveness of peer review versus that of discretion in identifying successful scientists. Using data from administrative records on applicants and awardees of a prestigious fellowship program at the NIH between 1996 and 2008, we compare the outcomes of applicants chosen with discretion with those of applicants chosen through peer review. In many cases, the NIH uses discretion in awarding grants, and this is particularly true when awarding the Ruth L. Kirschstein National Research Service Award (NRSA) F32 fellowships. As a result, we can categorize individuals who were "skipped over" and those "reached for" in discretionary decisions made by program and institute staff and identify how the peer review system performs compared with discretion. We use the scoring of fellowship applications and matching techniques that allow us to analyze a limited set of outcomes for individuals who received the fellowship and for those whose application scores and observable characteristics are similar, but who did not receive an award.

Our study makes two contributions. First, by comparing the applications of individuals who are deemed high quality by peer review but "skipped" in the funding process with those of individuals who are "reached for" through discretionary decisions by program officers and institute leadership, we show that peer review predicts future NIH research awards at a higher rate than discretion. Second, although in theory the method of awarding funding appears consistent with a regression discontinuity design (RDD), we demonstrate that in practice RDD is not an appropriate method to study scientific awards where high levels of discretion exist in the decision-making process. RDD is not an appropriate method in this environment because discretion drives the second stage of decision making. In addition, NIH institute budgets are semi-fluid. Within agencies, funds shift between first-, second-, and third-round funding opportunities and, in some instances, between programs as funders examine the quality and depth of projects in each round.

The paper proceeds as follows. Section II summarizes the historical use of discretion and peer review and describes why the federal grant-selection process does not always fit the mold of a discontinuous kink at an agency budget line for funding along some peer-evaluated score. Section III describes the methods we employ and data we use. Section IV analyzes the results comparing the career outcomes of those who were chosen with programmatic discretion versus those who were chosen through peer review. Section V is a discussion, and Section VI concludes.

II. Background

A. Scientific Funding, Discretion, and Peer Review

Scientific peer-review systems began to appear in the United States in the 1940s, driven by federal scientific agencies (Azoulay, Graff Zivin, and Manso 2013; Farrell, Farrell, and Farrell 2017). At its core, peer review provides a systematic process for scientific evaluation driven by the expertise and experience of an extramural community of experts. Its main goal is to provide guidance, direction, and leadership to federal agencies in selecting the most promising scientific research proposals for funding. Today, it is embedded within the scientific enterprise as a strong institutional norm and has been driving scientific-funding decisions since the 1970s (Baldwin 2018). While it is a foundation of much scientific activity, peer review "in theory" often looks different than peer review "in practice." Throughout history, scientific institutions have continually made discretionary decisions regarding which scientific proposals to fund and where to allocate scarce resources.

Discretion, however, may come at a cost. Past studies have shown that actuarial judgement outperforms clinical assessments (Dawes, Faust, and Meehl 1989). More recent studies have highlighted how human discretion in hiring can lead to less-than-average hiring outcomes (Hoffman, Kahn, and Li 2018) and how judicial discretion in criminal courts may lead to increased crime or increased jailing (Kleinberg et al. 2018).

Baldwin documents large amounts of discretion at the initiation of scientific peer review in the 1940s on the part of federal agencies like the National Institutes of Health:

> When the U.S. government formed the National Institutes of Health (NIH) in 1948, its Division of Research Grants initially evaluated grant applications with little or no consultation from outside referees. Instead, each application went first to a small "study section" composed of NIH-affiliated scientific experts in a particular field. From there, the study sections' recommendations were forwarded to an NIH council of scientists and laymen, which added its own recommendations. Final decision-making power rested in the hands of the institute directors, heads of NIH member institutions such as the National Cancer Institute and the National Eye Institute. While the directors took the earlier evaluations into account, they were not obligated to follow the recommendations of the study sections or the council. Furthermore, NIH applicants would receive little information about why their grants had been accepted or rejected. Deliberations about the grants were considered confidential and internal to the NIH. (2018, 545)

With time and criticism against agencies mounting, peer review gained momentum and played an increasingly dominant role in selection of awards, but always within the bounds of discretion and the agency's ability and desire to fund promising, high-priority research. Baldwin notes that some continued to endorse "the older system of using . . . peer review reports as advisory documents, saying . . . peer review has its uses as a first round of proposal screening, but it does not absolve the Government program manager from full responsibility for the decision to fund or reject a proposal. . . . There are some things that we should not ask of peer review. We should not ask it to take Government agencies off the hook on the question of protecting the public purse" (556).

The back-and-forth discussion of the peer review's role in federal funding ebbs and flows. Peer review's relevance continued to increase from the 1940s to the 1970s—when most, if not all, agencies established a formalized peer review system. While congressional hearings in 1975 concluded that peer review was the fairest and best way to allocate scientific funding (compared with agency discretion), some scientists were seeing a challenge to the full use of peer review in resource allocation. Gustafson (1975) writes, "Even in the programs in which external peer review panels have the greatest sway, there appears to be the opportunity for the agency staff to influence the process of proposal evaluation by shaping the agenda, channeling the flow of information to and from the outside advisers, or actually altering or overriding their decisions. In the NIH this influence is discreet and informal, while in the NSF it is usually more important than that of the external advisers. In both agencies, the importance of the professional staff appears to be growing" (1064). In fact, within the NIH, a wide range of variation in discretion exists depending on both the grant mechanism and the institute funding the research.

Smith (2006), decades later, argued against any reality of peer review being independently operational, highlighting challenges even in the existence of an agreed-upon operational definition of peer review. If, as Smith states, peer review is challenging to operationally define, how might we expect it to be the primary and sole driver of a grant-award system in practice? Regardless of the challenges, the NIH began enhancing the peer review program in 2007. The measures it took included incorporating process and change phases and shortened, restructured applications in an effort to "fund the best science, by the best scientists, with the least amount of administrative burden" (NIH 2011). Below, we further highlight how federal agencies handle peer review and the allocation of resources, which we argue always includes a peer-review process mixed with the discretionary priorities of the agency and program staff.

B. The Value of Peer Review

Studies have attempted to detangle the value of peer review in selecting the best, most innovative science. Li and Agha (2015) studied risk aversion in peer review. Specifically, they examined whether peer review selects projects already demonstrating success or chooses big-name scientists for continued funding regardless of the novelty of their current grant proposal idea. They established that proposals with better scores have higher numbers of publications and citations, a finding that supports the idea that peer review is working. This result was also confirmed by Gallo et al. (2014) in another funding context. However, when Fang, Bowen, and Casadevall (2016) compared percentile scores of 20 or better, they found that peer-review scoring had limited predictive power for future publications and citations.

In somewhat related research, Pier et al. (2018) uncovered little agreement on proposal quality in an experiment designed to mimic the NIH peer-review process. When Li (2017) examined expertise versus bias in NIH peer review, she discovered that expertise slightly outweighed the cost of bias. She also demonstrated that when reviewers and researchers shared expertise, they more harshly judged proposals. Gallo, Sullivan, and Glisson (2016) obtained a similar result in another funding context. Furthermore, Ayoubi, Pezzoni, and Visentin (2019) suggest that the very process of applying for research funding improves publications, regardless of whether the researcher receives an award.

Both a handful of smaller studies that randomized peer review and senior journal editors discussing their experience have suggested that peer review is little better than chance at selecting which science to fund or publish (Smith 2006). Taken together, prior work has come to different conclusions on the validity and efficiency of the peer-review process. These studies assess questions and concerns associated with the quality of the peer-review system and peer review's ability to predict bold and innovative science, but they do not evaluate specific differences between discretion and a systematic peer-review process. While Goldstein and Kearney (2018) show how federal-program directors use discretion to allocate funds in alignment with an agency mission, to our knowledge our study is one of the first within the context of federal funding for biomedical science to answer the comparative question of whether peer review or discretionary decisions more accurately identify scientists who will develop into independent researchers.

C. Two-Stage Institutional Behavior and Award Decisions

Most federal agencies have some style of two-stage review. At the NIH, there are 27 institutes and centers, and each one has complete independence from the others in allocating awards. Grant awards are generally decided based on some weighted mix of peer-review scores, the research priorities of institute leadership, and the discretionary behavior of staff. In terms of award type, these organizations greatly vary in their process for awarding a grant, with smaller grants like fellowships exhibiting much larger discretionary influence than large independent-research awards. For this study, we interviewed staff at four institutes and centers. These people provided contextual knowledge into the process of selection. Together, they encompassed a range of variation in institute size, disease focus, and training programs.

Our interviews with program staff indicate that the funding process for fellowships is indeed complex. Most institutes receive application-review scores as defined by a study section coordinated through the NIH's Center for Scientific Review (CSR). Once the institute receives the scores, program officers and staff assess the full application, including a summary statement from peer review, the quality of the applicant and his or her institution, and the alignment of the research proposal with the institute's priorities. Members of the institute's staff—specifically, program officers—then participate in a team meeting in which they defend the proposals that best match their defined priorities. Together, the program officers, the training director, and other institute staff make a joint decision for recommendations to the institute director. Either the institute director or his or her delegate makes the final decision and signs off on which proposals to fund. Institute directors vary in terms of their direct involvement in the consideration and final approval of proposals. Before making a final decision and informing the candidates, the budget office reviews and signs off on the final list of candidates, primarily making sure sufficient funds are available for the recommended awards.

D. Scientific Funding—Type and Evaluative Technique

i. Research Training—A Case Study

Since 1974, the U.S. government has formally committed to training high-potential, early-career scientists to carry out the nation's biomedical research agenda through congressionally mandated programs like the Ruth L. Kirschstein National Research Service Award (NRSA). Subject to periodic review (National Research Council 2011), large federally contracted studies have monitored the outcomes of those who received the award (Pion 2001; Mantovani, Look, and Wuerker 2006). Nevertheless, few studies have used more rigorous methods to estimate the award's unbiased impact on future career outcomes (Levitt 2010; Jacob and Lefgren 2011).

In this study, we focus on the NRSA F32 postdoctoral training award for two reasons. First, it allows us to capture scientists at the beginning of their career, when evidence of their potential success is not yet fully developed, thereby providing the strongest raw evidence of whether peer review can identify future successful scientists. Second, fellowship awards are relatively inexpensive compared with other major funding awards,[3] and as individual (non-institutional) training awards, they represent the federal government's best method for directly shaping the future generation of scientists. Both these facts drive NIH institutes and their leaders to impose even more discretion than is used with other awards. They use a higher level of discretion because the risks associated with making a selection error are relatively low, and the gains from influencing the future direction of science are potentially high. Leaving a lasting legacy influencing the next generation of researchers and future direction of science is an admirable goal that most institute leaders take seriously and are interested in pursuing. Altogether, this program gives our analysis a perfect mix of grants awarded based on both discretion and peer review.

ii. Evaluating the Validity of Regression Discontinuity

Numerous studies have used regression discontinuity design (RDD) to evaluate the impact of scientific R&D funding (Jacob and Lefgren 2011a, 2011b; Grilli and Murtinu 2011; Benavente et al. 2012; Li 2017; Howell 2017; Bol, de Vaan, and de Rijt 2018; Azoulay et al. 2019). An RDD works best when the level of discretion is minimal (as is the case with major independent-research grants like R01 awards) and budgets are fixed ahead of time. However, with awards like fellowships, high

[3] Individual training awards are generally around $60,000 each, whereas a standard R01 independent-research grant can run anywhere from five to ten times as much.

levels of staff discretion and congressionally mandated annual budgets that distribute awards in multiple annual cycles dilute the appropriateness of an RDD.

With fellowships, a budget line is generated showing how many grants the organization can fund in a particular council round within a particular year. Theoretically, the organization then funds the "best" proposals in each council round up to the point where it exhausts its budget. This scenario is seemingly appropriate for an RDD, with which one can compare applicants who "just" got funding with those who "just did not" get funding solely because of exogenous factors (e.g., the money was exhausted through no fault or manipulation of the applicant). These applicants would otherwise appear similar; therefore, following the logic of RDD, one could reasonably conclude that any difference observed between those funded and those not funded around a maximum budget level (called a "pay line") is due solely to the effect of the award.

The reality, however, is that funding and award-making decisions are complicated. What funding is available depends on the number of applicants in each council round and previous rounds and whether everyone who was previously offered a grant accepts. Institutions receive an annual budget but make grant-funding decisions by council round. Depending on the institute, there are anywhere from two to four council rounds in a given year. If enough high-quality applicants do not apply in council round one, the budget (and pay line) can be reduced allowing for more applications to be funded in future council rounds within the same year. The reverse is also true; budget offices may increase their budget (and pay line) if there is a large pool of candidates and funds are available.

Another complication is the fluidity of funds across programs and the ability to move allocations across scientific divisions. An institute must spend all the money appropriated to it by Congress, so if applications in other programs are light in a particular year, this could provide additional funds to increase a budget line in another program. The fluidity of a budget line and the fact that it could be influenced by the quality and quantity of applications violates the assumptions required for a valid RDD and even those required for a fuzzy RDD where limited discretion around the budget line takes place.

III. Data and Methods

A. The Data

Our analytical data include administrative records from the NIH's Information for Management, Planning, Analysis, and Coordination (IMPAC II) system from 1996 to 2008.[4] The NIH matches its administrative records to data from the National Science Foundation's Survey of Earned Doctorates (SED), an annual census of doctoral recipients from U.S. institutions. The NSF SED contains information on individual demographics, characteristics of graduate study, and future career plans. By linking these data sets, we are able to obtain missing data and add additional individual-level covariates on our sample. We use demographic variables before or at the point of PhD completion. These variables are extracted

[4] NIH administrative data is part of the IMPAC II grants data system National Institutes of Health, IMPAC II, http://era.nih.gov/. The data is restricted-use. Researchers interested in replicating our study or accessing the data for research can submit a request to the National Institutes of Health's Office of Extramural Research.
.

from the SED and include age at PhD completion, gender, race and ethnicity, marital status at PhD completion, PhD field of study, and type of doctorate-education funding. We use the two data sources to construct one large panel data set for analysis, limiting our data to those who applied for NRSA F32 funding between 1996 and 2008, and then observe these individuals' future NIH award-application and funding patterns through 2015.

We include application-review score, funded or non-funded status, time frame, the institute or center receiving applications or funding the award, and previous grant-funding or training affiliations. We further queried IMPAC II for subsequent applications for NIH funding and awards from these individuals. Similar to Jacob and Lefgren (2011a), we define our outcome variables to identify research-award application or receipt four or more years out from the individual's application year. Our control variables mimic those used in the Ginther et al. (2011) paper on research awards and race. In particular, we include controls for race and ethnicity, gender, marital status, age, degree, scientific field, and previous NIH training experience.

Our analytical sample is a subset of all applicants. We drop applications that are higher than the 60th percentile in each council round, because the NIH does not consistently save scores for these applications in the reporting database, and practically none of them get funding. Some institutes and centers have too few applicants for our matching method, so for this reason, we drop applicants from seven institutes and centers. Our final analytical sample contains 14,276 individuals.

B. Descriptive Statistics

We report descriptive statistics in Table 1. In our analytical sample, awardees and non-awardees do not differ in terms of age at application, marital status, or likelihood of having a prior T32 traineeship. Table 1 shows that awardees and non-awardees do differ across a number of observable characteristics. Awardees are significantly less likely to be black or Hispanic. Individuals with MD degrees are less likely to receive fellowship awards, whereas PhDs are more likely. Individuals with biomedical or social-science degrees are more likely to receive fellowship awards compared with those without a reported PhD field. Awardees are significantly more likely to aspire to and receive subsequent NIH funding, as measured by the number of Research Proposal Grant (RPG) applications and awards, the probability of an RPG award, and the probability of an R01 award. As expected, awardees have significantly lower (better) scores on their last observed application.

In Table A1 in the Appendix, we estimate the probability of receiving an NIH F32 award as a function of observable characteristics for the full sample and the analysis sample.

C. The NRSA and Discretionary Decision-Making

Using the NRSA F32 postdoctoral training fellowship program, we demonstrate the problems with RDD for federal-grant awards where discretion takes place. Figure 1 is an illustration using all applicants from 1996 to 2008. Panels A–C provide three typical institute-level examples of variation in processing the awarding of fellowships around a fictitious budget line similar to what is generated by a budget office each council round. Each panel shows a point for each F32 applicant. The points at the top signify applicants who received an award. The points along the bottom represent applicants who applied but did not receive an award. The *x*-axis ranks the applicants by peer-reviewed priority score from best to worst. In each panel, a vertical line represents a pseudo–pay line[5] imposed by a budget office suggesting to program officers and staff how many awards can be funded in that particular council round. Points at the top and left of the vertical line represent awards funded in order of peer-review scores. Points at the bottom and right of the vertical line are non-funded applicants in order of peer-review scores, meaning that given budget constraints, their scores fell outside of the range of feasible acceptance. Points at the top and to the right of the vertical line represent awards that were funded out of order. For each of these points, at the bottom and left of the vertical line, there is an equal point representing an applicant who had a priority score low (good) enough to be funded, but whom staff and institute discretion skipped over in order to fund an applicant in the top and to the right of the budget line.

If all funding were awarded based solely on the ranked order of priority scores, we would observe Panel A, and no discretion would creep into the award process. Panel A distributes awards based solely on the peer-reviewed score, allocated from

[5] A pay line is the end point of the budget, where all resources have been exhausted. If there are 10 applicants with scores ranging from 1 to 10, and the budget allows for the funding of 3 applicants, then the pay line is a score of 3. Anyone with a score lower than or equal to 3 would get funded if only peer review was followed in decision-making.

best score to worst score until the institute budget allocations for the fellowship are exhausted. If this were the case (and budgets were not fluid or exchangeable), a sharp RDD would be valid.

Panel B of Figure 1 illustrates an institute following the guidance of the peer-reviewed score for the best (lowest) scores. However, once the institute has funded a majority of applicants with meritorious scores, it uses discretion to distribute awards near the pay line. In this case, the institute is comfortable using discretion near the pay line to fund applications that best fit within its scientific priorities and where the institute staff believes the applicant has the best-case scenario for future success—perhaps because they consider the applicants to be more or less similar in quality. If this were the case (and budget lines were not fluid), a fuzzy RDD would be appropriate.

A third, more complicated case is Panel C. It demonstrates the most complex case of selection for fellowship awards. About two-thirds of proposals funded would be below the expected pay line if the institute were to fund based solely on peer-reviewed scores. Institutes represented here use a significant amount of discretion when selecting proposals for funding and, because of this, no real cutoff exists. For our purposes, a very valid question exists as to the frequency with which institutes engage in this third scenario and, to some extent, the second scenario as well.

We examine this question in Figure 1, Panel D. Since no real data are available for pay lines, we construct pseudo–pay lines by counting the number of awards funded and assuming that for each institute this number is equivalent to the pay line for that council round.[6] Each diamond represents the percent of applications funded within the pseudo-constructed pay line (e.g., the top and left of the budget line; funded in order of peer reviewed score) by year, institute, and council round. If an institute has three council rounds in one year, they will have three diamonds represented vertically for that year. In Panel D, the black, horizontal curved line represents an average yearly rate of funding strictly by peer-reviewed score among all institute-council rounds in that year. Between 1996 and 2002, overall NIH institute-council rounds funded approximately 40 percent of proposals in order based solely on the rank of scores. In other words, over half the time, institutes were reaching for applicants below the pay line within a council round and, equivalently, skipping a proportion of applicants above said pay line.

After the NIH doubling,[7] the rate of institute-council rounds funding solely in order dropped to a low of 28 percent in 2006 and increased to 35 percent by 2008. This finding suggests that when institutes have fewer resources, they use less discretion.

Between two-thirds to three-fourths of all institute-council rounds used discretion in award allocation in recent years. Although review scores assigned during study section are an important criterion in the selection of awardees, they are clearly not the only criteria. Over the entire period, only 37.5 percent of year, institute, and council round (YIC) units (N=701) complied with a sharp RDD

[6] This in and of itself is a false presumption because it does not take into account budget-office discretion in adding or reducing slots based on the applicant pool. However, we hold judgement on that piece in order to demonstrate our general point here.

[7] The NIH doubling occurred between 1998 and 2002. It was a period of 5 years during which the total federal budget allocated to the NIH doubled in size.

framework (ignoring the issue of budget fluidity). Not only were there few YIC units in compliance, but those that did comply had few applicants. Only 10.7 percent of individual applications were in YIC units that complied (data not shown). This result is worth emphasizing. *Only one in ten applicants* experienced a review process where peer review scores were strictly followed—the rest experience a two-stage process infused with institutional discretion. Extending to a fuzzy RDD,[8] around 61.5 percent of YIC units complied, translating into 47.9 percent of applicants in our sample (data not shown). Even under a fuzzy RDD, less than half of all applicants experienced a council round that met the criteria for some form of RDD.

In fact, institutes vary widely in how they implement the two-stage process. At one extreme, take the National Institute of General Medical Sciences, which explicitly delineates its two-stage process and discretionary actions. In a recent report on application and funding trends, the institute stated that “we do not use a strict percentile cut off (‘pay line’) to make funding decisions. Instead, we take a variety of factors into account, including peer review scores, summary statements, Institute priorities, overall portfolio diversity, and an applicant’s other research support” (Hechtman and Lorsch 2019).

We examine whether discretion varies by council round and year. If discretion depends solely on budgets, we would expect it to depend on the budget remaining in a given fiscal year. Thus, we examine whether the probability of reaching for or

[8] Here we assume that a YIC unit fits a fuzzy RDD if less than 10 percent of cases either skip a good review score or reach for a worse review score.

skipping a proposal depends on the timing of the council round (three per fiscal year) after control for institute and fiscal-year dummies. It could be that if fewer funds were systematically spent in the first council, then the amount of discretion would increase in the following two rounds. In Table 2, we show that there is not a statistically significant difference in the probability of discretion as a function of council round measured by reaching or skipping proposals. However, in times of tight budgets (after the NIH doubling ended in 2002), the probability of a proposal's being funded by discretion (reached) falls. As funding got tighter in the late 2000s, the probability of skipping proposals with good scores also fell. As with Figure 1, discretion falls when budgets become tight.

Figure 1 and Table 2 indicate that discretion is widely used in the allocation of the NRSA F32 fellowship. Since other factors besides review score drive the decision-making process (and pay line) in fellowship awards, the often-used RDD approach is not a valid methodology. We use more appropriate matching methods for our analysis. More importantly, however, since all applications in our analytical subsample have a review score (details described below), we can take advantage of the heavy use of discretion in the NRSA F32 postdoctoral training awards to evaluate whether discretion or peer review more often predicts future scientific success.

D. Using Matching to Identify a Causal Effect

As described, the multi-step selection process for grants first generates a review score via a systematic peer review process. Given this step and the fact that the groups of individuals applying to the awards are relatively homogenous, we use matching techniques. While any unobserved characteristics differing between funded and unfunded applicants could confound our results, we argue that matching is a feasible approach for the following reasons. First, we can account for unobserved differences by institute and council round by controlling for these factors. Additionally, selection is made at the institute level. Any unobserved differences among applicants are unobserved by the institute also and therefore not a driving component of the selection process. There is no self-selection of fellowship award offerings. Finally, the groups of individuals that apply for funding are relatively homogeneous within each institute. All of these people clearly excel in academics, have been encouraged to apply by their mentors (which means their mentors believe they have a chance of getting the award), and are typically intensely interested in biomedical research. If this unobservable variation does exist, we argue that it is minimal within this select group of applicants.

To estimate the causal effect of fellowship awards on subsequent NIH funding outcomes, we use the potential-outcomes framework employed in econometric analysis (Rubin 2004). To fix ideas, let $T_i = 1$ be the treatment when an individual's fellowship application is funded, and let $T_i = 0$ if the application is not funded. Each individual has two potential outcomes of subsequent NIH funding: $Y_i(1)$ if the individual receives the award treatment and $Y_i(0)$ if the individual is not treated. For each individual, the causal effect of the award on subsequent NIH funding is defined as the difference in potential outcomes $Y_i(1) - Y_i(0)$. However, each individual is observed only when they do or do not receive the award, and, in this case, we must estimate the counterfactual outcome using matching methods.

In order to implement matching methods, we assume that treatment is independent of the outcome conditional on covariates $T_i \perp \left(Y_i(0), Y_i(1)\right) | X_i$. This is the unconfoundedness assumption, which means that the treatment is conditionally independent of the outcome after conditioning on observable characteristics. Given unconfoundedness, we can define the average treatment effect in terms of potential outcomes as the expected value of potential outcomes:

$$ATE = E[Y_i(1) - Y_i(0)].$$

We use two matching methods to identify the ATE. First, we employ propensity-score matching, defining the propensity score as the probability of receiving treatment conditional on observed characteristics $e(X) = \Pr(T_i = 1 | X_i = x)$. In order to implement propensity-score methods, the propensity scores for the treated and untreated in our sample must overlap such that $0 < e(x) < 1$. Although the unconfoundedness assumption cannot be directly tested, we can examine whether the propensity score has a causal effect on a pseudo-outcome that was determined prior to the treatment. If the estimated effect of the treatment on the pseudo-outcome is significant, then unconfoundedness has likely been violated (Imbens 2015).

Propensity-score matching has been widely used in economics and other social sciences (Imbens 2015). However, King and Nielson (2019) and Imbens (2015) note that propensity-score estimates break down if the propensity-score model fits to the data too well. As a result, we cannot use the review score to estimate the propensity score related to fellowship funding. Thus, we use the coarsened exact matching (CEM) algorithm (Blackwell et al. 2009) to improve the balance of the data and nearest-neighbor methods to facilitate matching on the review score. We use propensity-score matching, and as a robustness check, we also use nearest-neighbor matching after reducing the data using the CEM algorithm (results in the Appendix).

In addition, given that we have information on a pseudo-constructed pay line (the total number of funded applicants in each council round), review score, and award, we construct four indicator categories: those funded and within a pseudo-constructed pay line based on review score (as expected), those not funded and outside a pseudo-constructed pay line based on review score (as expected), those not funded but within a pseudo-constructed pay line based on review score (*skipped),* and those funded outside a pseudo-constructed pay line based on review score (*reached*). We then use our matching methods to examine treatment effects by comparing outcomes. In particular, we compare outcomes for those who were (1) reached compared to not funded as expected; (2) funded as expected compared to reached; (3) reached compared to skipped; (4) skipped compared to not funded as expected; (5) skipped compared to those funded in order; and (6) funded as expected compared to those not funded as expected.

IV. Results of Peer Review versus Discretion

Table 3 reports the Average Treatment Effect (ATE) Propensity Score Matching (PSM) estimates for the analytical sample. If we include the review score in the propensity-score estimates, the propensity score becomes too precise, and the matching algorithm breaks down (King and Nielson 2019; Imbens 2015). Thus, our propensity-score estimates include institute and council round fixed effects and the covariates listed in column 1 of Table A1. Each column of Table 3 shows outcome variables: the number of research program grant (RPG) awards, the number of RPG applications, the probability of an RPG, the probability of an R01 and the probability of never applying for additional funding.

We identify four major categories of funding status based on scores and institute behavior: *funded in order*, *skipped*, *reached*, and *not funded in order*. Using the review score as a measure of the proposal's scientific merit, those proposals funded in order were applications judged as the most meritorious by the reviewers and institute staff. Some proposals with favorable review scores within the pseudo-constructed pay line (budget) were skipped in favor of proposals with worse review scores. In this case, the institute reached to fund a proposal out of the review score order. Proposals that were not funded in order had review scores in excess of the budget pay line and worse scores than those that were funded in order or skipped.

In Table 3, the first row compares the treated proposals that were reached compared with those that were not funded. Compared with those not funded, reached scientists secured around 0.17 more independent-research awards and had around 0.80 more applications in future years, more than the full sample (results not shown; see Heggeness et al. 2018 for a detailed analysis of the full sample). More importantly, they were between 7 to 8 percent more likely to receive independent awards conditional on applying, and they were 11.9 percent less likely to never apply for future independent awards. These estimates were slightly higher than those for the full analysis sample (see Heggeness et al. 2018).

Next, we compare proposals where the treatment was reached and had relatively higher (worse) review scores compared with those that were funded in order. The ATE estimates indicate that the subsequent NIH funding outcomes for reached proposals were significantly worse than those funded in order. Reached proposals received 0.096 fewer RPG awards, submitted 0.5 fewer RPG applications, had a 5.0 ppt lower probability of receiving an RPG award, had a 3.8 ppt lower probability of receiving an R01 award, and had a 6.3 ppt probability of never applying for additional NIH funding. These results suggest that those reached individuals do not perform as well as those who were funded in order.

We then compare the reached proposals that received fellowship funding with the skipped proposals that had better scores but were not funded. Reached proposals received .18 fewer RPG awards than skipped proposals, had a 5.1 ppt lower probability of receiving an RPG award, and had a 4.7 ppt lower probability of receiving an NIH award. Both skipped and not funded in order proposals did not receive the fellowship award. However, the skipped proposals were judged to have better scientific merit than the not funded in order proposals. On average, the skipped investigators have much better outcomes than the not funded in order proposals for all outcomes. In other words, discretion results in the selection of lesser-quality, less-productive scientists.

What is the opportunity cost of skipping meritorious proposals relative to comparable proposals that were found to be the best during the review process?

We find that skipped proposals submitted .342 fewer RPG applications and, as a result, had a 5.1 ppt lower probability of receiving an RPG compared with proposals funded in order. Skipped proposals were 6.9 ppt more likely to never submit a subsequent NIH application. Not getting the award hindered the future success of the skipped applicants, making them less productive. Even though skipped did not do as well as funded in order, they still did better than not funded in order. In other words, those skipped still thrived.

The last row of Table 3 compares the best proposals funded in order with those not funded in order. The estimated ATEs are between 10 to 41 percent larger for the reached compared with not funded in order. Compared with those not funded in order, four years out reached, skipped, and funded in order all do relatively better in achieving an independent R01 award: 7.8 ppt, 8.3 ppt, and 8.6 ppt, respectively. As a robustness check, we estimated these models using the coarsened exact matching (CEM) algorithm and nearest-neighbor matching in Appendix Table A2. The signs, magnitudes, and statistical significance of the estimated effects are similar to those found in Table 3.

Although we cannot test the unconfoundedness assumption directly, Imbens (2015) recommends using propensity score matching (PSM) on pseudo-outcomes that occur before the award treatment. Given the SED data, we evaluate whether the fellowship award predicts the probability that an applicant has a PhD degree, the applicant's highest degree is in biomedicine, and the applicant's doctoral funding was from a fellowship or scholarship. Table 4 presents these results and finds that the fellowship award has no significant impact on these pseudo-outcomes. These results indicate that the unconfoundedness assumption is not violated.

V. Discussion

Our findings provide evidence that within a pool of young, ambitious scientists, peer review more accurately identifies scientific “diamonds in the rough” than NIH-institute discretion. Our results indicate that scientific leaders and policymakers do in fact have a choice to make. If they want to fund in the most efficient way possible the expansion of the scientific frontier, they should encourage peer review and fund awards without discretion until funding is exhausted. These results echo findings by Li and Agha (2015) and Gallo et al. (2014), who found that evaluation scores are correlated with subsequent research publications, indicating that the peer-review process is efficient. While scientific staff members of federal agencies do their best to keep scientific innovation moving, our results hint that the priorities of program officers and institutes may come at a cost to the scientific enterprise in terms of advancing the best science.

Interestingly, our findings also indicate that for those with competitive applications but no funding (skipped), applying for the award favors future research success even if they do not receive the award. These results are consistent with those found by Ayoubi, Pezzoni, and Visentin (2019), who demonstrate that applying for research funding in a Swiss grant competition increased publications regardless of whether funding was obtained. For those on the margin (reached), however, our results show that the fellowship award can have an impact on keeping these young scientists engaged in science. For scientific organizations looking to retain talent of a particular nature, reaching for that talent does increase their ability to stay engaged.

VI. Conclusion

Our results have implications for the debate on the validity of the peer review process. Fang, Bowen and Casadevall (2016) and Pier et al. (2018) argued that peer review cannot distinguish between proposals of comparable quality. However, the fact that applicants with skipped proposals are more likely than those with reached proposals to receive subsequent NIH funding suggests that the peer-review process can identify small differences in research-proposal quality. Our results also indicate that review scores are a good predictor of subsequent NIH applications and awards and an efficient way to allocate research funding given an alternative option of discretion.

We described in detail why regression discontinuity design (RDD) is not appropriate and should not be used for studying the impact of scientific funding when discretionary decisions overriding peer-review rankings are common. The method of evaluation must fit with the idiosyncrasies of the setting. With matching methods, we found that the NRSA F32 award keeps postdoctoral researchers engaged in NIH-funded science at higher rates than they would have otherwise experienced. Regardless of the restrictions or matching methods we used, our estimates are robust. Overall, we have demonstrated the value of a peer-review system in selecting the best talent compared with discretionary decisions. We conclude by noting that even though it has been under intense scrutiny throughout its existence and has rarely been used in its purest form, peer review is an efficient option compared with institutional discretion if the goal is to maximize the advancement of the frontier of science.

What about other methods? Fang and Casadevall (2016) and others have argued for a two-step lottery, one where a subset of highly qualified applications are selected by peer review and then funded applicants are selected by lottery. Smith (2006) argues that peer review is basically equivalent to chance because reviewers differ, and selection decisions are random based on who one happens to get as a reviewer and study section.

Perhaps there is an argument for randomization after a preliminary peer review. If it produces results similar to peer review, cost efficiencies could be realized by randomizing funding decisions. Greenberg (2008) highlights this point and states that "reliance on chance wouldn't be inferior to what's happening now, which, as it turns out, is a game of chance in the guise of informed selection. Moreover, the [cost] savings from a lottery could be recirculated to research, providing many millions of dollars for projects that would otherwise go unsupported."

There is one additional reason why allocating funding via a lottery after initial peer review is a potentially wise managerial decision of federal scientific leaders. It would actually allow policy researchers to rigorously test the true effect of federal funding on both the system and scientific discovery. Randomized control trials are common in science. Everyone understands why they are needed, yet scientists still struggle to accept this simple, highly valued method as a mechanism to study their own productivity and innovation. As Smith (2006), who also noted little difference between peer review and randomization, stated when discussing challenges to incorporate randomization, “Peer review is . . . likely to remain central to science and journals because there is no obvious alternative, and scientists and editors have a continuing belief in peer review . . . *how odd that science should be rooted in belief*.”

While scientific leaders (and scientists) may be concerned by the random assignment of funding beyond a certain threshold, perhaps they would be open to running an experiment along those lines. Such an experiment would begin with traditional peer review, with study sections then randomly divided into three groups. Group one would allow unconstrained discretion by the part of staff. Group two would allow constrained discretion—where rules governing discretion would be documented and include an audit trail detailing deviations. The third group would randomize funding selection of those applications receiving the best peer-review scores that meet a certain threshold. This kind of experiment is already taking place in New Zealand. The Health Research Council allocates two percent of its budget for “explorer grants” that provide approximately $100,000 in funding. Short proposals are screened for eligibility and then funded at random until the budget is exhausted. A recent study reported that those who were funded by the lottery supported this system, however, it did not evaluate the impact of receiving funding on scientific output (Liu et al. 2020).

Understanding what we are trying to optimize with federal scientific funding is key. This type of experiment would allow us to understand two important factors.

First, when discretionary decisions are made specifically to achieve alternative institutional goals, are those goals met? Second, if we automate decision-making to a lottery, do outcomes such as publications and subsequent grants increase, decrease, or stay the same? If outcomes stay the same or increase, then perhaps using a lottery mechanism could, in fact, be more efficient, in the sense that resources currently allocated to discretionary decisions could be reallocated to other priorities. Regardless, it is clear that if we really care about figuring out the best way to allocate the limited funding available from the federal government for scientific advancements, we need to have access to application data from federal funding agencies and, with a clear understanding of the role discretion plays in research awards, evaluate the outcomes from funding.

REFERENCES

Alberts, Bruce, Marc W. Kirschner, Shirley Tilghman, and Harold Varmus. 2015. "Opinion: Addressing Systemic Problems in the Biomedical Research Enterprise." *PNAS* 112 (7): 1912–1913. doi: 10.1073/pnas.1500969112.

Ayoubi, Charles, Michele Pezzoni, and Fabiana Visentin. 2019. "The Important Thing Is Not to Win, It Is to Take Part: What If Scientists Benefit from Participating in Research Grant Competitions?" *Research Policy* 48 (1): 84–97. doi: 10.1016/j.respol.2018.07.021.

Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2013. "National Institutes of Health Peer Review: Challenges and Avenues for Reform." *Innovation Policy and the Economy* 13: 1–22.

Azoulay, Pierre, Joshua S. Graff Zivin, Danielle Li, and Bhaven N. Sampat. 2019. "Public R&D Investments and Private-Sector Patenting: Evidence from NIH Funding Rules." *Review of Economic Studies* 86 (1): 117–152.

Benos, Dale J., Edlira Bashari, Jose M. Chaves, Amit Gaggar, Niren Kapoor, Martin LaFrance, Robert Mans, David Mayhew, Sara McGowan, Abigail Polter, Yawar Qadri, Shanta Safare, Kevin Schultz, Ryan Splittgerber, Jason Stephenson, Cristy Tower, R. Grace Walton, and Alexander Zotov. 2007. "The Ups and Downs of Peer Review." *Advances in Physiology Education* 31 (2): 145–152.

Baldwin, Melinda. 2018. "Scientific Autonomy, Public Accountability, and the Rise of 'Peer Review' in the Cold War United States." *Isis* 109 (3): 538–558.

Benavente, José Miguel, Gustavo Crespi, Lucas Figal Garone, and Alessandro Maffioli. 2012. "The Impact of National Research Funds: A Regression Discontinuity Approach to the Chilean FONDECYT." *Research Policy* 41 (8): 1461–1475.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "CEM: Coarsened Exact Matching in Stata." *The Stata Journal* 9 (4): 524–546.

Blau, David M., and Bruce A. Weinberg. 2017. "Why the US Science and Engineering Workforce is Aging Rapidly." *PNAS* 114 (15): 3879–3884.

Bol, Thijs, Mathijs de Vaan, and Arnout van de Rijt. 2018. "The Matthew Effect in Science Funding." *PNAS* 115 (19): 4887–4890.

Costello, Leslie C. 2010. "Perspective: Is NIH Funding the 'Best Science by the Best Scientists?' A Critique of the NIH R01 Research Grant Review Policies. *Academic Medicine* 85 (5): 775–779.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical versus Actuarial Judgment." *Science* 243: 1668–1674.

Fang, Ferric C., and Arturo Casadevall. 2016. "Research Funding: The Case for a Modified Lottery." *mBio* 7 (2): e00422-16.

Fang, Ferric C., Anthony Bowen, and Arturo Casadevall. 2016. "NIH Peer Review Percentile Scores Are Poorly Predictive of Grant Productivity." *eLife* 5: e13323.

Farrell, P.R., Magida L. Farrell, and M.K. Farrell. 2017. "Ancient Texts to PubMed: A Brief History of the Peer-Review Process." *Journal of Perinatology*, 37: 13–15.

Gallo, Stephen A., Afton S. Carpenter, David Irwin, Caitlin D. McPartland, Joseph Travis, Sofie Reynders, Lisa A. Thompson, and Scott R. Glisson. 2014. "The Validation of Peer Review through Research Impact Measures and the Implications for Funding Strategies." *PLoS One* 9 (9): e106474.

Gallo, Stephen A., Joanne H. Sullivan, and Scott R. Glisson. 2016. "The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications." *PLoS One* 11 (10): e0165147.

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): 1015–1019.

Godlee, Fiona, Catharine R. Gale, Christopher N. Martyn. 1998. "Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports: A Randomized Control Trial." *JAMA* 280 (3): 237–240.

Goldstein, Anna P., and Michael Kearney. 2018. "Uncertainty and Individual Discretion in Allocating Research Funds." (February 28). Available at SSRN: https://ssrn.com/abstract=3012169 or http://dx.doi.org/10.2139/ssrn.3012169.

Greenberg, Dan. 2008. "Peer Review at NIH: A Lottery Would Be Better." *Brainstorm* (blog), *Chronicle of Higher Education*. http://chronicle.com/blogs/brainstorm/peer-review-at-nih-a-lottery-would-be-better/5696.

Grilli, Luca, and Samuele Murtinu. 2011. "Econometric Evaluation of Public Policies for Science and Innovation: A Brief Guide to Practice." In *Science and Innovation Policy for the New Knowledge Economy*, edited by Massimo G. Colombo, Luca Grilli, Lucia Piscitello, and Cristina Rossi-Lamastra. Northampton, MA: Edward Elgar Publishing.

Gustafson, Thane. 1975. "The Controversy over Peer Review." *Science* 190 (4219): 1060-1066.

Hechtman, Lisa, and Jon Lorsch. 2019. "Application and Funding Trends in Fiscal Year 2018." *National Institute of General Medical Sciences Feedback Loop Blog*, National Institute of General Medical Sciences . https://loop.nigms.nih.gov/2019/04/application-and-funding-trends-in-fiscal-year-2018/.

Heggeness, Misty L., Kearney T. Gunsalus, Jose Pacas, and Gary S. McDowell. 2016. "Preparing for the 21st Century Biomedical Research Job Market: Using Census Data to Inform Policy and Career Decision-Making." SJS Working Paper, http://www.sjscience.org/article?id=570.

Heggeness, Misty L., Kearney T. Gunsalus, Jose Pacas, and Gary S. McDowell. 2017. "The New Face of Science." *Nature* 541: 21–23.

Heggeness, Misty L., Donna K. Ginther, Maria I. Larenas, and Frances D. Carter-Johnson. 2018. "The Impact of Postdoctoral Fellowships on a Future Independent Career in Federally Funded Biomedical Research." NBER Working Paper No. 24508.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2018. “Discretion in Hiring.” *The Quarterly Journal of Economics* 133 (2): 765–800.

Howell, Sabrina T. 2017. “Financing Innovation: Evidence from R&D Grants.” *American Economic Review* 107 (4): 1136–1164.

Imbens, Guido W. 2015. “Matching Methods in Practice: Three Examples.” *Journal of Human Resources* 50 (2): 373–419.

Jacob, Brian A., and Lars Lefgren. 2011a. “The Impact of NIH Postdoctoral Training Grants on Scientific Productivity.” *Research Policy* 40 (6): 864–874. doi: 10.1016/j.respol.2011.04.003.

Jacob, Brian A., and Lars Lefgren. 2011b. “The Impact of Research Grant Funding on Scientific Productivity.” *Journal of Public Economics* 95: 1168–1177.

King, Gary, and Richard Nielson. 2019. “Why Propensity Scores Should Not Be Used for Matching.” *Political Analysis* 27 (4). Copy at http://j.mp/2ovYGsW. Accessed on January 11, 2019.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions.” *The Quarterly Journal of Economics* 133 (1): 237–293.

Levitt, David G. 2010. “Careers of An Elite Cohort of U.S. Basic Life Science Postdoctoral Fellows and the Influence of Their Mentor’s Citation Record.” *BMC Medical Education* 10, article number 80. doi: 10.1186/1472-6920-10-80.

Li, Danielle, and Leila Agha. 2015. “Big Names or Big Ideas: Do Peer-Review Panels Select the Best Science Proposals?” *Science* 348 (6233): 434–438. doi: 10.1126/science.aaa0185.

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

Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely T., and Pomeroy L. 2020. “The Acceptability of Using a Lottery to Allocate Research Funding: A Survey of Applicants.” *Research Integrity and Peer Review* (5):3 https://doi.org/10.1186/s41073-019-0089-z.

Mantovani, Richard, Mary V. Look, and Emily Wuerker. 2006. *The Career Achievements of National Research Service Award Postdoctoral Trainees and Fellows: 1975–2004*. Bethesda, MD: ORC Macro.

Marsh, Herbert W., Upali W. Jayasinghe, and Nigel W. Bond. 2008. "Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability." *American Psychologist* 63 (3): 160–168.

McNutt, Robert A., Arthur T. Evans, Robert H. Fletcher, and Suzanne W. Fletcher. 1990. "The Effects of Blinding on the Quality of Peer Review." *JAMA* 263 (10): 1371–1376.

National Institutes of Health. 2011. "Enhancing peer review." Retrieved May 9, 2017. https://enhancing-peer-review.nih.gov/index.html.

National Research Council. 2011. *Research Training in the Biomedical, Behavioral, and Clinical Research Sciences*. Committee to Study the National Needs for Biomedical, Behavioral, and Clinical Research. Washington, DC: National Academies Press.

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

Pion, Georgine M. 2001. *The Early Career Progress of NRSA Predoctoral Trainees and Fellows*. Prepared for U.S. Department of Health and Human Services, National Institutes of Health, NIH Publication No. 00-4900.

Roy, Rustum. 1985. "Funding Science: The *Real* Defects of Peer Review and an Alternative to It." *Science, Technology, & Human Values* 10 (3): 73–81.

Scarpa, Toni. 2006. "Peer Review at NIH." *Science* 311: 41.

Smith, Jeffrey, and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

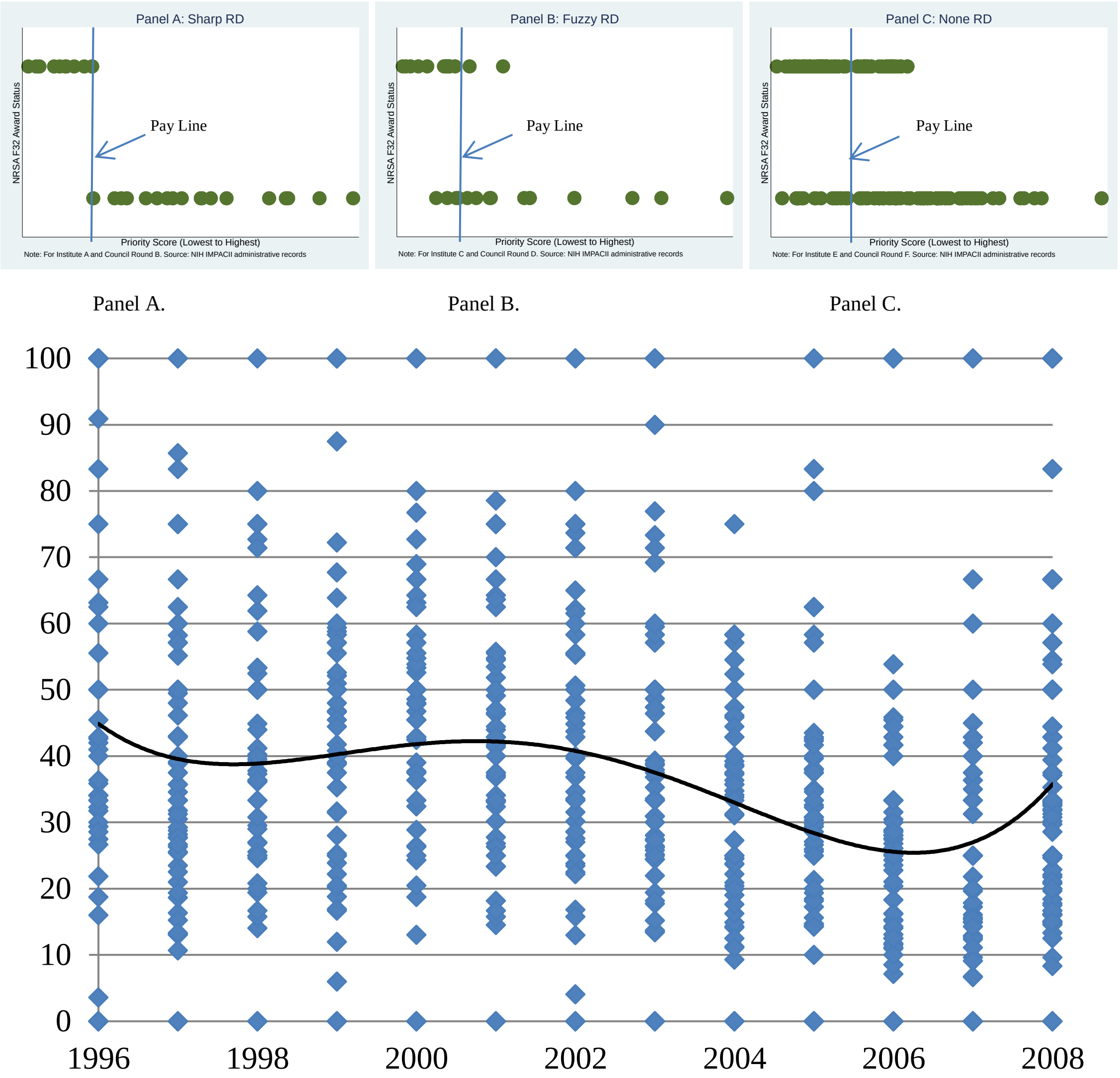
Smith, Richard. 2006. "Peer Review: A Flawed Process at the Heart of Science and Journals." *Journal of the Royal Society of Medicine* 99: 178–182.

Stahel, Philip, and Ernest E. Moore. 2014. "Peer Review for Biomedical Publications: We Can Improve the System." *BMC Medicine* 12: 179–182.

Tilghman, Shirley, and Sally Rockey. 2012. "Report of the Biomedical Research Workforce Working Group of the Advisory Committee to the NIH Director." Washington, DC: National Institutes of Health. http://acd.od.nih.gov/bwf.htm. Accessed 15 August 2014.

Travis, G. D. L., and H. M. Collins. 1991. "New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System." *Science, Technology, & Human Values* 16 (3): 322–-341.

Wessely, Simon. 1998. "Peer Review of Grant Applications: What Do We Know?" *Lancet* 352: 301–305. doi: 10.1016/S0140-6736(97)11129-1.



Panel A.

Panel B.

Panel C.

Panel D. Percent of NRSA F32 Applications Funded within Pay Line by Year, Institute, and Council Round, 1996 to 2008

FIGURE 1. THE NRSA F32 SELECTION PROCESS

Note: Figure excludes council rounds with an N<20.
Source: Authors’ calculations. National Institutes of Health IMPACII administrative records.

TABLE 1. DESCRIPTIVE STATISTICS OF APPLICANTS BY FUNDING STATUS, ANALYTICAL SAMPLE, 1996─2008

	All	F32 awarded	No F32 awarded	t-test	p-value
Review score	220.67	162.244	193.68	60.22	0.000
	(75.291)	(26.899)	(34.432)		
DEMOGRAPHICS					
Age at application	31.201	30.942	31.078	1.12	0.261
	(7.629)	(6.775)	(7.115)		
Age at application missing	0.042	0.034	0.037	0.98	0.328
	(0.200)	(0.181)	(0.189)		
Married at application	0.381	0.396	0.38	-1.84	0.066
	(0.486)	(0.489)	(0.485)		
Married at application missing	0.205	0.182	0.186	0.68	0.494
	(0.404)	(0.386)	(0.389)		
Female	0.417	0.412	0.424	1.30	0.194
	(0.493)	(0.492)	(0.494)		
Sex missing	0.051	0.038	0.055	4.90	0.000
	(0.219)	(0.190)	(0.228)		
White, non-Hispanic	0.344	0.386	0.319	-7.93	0.000
	(0.475)	(0.487)	(0.466)		
Black, non-Hispanic	0.009	0.006	0.01	2.64	0.008
	(0.095)	(0.075)	(0.098)		
Asian, non-Hispanic	0.086	0.085	0.079	-1.29	0.196
	(0.280)	(0.279)	(0.269)		
Other, non-Hispanic	0.002	0.002	0.002	-0.33	0.744
	(0.048)	(0.050)	(0.047)		
Hispanic	0.032	0.029	0.029	-0.07	0.946
	(0.175)	(0.168)	(0.167)		
Race missing	0.544	0.509	0.579	7.97	0.000
	(0.498)	(0.500)	(0.494)		
EDUCATION and TRAINING					
MD	0.086	0.084	0.08	-0.74	0.459
	(0.280)	(0.277)	(0.272)		
MD/PhD	0.032	0.035	0.03	-1.80	0.073
	(0.176)	(0.185)	(0.170)		
PhD	0.867	0.87	0.874	0.67	0.506
	(0.340)	(0.336)	(0.332)		
Other Degree	0.016	0.01	0.016	2.76	0.006
	(0.125)	(0.102)	(0.125)		
Biomedical degree	0.594	0.614	0.617	0.33	0.739
	(0.491)	(0.487)	(0.486)		
Physical Science degree	0.129	0.129	0.124	-0.72	0.471
	(0.335)	(0.335)	(0.330)		
Social Science degree	0.069	0.074	0.071	-0.67	0.506
	(0.253)	(0.261)	(0.256)		
Prior T32 Predoc Award	0.021	0.024	0.027	0.96	0.339
	(0.144)	(0.154)	(0.162)		
Prior T32 Postdoc Award	0.019	0.016	0.02	1.65	0.099
	(0.136)	(0.127)	(0.141)		
Prior NRSA Predoctoral Fellowship	0.001	0.001	0	-1.64	0.101
	(0.023)	(0.023)	0.000		
OUTCOME VARIABLES					
Number of RPG Awards	0.387	0.586	0.367	-10.24	0.000
	(1.071)	(1.281)	(1.095)		
Number of RPG Applications	1.94	2.752	1.698	-12.51	0.000
	(4.360)	(5.158)	(4.066)		
Probability of RPG	0.183	0.266	0.165	-13.69	0.000
	(0.386)	(0.442)	(0.372)		
Probability of R01	0.133	0.204	0.122	-12.25	0.000
	(0.339)	(0.403)	(0.328)		
Probability of Never Receiving an RPG	0.678	0.579	0.713	16.02	0.000
	(0.467)	(0.494)	(0.452)		
N	14,276	9,276	5,000		

Source: Authors' calculations. National Institutes of Health IMPACII and NIH/NSF Survey of Earned Doctorates.

TABLE 2. PROBABILITY OF NIH F32 PROPOSAL BEING FUNDED BY DISCRETION (REACHED) OR NOT FUNDED BY DISCRETION (SKIPPED)

	(1)	(2)
VARIABLES	Reached	Skipped
Second Council Round	-0.005	-0.006
	[0.006]	[0.006]
Third Council Round	-0.005	-0.004
	[0.005]	[0.005]
FY 1997	-0.001	-0.011
	[0.011]	[0.009]
FY 1998	0.012	0.007
	[0.012]	[0.011]
FY 1999	-0.006	-0.003
	[0.011]	[0.011]
FY 2000	-0.009	0.005
	[0.011]	[0.011]
FY 2001	-0.008	0.002
	[0.011]	[0.011]
FY 2002	-0.008	-0.015
	[0.012]	[0.010]
FY 2003	-0.012	-0.029***
	[0.011]	[0.009]
FY 2004	-0.012	-0.027**
	[0.010]	[0.009]
FY 2005	-0.015	-0.034***
	[0.010]	[0.008]
FY 2006	-0.025**	-0.038***
	[0.009]	[0.008]
FY 2007	-0.026**	-0.030***
	[0.009]	[0.008]
FY 2008	-0.030***	-0.035***
	[0.009]	[0.008]
N	14,276	14,276

Note: Robust Standard errors in brackets. *** p<0.001, ** p<0.01, * p<0.05

Source: Authors’ calculations. IMPACII and NIH/NSF Survey of Earned Doctorates, 1996 to 2008.

TABLE 3. PROPENSITY SCORE MATCHING (PSM) ESTIMATES OF THE IMPACT ON OUTCOMES BY COMPARATIVE FUNDING-STATUS TYPES

	Number	Number	Probability	Probability	Never
VARIABLES	RPG Awards	RPG Applications	RPG	R01	RPG
Reach vs. Not Funded	0.174***	0.801***	0.071***	0.078***	-0.119***
N = 5,215	[0.043]	[0.167]	[0.017]	[0.015]	[0.018]
Reach vs. In Order	-0.096*	-0.495**	-0.050**	-0.038*	0.063**
N = 9,062	[0.048]	[0.183]	[0.018]	[0.016]	[0.019]
Reach vs. Skip	-0.176**	-0.333	-0.051**	-0.047**	0.025
N=2,538	[0.063]	[0.220]	[0.019]	[0.017]	[0.022]
Skip vs. Not Funded	0.243***	0.655***	0.085***	0.083***	-0.083***
N=5,211	[0.053]	[0.181]	[0.019]	[0.016]	[0.021]
Skip vs. In Order	-0.047	-0.342*	-0.051**	-0.025	0.069***
N=9,058	[0.048]	[0.165]	[0.016]	[0.014]	[0.019]
In Order vs. Not Funded	0.246***	0.983***	0.106***	0.086***	-0.140***
N = 11,735	[0.034]	[0.153]	[0.012]	[0.010]	[0.014]

Note: Robust Standard errors in brackets. *** p<0.001, ** p<0.01, * p<0.05

Source: Authors’ calculations. IMPACII and NIH/NSF Survey of Earned Doctorates, 1996 to 2008.

TABLE 4. COUNTERFACTUAL TREATMENT EFFECTS WITH PSEUDO-TREATMENTS

	(1)	(3)	(5)
VARIABLES	PhD degree	Biomedical degree	Fellowship or scholarship PhD funding
ATE	0.007	0.001	0.005
	(0.007)	(0.010)	(0.010)
ATT	0.005	0.007	0.004
	(0.008)	(0.012)	(0.012)
Observations	14,273	14,273	14,273

Appendix.

APPENDIX TABLE A1. PROBIT REGRESSIONS ON EVER RECEIVING AN AWARD, 1996-2008

	(1)	(2)	(3)	(4)
	Full sample	Full sample	Analysis sample	Analysis sample
Review score		-0.006		-0.008
		(0.000)		(0.000)
Age (missing = <26)				
Age = 27	-0.107	-0.033	-0.053	0.021
	(0.024)	(0.030)	(0.036)	(0.039)
Age = 28	-0.040	-0.031	-0.034	-0.028
	(0.022)	(0.027)	(0.030)	(0.036)
Age = 29	-0.056	-0.022	-0.050	-0.016
	(0.021)	(0.026)	(0.029)	(0.035)
Age = 30	-0.061	-0.013	-0.041	0.002
	(0.021)	(0.026)	(0.029)	(0.034)
Age = 31	-0.072	-0.031	-0.035	0.024
	(0.020)	(0.026)	(0.029)	(0.033)
Age = 32	-0.088	-0.039	-0.060	-0.005
	(0.020)	(0.026)	(0.030)	(0.034)
Age = 33	-0.086	-0.017	-0.045	0.027
	(0.021)	(0.027)	(0.030)	(0.033)
Age = 34	-0.105	-0.031	-0.055	0.016
	(0.021)	(0.027)	(0.031)	(0.035)
Age = 35 or 36	-0.149	-0.035	-0.057	0.028
	(0.020)	(0.026)	(0.031)	(0.034)
Age = 37 or 38	-0.181	-0.068	-0.085	0.014
	(0.020)	(0.027)	(0.034)	(0.036)
Age > 38	-0.248	-0.104	-0.137	-0.019
	(0.017)	(0.026)	(0.034)	(0.037)
Marital status (missing = not married)				
Married	0.009	0.014	0.016	0.018
	(0.007)	(0.008)	(0.009)	(0.010)
Marital status missing	0.093	0.057	0.044	0.086
	(0.062)	(0.090)	(0.085)	(0.092)
Sex (missing = male)				
Female	-0.010	-0.008	-0.009	-0.006
	(0.006)	(0.007)	(0.009)	(0.009)
Sex missing	-0.120	-0.098	-0.115	-0.103
	(0.014)	(0.017)	(0.023)	(0.026)
Race and ethnicity (missing = White, non-Hispanic)				
Black, non-Hispanic	-0.156	-0.099	-0.146	-0.103
	(0.030)	(0.036)	(0.055)	(0.055)
Asian, non-Hispanic	-0.038	-0.021	-0.025	-0.016
	(0.011)	(0.013)	(0.016)	(0.017)
Other race, non-Hispanic	0.135	0.167	0.064	0.048
	(0.063)	(0.078)	(0.074)	(0.074)
Hispanic	-0.022	0.001	0.014	0.033
	(0.018)	(0.021)	(0.025)	(0.025)
Race missing	0.005	0.013	0.011	0.016
	(0.008)	(0.009)	(0.010)	(0.011)
MD	0.088	0.033	-0.009	-0.019
	(0.030)	(0.033)	(0.042)	(0.045)
MD/PhD	0.189	0.098	0.076	0.016
	(0.031)	(0.037)	(0.039)	(0.046)
PhD	0.126	0.093	0.060	0.049
	(0.024)	(0.027)	(0.039)	(0.041)
Biomedical science degree	0.145	0.065	0.041	0.073
	(0.059)	(0.086)	(0.088)	(0.101)
Physical science degree	0.107	0.054	0.052	0.094
	(0.062)	(0.090)	(0.084)	(0.089)
Social science degree	0.133	0.022	0.017	0.031
	(0.062)	(0.089)	(0.087)	(0.097)
Prior T32 Predoc Award	0.067	-0.004	-0.024	-0.045
	(0.022)	(0.025)	(0.026)	(0.028)
Prior T32 Postdoc Award	-0.061	-0.056	-0.048	-0.033

	(0.022)	(0.025)	(0.031)	(0.033)
Prior NRSA Predoc Award	-0.002	0.128		
	(0.140)	(0.260)		
Observations	27,504	25,719	14,268	14,268

Notes: Robust standard errors in parentheses. All specifications include controls for IC and council rounds.
Source: Authors’ calculations. IMPACII and NIH/NSF Survey of Earned Doctorates.

APPENDIX TABLE A2. Nearest Neighbor Estimates of Discretion versus Scientific Review on Career Outcomes

VARIABLES	Number RPG Awards	Number RPG Applications	Probability RPG	Probability R01	Never RPG
Reach vs. Not Funded	0.168**	0.709***	0.067***	0.071***	-0.093***
N = 5,160	[0.052]	[0.200]	[0.020]	[0.017]	[0.024]
Reach vs. In Order	-0.172**	-0.624**	-0.095***	-0.061**	0.099***
N = 8,502	[0.055]	[0.229]	[0.022]	[0.019]	[0.029]
Reach vs. Skip	-0.138	-0.101	-0.035	-0.031	0.003
N=2,471	[0.079]	[0.286]	[0.025]	[0.022]	[0.031]
Skip vs. Not Funded	0.166*	0.612**	0.058*	0.062**	-0.091**
N=5,151	[0.071]	[0.231]	[0.025]	[0.023]	[0.030]
Skip vs. In Order	0.040	-0.258	-0.017	-0.003	0.060**
N=8,493	[0.054]	[0.186]	[0.017]	[0.016]	[0.019]
In Order vs. Not Funded	0.201***	0.739**	0.090***	0.075***	-0.134***
N = 11,182	[0.049]	[0.238]	[0.019]	[0.016]	[0.023]

Source: IMPACII and NIH/NSF Survey of Earned Doctorates, 1996 to 2008
Note: Robust Standard errors in brackets. *** p<0.001, ** p<0.01, * p<0.05.