

NBER WORKING PAPER SERIES

NIH PEER REVIEW:
CHALLENGES AND AVENUES FOR REFORM

Pierre Azoulay
Joshua S. Graff Zivin
Gustavo Manso

Working Paper 18116
<http://www.nber.org/papers/w18116>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2012

This paper was prepared for the NBER meeting on Innovation Policy and the Economy. We are grateful to Scott Stern, Josh Lerner, Danielle Li, George Chacko, Stefano Bertuzzi, and Paula Stephan for helpful conversations and to the National Science Foundation for support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2012 by Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

NIH Peer Review: Challenges and Avenues for Reform
Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso
NBER Working Paper No. 18116
June 2012
JEL No. O31,O32

ABSTRACT

The National Institute of Health (NIH), through its extramural grant program, is the primary public funder of health-related research in the United States. Peer review at NIH is organized around the twin principles of investigator initiation and rigorous peer review, and this combination has long been a model that science funding agencies throughout the world seek to emulate. However, lean budgets and the rapidly changing ecosystem within which scientific inquiry takes place have led many to ask whether the peer-review practices inherited from the immediate post-war era are still well-suited to twenty first century realities. In this essay, we examine two salient issues: (1) the aging of the scientist population supported by NIH and (2) the innovativeness of the research supported by the institutes. We identify potential avenues for reform as well as a means for implementing and evaluating them.

Pierre Azoulay
MIT Sloan School of Management
100 Main Street, E62-482
Cambridge, MA 02142
and NBER
pazoulay@mit.edu

Gustavo Manso
University of California at Berkeley
manso@haas.berkeley.edu

Joshua S. Graff Zivin
University of California, San Diego
9500 Gilman Drive, MC 0519
La Jolla, CA 92093-0519
and NBER
jgraffzivin@ucsd.edu

NIH Peer Review: Challenges and Avenues for Reform

Pierre Azoulay, MIT & NBER
Joshua S. Graff Zivin, UCSD & NBER
Gustavo Manso, University of California at Berkeley

Executive Summary

The National Institute of Health (NIH), through its extramural grant program, is the primary public funder of health-related research in the United States. Peer review at NIH is organized around the twin principles of investigator initiation and rigorous peer review, and this combination has long been a model that science funding agencies throughout the world seek to emulate. However, lean budgets and the rapidly changing ecosystem within which scientific inquiry takes place have led many to ask whether the peer-review practices inherited from the immediate post-war era are still well-suited to twenty first century realities. In this essay, we examine two salient issues: (1) the aging of the scientist population supported by NIH and (2) the innovativeness of the research supported by the institutes. We identify potential avenues for reform as well as a means for implementing and evaluating them.

Introduction

The National Institutes of Health (NIH) is the primary organization within the United States government with responsibilities for health-related research. They are the single largest funder of research in the life sciences, spending approximately \$30 billion dollars per year on biomedical research. The principal goals of the NIH are to foster innovative scientific research and to support high-level training of the scientific workforce in order to protect and improve human health. While the NIH has long been the envy of science funding agencies the world over, lean budgets and the rapidly changing ecosystem within which scientific inquiry takes place have led many to ask whether the NIH is doing as well as it could be in meeting its goals.

In this essay, we examine two salient issues – the aging of the scientific workforce and the innovativeness of the research supported by the Institutes. We will not focus on the optimal mix of private and public funding of scientific research, although we will revisit the issue in our conclusions when we discuss institutional re-design in the context of the larger research ecosystem. For now, we will take public funding as given and focus our attention on the levers

available to policy makers to help the agency achieve its goals. We begin by extolling the virtues of the extraordinary system put in place to manage the extramural portion of this public funding, which supports scientific activities at universities and various research institutions.¹

The foundation of the NIH extramural program is the peer-review system that evaluates investigator-initiated applications. Copied by so many funding agencies that its features appear unremarkable, the notion that scientific inquiry should be driven from the bottom up, that individual scientists were in the best position to propose new and innovative research programs, and that those programs would be best evaluated by a panel of volunteer experts working in related fields, was nothing short of revolutionary when the basic system was launched in the late 1940s.

Equally impressive is the separation of funding decisions from scientific evaluation, allowing reviewers to focus their attention on scientific merit, thereby creating a strong meritocratic ethos. This program has been immensely successful. More than 75 Nobel prizes have been awarded for extramural NIH-supported research. In the process, our understanding of the human body and its interaction with the environment has expanded tremendously. Vaccines have been developed that eliminate the threats from certain diseases, deaths from cardiovascular disease have been greatly reduced, and some cancers have been cured. These accomplishments are all the more astonishing considering the significant efforts put forth by Congress, disease advocates, and disciplinary organizations to influence the funded research agenda (Hegde and Sampat 2011).

Yet, there is an increasing sense that the institutional structure that served the life sciences community of the 20th century so well may be less transformative in the 21st century. Science is becoming increasingly specialized and emerging fields may find it difficult to identify an appropriate venue to evaluate their research. At the same time, legacy fields benefit from study sections and commensurate resource allocations based on historical scientific contributions. This mismatch of expertise between applicants and reviewers, coupled with a strong emphasis on the

feasibility of the research proposal embodied in requirements for preliminary data, has led many to worry that the system rewards incremental research at the expense of work that could be more innovative. The rising age at which the average scientist first becomes the principal investigator on a major NIH grant is also a troubling sign of a system that has not evolved with the field. The primary goal of this essay is to identify potential avenues for reform as well as a means for implementing and evaluating them.

The remainder of this essay is structured as follows. We will begin with a brief overview of the NIH funding system. This will be followed by a more detailed discussion of the two principal avenues open for a meaningful reform of peer review as it applies to grant funding. We also re-examine the role of NIH grants in a more diverse and fragmented research funding environment. Finally, we conclude with some thoughts on the returns to adopting a more systematic approach for policy experimentation and evaluation.

An Overview of the NIH

The National Institutes of Health is made up of 27 different units called Institutes and Centers,² each with a distinct, though sometimes overlapping, research agenda. For example, the National Institute for Mental Health, as the name suggests, focuses on mental health related research. It clearly shares interests with the National Institute of Aging on issues related to dementia. Nearly all Institutes receive their funding directly from Congress, and manage their own budgets.

Approximately ten percent of the overall NIH budget is dedicated to the intramural research program, with almost all Institutes providing some support. The program directly supports about 6,000 scientists working within the federal laboratories on NIH Campuses. The NIH main campus in Bethesda, Maryland is also home to the largest hospital in the world that is exclusively dedicated to clinical research – the NIH Clinical Center. The intramural program provides long-term funding and considerable intellectual freedom with the intention of supporting high-risk and high-reward research that would be difficult to support in the broader research community through the extramural program (NIH 2009). While the intramural program is responsible for

numerous groundbreaking discoveries, opacity with regards to the allocation of money and research effort within the program makes a careful examination of the strengths and weaknesses of this operation quite challenging.

The operations of the extramural program, in contrast, are quite transparent. Moreover, it is a much more significant piece of the NIH budget – more than 80% of the total budget supports extramural research through competitive grants that are awarded to universities, medical schools, and other research institutions, primarily in the United States. The largest and most established of these grant mechanisms is the R01, a project-based renewable research grant which 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 dollars spread over 3 to 5 years and the application success rate is approximately 20 percent (Li 2012).

Requests for proposals identify priority areas, but investigators are also free to submit applications on unsolicited topics under the extramural research program. All applications are assigned to a review committee comprised of scientific peers, generally known as a study section. Reviewers are asked to ignore budgetary issues, limiting their attention to scientific and technical merit on the basis of five criteria: (1) Significance [does the project address an important issue?]; (2) Approach [is the methodology sound?]; (3) Innovation [is the research novel?]; (4) Investigator [are the skills of the research team well matched to the project?]; and (5) Environment [is the place in which the work will take place conducive to project success?].

The review process within a study section proceeds in the following manner. Each application undergoes an initial review by three members of the section. Based on these preliminary evaluations, the strongest grants (typically half) are discussed in a full committee meeting. After an open discussion before the full complement of section members, everyone is given the opportunity to revise their initial scoring based on the group deliberations before anonymously submitting their final scores. The overall priority score for the proposal is based on the average

across all study section members. Scores are then normalized within review groups through the assignment of percentile scores to facilitate funding decisions.³

As noted earlier, funding decisions are decoupled from the scientific review and determined by program areas within the Institutes. In essence, each decision making unit (e.g., Division, Program, Branch) within an Institute is allocated a fixed annual budget. Units then fund new projects in order of their priority score until their budget, net of encumbered funds for ongoing grants awarded in previous years, is exhausted. A grant's score is generally the sole determinant of the funding decision, irrespective of proposal costs (assuming they are deemed reasonable).⁴ Researchers who do not receive funding are given the opportunity to respond to reviewer criticisms and submit an amended application. A general overview of the extramural research application and review process is summarized in Figure 1.

This decentralized, investigator-initiated, extramural grant system is one of the distinctive hallmarks of the US national system of innovation. Relative to other funding mechanisms such as block grants, the system provides high-powered incentives to individual scientists while maintaining intellectual freedom. Of course, peer review also consumes a great deal of resources – for both the creation and the evaluation of proposals – and this effort is wasted for those proposals that are not eventually funded.

We turn next to the examination of two challenges that have become increasingly salient. First, recent evidence points to a “graying” of the scientific workforce and a worry that young scientists are increasingly locked out of the system. Second, scientists have voiced the concern that NIH study sections are notoriously risk-averse and often insist on a great deal of preliminary evidence before deciding to fund a project. This often leads researchers to resubmit their applications several times and to multiply the number of applications, taking time away from productive research activities. More importantly, the evaluation criteria might lead investigators to choose incremental research topics with sure and short-term payoffs over higher-risk explorations into the scientific unknown, which will often result in failure.

Aging of the NIH-supported Scientific Workforce

Figure 2 displays the distribution of career age (years elapsed since award of the highest doctoral degree) at the time of first R01 grant award, and how it evolved over time. It is immediately apparent that the distribution has steadily shifted to the right over time. In fact, the average career age of new NIH grantees almost doubled between 1965 and 2005, going from 7.25 years to 12.8 years. What explains this dramatic increase? Several interpretations of this stylized fact can be offered, and each of these interpretations have very different policy implications.

Burden of knowledge. The most benign interpretation is due to Jones (2009;2010) and Jones and Weinberg (2011). These authors emphasize that because the scientific frontier continually shifts outward, one possible response of would-be scientists is to lengthen the time they spend in training so as to bring themselves to the frontier. If the graying trend documented above reflects mostly this mechanism, this poses a serious challenge for long-run economic growth. Moreover, it may induce the best and the brightest to select careers in which they can become productive at a younger age, such as finance or consulting. But if increased knowledge burden is the most relevant prism through which we should interpret the pattern documented in Figure 2, there would seem to be very little that scientific agencies in general, and the NIH in particular, could or should do to counteract this trend.

Postdoc positions as “holding tanks.” Stephan and Ma (2005) focuses on the market for postdoctoral fellows and note that the length of the postdoctoral training period has increased over time. A specific worry is that this lengthening stems from the fact that postdocs provide an army of low-paid “worker bees” for the laboratories of tenure-track faculty, rather than a period during which postdoctoral fellows will acquire skills necessary to be productive, independent scientists. Relative to graduate students, postdocs are cheap since they do not pay tuition, and are not distracted by the need to take advanced doctoral courses. In the life sciences, most postdocs salaries are funded from research grants (such as R01s) rather than fellowship or training programs. As a result, postdocs have fairly little professional autonomy and are greatly

dependent on their faculty sponsor (Stephan 2012). If this “holding tank” interpretation helps explain the lengthening of the postdoctoral fellowship period, there may be ample scope to alter the incentives of NIH investigators through the grant system to better align their interests with those of their postdocs. Furthermore, postdoctoral fellowships that genuinely impart needed research skills will help ensure that the pipeline of trained scientists will not leak prematurely with departures to industry or to non-scientific careers.

Stepping on the grant “treadmill.” Most relevant to this essay, however, is the hypothesis that the structure of peer review, in and of itself, contributes to the graying of the scientific workforce. Most observers of the NIH agree that study sections put a great amount of weight on preliminary results when evaluating grant applications. Over time, this has led to a situation in which grants de facto function more like a prize system that recognizes competent investigations that have already been concluded, rather than as inputs into the production of novel research results (Lazear 1997). In a nutshell, it is not atypical for a grant applicant to have completed (though not published) two out of three aims that constitute the typical grant application.

One view is that these intertemporal logrolling strategies (whereby funding received today will support the preparation of a competing continuation for the same grant) are innocuous. This might well be the case for principal investigators whose laboratories are well funded and whose careers are already established. From the point of view of new investigators, the backward-looking nature of peer review creates a situation in which they might struggle to step onto the grant treadmill for the first time.

In addition, there is ample anecdotal evidence that “grantsmanship” – the craft of writing grant applications such that they will appeal to study section members, holding the quality of the underlying ideas constant – has increased in importance over time. Telltale signs of this phenomenon include the rise of an entire grant writing ecosystem complete with its consultants, seminars, and blogs. Most practicing scientists suspect that the returns to experience in this craft are extremely high, once again putting younger investigators at a disadvantage.

Combined with the fact that many life science academics are on “soft money” contracts (such that their department have no obligations to unfunded faculty beyond the traditional start-up package), the incentives provided to scientists by NIH peer review are extremely high-powered. The system provides little opportunity for second chances, reinvention, or change in scientific direction. Its harshness may be a virtue, provided that what is rewarded is the quality of ideas. The case is less compelling if adherence to norms of grantsmanship and scientific fads increases the likelihood of funding.

But is it in fact the case that, conditional on getting a permanent position in the academic life sciences, younger scientists find it more difficult to achieve funding independence (by securing their first R01 grant)? The aging trend documented in Figure 2 is consistent with this hypothesis, but not entirely convincing since it essentially selects the sample on the outcome of interest. To shed light on this question, we use the faculty roster of the Association of American Medical Colleges (AAMC), which we match with the NIH Compound Grant Applicant File (CGAF). In particular, we focus on 97,034 medical school faculty members who began their careers between 1981 and 2003. 14.14% (13,720) of these faculty members have received at least one R01 grant by 2007, the final year of the analysis. Controlling for gender, degree (MD vs. PhD vs. MD/PhD), and main department affiliation (internal medicine, pathology, surgery, etc.), we estimate a simple Cox proportional hazard model where the key explanatory variables are indicator variables for the year in which each of these scientists achieved their highest doctoral degrees (the dates range from 1978 to 1999).⁵ Figure 3 graphs the coefficient estimates corresponding to the degree year indicator variables, along with the 95% confidence interval around the estimates.

The results certainly suggest that the “treadmill” hypothesis is empirically relevant, and are even more startling if one recalls that many of the scientists who graduated between 1986 and 1995 would have benefited from the doubling of the NIH budget that occurred between the years 1998 and 2003.

The grant system and scientific risk-taking

In 1980, a scientist from the University of Utah, Mario Capecchi, sought R01 funding with a grant proposal containing three distinct projects. The NIH peer-reviewers liked the first two projects, which were building on Capecchi's past research efforts, but they were unanimously negative in their appraisal of the third project, in which he proposed to develop gene targeting in mammalian cells. They deemed the probability that the newly introduced DNA would ever find its matching sequence within the host genome vanishingly small, and the experiments not worthy of pursuit. The NIH funded the grant despite this misgiving, but strongly recommended that Capecchi drop the third project. In his retelling of the story, the scientist writes that despite this unambiguous advice, he chose to put almost all his efforts into the third project: *"It was a big gamble. Had I failed to obtain strong supporting data within the designated time frame, our NIH funding would have come to an abrupt end and we would not be talking about gene targeting today"* (Capecchi 2008). Fortunately, within four years, Capecchi and his team obtained strong evidence for the feasibility of gene targeting in mammalian cells, and in 1984 the grant was renewed enthusiastically. Dispelling any doubt that he had misinterpreted the feedback from reviewers in 1980, the critique for the 1984 competitive renewal started, *"We are glad that you didn't follow our advice."* The story does not stop there. In September 2007, Capecchi shared the Nobel prize for developing the techniques to make knockout mice with Oliver Smithies and Martin Evans. Such mice have allowed scientists to learn the roles of thousands of mammalian genes and provided laboratory models of human afflictions in which to test potential therapies.

In the anecdote reported above, the scientist was undeterred by his peers' advice to "play it safe," and eventually saw his bold ideas prevail. But is this heroic story (with a happy ending) typical? Should we worry that peer review, as practiced by NIH, provides strong incentives to perform normal science, but discourages the exploration of new, untested approaches? It is an often-heard complaint among academic biomedical researchers that study sections' prickliness encourages them to pursue relatively safe avenues that build directly on previous results, at the expense of truly exploratory research (Kaplan 2005; Kolata 2009; McKnight 2009). Do these

complaints constitute more than self-interested grumblings from scientists NIH declined to support? And if there is merit to these criticisms, what are the particular design features of the current grant system that are most likely responsible for the emphasis on projects with near-term payoffs?

Azoulay et al. (2011) study the careers of researchers who can be funded through two very distinct mechanisms: investigator-initiated R01 grants from the NIH, or support from the Howard Hughes Medical Institute (HHMI) through its investigator program. HHMI, a non-profit medical research organization, plays a powerful role in advancing biomedical research and science education in the United States. The Institute commits almost \$700 million a year – a larger amount than the NSF biological sciences program, for example. HHMI's stated goal is to push the boundaries of knowledge in some of the most important areas of biological research. The program urges its researchers to take risks, explore unproven avenues, and embrace the unknown – even if it means uncertainty or the chance of failure. Once selected, researchers continue to be based at their institutions, typically leading a research group of 10 to 25 students, postdoctoral associates and technicians.

HHMI has adopted practices that, in theory (Manso 2011), should provide strong incentives for breakthrough scientific discoveries: the award cycles are long – five years, and typically renewed at least once;⁶ the review process provides detailed, high-quality feedback to the researcher; and the program selects “people, not projects,” which allows (and in fact encourages) the quick reallocation of resources to new approaches when the initial ones are not fruitful (Ioannidis 2011). In contrast, the typical R01 grant cycle lasts three years, and renewal is not very forgiving of failure (see Figure 4 for a comparison between the typical time horizon for the projects funded by NIH and HHMI, respectively). Feedback on performance is limited in its depth. Importantly, the NIH funds projects with clearly defined deliverables, not individual scientists.⁷

The contrast between the HHMI and NIH grant mechanisms naturally leads to the question of whether HHMI-style incentives result in a higher rate of production of particularly valuable ideas. To answer this question, the authors identify a group of NIH-funded scientists that are

appropriate controls for the researchers selected into the HHMI program. In the absence of a plausible source of exogenous variation for HHMI appointment, they estimate the treatment effect of the program by contrasting HHMI-funded scientists' output with that of a group of NIH-funded scientists who focus their research on the same subfields of the life sciences as HHMI investigators, and received prestigious early career prizes. In addition, using an in-depth understanding of the HHMI appointment process, they cull from this control group scientists who look similar to the HHMI investigators on the observable factors that they know to be relevant for selection into the HHMI program. Finally, they are able to distinguish particularly creative contributions from incremental advances. The bulk of their analysis focuses on the number of publications that fall into different quantiles of the vintage-specific, article-level distribution of citations.

Their results provide support for the hypothesis that the HHMI investigator program appropriately designed incentives stimulate exploration. In particular, they find that the effect of selection into the HHMI program increases as they examine higher quantiles of the distribution of citations. Figure 5 displays the time path of the average number of publications in the top 5% of the citation distribution, for HHMIs and NIH-supported controls. Loosely, the figure depicts the difference between the change in outcomes for the HHMIs and for a pseudo-population of control scientists matched on observables. A necessary condition for the plausibility of this exercise is that the treated and control groups display parallel output trends prior to the appointment event. This appears to be the case here.

Interestingly, for three years after appointment, the outcomes for treated and control scientists continue to track each other closely. The figure even suggests that the control group (appropriately selected on observables) briefly outpaces the treatment group following the appointment, consistent with Manso's (2011) theory which predicts both slower and more variable returns under an exploration incentive scheme. This difference is not statistically significant, however, which is perhaps unsurprising given the sample's relatively small size. HHMI investigators' output begins to diverge from that of the controls only four to five years after appointment.

Success is also more frequent among HHMI investigators when assessed with respect to scientists' own citation impact prior to appointment, rather than relative to a universal citation benchmark. Symmetrically, the paper presents evidence that HHMI-supported scientists “flop” more often than NIH-funded controls: they publish 35% more articles that fail to clear the (vintage-adjusted) citation bar of their least well-cited pre-appointment work. This provides suggestive evidence that HHMI investigators are not simply rising stars anointed by the program. Rather, they appear to place riskier scientific bets after their appointment.

The authors bolster the case for the exploration hypothesis by focusing on various attributes of these scientists' research agenda. They show that the work of HHMI investigators is characterized by more novel keywords than controls. These keywords are also more likely to change after their HHMI appointment. Moreover, their research is cited by a more diverse set of journals, both relative to controls and to the pre-appointment period.

Policy Responses to Date

Support for investigators at the early-career stage. The aging of the NIH-funded scientist population has not escaped the attention of policy makers. NIH's Office of Extramural Research, in its statement of commitment to new and early stage investigators, states:

“New investigators are the innovators of the future – they bring fresh ideas and technologies to existing biomedical research problems, and they pioneer new areas of investigation. Entry of new investigators into the ranks of independent, NIH-funded researchers is essential to the health of this country's biomedical research enterprise.”

Over time, NIH administrators' standard response has been to create new grant mechanisms, specifically targeted to young investigators. The R23 “New Investigator Research Awards” were inaugurated in 1971, before being supplanted by R29 “First Independent Research Support and Transition Awards” in 1986.⁸ The newest support mechanism for young investigators is the K99/R00 Award (known as “kangaroo grants” in the vernacular of academic life scientists), an

hybrid of traditional training and research grants.⁹ Then as now, the hope is that the investigators supported through this mechanism will be in a good position to secure R01 funding.

More recently, NIH refined its new investigator policies involving the identification of Early Stage Investigators (ESIs). ESIs are scientists who are within 10 years of completing their terminal research degree or within 10 years of completing their medical residency at the time they apply for R01 grants. In order to encourage a reduction in the period of training leading to independence, the NIH Institutes and Centers are supposed to monitor their new investigator pool to make sure that approximately half have ESI status. Applications from ESIs, like those from all new investigators, are given special consideration during peer review and at the time of funding. Peer reviewers are instructed to focus more on the proposed approach than on the track record, and to expect less preliminary information than might be provided by an established investigator (NIH 2008). It is too early to gauge whether these efforts have in fact resulted in a slowdown of the aging trend documented above.

Support for high-risk scientific exploration. NIH administrators have long been aware of the need to encourage risk-taking on the part of their investigator population. Initiated in 1987, the MERIT (Method to Extend Research in Time) R37 Award program extends funding for up to 5 years (but typically 3 years) to a select number of NIH-funded investigators *“who have demonstrated superior competence, outstanding productivity during their previous research endeavors and are leaders in their field with paradigm-shifting ideas.”* The specific details vary across the component institutes of the NIH, but the essential feature of the program is that only researchers holding an R01 grant in its second or later cycle are eligible. Further, the application for renewal must be scored in the top percentile in a given funding cycle. While the MERIT designation is a prestigious award for mid-career investigators, it pertains to a particular project, not to the scientists' overall portfolio of projects. Moreover, it is not clear whether the longer time-horizon stimulates exploratory research, both because the award targets those investigators who are the most ensconced in the current system, and because MERIT status is conferred only once, and therefore cannot be a source of ongoing incentives for the recipients.

Perhaps stung by the recent attention garnered by the HHMI investigator program, NIH has sought to emulate its success by creating a grant mechanism, the NIH Director's Pioneer Award (or DP1 mechanism in the NIH vernacular), which espouses the same "transformative" rhetoric. Every year since 2006, NIH selects a cohort of approximately 15 scientists "*who propose pioneering – and possibly transforming approaches – to major challenges in biomedical and behavioral research.*" In contrast with HHMI investigator status, however, Pioneer Award status is not renewable and does not really alter the time horizon of the evaluation, since Pioneer awardees will need to compete for R01 grants once their pioneer status expires after 5 years.

Prospects for Reform

We offer some conjectures regarding the eventual success of the policies designed to address these two challenges. With respect to the aging of the NIH-supported investigator population, we begin with the cautionary note that, from a theoretical standpoint, it is not entirely clear whether scientific progress will be furthered by a policy that evaluates the merits of grant applications from seasoned and inexperienced scientists using different criteria.

Yet, there is reason to believe that the "burden of knowledge" explanation falls short of accounting fully for the pattern documented in Figure 2. Certainly, the fact that the hazard of securing NIH support for the first time is sharply increasing in investigator vintage suggests that institutional factors are at least partly responsible. Further, policy makers are surely correct to worry that a prolongation of the period during which young investigators are in a limbo where they are neither trainees, nor master of their own destiny and agenda, could act as a powerful incentive for younger generations to choose careers in industry rather than in academia (Shu 2012).

We are skeptical that current efforts aimed at supporting new investigators will achieve their objectives. Perhaps the most straightforward way to improve the odds that investigators inexperienced in the art of grantsmanship will secure a modicum of career-independence would be to dedicate to young scientists a share of the available pool of R01 funding, and explicitly

segregate the evaluation of their applications from that of the larger pool. Such a radical step has not been countenanced by NIH administrators and the most recent approach leaves a great deal of discretion to NIH's component institutes; as a corollary there is little transparency regarding the way in which the rebalancing of the NIH portfolio in favor of young scientists will be achieved.

With respect to scientific risk-taking, the design of funding mechanisms has not been grounded in a nuanced understanding of the ingredients necessary for the provision of incentives for exploration (e.g., Manso 2011). The MERIT awards, for example, do not select investigators on the basis of the boldness of their proposals, but on the basis of the priority scores which reflect NIH's traditional evaluation criteria. And since investigators know in advance that their applications will revert to traditional R01 status upon the completion of their MERIT award, incentives to branch out in new directions are dampened.

In contrast, the Pioneer Awards program uses evaluation criteria that are well matched to their stated goal of stimulating transformative research, but it does not alter the time horizon that investigators must confront before the outcomes of their scientific efforts are evaluated. And just as in the case of MERIT awards, the lack of a renewal option implies that pioneer awardees must fear that failure to achieve their goals could be severely punished in the relatively near future.

The Pioneer Awards borrows from HHMI a focus on "people, not projects" which could in principle encourage exploration. However, one must recognize that NIH operates under political constraints that a private foundation like HHMI can safely ignore. For instance, all public research agencies need to spread their support across many institutions, including those of lesser renown. Similarly, supporting individual projects, rather than individual scientists, introduces a veil of impersonality in the funding decisions that may make them easier to defend vis-à-vis congressional appropriators or disease-specific interest groups.

Finally, the results of the Azoulay et al. (2011) study may have more nuanced implications for the design of NIH funding programs than is apparent at first blush. While HHMI-style incentives appear to stimulate scientific creativity, it is unclear how easily, and at what cost, the program could be scaled up. Only scientists showing exceptional promise are eligible for HHMI

appointment, and the findings may not generalize to the overall population of scientists eligible for grant funding, which include gifted individuals as well as those with more modest talent.

While there is an argument for taming the “mechanismitis” that has caused the number of distinct, but overlapping grant funding schemes to proliferate at NIH, there exists a case for the creation of a mechanism that combines the best features of the MERIT and Pioneer Award mechanisms. In a nutshell, exploration grants could combine a longer time horizon (e.g., ten years with an intermediate review after five years to ensure that investigators indeed step out of the grant writing treadmill), a different set of evaluation criteria that emphasize risk-taking at the expense of preliminary results, and a renewal mechanism such that awardees will be motivated to transform their fields over the long term.

Conclusions

Despite our focus in this essay on current challenges in the NIH extramural funding program, we remain impressed by the system as a whole. Moreover, we have thus far afforded ourselves the comfortable position of observer, while eschewing the harder role of architect for change. It is not for a lack of want that we avoid laying out a roadmap for a 21st century NIH, but rather that we are no better positioned than others, both inside and outside of the Institutes, to identify the best strategies to improve the program going forward. We would, however, like to advocate a systematic approach for identifying them.

The randomized clinical trial is the gold standard for evaluating scientific results and the time has come for us to apply this approach to the analysis of potential reforms within the NIH. Our current system of programmatic experimentation followed by retrospective analyses on selected samples is quite simply below the methodological standards that the Institutes have worked so hard to establish. Moreover, they engender faulty decision making by painting partial pictures that are poorly designed to identify causal pathways. One way to address this issue would be for proposed policy and programmatic changes to be held to a standard of success (however defined) after a prospective, randomized controlled experiment.

How could such an experiment be designed? While no single design will work for all potential reforms, a simple model would assign experimental review and funding reforms to a random subset of study sections, with the remainder proceeding under business as usual and thus serving as a control group in the experiment. To be concrete, imagine our earlier suggestion for a separate evaluation of first-time applicants. In order to test whether this makes a meaningful impact on young scholars' scientific trajectories, we could randomly allocate a subset of first-time applicants to a special review panel, whereby newcomers are only competing against each other for a predetermined pot of funding. Scholars from this experiment as well as those funded under the usual rules could then be followed for a period of years and then compared based on the number and quality of their publications as well as other pre-determined success metrics. Such an experiment would then provide a clear read on whether the 'new' program should be abandoned or expanded.¹⁰ It might also help elucidate our thinking about whether the increasing age at which a scientist receives their first grant as a principal investigator is a problem in the first place. After all, other, more benign mechanisms could also explain this phenomenon (Jones 2010; Jones and Weinberg 2011).

As with any research study, the details of the design will depend on the question. Two overarching issues, however, merit some additional discussion. First, a common criticism of this approach is that randomized policy experiments of the sort described above gamble with scientists' careers. To this we retort that the nature of scientific funding, which necessarily relies on imperfect evaluations of research proposals and researchers, implies that all funding systems gamble with scientific careers. The principal advantage of randomization is that, in contrast with the haphazard experiments currently performed, it greatly improves our ability to assess which programs and processes are having their desired effect and which are not.

Of course, evaluating whether programs are successful requires measures of success, which is the second issue that is oft raised as sufficiently problematic as to undermine any efforts to systematically analyze the impacts of research funding programs and their various features. If this is true, then it must also be true that any retrospective analysis would be equally uninformative since it will also require success metrics, yet we frequently engage in this exercise. While we

certainly agree that any measure of success – whether it is based on publication or citation counts, scientific awards, or the accomplishments of one’s trainees – will be an imperfect proxy for what we actually care about, it seems quite unlikely that there would be no informational content in a reasonably defined noisy signal. From an experimental perspective, defining measures of success *ex ante* is considerably more important than which specific measures are chosen. Moreover, the requirement that metrics be determined in anticipation of any experiment forces a clear delineation of goals that imposes discipline on the decision making process that often suffers from mission creep, whereby a small set of changes is meant to satisfy a sprawling set of agency objectives.

It is also important to note that these experiments and NIH activities more broadly will not occur in a scientific funding vacuum. The ecosystem that supports scientific research has evolved considerably over the past half-decade. The Howard Hughes Medical Institute provides considerable funding to life sciences researchers as does the Bill and Melinda Gates Foundation. Both emphasize out-of-the-box thinking and breakthrough innovation (Murray, this volume). The private sector has also considerably expanded both the size and scope of its R&D spending in areas that overlap with federal agency priorities and interests but with a logical focus on long-term commercial potential. Deliberative decisions about reforms of the funding process might benefit from a re-assessment of the Institutes’ role vis-à-vis this changed research environment. That is, despite our earlier protestations regarding an inherently conservative funding system, the degree to which NIH funding needs to support more explorative research when others are offering this support and talent for such research is likely scarce, is unclear. Similarly, funding decisions in research areas that are likely to generate large private investments should be mindful of the potential for spending crowd-out. Further complicating matters, many investigators receive funding from many of these sources simultaneously. Clearly, attending to all details of the larger funding ecosystem is impossible, but it does suggest a slight re-framing of policy reforms such that they engage questions regarding the optimal mix of funding programs and associated incentive structures rather than individual activities in isolation.

Institutional change is always difficult, but we remain cautiously optimistic about the NIH of the 21st century. Experimenting on ourselves will certainly identify some deficiencies in the current governing structures within the scientific community, in turn opening up the process to additional scrutiny from politicians and taxpayers keen to cut funding for science. This is entirely analogous to the clinical trial that reveals a new treatment to be inferior to current standards of care. Disappointing, but not cause to abandon methodological rigor. The only alternative to (controlled) experimentation is the gradual stultification of our most cherished scientific institutions. The institution that placed the randomized experiment on the world's stage could play an equally leading role in the scientific inquiry of the management and production of science itself.

This paper was prepared for the NBER meeting on Innovation Policy and the Economy. We are grateful to Scott Stern, Josh Lerner, Danielle Li, George Chacko, Stefano Bertuzzi, and Paula Stephan for helpful conversations and to the National Science Foundation for support.

¹As we will describe later, the NIH also has a sizable intramural research program, which includes approximately 6,000 scientists serving as direct employees of the government, primarily on the NIH main campus in Bethesda, Maryland.

²For ease of exposition, we will refer to all Institutes and Centers as "Institutes" throughout the remainder of this document.

³This practice of percentiling implies that the quality of the marginal grant (i.e., the funded grant with the lowest priority score) could differ widely across study sections. A promising avenue for research is to increase our understanding of the factors that impinge on study section "value added" (cf. Li 2012).

⁴Institute directors have the discretion to fund applications out of order if, for example, they are especially important to the institute mission. Since applications can only be submitted three times, Institutes may also choose to fund applications on their last evaluation cycle instead of newly submitted applications that can be reconsidered later. These exceptions appear rare (Jacob and Lefgren 2011).

⁵This modeling approach controls for right-censoring, i.e., all other things being equal, younger scientists have had less opportunity to be funded.

⁶In the case of termination, there is a two-year phase-down period during which the researcher continues to be funded, allowing her to search for other sources of funding without having to close down her lab.

⁷In practice, this means that the scientific agenda that enabled scientists to secure an HHMI appointment is not binding, and can be renegotiated with the foundation; HHMI investigators face almost no adjustment/transaction costs when the natural evolution of their scientific investigations leads them into directions that were unforeseen at the time of initial funding. In contrast, NIH-funded scientists would put their career in jeopardy if they used the funding associated with a particular project to cross-subsidize activities related to a substantially different project.

⁸The R29 mechanism was itself discontinued in 2004.

⁹It begins with up to 2 years of \$90,000 per year for training while the postdoc finishes research with a mentor. Then, assuming the postdoc has gotten a position as an assistant professor, he or she will receive up to \$249,000 a year for as many as 3 years for independent research.

¹⁰One difficulty would be to structure the experiment such that grant applicants cannot engage in forum shopping. As is customary with the design of social field experiments, the devil is in the details, but it is our impression that scientists have very little leeway to choose the study section that is most appropriate to evaluate their grant proposals, conditional on having chosen a particular area of science, disease focus, or methodological approach.

Figure 1: Overview of the Extramural Research Review and Funding Process

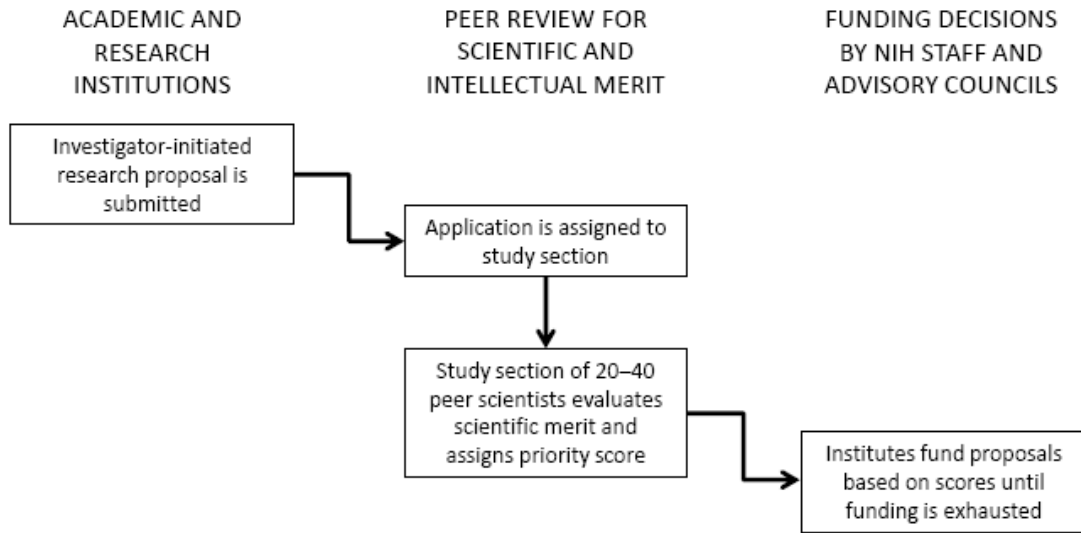
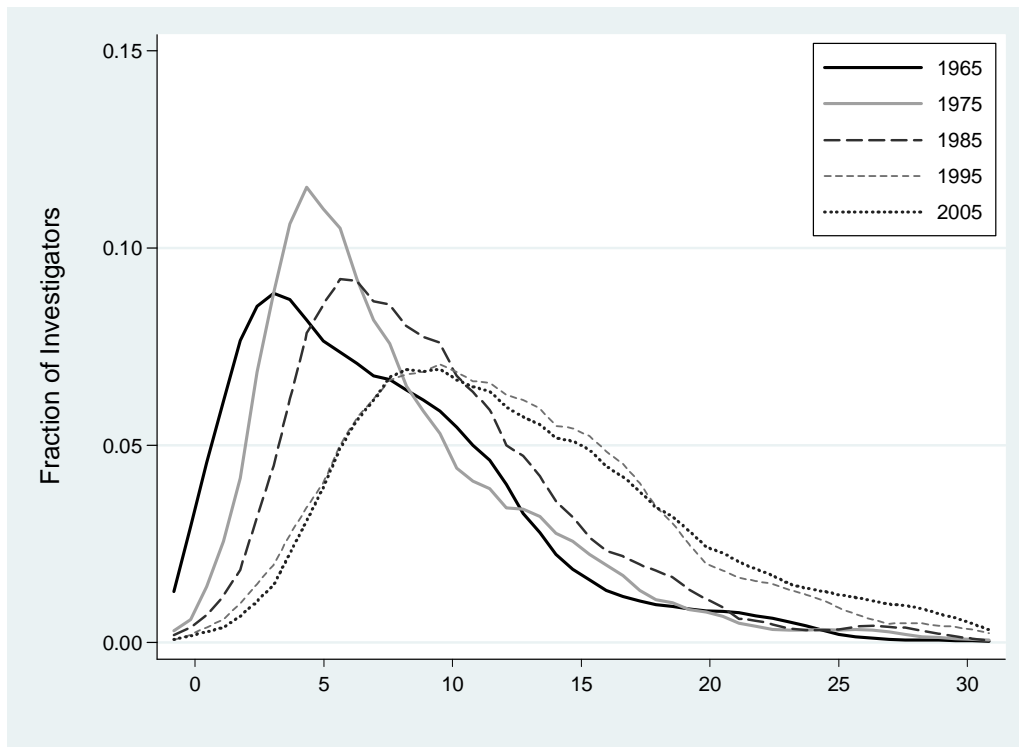
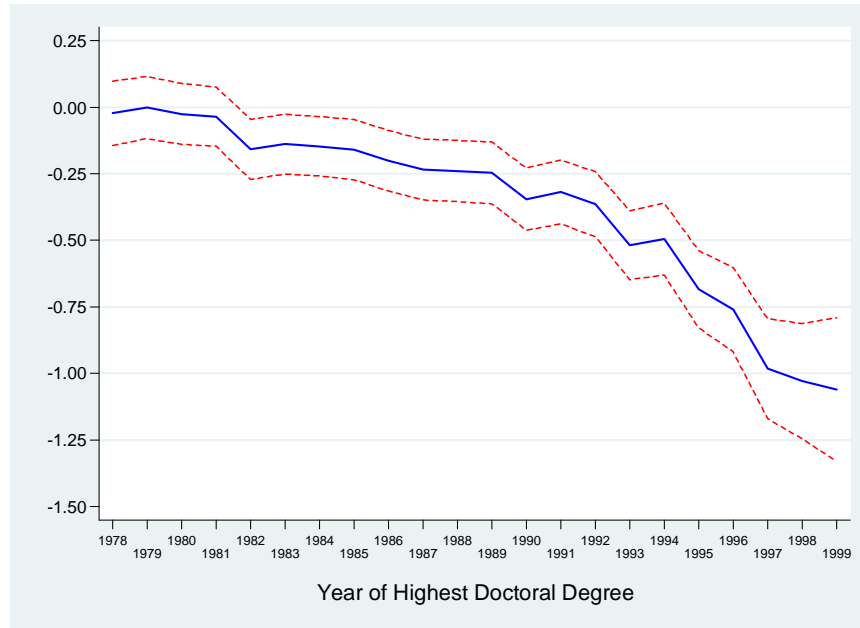


Figure 2: Distribution of Career Age for First-time NIH R01 Awardees



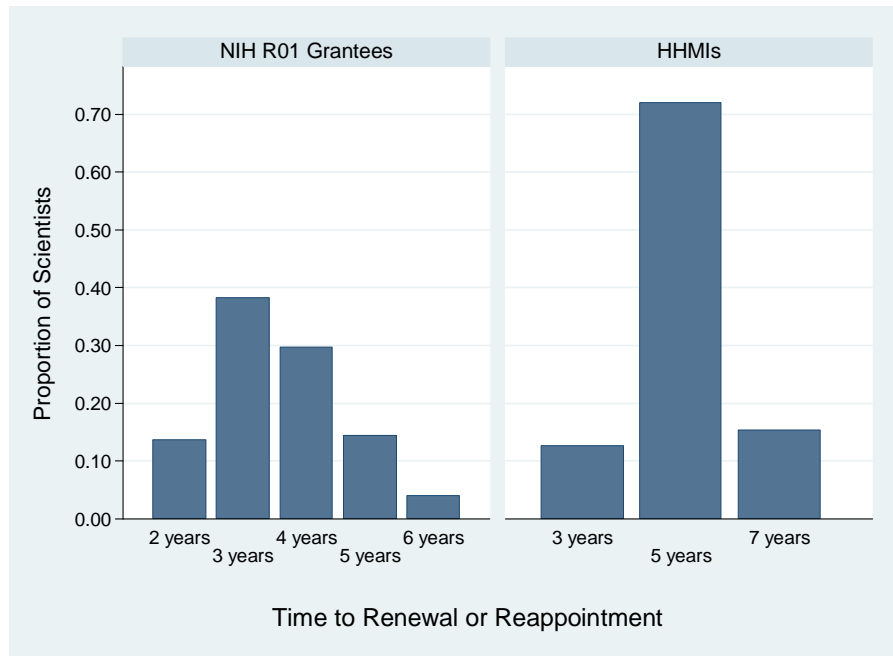
Note: Career age is the number of years elapsed since the award of a scientist’s highest doctoral degree. The sample used to draw these empirical distributions comprises 5,867 scientists who secured R01 funding for the first time in 1965, 1975, 1985, 1995, and 2005.

Figure 3: R01 Grantee Status and Investigator Vintage



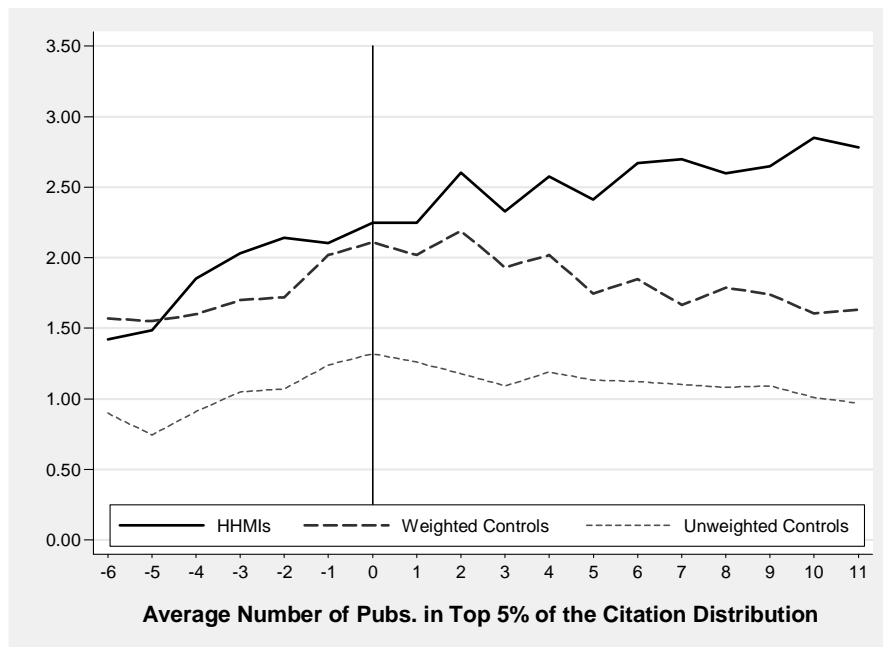
Note: Hazard of first R01 grant, conditional on medical school academic position. The data stems from the Faculty Roster of the Association of American Medical Colleges (AAMC), which we match with the NIH Compound Grant Applicant File (CGAF). In particular, we focus on 97,034 medical school faculty members who began their careers between 1981 and 2003. 14.14% (13,720) of these faculty members have received at least one R01 grant by 2007, the final year of the analysis. Controlling for gender, degree (MD vs. PhD vs. MD/PhD), and main department affiliation (internal medicine, pathology, surgery, etc.), we estimate a simple Cox proportional hazard model where the key explanatory variables are indicator variables for the year in which each of these scientists achieved their highest doctoral degrees (the dates range from 1978 to 1999). The solid line correspond to the coefficient estimates for the degree year indicator variables, while the dashed lines denote the 95% confidence interval around these estimates.

Figure 4: Length of NIH R01 Grants vs. HHMI Appointments



Note: NIH tabulations stem from the Compound Grant Applicant File (CGAF). The grants considered are R01 and equivalent whose first cycle began later than 1970, but earlier than 2002.

Figure 5: Dynamics of HHMI Appointment on the Number of Publications



Note: The dashed and solid black lines correspond to the average yearly number of articles in the Top 5% of the citation distribution for early career prize winner controls and HHMI investigators, respectively. The averages for the control scientists are weighted by each researcher’s inverse probability of selection into the HHMI program. The dashed light gray line corresponds to the unweighted average yearly number of articles for the control scientists.

References

- Azoulay, Pierre, Joshua Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* **42**(3): 527-554.
- Capecchi, Mario R. 2008. Response. *Science* 319(5865): 900-1.
- Hegde, Deepak, and Bhaven N. Sampat. 2011. "Lobbying, Congressional Oversight, and Agency Allocations in U.S. Science Policy: Evidence from Federal Funding for Rare Diseases." Working Paper, Columbia University.
- Jones, Benjamin F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* **76**(1): 283-317.
- Jacob, Brian A., and Lars Lefgren. 2011. "The Impact of Research Grant Funding on Research Productivity." *Journal of Public Economics* **95**(9-10): 1168-1177.
- Kaplan, David. 2005. "How to Improve Peer Review at NIH." *The Scientist* 19(17): 10.
- Kolata, Gina. 2009. "Grant System Leads Cancer Researchers to Play it Safe." *The New York Times*, June 28.
- Ioannidis, John P.A. 2011. "Fund People Not Projects." *Nature* **477**(7366): 529-531.
- Jones, Benjamin F. 2010. "Age and Great Invention." *Review of Economics and Statistics* **92**(1): 1-14.
- Jones, Benjamin F., and Bruce A. Weinberg. 2011. "Age Dynamics in Scientific Creativity." *Proceedings of the National Academy of Sciences* **108**(47): 18910-18914.
- Lazear, Edward P. 1997. "Incentives in Basic Research." *Journal of Labor Economics* **15**(1): S167-S197.
- Li, Danielle. 2012. "Information, Bias, and Efficiency in Expert Evaluation: Evidence from the NIH." Working Paper, Massachusetts Institute of Technology.
- Manso, Gustavo. 2011. "Motivating Innovation." *Journal of Finance* **66**(5): 1823-1860.
- McKnight, Steven L. 2009. "Unconventional Wisdom." *Cell* 138(5): 817-819.
- National Institutes of Health Office of Extramural Research. 2008. "Revised New and Early Stage Investigator Policies." Notice NOT-OD-09-013. Accessed at <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-09-013.html>
- National Institutes of Health. 2009. "NIH Intramural Research at the Threshold of a New Era." Report, Bethesda, MD.
- 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): 332-339.
- Shu, Pian. 2012. "The Long-Term Impact of Business Cycles on Innovation: Evidence from the Massachusetts Institute of Technology." Working Paper, Massachusetts Institute of Technology.
- Stephan, Paula, and Jennifer Ma. 2005. "The Increased Frequency and Duration of the Postdoctorate Career Stage." *American Economic Review Papers & Proceedings* **95**(2): 71-75.
- Stephan, Paula E. 2012. *How Economics Shape Science*. Cambridge, MA: Harvard University Press.