NBER WORKING PAPER SERIES

INCENTIVES AND CREATIVITY: EVIDENCE FROM THE ACADEMIC LIFE SCIENCES

Pierre Azoulay
Joshua S. Graff Zivin
Gustavo Manso

Working Paper 15466
http://www.nber.org/papers/w15466

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2009

We gratefully acknowledge the financial support of the Kauffman Foundation and the National Science Foundation through its SciSIP Program (Award SBE-0738142). We thank Thomas Cech, Purnell Choppin, David Clayton, Nico Lacetera, Antoinette Schoar, Scott Stern, and Heidi Williams for useful comments, and Sherry White and Terry Wood at HHMI for facilitating access to funding data. The usual disclaimer applies. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Incentives and Creativity: Evidence from the Academic Life Sciences
Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso
NBER Working Paper No. 15466
October 2009
JEL No. O31,O32

ABSTRACT

Despite its presumed role as an engine of economic growth, we know surprisingly little about the drivers of scientific creativity. In this paper, we exploit key differences across funding streams within the academic life sciences to estimate the impact of incentives on the rate and direction of scientific exploration. Specifically, we study the careers of investigators of the Howard Hughes Medical Institute (HHMI), which tolerates early failure, rewards long-term success, and gives its appointees great freedom to experiment; and grantees from the National Institute of Health, which are subject to short review cycles, pre-defined deliverables, and renewal policies unforgiving of failure. Using a combination of propensity-score weighting and difference-in-differences estimation strategies, we find that HHMI investigators produce high-impact papers at a much higher rate than two control groups of similarly-accomplished NIH-funded scientists. Moreover, the direction of their research changes in ways that suggest the program induces them to explore novel lines of inquiry.

Pierre Azoulay
MIT Sloan School of Management
50 Memorial Drive E522-555
Cambridge, MA 02142
and NBER
pazoulay@mit.edu

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

Gustavo Manso
MIT
Sloan School
manso@mit.edu

In 1980, a scientist from the University of Utah, Mario Capecchi, applied for a grant at the National Institutes of Health (NIH). The application contained three 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.

Across all of the social sciences, researchers often model the creative process as the cumulative, interactive recombination of existing bits of knowledge in novel ways (Weitzman 1998; Burt 2004; Simonton 2004). But the combinatoric metaphor does not speak directly to the important trade-off illustrated by the anecdote above. Some discoveries are incremental in nature, and reflect the fine-tuning of previously available technologies, or the exploitation of established scientific trajectories. Others are more radical and require the exploration of new, untested approaches. Both forms of innovation are valuable, but we still have a poor understanding of what drives radical innovation. One view is that radical innovation

happens by accident. From Archimedes' *eureka* moment to Newton's otherworldly contemplation interrupted by the fall of an apple, luck (and sometimes talent) play an essential role in lay theories of breakthrough innovation. Of course, if luck and talent exhaust the list of ingredients necessary to produce breakthroughs, then there is little for economists to contribute.

As argued by Manso (2009), however, incentives may play an important role in the production of ideas, particularly novel ones. In this paper, we build upon his theoretical framework to provide empirical evidence that nuanced features of incentive schemes embodied in the design of research contracts exert a profound influence on the subsequent development of breakthrough ideas. The challenge is to find a setting in which (1) radical innovation is a key concern; (2) agents are at risk of receiving different incentive schemes; and (3) it is possible to measure innovative output and to distinguish between incremental and radical ideas. We argue that the academic life sciences in the United States provides a near-ideal testing ground.

Specifically, we 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 nonprofit 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. To do so, the HHMI program has adopted practices that according to Manso (2009) should provide strong incentives for breakthrough scientific discoveries: the grant cycles are long (five years, and typically renewed at least once); 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. This stands in sharp contrast with the incentives faced by life scientists funded by the NIH. The typical R01 grant cycle lasts only three years, and renewal is not very forgiving of failure. Feedback on performance is

limited in its depth. Importantly, the NIH funds projects with clearly defined deliverables, not individual scientists, which could increase the costs of experimentation.

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. Three significant hurdles must be overcome to answer this question.

First, we need to identify a group of NIH-funded scientists that are appropriate controls for the researchers selected into the HHMI program. Given the high degree of accomplishment exhibited by HHMI investigators at the time of their appointment, a random sample of scientists of the same age, working in the same fields, would not be appropriate. We estimate the treatment effect of the program by contrasting HHMI-funded scientists' output with that of two groups of NIH-funded scientists. The first control group is composed of scientists who received prestigious early career prizes that focus on the same subfields of the life sciences as HHMI. The second control group is composed of scientists whose NIH grants received the coveted MERIT designation, which is only available for scientists whose application for a grant renewal was scored in the top percentile by NIH peer reviewers. Furthermore, using an in-depth understanding of the HHMI appointment process, we cull from these two control groups scientists who look similar to the HHMI investigators on the observable factors that we know to be relevant for selection into the HHMI program.

Second, we must be able to distinguish particularly creative contributions from incremental advances. While we investigate the effect of the program on the raw number of original research articles published, the bulk of our analysis focuses on the number of publications that fall into different quantiles of the vintage-specific, article-level distribution of citations (see Figure 1): top quartile, top ventile, and top percentile. Another prong in our attempt to measure creativity is to measure explorative behavior directly. Specifically, we examine whether the research agenda of HHMI investigators changes after their appointment; we measure the novelty (both relative to the universe of published research and to the scientists themselves) of the keywords tagging their publications; and we also assess whether the impact of their research broadens, as inferred by the range of journals that cite it.

Third, we need to ascertain whether it is the incentive features of the program that explain its effects, or some alternative mechanism, such as increased resources, ascription dynamics (whereby HHMI investigators get cited more due to their enhanced status), or peer effects. We tackle these issues (to the extent possible) in the discussion.

Our results provide support for the hypothesis that appropriately designed incentives stimulate exploration. In particular, we find that the effect of selection into the HHMI program increases as we examine higher quantiles of the distribution of citations. When compared to Early Career Prize Winners (ECPWs), our preferred econometric estimates imply that the program increases overall publication output by 34%; the magnitude jumps to 98% when focusing on the number of publications in the top percentile of the citation distribution. The main message is qualitatively similar when we compare HHMI researchers and MERIT awardees, though the magnitudes of the effects are smaller and often imprecisely estimated.

We bolster the case for the exploration hypothesis by focusing on various attributes of these scientists' research agenda. We 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. We find weaker evidence that HHMI scientists fail more often than their NIH-funded counterparts, using the rate of publications in the bottom quartile of the citation distribution as an indicator of failure. Though the effects are sometimes large in magnitude (especially relative to the ECPW comparison group), they are also imprecisely estimated. This provides weak evidence that HHMI scientists place more risky scientific bets after their appointment.

The rest of the paper proceeds as follows. In the next section, we present the theoretical motivation for our hypothesis. Section 2 describes the construction of the sample and presents descriptive statistics. Section 3 lays out our econometric methodology. Section 4 reports and discusses the results of the analysis. Section 5 concludes.

1 Theoretical Background

The bulk of the literature on incentives for innovation has focused on the problems inherent to the measurement and contractability of output that plagues most innovative activities. For example, Holmström (1989) observes that most innovation projects are risky, unpredictable, long-term, labor-intensive, and idiosyncratic. In such settings, performance measures are likely to be extremely noisy, and contracting particularly challenging. This leads him to see virtue in the adoption of low-powered incentives when creativity is what is required of the agent, for salary is less likely to distort the agent's attention away from the less-easily measurable tasks that compete for her attention. This view stands in sharp contrast with the standard prescription to adopt piece rates whenever agent's individual contributions are easy to measure, such as in the case of the windshield installers studied by Lazear (2000). A substantial body of experimental and field research in psychology reaches a similar conclusion, but for different reasons: the worry is that pay-for-performance might encourage the repetition of what has worked in the past, at the expense of the exploration of untested approaches (Amabile 1996).

In contrast, Manso (2009) explicitly models the innovation process as the result of learning through experimentation. In this setting, the trade-off between the exploitation of well-known approaches and the exploration of new untested approaches first emphasized by March (1991) arises naturally. The main insight of his contribution is that the optimal incentive scheme to motivate exploration exhibits substantial tolerance for early failure and rewards for long-term success. Tolerance for early failure allows the agent to explore in the early stages of the contractual relationship without incurring the usual negative consequences of lower pay or termination. At the same time, reward for long-term success prevents the agent from shirking early on and induces the agent to explore new ideas that will allow him to perform well in the long-run. Another important ingredient of Manso's model is timely feedback on performance. Providing information to the agent about how well he is doing allows the agent to explore more efficiently, reducing the costs of experimentation. An agent who does not get feedback on performance may waste more time on unfruitful ideas.

Empirical evidence on the effects on long-term incentives is scant. Most relevant to the findings presented below is Lerner and Wulf's (2007) study of corporate R&D lab heads. They show that higher levels of deferred compensation are associated with the production of more heavily cited patents, while short-term incentives bear no relationship to firm innovative performance. The present paper presents the first attempt to test Manso's theoretical arguments in the field (see Ederer and Manso [2008] for experimental evidence with a similar flavor). We believe that the academic life sciences in the United States provide an appropriate setting, first and foremost because it provides naturally-occurring variation in incentives that closely matches the contrast between pay-for-performance and exploration-type schemes emphasized by Manso (2009).

Most academic life scientists must rely on grants from the National Institutes of Health (NIH), the largest public funder of biomedical research in the United States. With an annual budget of $28.4 billion in 2007, support from the NIH dwarfs that available from other public or private funders, including the National Science Foundation ($6 billion in 2007) or the American Cancer Society ($147 million in 2007). The most common type of NIH grant for investigator-initiated projects is the R01 grant. In 2007, their average amount was $225,000 in annual direct costs, and the awards last for a typical three to five years before coming up for renewal (see Figure 2). The NIH "study sections," or peer-review panels in charge of allocating awards, 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. 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).

An alternative funding mechanism is provided by the investigator program of the Howard Hughes Medical Institute (HHMI). This program *"urges its researchers to take risks, to explore unproven avenues, to embrace the unknown – even if it means uncertainty or the chance*

of failure."[1] New appointments are based on nominations from research institutions; once selected, researchers continue to be based at their institutions, typically leading a research group of 10 to 25 students, postdoctoral associates and technicians. In its stated policies, HHMI departs in striking fashion from NIH's funding practices, in ways that should bring incentives in line with the type of schemes suggested by Manso (2009). HHMI Investigator are initially appointed for 5 years,[2] and in 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.

Moreover, HHMI investigators appear to share the perception that their first appointment review is rather lax, with reviewers more interested in making sure that they have taken on new projects with uncertain payoffs, rather than insisting on achievements. Below, we validate this perception by showing that the second review is much more sensitive to performance than the first. The review process is also streamlined, lasting a mere six weeks. Investigators are asked to submit a packet containing their five most notable papers in the past five years, along with a short research proposal for the next five years. In contrast, NIH grants take at a minimum three months to be reviewed, and success typically depends upon a rather exhaustive list of accomplishments by the primary research team members.

Since HHMI researchers publish 29 articles on average in the five years that follow their initial appointment (the median is 25), constraining their renewal packet to contain only five papers ensures that only what they see as their most meaningful achievements matters for the renewal decision. The review process culminates in an oral defense in front of an elite panel especially convened for the occasion. The reviewers must not be HHMI researchers, and are of very high caliber (e.g., members of the National Academies). The richness of the feedback is yet another point of departure between HHMI and NIH practices. Besides the intensity and quality of the advice generated by the review process, HHMI-funded scientists participate in annual science meetings during which they can interact with other HHMI investigators. This gives them access to a deep level of critique, encouragement, ideas,

[1] See `http://www.hhmi.org/research/investigators/`

[2] Appointment lengths have varied over the history of the program, more detailed information will be provided in the data section.

and potential collaborations. While NIH-funded researchers receive a critique of their grant applications, these vary widely in quality and depth. Furthermore, the federal agency does not provide any meaningful feedback between review cycles.

Finally, an important distinction between the two sources of funding is the unit of selection. The NIH funds specific projects. Applicants need to map out experiments far into the future, and have limited flexibility to change course between funding cycles. Together with study sections' insistence on preliminary results, this has led many NIH grantees to submit research that is already quite developed. In contrast, HHMI insists on funding "people, not projects." This allows HHMI researchers to quickly reallocate effort and resources away from avenues that do not bear fruit. The economics literature (e.g., Aghion, Dewatripont, and Stein 2008) views unfettered control over one's research agenda as the key distinguishing feature of innovative activities performed in academia (relative to the private sector). Variation in the unit of selection reminds us that the degree of *effective* control experienced by academic researchers often depends on the arcane details of funding mechanisms. Though not part of Manso's (2009) initial analysis, we extend his model in the Appendix to show that providing the researcher greater latitude in her search activities encourages exploration. Table 1 summarizes the main differences between the two sources of funding.

2 Data and Sample Characteristics

This section provides a detailed description of the process through which the data used in the econometric analysis were assembled. In order, we describe (1) the Howard Hughes Medical Investigator sample; (2) the set of control investigators against which the HHMI scientists will be compared; and (3) our metrics of scientific creativity. We also present relevant descriptive statistics.

2.1 HHMI Sample

We begin with a basic description of the criteria necessary for nomination and appointment as an HHMI investigator. To be eligible, a scientist must be tenured or on the tenure-track

at a major research university, academic medical center, or research institute. The subfields of the life sciences of interest to HHMI are not cast in stone, but in the recent past the Institute has shunned clinical and epidemiological research, and concentrated on the fields of cell and molecular biology, neurobiology, immunology, and biochemistry. Career-stage considerations have varied over time. Traditionally, HHMI has focused its appointments on relatively early and mid-career investigators, but not so early that their output would be hard to distinguish from that of their postdoctoral or graduate school adviser.

Upon receipt of nominations from participating institutions, HHMI empanels a jury that reviews these nominations in two sequential steps. In a first step, the number of nominees is whittled down to a manageable number, mostly based on observable characteristics. For example, NIH-funded investigators have an advantage because the panel of judges interprets receipt of federal grants as a signal of management ability. The jury also looks for evidence that the nominee has stepped out of the shadow cast by his/her mentors: an independent research agenda, and a "big hit," i.e., a high-impact publication in which the mentor's name does not appear on the coauthorship list. In a second step, each remaining nominee's credentials and future plans are given an in-depth qualitative look.[3] Finally, until recently, appointment contracts varied in their initial length. Assistant Investigators (Assistant Professors in their home institution) were appointed for three years; Associate Investigators, for five years; and Investigators, for seven years.[4]

Our analysis focuses on HHMI investigators appointed in 1993, 1994, and 1995. We exclude the three researchers that withdrew from the program voluntarily, leaving us with a sample of 73 scientists.[5]

[3]While an input into this process is a letter grade, the review does not provide a continuous score that could be used in a regression discontinuity-type framework. Moreover, the cutoff that separates successful from unsuccessful nominees is endogenous in the sense that it depends on the overall quality of the applicant pool.

[4]In our sample, these categories respectively account 15%, 70%, and 15% of the total number of scientists in the treatment group. Of course, such variation raises the specter that appointment length might be endogenous. In fact, the length of the initial term is purely a function of the scientist's academic rank in his/her home institution.

[5]One accepted a top administrative position in her university (HHMI rules prevent investigators to hold major administrative posts), one moved to the Scripps Research Institute in La Jolla, CA, an institution that has no relationship with HHMI. Yet another wished to move to a different institution during his first

2.2 Control Samples

In the absence of information on the runners-up of the HHMI competitions, we must rely on observable characteristics to create viable control groups. The main challenge is that HHMI investigators are extremely accomplished at the time of their appointment. Controls should not only be well-matched with HHMI investigators in terms of fields, age, gender, and host institutions; their accomplishments should also be comparable at baseline. In practice, we draw on two set of academic life scientists to construct our control groups: early career prize winners on the one hand; MERIT awardees from the NIH on the other hand.

Early Career Prize Winners. The Pew, Searle, Beckman, Packard, and Rita Allen Scholarships are early-career prizes that target scientists in the same life science subfields and similar research institutions as HHMI. Every year, these charitable trusts provide seed funding to around 60 life scientists in the first two years of their independent careers. These scholarships are among the most prestigious accolades that young researchers can receive as they are building a laboratory, but they differ from HHMI investigatorships in one essential respect: they are structured as *one-time grants* (e.g., $60,000 a year over 4 years for the Pew Scholarship; $80,000 a year for 3 years for the Searle Scholarship, etc.). These amounts are relatively small, roughly corresponding to 35% of a typical NIH R01 grant. As a result, these scholars must still attract grants from other funding sources (especially NIH) if they intend to further their independent research career. After a screen to eliminate investigators whose age place them outside the age range of the treatment group, a second screen to exclude researchers that go on to be appointed HHMI investigators, and a final screen to eliminate researchers working in idiosyncratic fields, we are left with 393 early career prize winning (ECPW) scientists awarded one of these scholarships.

MERIT awardees. 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, outstand-*

appointment. To prevent the eruption of bidding wars over HHMI investigators, the Institute forces such investigators to resign their appointment.

ing 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 round of funding. While the MERIT designation is a prestigious award for mid-career investigators, it pertains to a particular project, not to the scientists's overall portfolio. To construct the MERIT control group, we start from the set of all scientists whose R01 grants receive the R37 designation in 1993, 1994, and 1995. We eliminate from this group scientists whose highest doctoral degree was obtained prior to 1974; scientists employed in institutions that are not HHMI host institutions; scientists working in fields not targeted by HHMI; and scientists who are eventually appointed HHMI investigators. There are 92 scientists meeting these criteria in the final sample.

Before presenting descriptive statistics, it is useful to discuss broad features of these control groups that will influence the interpretation of the treatment effect. The ECPW sample comprises scientists who show great promise at the very start of their independent career, when it is difficult to distinguish their output from that of their postdoctoral mentor. In contrast, the modal HHMI investigator stands at the cusp of the tenure decision when s/he is appointed. As a result, there is more variability in the expected performance of ECPW scholars than is the case among HHMI investigators, but as we will show, it is possible to cull from this group a subsample of scientists whose characteristics match well those of HHMI scientists at baseline.

Conversely, MERIT awardees must have, by design, completed successfully at least one R01 funding cycle when they receive the R37 designation. This means that they are older on average, relative to HHMI investigators. Furthermore, one should think of the HHMI/MERIT contrast as a comparison between two programs, rather than an evaluation of the effectiveness of the HHMI program in fostering exploration. The HHMI and MERIT program have at their core a common feature: the extension of the time horizon available to scientists before they must show evidence of actual accomplishments. But since the R37 designation is specifically attached to particular project, this program limits researchers'

ability to branch out in novel directions. As well, it has has no impact on the depth and quality of the feedback provided to the researcher. Therefore, comparing MERIT awardees with HHMI investigators can help us ascertain whether shifting the time horizon is enough on its own to foster exploration and scientific creativity.

2.3 Measuring Scientific Creativity

Creativity is a loaded term. The *wikipedia* entry informs us that more than 60 different definitions can be found in the psychological literature, none of which is particularly authoritative. Furthermore, there exists no agreed-upon measurement metrics or techniques to measure creative outputs.

The perspective adopted in this paper is very pragmatic, and guided by the constraints put on us by the availability of data. Amabile (1996) suggests that while innovation *"begins with creative ideas...creativity by individuals and teams is a starting point for innovation; the first is a necessary but not sufficient condition for the second."* While we certainly agree with this view at a conceptual level, the measurement of scientific productivity — an already well-established discipline — makes it hard to recognize this nuance. A crucial development in the bibliometric literature has been the use of citation information to adjust raw publication counts for quality. Such an approach is not entirely satisfying here, as both "humdrum" and "breakthrough" research generate publications and citations. Moreover, some types of publications, like review articles, tend to generate a number of citations not commensurate with their degree of originality. It has long been noted that the distributions of publications and citations at the individual level is extremely skewed, and typically follows a power law (Lotka 1926). The distribution of citations *at the article level* exhibits even more skewness. In this paper, we make use of the wide variation in impact across the publications of a given scientist to compute measures of creative output. Specifically, we sum the number of distinct contributions that fall into the higher quantiles (top quartile, top ventile, or top percentile) of the article-level distribution of citations, for an individual scientist in a given time period.

One practical hurdle is truncation: older articles have had more time to be cited, and hence are more likely to reach the tail of the citation distribution. To overcome this issue, we compute a different empirical cumulative distribution function in each year.[6] For example, in the life sciences broadly defined, an article published in 1980 would require at least 98 citations to fall into the top ventile of the distribution; an article published in 1990, 94 citations; and an article published in 2000, only 57 citations (this is illustrated in Figure 1). With these empirical distributions in hand, it becomes meaningful to count the number of articles that fall, for example, in the top percentile over a scientist's career. Counting the number of contributions that fall "in the tail" is predicated on the idea that exploration is more likely to result in high-impact publications, relative to exploitation.[7]

We rely on two additional metrics of scientific excellence. We tabulate elections to two of the most prestigious scientific societies: the Institute of Medicine and the National Academy of Sciences. We also measure the number of students and fellows trained in a scientist's lab that go on to win a Pew, Searle, Beckman, Packard, or Rita Allen scholarship.[8]

Besides the higher likelihood of right-tail outcomes, increased exploration has additional implications we can examine empirically. First, it might also fatten the left-hand tail of the outcome distribution, since pushing the boundaries of one's field is a riskier endeavor than cruising along an already-established scientific trajectory. Second, it seems intuitive that scientists would need to change the direction of their research endeavors when they take on higher-risk projects, independently of the success or failure of these projects. To test the first prediction, we compute the number of contributions that fall in the bottom quartile of the vintage-specific, article-level distribution of citations (about three citations or less).[9]

[6] We thank Stefan Wuchty and Ben Jones from Northwestern University for performing these computations.

[7] We exclude review articles, editorials, and letters from the set when computing these measures. We also eliminate articles with more than 20 authors.

[8] We do not emphasize the results pertaining to these outcomes, because they seem particularly subject to alternative interpretations: NAS and IoM members are elected, and the large contingent of HHMI investigators among the incumbent membership might skew the results in favor of the treated scientists; similarly, it is plausible that better students match with HHMI PIs after their appointment.

[9] Too few investigators exit science altogether to make exit a useful indicator of failure.

To address the second prediction, we construct a battery of measures designed to capture potential changes in the scientists' research trajectories. Most of these measures use MeSH keywords as an essential input.[10] First, we calculate the average age of MeSH keywords, and the average age of 2-way MeSH keyword combinations for the published research of every scientist in the sample, separately for each year of their independent career. A keyword (or keyword combination) is said to be born the first year it appears in any paper indexed by PubMed. These two measures capture the extent to which a scientist's research is novel relative to the world's research frontier. Equally important is to document the extent to which scientists place new scientific bets in the post-appointment period (1998-2006) relative to the pre appointment-period (1986-1994).[11] We do so by (a) computing the degree of overlap in MeSH keywords corresponding to articles published in both periods; computing the Herfindahl index of MeSH keywords in both periods (a proxy for variety in topic choice); and (c) computing one minus the Herfindahl index of citing journal diversity in both periods (a measure of impact breadth, rather than impact depth as with the citation quantiles). If HHMI investigators are induced to explore novel approaches following their appointment, we would expect this behavior to be reflected in these measures.

2.4 Descriptive Statistics

For each scientist, we gathered employment and basic demographic data from CVs, sometimes complemented by Who's Who profiles or faculty web pages. We record the following information: degrees (MD, PhD, or MD/PhD); year of graduation; mentors during graduate school or post-doctoral fellowship; gender; and department(s).

We obtain publication and citation data from PubMed and Thomson Scientific's *Web of Science*, respectively. Funding information stems from HHMI, NIH's Compound Applicant Grant File (CGAF), as well as a number of non-profits that are active in the funding of

[10] MeSH is the National Library of Medicine's controlled vocabulary thesaurus; it consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. There are 24,767 descriptors in the 2008 MeSH.

[11] We do not take into account papers published in 1995, 1996, or 1997 to compute these measures because we assume that HHMI's previous research agenda will be somewhat "sticky."

basic biomedical research: The Pew, Searle, Beckman, MacArthur, March of Dimes, and Sloan foundations. We sum the direct costs awarded to each scientist across funding source to create our measure of overall research funding.[12]

Finally, we categorize the type of laboratory run by each scientist into four broad types: *macromolecular* labs, *cellular* labs, *organismal* labs, and *translational* labs. For the first three types, the taxonomy is based on the level of analysis at which most of the research is performed in the lab. Some scientists work mostly at the molecular level (i.e., in test tubes). This type of research does not require living cells, and includes fields such as molecular biology, biochemistry, and structural biology. Others do most of their research at the cellular level (i.e., in petri dishes), and ask questions that require living cells. Prominent subfields include subcellular trafficking, cell morphology, cell motility, and some aspects of cell signaling. Yet others work with model organisms (mice, flies, monkeys, worms, etc.), asking questions that require, if not a whole organism, at least the interaction of multiple cells. The translational label is given to labs run by physician-scientists whose research have both a laboratory and a clinical component.

Demographic characteristics. The distributions of fixed traits for control and treatment scientists are presented in Tables 2 and 3. The HHMI sample is more gender-balanced; it also contains more scientists with dual MD/PhD degrees. Turning our attention to laboratory type, HHMI investigators are slightly less likely to work primarily at the cellular level, and slightly more likely to be engaged in macromolecular research.

HHMI and control samples at baseline. Table 4 presents baseline descriptive statistics. Approximately 37% of the HHMI sample is female, versus 20% of the ECPW sample, and 22% of the MERIT sample. HHMI investigators are of the same career age on average as the ECPW scholars, but are significantly younger than the MERIT awardees. They are better funded than ECPW scholars at baseline ($1.45 million vs. $1.10 million on average), but much less well funded than MERIT awardees ($4.21 million on average). In terms of raw

[12]This measure fails to capture industrial sources of funding or start-up funds provided by the employing institution, if any.

publication output, the pattern is very similar, with HHMI investigators lagging significantly MERIT awardees, but leading ECPW scholars. As one looks at the rate of high-impact publications, the differences between the HHMI and MERIT samples disappear, indicating that HHMI investigators have a much higher "hit rate" than MERIT awardees, even at baseline. The breadth of impact and diversity of topics studied by these scientists appears similar for the three groups of scientists. Controls (whether ECPWs or MERITs) and HHMIs appear to be drawn from a similar set of academic employers in a dimension relevant for selection: the number of slots allocated to their institution at the nomination stage.

Of course, these averages tell only part of the story. Figures 3A and 3B plot the distributions of baseline publications in the Top 5% of the citation distribution. Note that we are only including here publications for which the scientist is the senior author, i.e., where s/he appears in last position on the authorship list. The distribution for ECPW scholars appears significantly more skewed than that for HHMI investigators. The contrast with MERIT awardees is less sharp: the two distributions have roughly the same shape. Similarly, Figures 4A and 4B plot the distributions of NIH funding at baseline for treatment and control scientists. The distribution for the MERIT sample is, almost by construction, less skewed than that of the HHMI sample. In contrast, the shape of the distributions for HHMIs and ECPW scholars are very similar.

In summary, characteristics that determine selection into the HHMI program are not especially well-balanced at baseline between treatment and control scientists. However, the region of common support is wide, indicating that it should be possible to create "synthetic" control scientists that will be good matches for HHMI investigators on these important dimensions.

Career achievement. While the differences between treatment and control samples are relatively modest at baseline, their magnitude increases when we examine achievements over the entire career (up to 2006). In Table 5 we see that HHMI scientists publish many more papers than ECPW scientists, with this output of higher quality, regardless of the quantile threshold one chooses to focus on. Of course, these accomplishments should be

viewed in light of their tremendous funding advantage: HHMI scientists receive on average $14.5 million in funding, versus $5.5 million on average for the ECPWs. HHMIs' funding advantage over MERIT awardees is noticeable, but not as stark. Publication outcomes are also more comparable, though HHMI investigators produce blockbuster papers (those falling in the top ventile or top percentile of the citation distribution) at a much higher rate. The average level of normalized keyword overlap appears to be lower for HHMIs, compared with both the MERIT and ECPW controls.

When we focus on discrete career accolades (Table 6), we observe an even greater contrast between HHMI and control scientists. Approximately a third of the HHMI investigators are elected members of the National Academy of Sciences; 16% are elected members of the Institute of Medicine. This is in contrast to 6% and 3%, respectively, for the control sample. Our 73 HHMI investigators collectively train 83 future early career prize winners (an average of 1.13 per scientist), whereas the control investigators are mentors to 118 such "young superstars" (an average of 0.24 per scientist).

3 Econometric Considerations

In order to estimate the treatment effect of the HHMI investigator program, we must confront a basic identification problem: appointments are driven by expectations about the creative potential of scientists, and selected investigators might have experienced very similar outcomes had they not been appointed. As a result, traditional econometric techniques, which assume that assignment into the program is random, cannot recover causal effects.

Propensity-score weighting. As an attempt to overcome this challenge, we estimate the effects of the program using inverse probability of treatment weighted estimation (Robins and Rotnitzky 1995; Hirano and Imbens 2001; Busso, DiNardo, and McCrary 2008). Suppose we have a random sample of size N. For each individual i in this sample, let $TREAT_i$ indicate whether s/he received treatment. Using the counterfactual outcome notation (e.g., Rubin 1974), let y_i^1 be the value of the outcome y that would have been observed had i received

treatment, and y_i^0 the value of the outcome had i been assigned to the control arm of the experiment. In addition, we will assume that we observe a vector of covariates denoted by $X = (W, Z)$. The variables included in W are assumed to be strictly exogenous; in contrast, the vector Z include pre-treatment variables such as lagged outcomes.

For each individual i, the treatment effect is $y_i^1 - y_i^0$. For the population as a whole, we are interested in two distinct estimands, the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT). Formally:

$$\beta^{ATE} = E[y_i^1 - y_i^0]$$
$$\beta^{ATT} = E[y_i^1 - y_i^0 | TREAT_i = 1]$$

While ATE elucidates what would be the average effect of treatment for an individual picked at random from the population, ATT measures the average effect for the subpopulation that is likely to receive treatment. The difficulty in identifying these coefficients is identical, however: for a given individual, we observe y^1 or y^0, but never both.

Following Rosenbaum and Rubin (1983), we make the "selection on observables" or unconfoundedness assumption:

$$TREAT \amalg (y^1; y^0; Z) | X$$

where the $\amalg$ sign denotes statistical independence. Let the propensity score, the conditional probability of treatment, be denoted by $p(x) = Prob(TREAT_i = 1 | X_i = x)$; further, we assume that $0 < p(x) < 1$. These admittedly strong assumptions enable the identification of ATE and ATT; the two effects can be recovered by a two-step procedure relying on a first-step estimate of the propensity score $\hat{p}(x)$. In the second step, the outcome equation

$$y_i = \beta_0 + \beta_1' W_i + \beta_2 TREAT_i + \varepsilon_i \tag{1}$$

is estimated by weighted least squares or weighted maximum likelihood (depending on the type of dependent variable), where the weights are simple functions of the estimated propen-

sity score:

$$w_i^{ATE} = \frac{TREAT_i}{\hat{p}(x_i)} + \frac{1 - TREAT_i}{1 - \hat{p}(x_i)}$$
$$w_i^{ATT} = TREAT_i + (1 - TREAT_i) \cdot \frac{\hat{p}(x_i)}{1 - \hat{p}(x_i)}$$

In order to develop the intuition for this weighting strategy, we examine the formula corresponding to w^{ATE} a bit more closely. Each factor in the denominator is the probability that an individual received her own observed treatment, conditional on her past history of "prognosis factors" for treatment. Suppose that all relevant variables are observed and included in X. Then, weighting effectively creates a pseudo-population in which X no longer predicts selection into treatment and the causal association between treatment and outcome is the same as in the original population.[13]

Assessing unconfoundedness. Propensity-score weighting relies on the assumption that selection into treatment occurs solely on the basis of factors observed by the econometrician. This will appear to many readers as a strong assumption — one that is unlikely to be literally true. Despite the strength of the assumption, we consider it a useful starting point. Past research in the program evaluation literature has shown that techniques that assume selection on observables perform well (in the sense of replicating an experimental benchmark) when (1) researchers use a rich list of covariates to model the probability of treatment; (2) units are drawn from similar labor markets, and (3) outcomes are measured in the same way for both treatment and control groups (Dehejia and Waba 2002; Smith and Todd 2005). Conditions (2) and (3) are trivially satisfied here, but one might wonder about condition (1), namely the extent to which the analysis accounts for the relevant determinants of HHMI appointment.

Through interviews with HHMI senior administrators, we have sought to identify the criteria that increase the odds of appointments, conditional on being nominated. As described earlier, the Institute appears focused on making sure that its new investigators have stepped out of the shadow cast by their graduate school or postdoctoral mentors. They also

[13]One might worry about statistical inference, since the weights used as inputs to estimate the outcome equation are themselves estimated. In contrast to two-step selection correction methods, the standard errors obtained in this case are conservative (Wooldridge, 2002).

want to ensure that these investigators have the leadership and managerial skills required to run a successful laboratory, and interpret receipt of NIH funding as a important signal of possessing these skills. In practice, we capture the "stepping out" criteria by counting the number of last-authored, high-impact contributions the scientist has made since the beginning of his/her independent career.[14] We proxy PI leadership skills with an measure of cumulative R01 NIH funding at baseline. Of course, our selection equation also includes important demographic characteristics, such as gender, laboratory type, degree, and career age.

Semi-Parametric Difference in Differences. An alternative methodology is to rely on within-scientist variation to identify the program's treatment effect. Scientist fixed effects purge estimates from any influence of unobserved heterogeneity that is constant over time. However, for difference-in-differences (DD) estimation to be valid, it must be the case that the average outcome for the treated and control groups would have followed parallel paths over time in the absence of treatment. This assumption is implausible if pretreatment characteristics that are thought to be associated with the dynamics of the outcome variable are unbalanced between treatment and control units. Below, we provide strong evidence that selection into the program is influenced by transitory shocks to scientific opportunities: HHMI scientists have higher output in the years immediately preceding their appointment. As a result, the fixed effect estimator is likely to *underestimate* the effect of the program on scientific achievement.

In such a case, Abadie (2005) proposes a semiparametric difference-in-differences (SDD) estimator that combines the advantages of adjustment for observed heterogeneity with differencing. The idea is to apply propensity score reweighting not to the *levels* of outcome y as above, but to the *differences* in outcome between the post- and pre-treatment periods. Under some additional regularity conditions, the Average Treatment Effect on the Treated

[14] A robust social norm in the life sciences systematically assigns last authorship to the principal investigator, first authorship to the junior author who was responsible for the actual conduct of the investigation, and apportions the remaining credit to authors in the middle of the authorship list, generally as a decreasing function of the distance from the extremities of the list.

(ATT) is identified and can be recovered by weighting $y^{post} - y^{pre}$ using:

$$w_i^{SDD} = \frac{TREAT_i - \hat{p}(x_i)}{\pi \cdot (1 - \hat{p}(x_i))}$$

where π denotes the unconditional odds of treatment $Prob(TREAT_i = 1)$. Intuitively, the weights create a pseudo-population of untreated scientists that follow similar dynamics to the treated group in the pre-treatment period. The SDD estimator then subtracts the change in outcomes for treated scientists with the change in outcome for this pseudo-population of control scientists. Inference is performed using a non-parametric pairwise bootstrap procedure with 500 replications.

The SDD estimates are still vulnerable to the critique that time-varying sources of unobserved heterogeneity could bias the effects. In particular, we are hard-pressed to control for the true creative potential of the research agenda laid out by these scientists in their grant applications, regardless of funding source. In combination, however, the three estimation methods proposed in this paper are likely to bound the true impact of treatment, with the DD estimate providing a lower bound, propensity-score weighting in the cross sectional dimension of the data providing an upper bound, and propensity-score weighting in the within-scientist dimension of the data (our preferred methodology) falling somewhere in between the first two estimates. The evidence presented below will show that, in at least some cases, these bounds are sufficiently tight to pin down the sign of the treatment effect.

4 Results

Our presentation of results is organized in three sets of tables. Tables 7 and 8 pertain solely to HHMI investigators, and validate empirically some of the purported distinctive features of the program. Table 9 presents evidence on the determinants of HHMI appointment. Finally, Tables 10, 11, and 12 present estimates of the program's effects.

HHMI appointments: rhetoric and practice. We begin by validating our claims about the terms of the HHMI investigator award. Table 7 reports unconditional probabilities of termination at the end of the first and second appointment terms for HHMI investigators.

Conditional on being renewed once, the odds of termination at the end of the second term are twice as high as the odds of termination at the end of one's initial appointment. However, our contention that the first review is laxer than the second has implications for the *conditional* probability of first and second reappointment. Specifically, if the perception of the program's administrators and investigators is accurate, the probability of second reappointment should be more responsive to achievements during the preceding term than the probability of first reappointment. Table 8 provides evidence consistent with this hypothesis. It reports estimates from logit models of reappointment as explained by various indicators of achievement during the preceding term. We find a consistent pattern, regardless of the achievement variable on the right-hand side: higher achievement significantly increases the likelihood of renewal at the end of the second term, but not at the end of the first term. Moreover, the marginal effect for blockbuster papers produced in the previous period is twice as large as the marginal effect for total publication output. This is consistent with the idea that HHMI review panels care more about whether investigators "transform their fields" than they care about counting lines on their CVs.

From these results, we conclude that the HHMI program conforms both in its *stated* and *actual* practices with the features that Manso (2009) predicts should encourage exploration.

Determinants of HHMI appointment. We now turn to the observable determinants of selection into the HHMI program (Table 9). We break down the analysis by control group (ECPW scholars and MERIT awardees). For each, we present the results from logit specifications that include demographic characteristics as controls, as well as cumulative NIH funding at baseline, and achievements as PI in the pre-appointment period. Among the demographic characteristics, the only consistent pattern is the higher appointment probability of female scientists. Consistent with the qualitative evidence on the selection process, we find that the number of "hit papers" at baseline is highly predictive of appointment. The effect of funding depends on the control group; as was already evident from the descriptive statistics presented in Table 4, MERIT awardees are much better funded at baseline, almost

by construction. In contrast, when compared to ECPW scientists, funding plays no role in the odds of appointment to the program.

Effects of HHMI appointment on Citation Impact. We report the effect of the program on the rate of publication output falling in four distinct citation quantile bins: all publications, publications in the top quartile, in the top ventile, and in the top percentile. For each outcome variable, we present five coefficients corresponding to different ways of assessing the program's effects. The first column reports naive cross-sectional results, which ignore the selection process. The second and third columns weight the outcome equations by the inverse probability of treatment so as to recover the Average Treatment Effect (ATE), and the Average Treatment Effect on the Treated (ATT) under unconfoundedness. The fourth column reports simple conditional fixed effects estimates, a naive difference-in-differences (DD). Finally, the fifth column reports results corresponding to semi-parametric difference-in-differences (SDD) estimates as in Abadie (2005). Since the SDD estimator adjusts the treatment effect for selection on observables while purging the estimates of time-invariant unobserved heterogeneity, it is our preferred specification.

Following a long-standing tradition in the study of scientific and technical change, the cross-sectional, ATE, ATT, and DD effects are estimated on the full panel using quasi-maximum likelihood Poisson.[15] In contrast, the SDD effects stem from a two-step procedure detailed in Appendix III.

Panel A of Table 10 compares outcomes for HHMIs and ECPW scholars. The naive cross-sectional estimate is always the largest in magnitude, and using propensity-score weighting reduces the magnitude of the effect by approximately a third. In contrast, the DD estimate is systematically lower than the SDD estimate, consistent with our conjecture that HHMIs and controls are on different output trends even before appointment. The magnitudes of the effects are large. For instance, the SDD estimates imply that HHMI increase the rate

[15]Because the Poisson model is in the linear exponential family, the coefficient estimates remain consistent as long as the mean of the dependent variable is correctly specified (Wooldridge 1996; Santos Silva and Tenreyro 2006). Further, robust standard errors are consistent even if the underlying data generating process is not Poisson.

at which they produce publications by $e^{.298} - 1 = 34\%$; the figure for papers in the Top 5% of the citation distribution is 47%; and for papers in the Top 1%, a 98% increase. The observed pattern is that the program has a bigger effect on the upper tail of the distribution of accomplishments, regardless of the estimation method used.

Figure 5 display the time path of the average "Top 5%" outcome, for HHMIs and ECPWs separately. In Panel A, we simply use the raw data (a "classic DD" picture); Panel B reweights each control scientist's outcome by his/her inverse probability of being selected into the program, while leaving the treated scientists' outcomes unchanged. In Panel A, there is no discernible break in the trend of output for HHMIs after 1994; in contrast, ECPW's output increases only modestly. Panel B tells a very different story. The trends for controls and treated scientists are very similar up until the time appointment; HHMIs' output begins to diverge sharply away from ECPW's only after 1998, consistent with the idea that scientists face adjustment costs when trying to alter the risk profile of their portfolio of projects. Loosely, Panel B of Figure 5 provides intuition for interpreting the SDD estimates: they correspond to the difference between the change in outcomes for the HHMIs and the change in outcomes for a pseudo-population of control scientists who are matched on observables with treated scientists, and in particular display parallel pre-appointment output trends.

We now turn to Panel B of Table 10, which assesses the program's effect on publication output relative to the group of MERIT awardees. Recall that the sample of MERIT scientists is much smaller than the sample of ECPW scholars, so that it is more difficult to cull from that population good matches for the HHMIs. Furthermore, MERIT scientists also receive a treatment, since they benefit from longer time horizons, before they must show evidence of achievements, relative to other NIH grantees. The broad patterns observed in Panel A are also observed in Panel B, but there are important exceptions. First, the effects are smaller in magnitude across the board. Second, there is little consistent evidence of effects of HHMI appointment, unless one focuses on outcomes in the tail. MERIT and HHMI scientists appear to produce "humdrum" publications at about the same rate, but the SDD estimates imply that HHMIs increase the number of "blockbuster" papers (papers in the Top 5% of the citation distribution) by about 24%, relative to MERIT awardees. This evidence suggests

that the large effects of the HHMI program found in Panel A should not be solely attributed to longer time horizons to plan and execute research projects. Some of the program's other features (rich and detailed feedback; "people, not projects") might also be important.

Effects of HHMI Appointment on Failure. It seems intuitive that exploration would lead scientists to "strike out" more often. Measuring failure is difficult, since it might lead researchers to abort projects altogether. Here we ask whether HHMIs produce more papers of little import, relative to controls. To answer this question, we examine whether HHMI appointment increases the rate of publications that fall in the bottom quartile of citations. The evidence (also presented in Table 10) is mixed. Relative to ECPW scholars, HHMIs indeed fail more often, regardless of estimation method; some of these estimates are large in magnitude, but they are also imprecisely estimated. Relative to MERIT scholars, the average treatment on the treated is negative and statistically significant. In contrast, the DD estimate is positive, and the SDD estimate negative but very small in magnitude. The lack of consistent result is perhaps not surprising if one remembers that relatively few of the papers produced by these elite scientists will fail to garner the three citations that correspond to the 25^{th} percentile of the citation distribution in most years.

Effects of HHMI appointment on the direction of research. So far, our presentation of results has conflated intensity of exploration with the rate at which tail outcomes are produced. But taken literally, the Manso (2009) model does not predict that "pay-for-future performance" incentives will result in better outcomes; it simply asserts that agents subject to those incentives will increase their rates of exploration, relative to agents who receive piece rates. Choosing less traveled scientific avenues could also leave trails in the content of what scientists publish, and in particular affect the keywords that tag their publications. This is the evidence we examine in Tables 11 and 12. We first focus on whether HHMI investigators are prone to define the scientific frontier, by examining the vintage of the MeSH keywords in their output. We do so both at the level of individual keywords, and for keyword pairs; the latter measure is justified by the presumption that the creative process is combinatoric in nature (Weitzman, 1998; Fleming et al. 2007) — few valuable ideas emerge from the

scientist's mind *ex nihilo.* In our analysis, a keyword or keyword pair is born the earliest year in which it appears in any publication indexed by PubMED. We then compute the average age of all keywords or all keyword pairs in each scientist's yearly output. Table 11 shows that HHMIs indeed tackle more novel topics; the coefficient estimates are negative regardless of estimation method. The magnitudes are relatively modest, and comparable for the two reference groups.

In Table 12, we ask whether evidence exists that HHMIs alter the direction of their scientific trajectory following their appointment. We first examine the program's effect on the number of unique publication keywords that overlap between the set of articles published in the "before period" (1986-1994) and the "after" period (1998-2006). This measure is then normalized by the number of unique keyword used in the after period. For each control group, we report both the results of a "naïve" specification, and the results of two specifications which incorporate inverse probability of treatment weights corresponding to ATE and ATT, respectively. Since the dependent variable is a proportion, we estimate these model using the quasi-maximum likelihood fractional logit estimator of Papke and Wooldridge (1996). Relative to ECPW scholars, HHMIs exhibit unambiguously lower overlap in keyword use (Panel A). The effects are statistically significant, and imply that HHMI appointment is associated with about a 10% lower rate of overlap. The evidence pertaining to the comparison with the MERIT awardees is more tenuous.

In a similar fashion, we measure how eclectic these scientists are in their choice of topic in the 1998-2006 period, by computing an Herfindahl index of MeSH keyword concentration. These models are estimated using the fractional logit estimator, and the specifications include (in addition to the treatment effect and demographic controls) the Herfindahl index computed using the output in the 1986-1994 period. The results are consistent with the idea that HHMIs broaden their research agenda in the post-appointment period, a necessary condition for exploration.

Our last test focuses on the *breadth* — rather than the *depth* — of impact for these scientists's publications. To do so, we examine the journals in which citing articles appear,

and compute the Herfindahl of journal concentration H. We find that HHMI's exhibit higher levels of $(1 - H)$ in the post-appointment period, relative to ECPWs. The comparison with MERIT awardees yields estimates of the same sign, but small in magnitude and imprecisely estimated.

Incentives vs. Alternative Mechanisms. Overall, the evidence points to program effects in line with our main hypotheses. Even if our estimates of HHMI appointments' treatment effects can be given a causal interpretation, ascribing them to the program's *incentive* features requires an interpretive leap.

First, we are unable to ascertain the extent to which the program increases productivity, rather than output. A quick glance at the descriptive statistics suffices to show that, per dollar of funding, HHMI investigators do not publish more papers than researchers funded by the NIH. Of course, if the supply of genuinely creative ideas is very inelastic, then publications per dollar of funding will not adequately measure researchers' productivity.[16] We also note that the results pertaining to the diversity of experimentation are less vulnerable to this critique, since they essentially hold output constant.

Second, the prestige conferred by HHMI appointment might have independent effects on scientists' achievements, either by increasing exposure to their research, or through a dynamic of ascription that has long been the focus of sociologists of science (Merton 1968). Azoulay, Stuart, and Wang (2009) provide estimates of the HHMI investigator program's "anointment effects" by examining whether appointment shifts the citation rate of articles written in the pre-appointment period; their evidence points to effects of very modest magnitude. As such, an interpretation of our results that emphasizes the status benefits of HHMI appointment appears unwarranted.

Third, collaboration between scientists in the treatment and control groups might threaten the validity of the comparisons drawn in the analysis. Fifty nine out of the seventy three

[16] In a recent paper, Jacob and Lefgren (2007) estimate that the elasticity of citations with respect to NIH R01 grant funding is quite small in magnitude, and often insignificantly different from 0. Given the regression-discontinuity design followed in their paper, it would be hazardous to import their estimate for the analysis of the scientist population analyzed in the present study.

HHMI investigators have at least one control collaborator; 10 have five or more. However, peer effects from coauthorship (e.g., Azoulay, Graff Zivin, and Wang 2009), which enhance the accomplishments of the control group, would tend to dampen the magnitudes of the effects estimated above.

In summary, we argue the differences observed between HHMI investigators and controls are primarily driven by the program's distinct incentive structure, as opposed to other potential effects of HHMI appointment.

5 Conclusion

In this paper, we exploit key differences across funding streams within the academic life sciences to examine the impact of incentives embodied in research contracts on the rate of scientific exploration. We find that selection into the HHMI investigator program — which rewards long-term success, encourages intellectual experimentation, and provides rich feedback to its appointees — leads to higher levels of breakthrough innovation, compared with NIH funding — which is characterized by short grant cycles, pre-defined deliverables, and unforgiving renewal policies. Moreover, the magnitudes of these effects are quite large.

Our findings are important for at least two reasons. First, they demonstrate the impact of nuanced features of research contracts for the rate and direction of scientific progress. Given the prominent role that scientific change is presumed to play in the process of economic growth (e.g., Mokyr 2002), this has important implications for the organization of public and private research institutions. Second, they offer empirical support for the theoretical model developed by Manso (2009), and as such may provide insights relevant to a wider set of industries that rely on creative professionals, ranging from advertising and computer programming to leadership roles at the upper echelons of the corporate world.

Of course, all our results depend on the maintained assumption that selection into the program operates solely on the basis of observable factors. As in many observational studies, this assumption cannot be tested from the data. In the absence of random assignment, our conclusions must therefore remain guarded.

Moreover, it is possible that NIH grant and HHMI funding mechanisms complement one another in ways we cannot account for. HHMI investigators remain eligible for NIH grants, and only four of those in the sample never become NIH grantees during the observation period. As such, our results are most informative about funding decisions at the margin, those that help determine the optimal mix of "exploration" and "exploitation" incentives necessary to stimulate the creation of particularly valuable scientific knowledge.

Finally, our results should not be interpreted as a critique of NIH and its funding policies. While "exploration" incentive contracts appear to stimulate creativity in this setting, 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 our results 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. Moreover, HHMI provides detailed evaluation and feedback to its investigators. The richness of this feedback consumes a great deal of resources, particularly the time of the elite scientists that serve on review panels, and its quality might degrade if the program was expanded drastically.

Much more could be done to explore the impacts of contract design on research output in this setting. For example, do the quality of peers at these investigators' institution temper or magnify these effects? Do the effects of exploration-style incentives exhibit hysteresis, i.e., do they lead scientists to be more creative under more conventional contractual arrangements? Answering these questions are the next steps of our research agenda.

References

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72(1): 1–19.

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. 2008. "Academic Freedom, Private Sector Focus, and the Process of Innovation." *RAND Journal of Economics* 39(3): 617–35.

Amabile, Teresa M. 1996. Creativity in Context. Boulder, CO: Westview Press.

Azoulay, Pierre, Andrew Stellman, and Joshua Graff Zivin. 2006. "PublicationHarvester: An Open-Source Software Tool for Science Policy Research." *Research Policy*, 35:7: 970-974.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2009. "Superstar Extinction." Forthcoming, *Quartely Journal of Economics.* Also NBER Working Paper #14577.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. 2009. "Matthew: Effect or Correlation?" Working Paper, MIT Sloan School of Management.

Burt, Ronald S. 2004. "Structural Holes and Good Ideas." *American Journal of Sociology* 110(2): 349–99.

Busso, Matias, John DiNardo, and Justin McCrary. 2008. Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects. Working Paper, University of Michigan.

Cappechi, Mario R. 2008. Response. S*cience* 319(5865): 900-1.

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics,* 84(1): 151-61.

Ederer, Florian and Gustavo Manso. 2008. "Are Financial Incentives Detrimental to Innovation?" Working Paper, MIT Sloan School of Management.

Fleming, Lee, Santiago Mingo, and David Chen. 2007. "Collaborative Brokerage, Generative Creativity, and Creative Success." *Administrative Science Quarterly* 52(3): 443–75.

Hirano, Keisuke and Guido W. Imbens. 2001. "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catherization." *Health Services & Outcomes Research Methodology,* 2(3-4): 259-278.

Holmström, Bengt. 1989. "Agency Costs and Innovation." *Journal of Economic Behavior and Organization* 12(3): 305-327.

Jacob, Brian and Lars Lefgren. 2007. The Impact of Research Grant Funding on Research Productivity. NBER Working Paper #13519.

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.

Lazear, Edward. 2000. "Performance Pay and Productivity." *American Economic Review* 90(5): 1346-1361.

Lerner, Josh and Julie Wulf. 2007. "Innovation and Incentives: Evidence from Corporate R&D." *Review of Economics and Statistics* 89(4): 634–644.

Lotka, Alfred J. 1926. The Frequency Distribution of Scientific Productivity. *Journal of the Washington*

Academy of Sciences 16(12): 317-323.

Manso, Gustavo. 2009. "Motivating Innovation." Working Paper, MIT Sloan School of Management.

March, James G. 1991. Exploration and Exploitation in Organizational Learning. *Organization Science* 2(1): 71-87.

Merton, Robert K. 1968. The Matthew Effect in Science. *Science* 159(3810): 56-63.

Mokyr, Joel. 2002. The Gifts of Athena. Princeton, NJ: Princeton University Press.

Papke, Leslie E. and Jeffrey M. Wooldridge. 1996. "Econometric Methods for Fractional Responses with an Application to 401(k) Plan participation Rates." *Journal of Applied Econometrics* 11(6): 619–632.

Robins, James M., and Andrea Rotnitzky. 1995. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association* 90(429): 122- 129.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

Rubin, Donald B. 1974. "Characterizing the Estimation of Parameters in Incomplete-Data Problems." *Journal of the American Statistical Association* 69(346): 467-474.

Santos Silva, J.M.C. and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics* 88(4): 641-658.

Simonton, Dean Keith. 2004. Creativity in Science: Chance, Logic, Genius, and Zeitgeist. New York: Cambridge University Press.

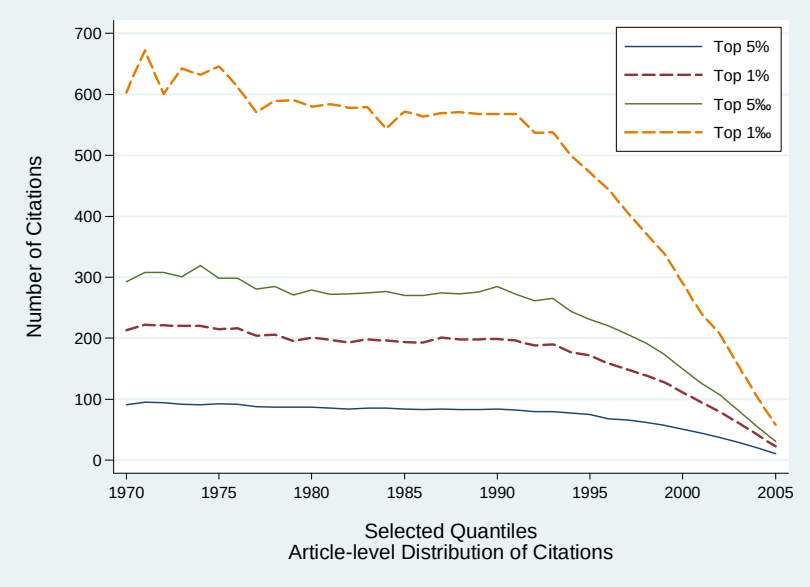
Smith, Jeffrey A. and Petra E. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics,* 125(1-2): 305-353.

Weitzman, Martin. 1998. Recombinant Growth. *Quartely Journal of Economics* 113(2): 331-360.

Wooldridge, Jeffrey M. 1996. Quasi-Likelihood Methods for Count Data. pp. 352-406 in Handbook of Applied Econometrics, Vol. 2. M.H. Pesaran and P. Schmidt (eds.). Oxford: Blackwell.

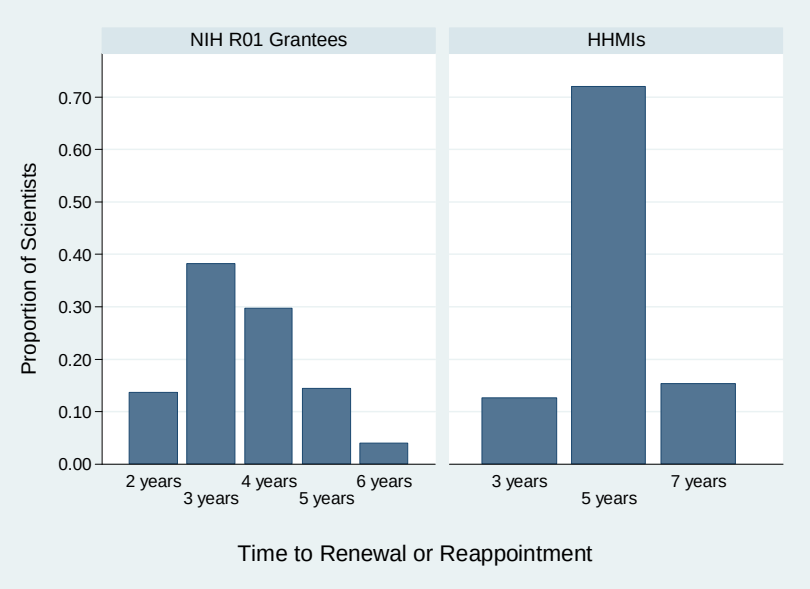
Wooldridge, Jeffrey M. 2002. "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification." *Portuguese Economic Journal,* 1(2): 117–139.

Figure 1
Measuring the Tail of the Distribution of Citations



Notes: Selected quantiles (0.050, 0.010, 0.005, & 0.001) for the vintage-specific empirical distribution of the number of citations at the article level. These quantiles were computed in early 2008 using the universe of all articles indexed by ISI/Web of Science that appeared in life science journals.

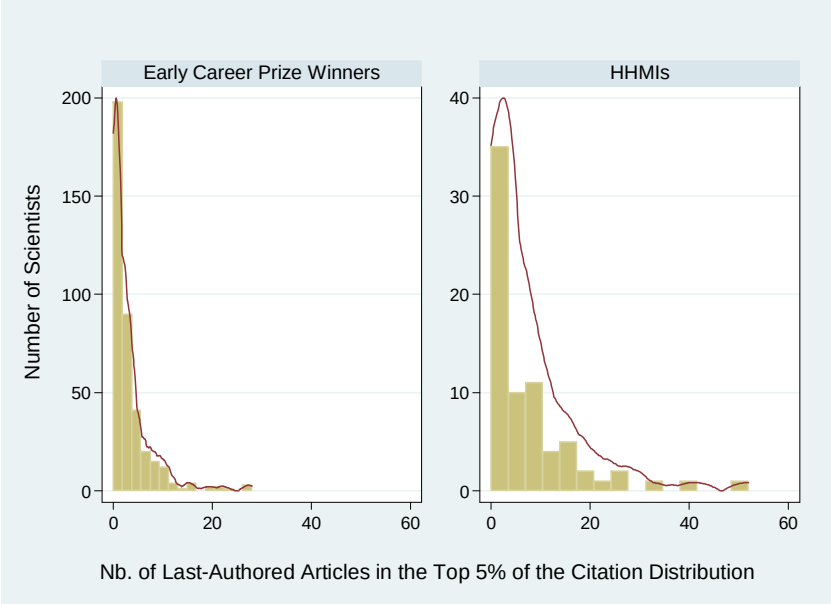
Figure 2
Length of NIH R01 Grants vs. HHMI Appointments



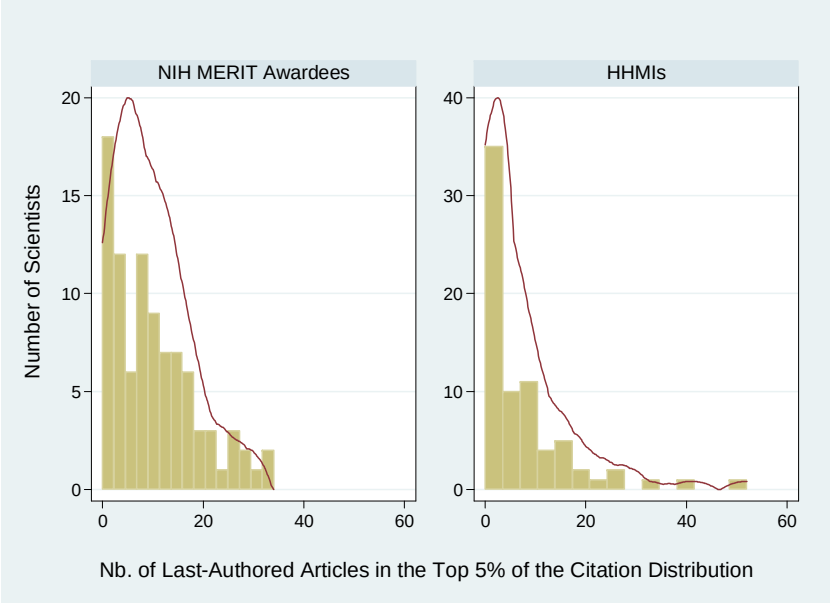
Notes: NIH tabulations stem from the Compound Grant Applicant File (CGAF). The grants considered are R01 and equivalent whose first cycle began later than 1970, but earlier than 2002.

Figure 3
Baseline Number of "Hits" as PI

A. Early Career Prize Winners vs. HHMIs

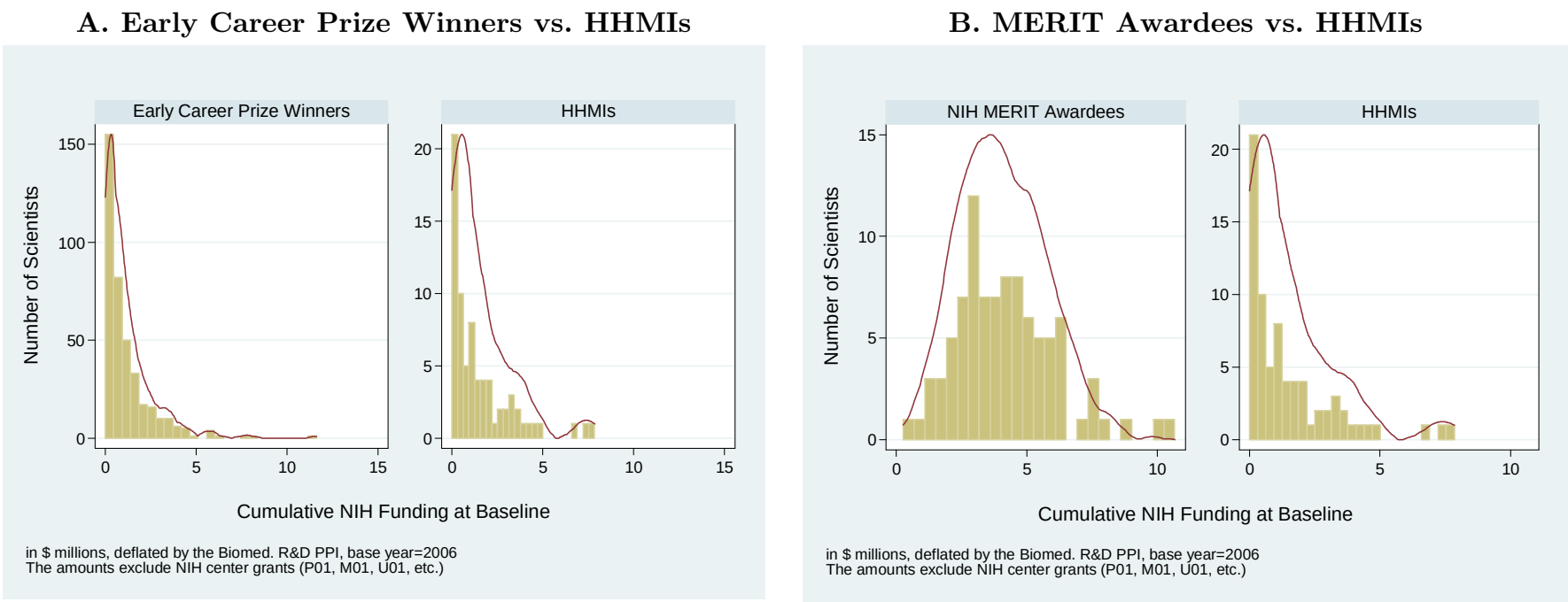


B. MERIT Awardees vs. HHMIs



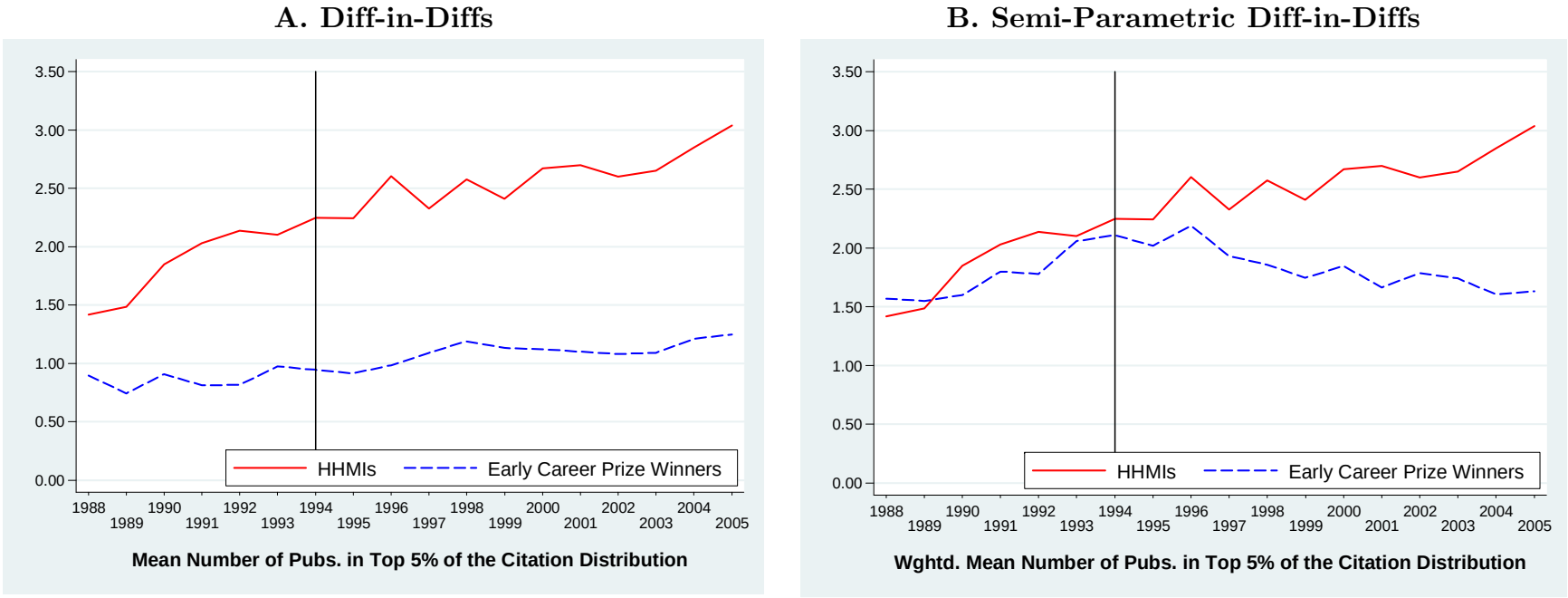
Notes: We focus on articles in which the scientist appears in last position on the authorship list, because this is the set that clearly identifies treated and control scientists as principal investigators in the pre-appointment period.

Figure 4
Baseline NIH Funding



Notes: We exclude research center grants when computing our funding measures because these grants are less likely to correspond to individual effort; in some cases, deans or department chairs serve as pro-forma PIs on such grants, making it a less useful measure for our purposes.

Figure 5
Dynamics of HHMI Appointment on "Hits"



Notes: The dashed blue and solid red lines in the above plots correspond to the average yearly number of "hits" (publications that fall in the top ventile of the vintage-specific article-level distribution of citations) for early career prize winners and HHMIs, respectively. In Panel B, the averages for the control scientists are weighted by each researcher's inverse probability of treatment, where the weights are computed using fitted values of the logit specification in Table 9A, Model (3).

Table 1
Comparison Between the Two Sources of Funding

NIH R01 Grants	HHMI Investigator Program
3-year funding	5-year funding
first review is similar to any other review	first review is rather lax
funds dry up upon non-renewal	two-year phase-down upon non-renewal
some feedback in the renewal process	feedback from renowned scientists
funding is for a particular project	"people, not projects"

Table 2
Demographic Characteristics

		Female	MD	PhD	MD/PhD	US Born
Controls	N=485	0.204	0.0825	0.792	0.126	0.827
HHMIs	N=73	0.342	0.0822	0.753	0.164	0.863
Total	N=558	0.222	0.0824	0.787	0.131	0.832

Table 3
Lab Type

		Macromolecular	Cellular	Organismal	Translational
Controls	N=485	0.245	0.373	0.272	0.105
HHMIs	N=73	0.288	0.329	0.274	0.110
Total	N=558	0.251	0.367	0.272	0.106

Table 4
Descriptive Statistics at Baseline

	Mean	Median	Std. Dev	Min.	Max.
Early Career Prize Winners (N=393)					
Degree Year	1983.689	1984	3.738	1974	1991
Female	0.199	0	0.400	0	1
Nb. of Nomination Slots	2.179	2	1.296	0	8
Cum. NIH Funding	$1,106,790	$676,249	$1,375,588	0	$11,634,552
Cum. Nb. of Pubs.	24.775	20	20.764	2	200
Cum. Nb. of Pubs. in the Bttm. 25%	0.647	0	1.410	0	15
Cum. Nb. of Pubs. in the Top 25%	18.718	15	14.146	0	123
Cum. Nb. of Pubs. in the Top 5%	9.647	8	7.822	0	51
Cum. Nb. of Pubs. in the Top 1%	3.712	3	3.875	0	27
Average MeSH Age	23.376	23	2.808	18	35
Average MeSH 2-way Combo Age	11.970	12	2.234	8	22
Herf. Index of Focus, 1986-1994	0.024	0	0.013	0.012	0.167
Citing Journal Diversity, 1986-1994	0.963	1	0.020	0.837	0.992
NIH MERIT Awardees (N=92)					
Degree Year	1976.978	1977	2.361	1974	1983
Female	0.222	0	0.418	0	1
Nb. of Nomination Slots	1.989	2	1.426	0	8
Cum. NIH Funding	$4,212,431	$4,017,763	$1,936,989	$265,601	$10,678,306
Cum. Nb. of Pubs.	69.222	62	36.670	4	228
Cum. Nb. of Pubs. in the Bttm. 25%	2.133	1	2.904	0	20
Cum. Nb. of Pubs. in the Top 25%	49.167	46	25.806	2	168
Cum. Nb. of Pubs. in the Top 5%	21.989	20	15.496	1	89
Cum. Nb. of Pubs. in the Top 1%	6.711	5	6.816	0	32
Average MeSH Age	23.433	23	2.796	17	30
Average MeSH 2-way Combo Age	12.228	12	2.048	9	18
Herf. Index of Focus, 1986-1994	0.019	0	0.005	0.009	0.037
Citing Journal Diversity, 1986-1994	0.969	1	0.020	0.904	0.992
HHMIs (N=73)					
Degree Year	1983.723	1984	4.002	1974	1991
Female	0.369	0	0.486	0	1
Nb. of Nomination Slots	2.194	2	1.222	0	8
Cum. NIH Funding	$1,502,810	1,005,176	$1,768,341	0	$7,852,110
Cum. Nb. of Pubs.	32.657	23	27.399	3	172
Cum. Nb. of Pubs. in the Bttm. 25%	0.627	0	0.902	0	4
Cum. Nb. of Pubs. in the Top 25%	26.866	19	23.398	3	148
Cum. Nb. of Pubs. in the Top 5%	16.910	13	16.889	1	119
Cum. Nb. of Pubs. in the Top 1%	8.478	5	10.224	0	73
Average MeSH Age	22.824	23	2.253	17	29
Average MeSH 2-way Combo Age	11.453	11	1.735	9	17
Herf. Index of Focus, 1986-1994	0.021	0	0.008	0.011	0.051
Citing Journal Diversity, 1986-1994	0.965	1	0.018	.921	.992

Table 5
Descriptive Statistics — Career Achievement

	Mean	Median	Std. Dev	Min.	Max.
Early Career Prize Winners (N=393)					
Career Total Funding	$5,569,900	$5,064,601	$3,468,376	$258,991	$23,668,086
Career Nb. of Articles	65.003	53	43.444	11	314
Career Nb. of Citations	4,489	3,504	3,489	242	21,448
Career Nb. of Articles in the Top 25%	47.952	40	30.829	7	212
Career Nb. of Articles in the Top 5%	22.214	18	15.760	0	96
Career Nb. of Articles in the Top 1%	7.926	6	7.410	0	38
Herf. Index of Focus, 1998-2006	0.016	0	0.009	0.006	0.143
Citing Journal Diversity, 1998-2006	0.968	1	0.025	0.667	0.992
Normalized MeSH Kwd. Overlap	0.104	0	0.062	0	0.462
NIH MERIT Awardees (N=92)					
Career Total Funding	$11,449,776	$10,649,311	$5,669,283	$3,169,828	$36,202,848
Career Nb. of Articles	140.174	122	73.496	14	463
Career Nb. of Citations	9,294	7,560	7,145	260	37,696
Career Nb. of Articles in the Top 25%	100.076	89	56.401	4	344
Career Nb. of Articles in the Top 5%	44.196	40	33.997	2	168
Career Nb. of Articles in the Top 1%	14.533	12	15.951	0	81
Herf. Index of Focus, 1998-2006	0.014	0	0.004	0.007	0.023
Citing Journal Diversity, 1998-2006	0.971	1	0.017	0.914	0.994
Normalized MeSH Kwd. Overlap	0.109	0	0.034	0	0.232
HHMIs (N=73)					
Career Total Funding	$14,652,401	$13,099,974	$5,600,108	$6,282,796	$31,989,716
Career Nb. of Articles	95.521	83	56.126	17	321
Career Nb. of Citations	10,550	6,672	14,542	798	117,401
Career Nb. of Articles in the Top 25%	78.219	69	48.843	10	284
Career Nb. of Articles in the Top 5%	45.562	38	33.863	4	224
Career Nb. of Articles in the Top 1%	21.014	16	21.270	0	144
Herf. Index of Focus, 1998-2006	0.014	0	0.004	0.005	0.026
Citing Journal Diversity, 1998-2006	0.975	1	0.013	0.921	0.993
Normalized MeSH Kwd. Overlap	0.085	0	0.037	0	0.188

Table 6
Accolades

		Early Career Prize Winners Trained	Nobel Prize Winners	Elected NAS Member	Elected IoM Member
Controls	N=485	118 (0.243 per scientist)	1 (0.20%)	28 (5.80%)	12 (2.50%)
HHMIs	N=73	83 (1.137 per scientist)	1 (1.40%)	24 (32.90%)	12 (16.40%)
Total	N=558	201 (0.360 per scientist)	2 (0.40%)	52 (9.30%)	24 (4.30%)

Table 7
Univariate Termination Data for HHMI Investigators

	End of First Appointment	End of Second Appointment
Reappointed	60 (84.51%)	43 (71.16%)
Terminated	11 (15.49%)	17 (28.33%)

Table 8
Sensitivity of HHMI Reappointment to Scientific Output

	First Reappt. (1a)	Second Reappt. (1b)	First Reappt. (2a)	Second Reappt. (2b)	First Reappt. (3a)	Second Reappt. (3b)	First Reappt. (4a)	Second Reappt. (4b)
Pubs in the Elapsed Period	-0.001	0.024**						
	(0.001)	(0.005)						
Pubs in the Top 25% in the Previous Period			-0.002	0.027**				
			(0.002)	(0.007)				
Pubs in the Top 5% in the Elapsed Period					-0.003	0.027**		
					(0.003)	(0.010)		
Pubs in the Top 1% in the Elapsed Period							-0.003	0.053**
							(0.006)	(0.020)
Female	0.039	0.022	0.040	0.035	0.036	0.053	0.045	0.086
	(0.100)	(0.114)	(0.102)	(0.115)	(0.105)	(0.121)	(0.107)	(0.119)
Associate	0.028	0.096	0.029	0.076	0.023	0.128	0.027	0.153
	(0.100)	(0.104)	(0.100)	(0.117)	(0.097)	(0.121)	(0.099)	(0.119)
Full	0.070	0.001	0.066	-0.026	0.059	0.074	0.057	0.098
	(0.114)	(0.146)	(0.112)	(0.192)	(0.110)	(0.206)	(0.110)	(0.213)
Nb. Scientists	71	60	71	60	71	60	71	60
Log Quasi-Likl.	-27.497	-19.841	-27.653	-21.251	-27.674	-24.150	-27.895	-24.176
Pseudo-R^2	0.102	0.338	0.097	0.291	0.096	0.194	0.089	0.193

Note: The dependent variable is the probability of being reappointed, whether at the end of the first term (Models 1a, 2a, 3a, & 4a), or at the end of the second term (Models 1b, 2b, 3b, & 4b), among 71 HHMI investigators who did not terminate their appointment voluntarily. The sample relevant to specifications 1b, 2b, 3b, & 4b comprises only 60 observations since 11 investigators were either not renewed at the end of the first appointment period, or resigned their posts voluntarily. Estimates correspond to marginal effects from logit specifications, with robust standard errors in parentheses.

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 9
Determinants of Selection into the HHMI Program

	A. Early Career Prize Winners Control Group				B. MERIT Awardees Control Group			
	(1a)	(2a)	(3a)	(4a)	(1b)	(2b)	(3b)	(4b)
Cum. Nb. of Pubs as PI	0.006**				0.003**			
	(0.002)				(0.001)			
Pubs in Top 25% as PI		0.013**				0.006**		
		(0.002)				(0.002)		
Pubs in Top 5% as PI			0.023**				0.015**	
			(0.004)				(0.003)	
Pubs in Top 1% as PI				0.039**				0.029**
				(0.008)				(0.004)
NIH Funding	0.004	-0.018	-0.015	-0.001	-0.039*	-0.048**	-0.073**	-0.076**
	(0.024)	(0.019)	(0.018)	(0.015)	(0.016)	(0.016)	(0.021)	(0.016)
Female	0.121**	0.123**	0.119**	0.122**	0.088	0.101†	0.117*	0.100*
	(0.036)	(0.035)	(0.034)	(0.034)	(0.055)	(0.053)	(0.051)	(0.047)
PhD	-0.082	-0.078	-0.058	-0.032	-0.210**	-0.201**	-0.192**	-0.131**
	(0.087)	(0.096)	(0.104)	(0.100)	(0.059)	(0.061)	(0.057)	(0.049)
MD/PhD	-0.048	-0.053	-0.022	0.007	-0.086	-0.090	-0.101	-0.041
	(0.082)	(0.087)	(0.092)	(0.089)	(0.087)	(0.088)	(0.080)	(0.081)
Nb. of Nomination Slots	-0.010	-0.011	-0.008	-0.006	0.028†	0.026†	0.024†	0.025*
	(0.014)	(0.013)	(0.012)	(0.012)	(0.015)	(0.014)	(0.013)	(0.012)
Macromolecular Lab	-0.039	-0.041	-0.024	-0.030	0.047	0.036	0.073	0.095†
	(0.043)	(0.042)	(0.041)	(0.042)	(0.063)	(0.063)	(0.051)	(0.050)
Organismal Lab	0.002	0.004	0.002	-0.004	0.028	0.026	0.016	0.025
	(0.046)	(0.045)	(0.044)	(0.044)	(0.071)	(0.069)	(0.059)	(0.062)
Translational Lab	-0.014	-0.005	0.008	0.013	0.004	0.002	-0.056	0.007
	(0.085)	(0.087)	(0.090)	(0.083)	(0.093)	(0.097)	(0.083)	(0.069)
Pseudo-R^2	0.074	0.111	0.143	0.133	0.585	0.610	0.671	0.715
Nb. of Scientists	466	466	466	466	165	165	165	165

Note: The dependent variable is the probability of being appointed as an HHMI investigator. Estimates correspond to marginal effects from logit specifications, with robust standard errors in parentheses. Achievement at baseline is measured as the cumulative number of publications that fall in a particular citation bin, considering only these articles in which the scientist appears in last position on the authorship list, i.e., is clearly identified as the principal investigator of a laboratory. All models also include year of highest degree indicator variables (coefficients not reported).

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 10
Effects of HHMI Appointment on Citation Impact

Achievement Metric	A. Early Career Prize Winners Control Group					B. MERIT Awardees Control Group				
	"Naïve" X-Sect.	ATE	ATT	DD	SDD	"Naïve" X-Sect.	ATE	ATT	DD	SDD
All Pubs	0.462**	0.290**	0.227*	0.105	0.298**	-0.076	-0.068	-0.146	0.199**	0.148
	(0.073)	(0.075)	(0.089)	(0.072)	(0.099)	(0.105)	(0.117)	(0.133)	(0.072)	(0.093)
Top 25%	0.573**	0.361**	0.319**	0.144†	0.195†	0.057	0.050	-0.110	0.198*	0.159†
	(0.078)	(0.084)	(0.086)	(0.075)	(0.100)	(0.111)	(0.122)	(0.128)	(0.077)	(0.097)
Top 5%	0.828**	0.560**	0.544**	0.197†	0.384*	0.290*	0.261†	0.008	0.140	0.212†
	(0.092)	(0.110)	(0.102)	(0.101)	(0.187)	(0.140)	(0.152)	(0.143)	(0.098)	(0.123)
Top 1%	1.103**	0.777**	0.899**	0.254†	0.684**	0.569**	0.427*	0.347*	0.011	0.035
	(0.133)	(0.135)	(0.136)	(0.138)	(0.231)	(0.212)	(0.169)	(0.167)	(0.147)	(0.208)
Bttm. 25%	0.183	0.115	0.081	0.045	0.352	-0.398*	-0.527*	-0.239	0.132	-0.005
	(0.122)	(0.130)	(0.140)	(0.266)	(0.766)	(0.164)	(0.224)	(0.228)	(0.298)	(0.327)
Nb. of Obs.	8,767	8,767	8,767	8,767		3,832	3,832	3,832	3,832	
Nb. of Scientists	466	466	466	466	466	165	165	165	165	165

Note: Each coefficient corresponds to the treatment effect of HHMI appointment in a specification that regresses output on treatment status, five age indicator variables (5 to 10 years of career age, 10 to 15 years, 15 to 20 years, 20 to 25 years, and 25 years and more of career age), and year indicator variables in all models. The cross-sectional models (corresponding to the first three columns in each panel) also include three lab indicator variables, a gender indicator variable, and two degree type indicator variables (coefficients not reported). Estimates derive from QML Poisson estimation, with robust standard errors in parentheses, clustered around scientist (X-section, ATE, ATT, and DD columns); bootstrapped standard errors are reported for the semi-parametric diff-in-diffs estimates. All specifications except the naïve cross-sections and the plain diffs-in-diffs include regression weights computed using fitted values for the probability of HHMI appointment estimated in Table 9. The weights differ depending on whether ATT or ATE is the effect of interest, and whether the focus is on generating a between-scientist comparison (ATE & ATT columns), or a within-scientist comparison (SDD column). See section 3 in the text for more details.
† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 11
Effects of HHMI Appointment on Topic Novelty

Depndt. Variable	A. Early Career Prize Winners Control Group					B. MERIT Awardees Control Group				
	X-Sect.	ATE	ATT	DD	SDD	X-Sect.	ATE	ATT	DD	SDD
Avg. MeSH Kwd. Age	-0.035**	-0.021*	-0.024**	-0.015	-0.030*	-0.042**	-0.029†	-0.021	-0.007	-0.010
	(0.009)	(0.010)	(0.009)	(0.013)	(0.012)	(0.015)	(0.017)	(0.017)	(0.015)	(0.016)
Avg. 2-way MeSH Kwd. Combo Age	-0.056**	-0.029†	-0.037*	-0.017	-0.033	-0.071**	-0.062*	-0.048†	-0.023	-0.041
	(0.014)	(0.016)	(0.015)	(0.021)	(0.023)	(0.023)	(0.029)	(0.029)	(0.025)	(0.030)
Nb. of Obs.	8,767	8,767	8,767	8,767		3,832	3,832	3,832	3,832	
Nb. of Scientists	466	466	466	466	466	165	165	165	165	165

Note: Each coefficient corresponds to the treatment effect of HHMI appointment in a specification that regresses measures of scientific novelty on treatment status, five age indicator variables (5 to 10 years of career age, 10 to 15 years, 15 to 20 years, 20 to 25 years, and 25 years and more of career age), and year indicator variables in all models. The cross-sectional models (corresponding to the first three columns in each panel) also include three lab indicator variables, a gender indicator variable, and two degree type indicator variables (coefficients not reported). Estimates derive from QML Poisson estimation, with robust standard errors in parentheses, clustered around scientist (X-section, ATE, ATT, and DD columns); bootstrapped standard errors are reported for the semi-parametric diff-in-diffs estimates. All specifications except the naïve cross-sections and the plain diff-in-diffs include regression weights computed using fitted values for the probability of HHMI appointment estimated in Table 9. The weights differ depending on whether ATT or ATE is the effect of interest; and whether the focus is on generating a between-scientist comparison (ATE & ATT columns), or a within-scientist comparison (SDD column). See section 3 in the text for more details.

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Table 12
Effects of HHMI Appointment on the Direction of Research

Depndt. Variable	A. Early Career Prize Winners Control Group			B. MERIT Awardees Control Group		
	X-Sect.	ATE	ATT	X-Sect.	ATE	ATT
Normalized MeSH Keyword Overlap	-0.209**	-0.185**	-0.209**	-0.028	0.007	-0.066
	(0.061)	(0.058)	(0.061)	(0.095)	(0.079)	(0.087)
Herfindahl Index of Scholarly Focus	-0.174**	-0.153**	-0.125**	-0.007	-0.068	-0.031
	(0.042)	(0.041)	(0.042)	(0.061)	(0.055)	(0.055)
Citing Journal Diversity Index	0.253**	0.194**	0.216**	0.110	0.137	0.094
	(0.072)	(0.065)	(0.070)	(0.139)	(0.127)	(0.123)
Nb. of Scientists	466	466	466	165	165	165

Note: Each coefficient corresponds to the treatment effect of HHMI appointment on various measures of an investigator's scientific direction in the "after period" (1998-2006). All models include as independent variables year of highest degree indicator variables, three lab indicator variables, a gender indicator variable, and two degree type indicator variables (coefficients not reported). Also included is an offset for the dependent variable in the "before period" (1986-1994). Because all of the dependent variable are bounded inclusively by 0 and 1, estimates derive from a QML fractional logit procedure (Papke & Wooldridge 1996), with robust standard errors in parentheses. The specifications for ATE and ATT include regression weights computed using fitted values for the probability of HHMI appointment (See section 3 in the text for more details).

† $p < 0.10$, * $p < 0.05$, ** $p < 0.01$.

Appendix I
Funding People vs. Funding Specific Projects

We develop a simple model to contrast the "specific project" and the "people, not projects" approaches to scientific funding. The researcher lives for two periods. In each period, she chooses a project $i \in \mathcal{I}$, producing output S ("success") with a probability p_i or output F ("failure") with probability $1-p_i$. The probability p_i of success when the researcher chooses project i may be unknown. To obtain information about p_i, she must engage in experimentation. We let $E[p_i]$ denote the unconditional expectation of p_i, $E[p_i|S,j]$ denote the conditional expectation of p_i given a success on project j, and $E[p_i|F,j]$ denote the conditional expectation of p_i given a failure on project j.

When the researcher chooses project $i \in \mathcal{I}$, she only learns about the probability p_i, so that

$$E[p_j] = E[p_j|S,i] = E[p_j|F,i] \quad \text{for} \quad j \neq i.$$

The central concern that arises is the tension between exploration of new ideas and the exploitation of already existing ideas along conventional lines (March, 1991). To focus on the tension between exploration and exploitation, we assume that in each period the researcher chooses between two projects. Project 1, the "conventional" research project, has a known probability p_1 of success, such that

$$p_1 = E[p_1] = E[p_1|S,1] = E[p_1|F,1].$$

Project 2, the innovative research project, has an unknown probability p_2 of success such that

$$E[p_2|F,2] < E[p_2] < E[p_2|S,2].$$

We assume that the innovative research project is of an exploratory nature. This means that when the researcher experiments with the innovative research project, she is initially not as likely to succeed as when she treads a well-trodden path, as is the case with the conventional research project. However, if she succeeds with the innovative project, she updates her beliefs about p_2, so that choosing the innovative project becomes perceived as superior to choosing the conventional project. This is captured by:

$$E[p_2] < p_1 < E[p_2|S,2]. \tag{1}$$

The researcher is risk-neutral and has a discount factor normalized to one. Her objective function R assigns some weight α to the outcome produced by his research as well as some weight to her private preferences between the two projects. These private preferences are represented with a cost c_i that is incurred by the researcher whenever she pursues project i. The researcher thus chooses an action plan $\langle {}_i{}^j_k \rangle$ to maximize her total expected payoff

$$\begin{aligned} R(\langle {}_i{}^j_k \rangle) = {} & \{E[p_i]S + (1-E[p_i])F - c_i\} \\ & + E[p_i]\{E[p_j|S,i]S + (1-E[p_j|S,i])F - c_j\} \\ & + (1-E[p_i])\{E[p_k|F,i]S + (1-E[p_k|F,i])F - c_k\} \end{aligned} \tag{2}$$

where i is the first-period action, j is the second-period action in case of success in the first period, and k is the second-period action in case of failure in the first period. We assume that the researcher gets enough funding to perform research during the two periods. We consider two funding mechanisms: the "specific project" approach and the "people, not projects" approach.

The "Specific Project" approach. Under this approach the researcher must choose one project to submit for funding and must work on that project during the two periods. Two action plans need to be considered:

$\langle {}_1{}^1_1 \rangle$ and $\langle {}_2{}^2_2 \rangle$. If the researcher chooses action plan $\langle {}_1{}^1_1 \rangle$ his total expected payoff is

$$\begin{aligned} R(\langle {}_1{}^1_1 \rangle) = &\{E[p_1]S + (1 - E[p_1])F - c_1\} \\ &+ E[p_1]\{E[p_1]S + (1 - E[p_1])F - c_1\} \\ &+ (1 - E[p_1])\{E[p_1]S + (1 - E[p_1])F - c_1\} \end{aligned} \quad (3)$$

If the researcher chooses action plan $\langle {}_2{}^2_2 \rangle$ his total expected payoff is

$$\begin{aligned} R(\langle {}_2{}^2_2 \rangle) = &\{E[p_2]S + (1 - E[p_2])F - c_2\} \\ &+ E[p_2]\{E[p_2|S,2]S + (1 - E[p_2|S,2])F - c_2\} \\ &+ (1 - E[p_2])\{E[p_2|F,2]S + (1 - E[p_2|F,2])F - c_2\} \end{aligned} \quad (4)$$

From Bayes' rule, the payoff $R(\langle {}_2{}^2_2 \rangle)$ is higher than the payoff $R(\langle {}_1{}^1_1 \rangle)$ if and only if

$$\alpha(E[p_2] - p_1)(S - F) \geq (c_2 - c_1). \quad (5)$$

The "People, not Projects" approach. Under this approach, the researcher can choose any of the two projects in each period. Two action plans need to be considered: $\langle {}_1{}^1_1 \rangle$, and action plan $\langle {}_2{}^2_1 \rangle$. If the researcher chooses action plan $\langle {}_1{}^1_1 \rangle$, her total expected payoff is

$$\begin{aligned} R(\langle {}_1{}^1_1 \rangle) = &\{E[p_1]S + (1 - E[p_1])F - c_1\} \\ &+ E[p_1]\{E[p_1]S + (1 - E[p_1])F - c_1\} \\ &+ (1 - E[p_1])\{E[p_1]S + (1 - E[p_1])F - c_1\} \end{aligned} \quad (6)$$

If the researcher chooses action plan $\langle {}_2{}^2_1 \rangle$, her total expected payoff is

$$\begin{aligned} R(\langle {}_2{}^2_1 \rangle) = &\{E[p_2]S + (1 - E[p_2])F - c_2\} \\ &+ E[p_2]\{E[p_2|S,2]S + (1 - E[p_2|S,2])F - c_2\} \\ &+ (1 - E[p_2])\{E[p_1]S + (1 - E[p_1])F - c_1\} \end{aligned} \quad (7)$$

The payoff $R(\langle {}_2{}^2_1 \rangle)$ is higher than $R(\langle {}_1{}^1_1 \rangle)$ if and only if

$$\alpha\{(E[p_2](E[p_2|S,2] - p_1) + (E[p_2] - p_1)\}(S - F) \geq (1 + E[p_2])(c_2 - c_1) \quad (8)$$

The following proposition contrasts exploration under "specific project" funding and "people, not projects" funding.

Proposition 1 *If the agent explores under "specific project" funding, he also explores under "people, not projects" funding. However, there are situations in which the agent explores under "people, not projects" funding, but exploits under "specific project" funding.*

Proof The first statement follows from the fact that (5) implies (8). For the second statement, we construct the following example. If $c_2 > c_1$, (5) implies that the agent never explores under the "specific project" approach. However, from (8), if the payoff from exploration is sufficiently high, the agent will explore under the "people, not projects" approach. ■

Appendix II
Career & Output Data

For every scientist in the control or treatment group, we collected career information from three sources: original CVs/NIH biosketches; Who's Who profiles, and Google searches. In practice, the combination of these approaches enabled us to find employment and demographic data for all the investigators considered in the paper. Matching these individuals with NIH grant information is not challenging since both full names and institutional affiliations can be used. Getting a precise tally of publications at the individual level is more involved. We will describe this process using as an example Mario Capecchi, the Nobel Prize winner (and HHMI) mentioned in the introduction.

The matching process begins with the creation of a customized PubMED search query for each scientist. In the case of Capecchi, the query is `(("capecchi mr"[au] OR "capecchi m"[au]) NOT 7816017[pmid] AND 1966:2006[dp])`, and it returns 122 original publications (the query also returns 19 letters, editorials, interviews, reviews, etc., which we ignore). The process of harvesting bibliomes from PubMED using name variations and queries as inputs is facilitated by the use of PubHarvester, a software program we specifically designed for this purpose (Azoulay, Stellman, and Graff Zivin 2006).

Capecchi's PubMED query accounts for his inconsistent use of the middle initial, but is otherwise quite simple. For other scientists, queries might factor in their inconsistent use of the suffix "Jr.," or name variations coincident with changes in marital status. For yet many others with frequent names, the queries are more involved, and make use of CV information such as scientific keywords, institutional affiliation, frequent coauthors' names, etc. This degree of labor-intensive customization ensures that a scientist's bibliome excludes publications belonging to homonymous scientists.

Appendix III
Estimation Procedure for the Semiparametric DD Estimates

The ATE, ATT, and DD effects stem from panel specifications; the sample size is equal to the total number of independent career years for each scientist($N \times T = 8,767$ or $N \times T = 3,832$, depending on the control group). The procedure followed to estimate the SDD effects is slightly different. We first regress the various measures of output on calendar year and age indicator variables using the full panel, and compute the residuals ε_{it}. In a second step, we sum the residuals corresponding to the pre-appointment (1986-1994) and post-appointment (1998-2006) periods separately for each scientist. In the final step, the SDD effects are obtained by regressing $\sum_{t=1986}^{1994} \varepsilon_{it} - \sum_{t=1998}^{2006} \varepsilon_{it}$ on treatment status, weighting these differences as described in section 3. Note that the sample size corresponds in this case to the number of scientists ($N = 466$ or $N = 165$, depending on the control group), not the number of scientist-year observations.