

Incentives and Creativity: Evidence from the Academic Life Sciences

Pierre Azoulay
MIT and NBER
Sloan School of Management
50 Memorial Drive, E52-555
Cambridge, MA 02142

Joshua Graff Zivin
Columbia University and NBER
Mailman School of Public Health
600 West 168th Street, Room 608
New York, NY 10032

Gustavo Manso
MIT
Sloan School of Management
50 Memorial Drive, E52-446
Cambridge, MA 02142

May 19, 2008

Abstract

This paper tests the hypothesis that freedom to experiment, tolerance for early failure, long time horizons to evaluate results, and detailed feedback on performance stimulate creativity and innovation in scientific research. We do so by studying the careers and output of U.S. academic life scientists funded through two very distinct mechanisms: investigator-initiated R01 grants from the NIH, or appointment as an investigator of the Howard Hughes Medical Institute (HHMI), whose funding practices embody many of the elements mentioned above. Using careful adjustment for selection on observables, we find that HHMI investigators produce high-impact papers at a much higher rate than two control groups of similarly-accomplished NIH-funded scientists. In contrast, the program does not appear to have much effect on the raw number of articles published. We also observe large effects on the probability of being elected to prestigious scientific societies or the training of students that go on to win early career prizes.

Keywords: creativity, economics of science, incentives.

*preliminary and incomplete. Send correspondence to pazoulay@mit.edu. We gratefully acknowledge the financial support of the Kauffman Foundation. We thank Thomas Cech and Scott Stern for useful comments, and David Clayton, Sherry White, and Terry Wood at HHMI for facilitating access to funding data. The usual disclaimer applies.

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 (2007), however, incentives may play an important role in the production of radical innovation. In a model that incorporates the trade-off between the exploitation of well-known approaches and the exploration of new untested approaches, he shows that optimal incentive schemes for exploration depart in systematic ways from standard pay-for-performance contracts. In particular, schemes that motivate exploration exhibit substantial tolerance for early failure, and reward for long-term success.¹

The goal of our study is to test the predictions of this theory in the field. 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 do so by studying the careers and output of U.S. academic life scientists who can be funded through two very distinct mechanisms: investigator-initiated R01 grants from the NIH, or support from the Howard Hughes Medical Institute (HHMI) through its investigator program. HHMI, a non-profit medical research organization, plays a powerful role in advancing biomedical research and science education in the United States. The Institute commits almost \$700 million a year — a larger amount than the NSF biological sciences program, for example. HHMI's stated goal is to “push the boundaries of knowledge” in some of the most important areas of biological research. To do so, the HHMI program has adopted practices that according to Manso (2007) should provide strong incentives for breakthrough scientific discoveries: the grant cycles are long (from five to seven 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

¹In the same spirit, Garner (1979) provides a model in which scientists refrain from bold conjecture because objective criteria (publications, citations, etc.) enables them to signal at relatively low cost their competence (in the sense of being able to perform “normal science”).

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. Most 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. Two 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 selection, 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 Pew and Searle scholarships. The second control group is composed of scientists whose NIH grants received the coveted MERIT designation, which is only available for scientists whose grant application was scored in the top two or three percentiles by NIH peer reviewers. Furthermore, using the propensity score, 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 program.

Second, we use a variety of metrics 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. As well, we examine the effect of the program on future appointment to elite societies, such as the National Academy of Sciences and the Institute of Medicine. Finally, we contrast control and treat-

ment scientists by focusing on the quality of their predoctoral and postdoctoral trainees, using receipt of an early-career prize as a measure of quality.

Our results provide support for the hypothesis that appropriately designed incentives stimulate radical innovation. 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 Pew and Searle scholars, the average treatment effect on number of publications ranges from 10% to 25%, with the lower bound imprecisely estimated. The average treatment effect on the number of publications in the top percentile, in contrast, ranges from 60% to 110%. The main message is qualitatively similar when we compare HHMI researchers and MERIT awardees, though the magnitudes of the effects are smaller. Similarly, 5% of our controls are elected to the NAS, versus 30% of our HHMI scientists. During our period of observation, our control scientists train on average .15 students who go on to win an early-career prize. Among HHMI scientists, the average is .55.

However, we find no 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. We also look at evidence that HHMI scientists change the direction of their research efforts after appointment into the program by examining the extent of overlap between the keywords in their articles published after, versus before appointment. The evidence is mixed, with more overlap relative to Pew and Searle scholars, and less overlap relative to MERIT awardees.

In the absence of random assignment or data on the runner-ups of the HHMI competitions, our conclusions must remain guarded, since we rely mostly on observables to distinguish the treatment effects of the program from its pure selection effects. Beyond this major caveat, we face additional hurdles in interpreting our results through the lens of an incentive story. First, HHMI appointment could also confer prestige to selected scientists, and this “anointment” effect would be hard to separate from the program’s incentive effects. Second, to the extent that coauthorship exists among control and treated scientists, this will also muddy the effects’s interpretation, since appointment could generate peer effects onto

collaborators. Finally, HHMI- and NIH-funded scientists differ in the amount of resources available to their laboratories, but measuring the effects of the program on productivity rather than output would require an estimate of the elasticity of publication output with respect to funding. At this stage, we know of no credible estimate of this important policy parameter.

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 the results of the analysis, which are discussed in section 5.

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 (2007) 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 reward 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.

This paper presents the first attempt to test Manso's arguments in the field (see Manso and Ederer (2008) for experimental evidence with a similar flavor). A necessary ingredient for this exercise is to identify 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 incremental from radical ideas. We believe that the academic life sciences in the United States provide a well-suited 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.

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 for the NIH dwarfs that available to 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).

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.”*² New appointments are made every three years, 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 (2007). HHMI Investigator are initially appointed for 5 years,³ 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

²See <http://www.hhmi.org/research/investigators/>

³Appointment lengths have varied over the history of the program, more detailed information will be provided in the data section.

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, 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 et al., 2005) 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 (2007) 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.

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 1: Comparison 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 focus on the 1994 and 1997 cohorts of HHMI investigators, which include 131 researchers (65 in the 1994 cohort; 66 in the 1997 cohort). 94 scientists were appointed following national competitions; 37 were “out-of-competition” appointments. Until the early 1990s, HHMI allocated nomination slots to a small number of elite institutions, which then nominated researchers for HHMI appointment. While HHMI still reserved the right to reject a nominee, in practice the Institute often deferred to the host institutions’ choices. As a result, until a

relatively recent past, HHMI's selection practices were perceived as relatively opaque, with connections to administrators playing a role in shaping the nomination process. Formal competitions and independent evaluations changed the nature of the institute, bringing its practices more closely in line with meritocratic principles. Out-of-competition appointments were progressively phased out. This evolution reached its natural conclusion in 2007, when HHMI announced that in all future competitions, scientists would apply directly to the Institute, thus bypassing the initial nomination step.

It is important to lay out 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.⁴ The nominee must have an MD or a PhD. 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. While this is true of the investigators appointed after the 1994 and 1997 competitions, the out-of-competition nominees include more senior scientists as well as a few scientists appointed immediately upon the completion of their postdoctoral fellowship.

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 more manageable number of about 100 scientists, 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.

⁴There were 154 nominating institutions in the 1994 competition, and 226 in the 1997 competition.

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. In the words of a senior administrator of the program: “*We look for our own discontinuity.*”⁵

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. In our sample, each of these categories account for about a third of the total number of scientists in the treatment group. Of course, such variation raises the specter that appointment length might be endogenous. Senior HHMI administrators have assured us that it is not the case, and we verify this empirically below.

Six out of 131 appointed investigators withdrew from the program voluntarily.⁶ They do so for two reasons. Some accept a top administrative position such as a deanship; HHMI rules prevent investigators to hold major administrative posts. Others desire to move to an institution that has no relationship with HHMI, such as the Scripps Research Institute in La Jolla, CA. Still others wish to move to a different institution during their first appointment. To prevent the eruption of bidding wars over HHMI investigators, the Institute forces these investigators to resign their appointment.

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

⁵HHMI administrators recognize that at the margin, their appointment decisions may deviate from what a pure numerical ranking would imply in one respect: the Institute wants to maintain a cadre of physician-scientists who do “translational research,” i.e., physicians whose research attempts to evaluate the clinical implications of insights gleaned from laboratory research.

⁶In addition, one investigator passed away.

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: Pew and Searle Scholars on the one hand; MERIT awardees from the NIH on the other hand.

Pew and Searle Scholars. The Pew and Searle Scholarships are two early-career prizes that target scientists in the same life science subfields and similar research institutions as HHMI. Every year, the Pew and Searle charitable trusts provide seed funding to 40 life scientists in the first two years of their careers as independent investigators, the period during which they are building a laboratory. The Pew and Searle scholarships are among the most prestigious accolades that young researchers can receive at the start of their careers, but they differ from HHMI investigatorships in one essential respect: they are structured as *one-time grants* (\$60,000 a year over 4 years for the Pew Scholarship; \$80,000 a year for 3 years for the Searle Scholarship). These amounts are relatively small, roughly corresponding to 35% of a typical NIH R01 grant. As a result, Pew and Searle 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 277 scientists awarded a Pew or Searle scholarship between 1985 and 1998.

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, outstanding productivity during their previous research endeavors and are leaders in their field with paradigm-shifting ideas.*” The specific details vary across the component institutes of the NIH, but the essential feature of the program is that only researchers holding an R01 grant in its second or later cycle are eligible. Further, the application for renewal must be scored in the top percentile in a given funding cycle. While the MERIT designation is a prestigious award for mid-career investigators, it pertains to a particular project, not to the scientists’s

overall portfolio of projects. To construct the MERIT control group, we start from the set of all scientists whose R01 grants receive the R37 designation between 1993 and 1999. We eliminate from this set 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 134 scientists meeting these criteria in the final sample, 10 of whom are also Pew and Searle scholars.

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 Pew/Searle 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 Pew and Searle 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 curtails researchers' ability to branch out in novel directions. As well, it 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* is even more skewed, and this leads us to measure creative output by summing 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 for each year.⁷ 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).

⁷We thank Stefan Wuchty and Ben Jones from Northwestern University for performing these computations.

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.⁸ 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.⁹

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 or Searle scholarship early in their independent career.¹⁰

Conflating the creative process with creative outcomes in this way raises two objections. First, it is possible that exploration might also fatten the left-hand tail of the outcome distribution: 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 address the first objection, 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).¹¹ To address the second objection, we calculate the degree of overlap in the MeSH keywords corresponding to articles published after selection into the program, relative to before selection.¹² Our conjecture is

⁸The Top 5 life scientists according to this metric are Solomon Snyder, a neuroscientist at Johns Hopkins; Walter Willett and Meir Stampfer, two epidemiologists at the Harvard School of Public Health; Robert Lefkowitz, a molecular biologist at Duke University and an HHMI investigator; and Michael Karin, a pharmacologist at UCSD.

⁹We exclude review articles, editorials, and letters from the set when computing these measures. We also eliminate articles with more than 20 authors. However, we are unable to tell purely methodological contributions apart from more substantive ones. The former are said to receive a disproportionate amount of citations.

¹⁰We do not focus on the award of Prizes, such as the Nobel Prize, the Lasker Award, and the E.B. Wilson Medal (awarded by the American Society for Cell Biology), because there are simply too few of them to produce meaningful variation. We ignore approved pharmaceutical products with which a scientist may be associated for the same reason. Finally, we do not take advantage of patent data because many patents applied for during the period of this study have not been granted yet; many more still have not had the time to garner citations that would enable us to distinguish “blockbuster” patents from worthless ones.

¹¹Too few investigators exit science altogether to make exit a useful indicator of failure.

¹²MeSH 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. At the most general level of the hierarchical structure are very broad headings such as “Anatomy” or “Mental

that explorative behavior should result in a lower degree of keyword overlap for a given scientist.

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. We also collect patent data from the USPTO database. 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 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.¹³

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

Disorders.” More specific headings are found at more narrow levels of the eleven-level hierarchy, such as “Ankle” and “Conduct Disorder.” There are 24,767 descriptors in the 2008 MeSH.

¹³This measure fails to capture industrial sources of funding or start-up funds provided by the employing institution, if any.

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 “translational” research.

HHMI and control samples at baseline. Table 4 presents baseline descriptive statistics. Panel A pertains to the 1994 cohort; Panel B to the 1997 cohort. Focusing on Panel A, approximately 37% of the HHMI sample is female, versus 22% of the Pew/Searle sample, and 19.4% of the MERIT sample. HHMI investigators are of the same career age on average as the Pew/Searle scholars, but are significantly younger than the MERIT awardees. They are much better funded than Pew/Searle scholars at baseline (\$1.45 million vs. \$0.83 million on average), but much less well funded than MERIT awardees (\$1.45 million vs. \$2.93 million on average). In terms of raw publication output, the pattern is very similar, with HHMI investigators lagging significantly MERIT awardees, but leading Pew/Searle 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 patterns in Panel B (for the 1997 cohort) are very similar.

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 for the 1994 cohort (plots for the 1997 cohort paint a very similar picture). 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 Pew/Searle 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 Pew/Searle 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 2005). In Table 5 we see that HHMI scientists publish approximately 7 more career papers relative to control scientists, with this output of higher quality by all measures. Their work generates more than 9,000 citations on average, and many of their contributions belong in the right-hand tail of the citation distribution. We also compute the h index due to Hirsch (2005): h is the highest integer such that an individual has at least h publications cited at least h times. In the HHMI sample, h is approximately 42, relative to 35 among the controls. HHMI scientists also hold about twice as many career patents than control scientists, on average. Of course, these accomplishments should be viewed in light of their tremendous funding advantage: HHMI scientists receive approximately \$12 million in funding, versus \$5.5 million on average for the controls.

When we focus on discrete career accolades (Table 6), we observe an even greater contrast between HHMI and control scientists. Approximately 29% of the HHMI investigators are elected members of the National Academy of Sciences; 15% are elected members of the Institute of Medicine. This is in contrast to 4% and 3%, respectively, for the control sample. Our 131 HHMI investigators train 72 future Pew or Searle scholars. This is only one less than the number produced by 477 control scientists.

Since averages can mask a great deal of heterogeneity, we also plot histograms for the distributions of various metrics of scientific excellence. Figures 5 through 8 reveal a greater mass in the right tail of the distribution for HHMI scientists, relative to controls, in terms

of high-impact publications (top ventile, top percentile), overall funding, and cumulative number of citations, respectively. This is even more pronounced if one focuses solely on the “after” period, as in Figures 9A through 9D. These plot the distribution of HHMI and control scientists’ “hit rate,” i.e., the proportion of their articles published that fall in the top percentile of the article-level distribution of citations. In each case, we observe that many more HHMI investigators than controls belong to the right-hand tail of the distribution.

Finally, in Table 7, we present a descriptive assessment of the program’s aggregate outcomes. From the funder’s standpoint, the program could be valuable even if most agents fail to create real breakthroughs, as long as a few succeed in producing sufficiently valuable ideas. Both HHMI and control scientists become more successful in the post-selection period, but these improvements appear to be more dramatic for the HHMI investigators. Their magnitudes are also significantly larger for the 1994 cohort. Since our data end in 2005, it is certainly possible that an additional 3 years of follow-up data would have brought the results of the 1997 cohort in line with those of the 1994 cohort. Lastly, it is immediately apparent that, *per dollar of funding*, HHMI investigators do not exhibit greater output than control scientists. Of course, one should not interpret this result as providing evidence that the program is not cost-effective: if the additional exploration efforts induced were enough to produce even a single idea on the scale of, for instance, recombinant DNA technology, then the welfare benefits of the program might well dwarf the additional level of resources provided to HHMI investigators in the aggregate.

3 Econometric Considerations

In order to estimate the treatment effect of the HHMI investigator program, we must confront a basic identification problem: appointment might be mostly 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. The standard approach for this type of problem is instrumental variable estimation.

Yet, the credibility of IV estimates hinges on the validity of the associated exclusion restriction(s). At this juncture, the elite of academic life science is not a setting that provides many (or in fact *any*) sources of exogenous variation in the probability of selection. In what follows, we will simply assume that a good instrument is not available.

A second approach 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, it is well-known that for difference-in-differences 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. In this respect, estimating the causal effect of HHMI appointment on research output presents similar challenges to that of estimating the effect of a job training program on wages. In the job training example, treated individuals have lower earnings on average (relative to their pre-treatment average) in the year immediately preceding enrollment into the program; therefore, the fixed effects estimator is likely to overestimate the treatment effect. Conversely, we will show that 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.

As an attempt to overcome this challenge, we estimate the effects of the program using Inverse Probability of Treatment Weighted (IPTW) estimation (Robins et al., 2000; Hernán et al., 2001). This technique is closely related to propensity-score matching (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 2002). Both techniques make the (untestable) assumption that selection into the program is based on variables that are observable to the econometrician, but IPTW extends the idea of matching on observables to the case of time-varying treatments. In particular, IPTW estimation allows one to recover average treatment effects even in the presence of time-varying confounders, i.e., variables that predict selection into treatment and are correlated with future values of the dependent variable.

Inverse Probability of Treatment Weighted (IPTW) estimation: A Primer. Consider a program in which treatment decisions are made in $T + 1$ distinct periods $0, 1, \dots, T$. At each time t , for each individual i , an outcome of interest y_{it} is measured, and a discrete treatment $TREAT_{it} \in \{0, 1\}$ can begin. We will model initiation of treatment as an absorbing state, i.e., once begun, treatment is never interrupted. In practice, this means that we will ignore the termination decision when evaluating the program, focusing purely on the effect of the “intent to treat.” Also measured at each point in time are a set of exogenous covariates X_{it} and time-varying confounders Z_{it} . (X, Z) are prognostic factors for treatment. For any variable W , denote \widetilde{W}_{it} its history up to time t .

Let y_{it}^a be the value of y that would have been observed at time t had i chosen treatment sequence $\widetilde{a}_{it} = (a_{i0}, a_{i1}, \dots, a_{it})$ rather than his observed treatment history \widetilde{TREAT}_{it} .

By definition, the average treatment effect of treatment history \widetilde{a} on the outcome y is the difference $E[y^a] - E[y^0]$, the average difference between outcomes when following \widetilde{a} and outcomes when never treated. We model the mean of y^a conditional on treatment and exogenous covariates X as:

$$E[y_{it}^a | \widetilde{TREAT}_{it}, X_{it}] = \beta_0 + \beta_1' X_{it} + \beta_2 TREAT_{it} \quad (1)$$

where $TREAT_{it}$ is a regime indicator variables that switches from 0 to 1 upon treatment initiation. Following Robins (1999), we introduce the Sequential Conditional Independence Assumption (SCIA), which provides a formal way to extend the assumption of selection on observables to the case of dynamic treatments:

$$y_{it}^a \amalg TREAT_{it} | \widetilde{TREAT}_{i,t-1}, \widetilde{Z}_{i,t-1}, \widetilde{X}_{it}$$

for all i and t , where the \amalg sign denotes statistical independence. Robins (1999) shows that under SCIA, the average treatment effect β_2 is identified and can be recovered by estimating

$$y_{it} = \beta_0 + \beta_1' X_{it} + \beta_2 TREAT_{it} + \varepsilon_{it} \quad (2)$$

by weighted least squares, where the weights correspond to the inverse probability of following actual treatment history \widetilde{TREAT}_{it} for individual i . Note that (2) differs from (1) in that the observed outcomes y have been substituted for the counterfactual outcomes y^a .

Implementing IPTW estimation is relatively straightforward. Under SCIA, the selection bias can be removed by weighting the regression by:

$$w_{it} = \frac{1}{\prod_{\tau=0}^t \text{Prob}(TREAT_{i\tau} | \widetilde{TREAT}_{i,\tau-1}, \widetilde{Z}_{i,\tau-1}, \widetilde{X}_{i\tau})}$$

Each factor in the denominator is the probability that the researcher received her own observed treatment at time τ , conditional on her past treatment history and on her past history of “prognosis factors” for treatment, whether time-varying or fixed over time. Therefore, the denominator of w_{it} represents the conditional probability that an individual followed his or her own history of treatment up to time t . Suppose that all relevant time-varying confounders are observed and included in Z . Then, weighting by w_{it} effectively creates a pseudo-population in which Z no longer predicts selection into treatment and the causal association between treatment and outcome is the same as in the original population. We refer to $\widehat{\beta}_2$ when eqn. (2) is weighted by w_{it} as the Inverse Probability of Treatment Weighted (IPTW) estimator of β_2 .¹⁴

Estimation of the weights. Estimating the weights is particularly straightforward in the case examined here, for two reasons. First, appointment is a discrete regime change: $HHMI_{it}$ is uniformly equal to zero in the years preceding selection, and uniformly equal to one after selection. Second, a “before” and “after” period can be unambiguously defined for both control and treatment scientists, since all scientists are only at risk of appointment when an HHMI competition takes place (in practice, we will evaluate each competition separately). As a result, only the cross-section corresponding to the year in which the competition takes place is necessary to estimate the weights, and we can drop the index t from $sw_{it} \equiv sw_i$. We estimate a logit model for

$$\text{Prob}(HHMI_i = 1) = \alpha_0 + \alpha_1 Z_{i,baseline} + \alpha_2 X_i \tag{3}$$

¹⁴One might worry about performing statistical inference using “second stage” IPTW estimates, since the weights that are used as input in the outcome equation are themselves estimated. In contrast to two-step selection correction methods, Wooldridge (2002) has shown that the standard errors obtained in this case are conservative.

where the vector X includes exogenous characteristics of the scientists (such as career age at baseline, gender, degree, and laboratory type, etc.) and the vector Z includes two time varying confounders: NIH grantee status at baseline, and number of senior-authored, high-impact publications at baseline. Our estimate of sw_i is $\frac{1}{1-\hat{p}_i}$ if scientist i is a control, and $\frac{1}{\hat{p}_i}$ if scientist i is selected by HHMI as an investigator, where \hat{p}_i is the predicted probability of appointment implied by (3).

Sequential Conditional Independence: An Econometric Free Lunch? IPTW estimation 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 the IPTW estimates to be a useful component of our overall approach. 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. In particular, 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 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 an 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.¹⁵ We measure leadership skills with an indicator variable for principal

¹⁵A 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,

investigatorship on at least one R01 grant at baseline. Of course, our selection equation also includes important demographic characteristics, such as gender, laboratory type, degree, and career age.

Despite these efforts, the conclusion that our list of proxies for the determinants of selection is exhaustive, i.e., that there is no residual selection on unobservables, is premature without further evidence. 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. Unobserved heterogeneity of this type is likely to bias upwards IPTW estimates. As a result, we contrast results of cross-sectional IPTW specifications with those of specifications that include scientist fixed effects. Since HHMI scientists exhibit higher output in the years immediately preceding selection (relative to their pre-appointment average), fixed effects models likely *underestimate* the effect of HHMI appointment on scientific achievement. In combination, however, these two techniques implicitly define a confidence interval, with the fixed-effects estimate providing a lower bound, and the IPTW estimate providing an upper bound. 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 8 through 10 pertain solely to HHMI investigators, and validate empirically some of the purported distinctive features of the program. Table 11 presents evidence on the determinants of HHMI appointment. Finally, Tables 12 through 14 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 8 reports results of OLS regressions in which the dependent variable is the log of the number of days elapsed between initial appointment and the first renewal decision, for the 127 HHMI investigators who did not withdraw from the

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.

program voluntarily. The results demonstrate that variation in the length of initial HHMI appointment is determined by career age, rather than by accomplishments at the time of the appointment. This is true whether we measure accomplishments with the raw number of publications at baseline (column 1); the JIF-weighted number of publications at baseline (column 2); the total number of cites as of 2007 to all original research articles written up to the baseline year (column 3); or the number of publications in the top ventile and top percentile of the citation distribution (columns 4 and 5). Other important demographic characteristics, such as lab type, gender, degree, and whether the appointment was out-of-competition, also have no bearing on the length of the initial appointment. As a result, the rest of the paper will maintain the assumption that appointment length is exogenous, conditional on career age.

Table 9 confirms that the unconditional probability of reappointment is high: 84% of investigators are reappointed at the end of the first appointment; the figure is only 77% at the end of the second appointment. The latter figure is also less precisely estimated, since only a small number of HHMI investigators appointed following the 1997 competition had come up for second review at the time we collected the data. 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 10 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. This difference is all the more remarkable since, as previously explained, only 70 investigators have come up for second review in our data, as opposed to 127 for first review. Regrettably, the evidence does not enable us to isolate any observable factor that explains the decision to not renew an HHMI’s investigator at the end of his/her first term.

From these results, we conclude that the HHMI program conforms both in its *stated* and *actual* practices with the features that Manso (2007) predicts should encourage exploration.

Determinants of selection. We now turn to the observable determinants of selection into the HHMI program (Table 11). We break down the analysis by cohort (1994 and 1997), and by control group (Pew/Searle scholars and MERIT awardees). For each combination of cohort and control group, we present the results for logit specifications that include only demographic characteristics (columns 1a, 1c, 2a, 2c) and for specifications that add to these demographic controls time-varying confounders (columns 1b, 1d, 2b, 2d). Among the demographic characteristics, the only consistent pattern is the higher appointment probability of female scientists in the 1994 cohort, regardless of control group. Consistent with the qualitative evidence on the selection process, we find that the number of “hit papers” published as last author at baseline is highly predictive of appointment (except in the case of the 1997 cohort when MERIT awardees form the control group). We find inconsistent evidence regarding the effect of holding an R01 grant at baseline. There is no effect with the Pew/Searle control group, and a negative effect with the MERIT control group. The latter result simply reflects the fact that MERIT awardees must, by construction, already receive NIH funds in order to receive the R37 designation.

In the results presented below, we weight the observations corresponding to each scientist by his/her inverse probability of appointment as implied by the results of Table 11, columns 1b, 1d, 2b, and 2d.

An econometric side-note. The number of articles published, the number of keywords overlapping between articles published in two different periods, or NIH grants awarded are examples of count dependent variables — non-negative integers with many zeros and ones. For example, 8.25% of the scientist/year observations in the data correspond to years with no publication output; the figure climbs to 37.08% if one focuses on the count of publications in the top ventile of the citation distribution; and to 66.83% if one focuses on the count of publications in the top percentile. Following a long-standing tradition in the study of scientific and technical change, we present quasi-maximum likelihood estimates of the outcome

equations of interest. 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 (Gouriéroux et al., 1984). Further, “robust” standard errors are consistent even if the underlying data generating process is not Poisson. In fact QML Poisson estimation can be used with any non-negative dependent variables, whether of a count or continuous nature (Wooldridge, 1996; Santos Silva and Tenreyro, 2006), as long as the variance/covariance matrix is computed using the outer product of the gradient vector.¹⁶

Effects of HHMI appointment. Our main goal is to estimate the effect of HHMI appointment on scientific excellence — the outputs of the creative process. But taken literally, the Manso (2007) 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. As a result, we begin our investigation of the program’s effects by asking whether HHMI appointment increases the number of publication keywords that overlap between the set of articles published in the “before period” (i.e., the interval $[year_{bln} - 5; year_{bln}]$), and the “after” period (i.e., the interval $[year_{bln} + 3; 2005]$). We reason that if scientists alter the risk profile and time horizon of the projects they begin following appointment, this could translate in articles with somewhat different research foci, regardless of their ultimate scientific impact. We allow for a three-year “wash-out” period following the baseline year because we expect publications immediately following appointment will report the results of projects that began before the scientists were eligible for HHMI appointment.

For each combination of cohort/control group, we report both the results of a “naïve” specification in which the extent of keyword overlap between HHMI investigators and control scientists is assessed (columns 1a, 1c, 2a, 2c); and the results of a specification which incorporates inverse probability of treatment weights (columns 1b, 1d, 2b, 2d). In each model, we include an offset for the number of papers published in the “before” period, so that what

¹⁶In other words, inference on the coefficient estimates presented below do not make use of the Poisson variance assumption. We assume only that the conditional mean of our dependent variables can be written $E[y|X] = \exp(X'\beta)$, a requirement we consider fairly innocuous, given the non-negative nature of the outcomes considered here.

is being analyzed is a normalized rate of keyword overlap. All estimates are presented in the form of incidence rate ratios; the formula $(e^\beta - 1) \times 100\%$ (where β denotes an estimated coefficient) provides a number directly interpretable in terms of elasticity. Model (1b), for instance, implies that the rate of keyword overlap is $(1.114 - 1.000) \times 100 = 11.4\%$ higher for HHMI investigators, relative to Pew/Searle scholars in the 1994 cohort. In contrast, the rate of is overlap is $(1.000 - 0.907) \times 100 = 9.93\%$ lower for HHMI investigators, relative to MERIT awardees (column 2b). At this juncture, we do not have an explanation for these contradictory results.

In Tables 13A and 13B, we pool cohorts and contrast cross-sectional estimates of the program’s effect with estimates that include scientist fixed-effects, using the conditional fixed effect poisson model of Hausman et al. (1984). As noted earlier, these estimates are likely to understate the causal effect of HHMI appointment. Table 13A compares HHMI investigators with Pew/Searle scholars. We find that the fixed-effects estimates are always much smaller than the corresponding cross-sectional estimates. Relative to Pew/Searle scholars, HHMI investigators increase the raw number of publications by 11.5% after appointment, the number of publications in the top quartile by 11.5%, and the number of publications in the top ventile and top percentile by about 15% (statistically insignificant in the latter case). The magnitudes in Table 13B are smaller, with a relative increase of about of 10% in the number of publications in the top quartile (statistically significant only at the 10% level), and statistically insignificant increases of 6% and 3% for the number of publications in the top ventile and top percentile, respectively. In other words, the lower bounds for the average treatment effect provided by the fixed-effects estimates are not always informative.

Our final set of tables (Table 14, panels A through D) present inverse probability of treatment-weighted (IPTW) estimates, each time contrasting them with the naïve pooled cross-sectional estimates. In each case, inference is based on robust (QML) standard errors, clustered at the scientist level. Under the sequential conditional independence assumption, these estimates correspond to the average treatment effect of the program on the outcome of interest. Panels A and B use Pew and Searle scholars as a control group, for the 1994

and 1997 cohort respectively. Panels C and D perform a similar exercise but use MERIT awardees as the control group.

We discuss the results displayed in Panels A and C in detail. The substantive conclusions drawn from the results for the 1997 cohort are consistent with those for the 1994 cohort, but with smaller magnitudes.¹⁷ We observe a positive and statistically significant 26.4% increase for the raw number of articles published in Panel A. The corresponding number is an insignificant 11.9% in Panel B. As a whole, the results in these panels yield three important and relatively robust conclusions: (a) regardless of the outcome measured, the estimated effect of the HHMI investigator program is in most cases positive and statistically significant; (b) the magnitude of the effect is lower once we account for selection into the program based on observable characteristics; and (c) the program has a bigger effect on the upper tail of the distribution of accomplishments. These results strongly support our hypothesis that the HHMI program, and the incentives for exploration incentives it embodies, leads to a higher rate of path-breaking innovation. Recall that MERIT awardees benefit from longer time horizons, before they must show evidence of achievements, relative to other NIH grantees. The fact that even compared with this population, one can still find an effect of HHMI appointment implies that the program’s other features (rich and detailed feedback; “people, not projects”) are important as well.

However, we find no support for another implication of the model: that exploration incentives also result in higher rates of failure, relative to the effect of “exploitation incentives” embodied by the NIH grant funding mechanism. We measure failure by the rate of publications in the bottom quartile of the citation distribution. In this case, the treatment effects we measure fail to reach statistical significance, are relatively small in magnitude, and negative in sign, contrary to our expectations.

¹⁷Of course, there is only a relatively short period to evaluate the effects of the program in that case, which may explain some of the discrepancies we observe.

5 Discussion and Conclusion

We examine the creative outputs of two groups of scientists whose research was funded under mechanisms with rather distinct incentive features: HHMI investigatorship on the one hand; NIH funding on the other hand. We find that HHMI appointments are associated with higher rates of output, especially when we constrain our outcome measures to lie far in the right-hand tail of the achievement distribution. Because the HHMI program is comparatively tolerant of early failure, provides rigorous feedback to researchers, and lets them reallocate resources quickly from failed to promising avenues of enquiry, our results provide support for the main comparative statics found in Manso (2007).

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. It is obviously better to include a large set of potential confounders to model the probability of selection, but we recognize that the sequential conditional independence assumption may still not be precisely or even approximately tenable. We are currently examining the feasibility of using alternative methodologies, including using the nomination stage of the appointment process as a source of exogenous variation in the probability of selection.

Even if our estimates of HHMI appointments' treatment effects can be given a causal interpretation, ascribing them to the program's distinctive *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. In a recent paper, Jacob and Lefgren (2007) estimate that the elasticity of citations to 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. To our

knowledge, there exists no credible estimate of the parameters of an idea production function that would enable us to distinguish the effects of the program on scientific productivity, rather than output.

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). Of course, the magnitude of the program's "anointment effects" is an empirical question, best answered in a separate study. Third, collaboration between scientists in the treatment and control groups might threaten the validity of the comparisons drawn in the analysis. This will be the case if there are peer effects from coauthorship (Azoulay et al., 2008). Fourth, 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 82% (107) of those in our sample have at least some NIH funding at the end of 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 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 were 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

- Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. 2005. Academic Freedom, Private Sector Focus, and the Process of Innovation. NBER Working Paper #11542.
- Amabile, Teresa M. 1996. Creativity in Context. Boulder, CO: Westview Press.
- Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2008. Superstar Extinction. Working Paper, Massachusetts Institute of Technology.
- Burt, Ronald S. 2004. Structural Holes and Good Ideas. *American Journal of Sociology* 110(2): 349–99.
- Cappechi, Mario R. 2008. Response. *Science* 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, pp. 151-61.
- Ederer, Florian and Gustavo Manso. 2008. “Are Financial Incentives Detrimental to Innovation?” Working Paper, MIT Sloan School of Management.
- Garner, C. Alan. 1979. Academic Publication, Market Signaling, and Scientific Research Decisions. *Economic Inquiry* 17(4): 575-84.
- Gouriéroux, Christian, Montfort, Alain, and Alain Trognon. 1984. “Pseudo Maximum Likelihood Methods: Applications to Poisson Models.” *Econometrica*, 52:3, pp. 701-720.
- Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. “Econometric Models for Count Data with an Application to the Patents-R&D Relationship.” *Econometrica*, 52:4, pp. 909-38.
- Hirsch, Jules E. 2005. “An Index to Quantify an Individual’s Scientific Research Output.” *Proceedings of the National Academy of Sciences* 102: 16569–16572.
- Hernán, Miguel A., Babette Brumback, and James M. Robins. 2001. “Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments.” *Journal of the American Statistical Association*, 96:454, pp. 440-48.
- 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.
- Lazear, Edward. 2000. “Performance Pay and Productivity.” *American Economic Review* 90(5): 1346-1361.

- Lotka, Alfred J. 1926. The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences* 16(12): 317-323.
- Manso, Gustavo. 2007. "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* 159:56-63.
- Robins, James M. 1999. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." pp. 95-134 in Statistical Models in Epidemiology: The Environment and Clinical Trials. M.E. Halloran and D. Berry, eds. Springer-Verlag: New York.
- Robins, James M., Miguel A. Hernán, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology*, 11:5, pp. 550-60.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70:1, pp. 41-55.
- Santos Silva, J.M.C. and Silvana Tenreyro. 2006. "The Log of Gravity." *The Review of Economics and Statistics*, 88:4, pp. 641-658.
- Sawyer, R. Keith. 2006. Explaining Creativity. New York: Oxford University Press.
- 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, pp. 305-353.
- Weitzman, Martin. 1998. Recombinant Growth. *Quarterly 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, pp. 117-139.

Tables & Figures

Figure 1: Measuring the Tail of the Distribution of Citations

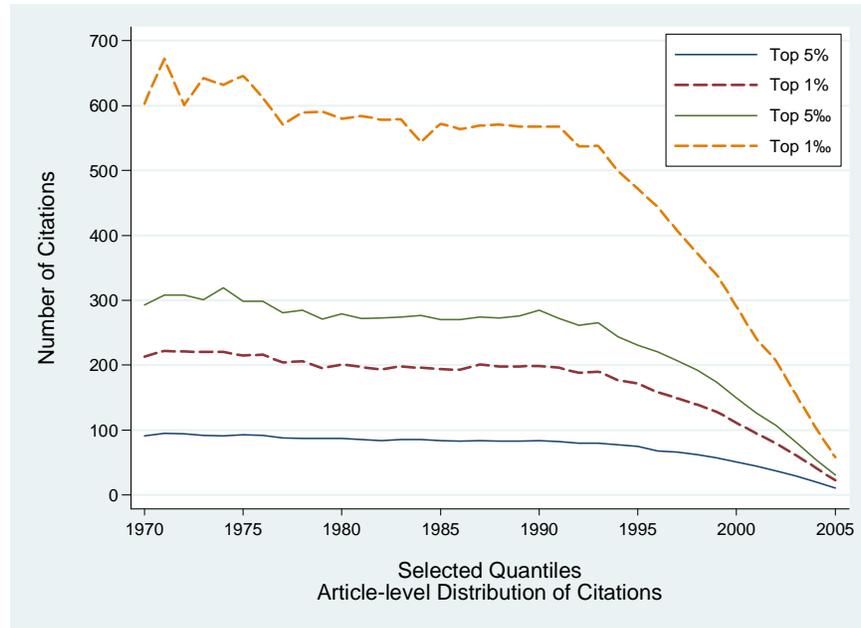
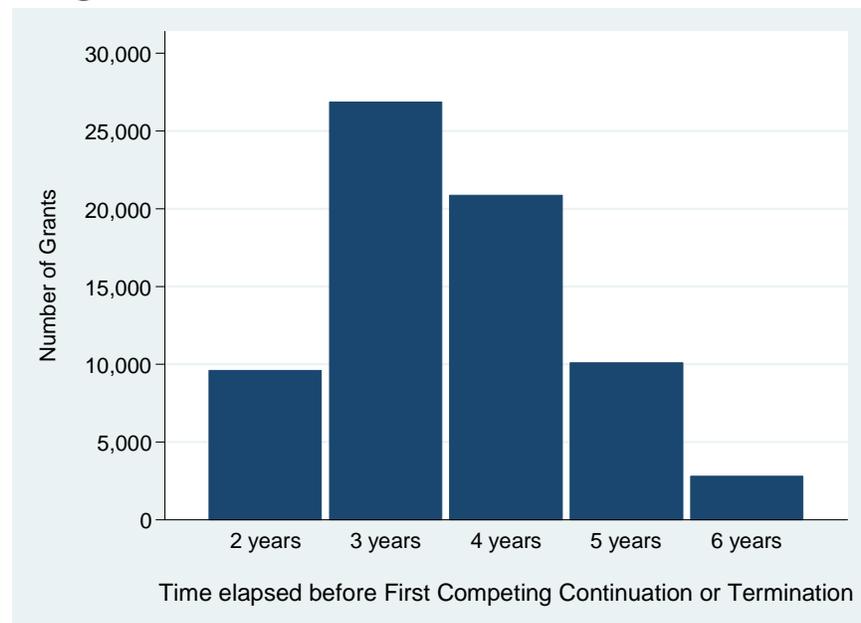
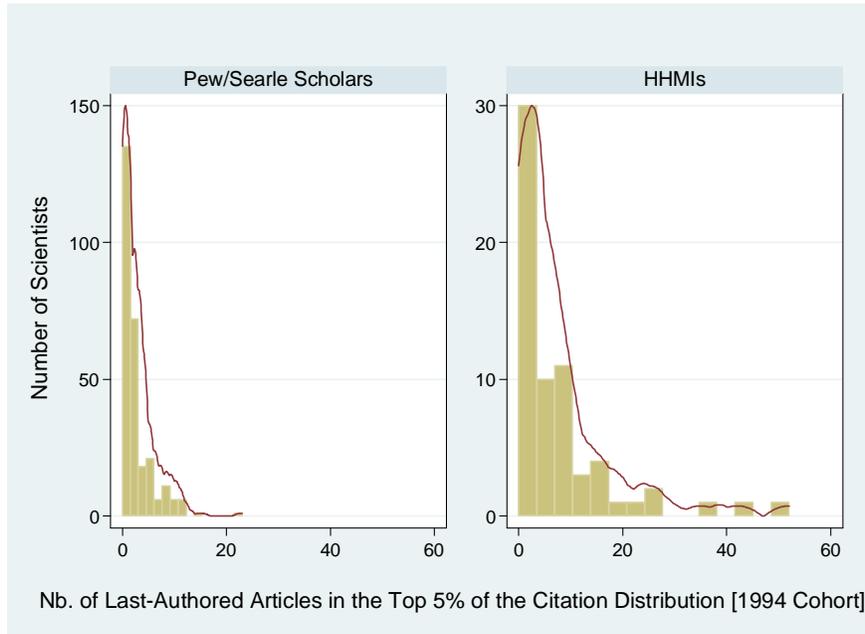


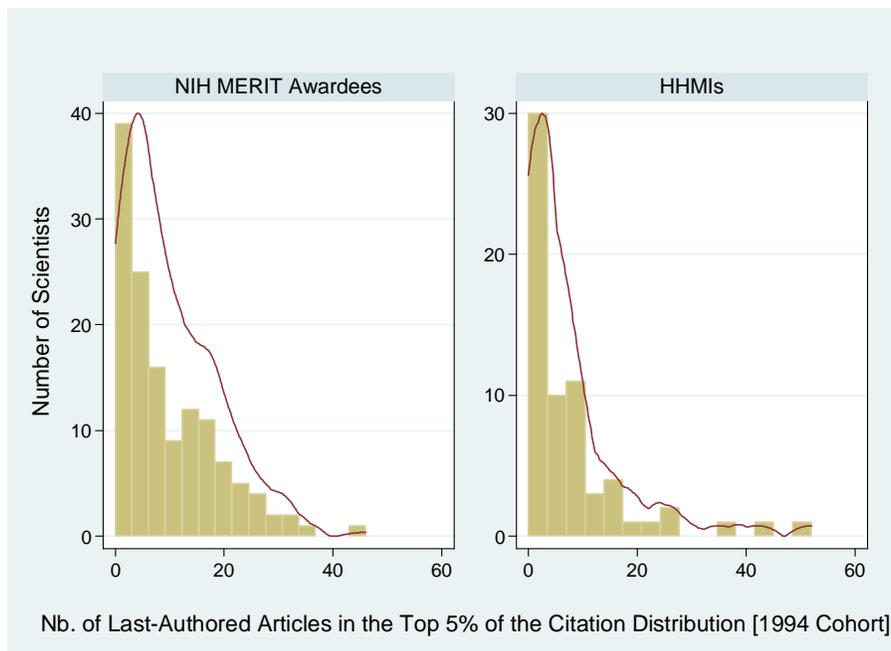
Figure 2: Length of NIH R01 Grants



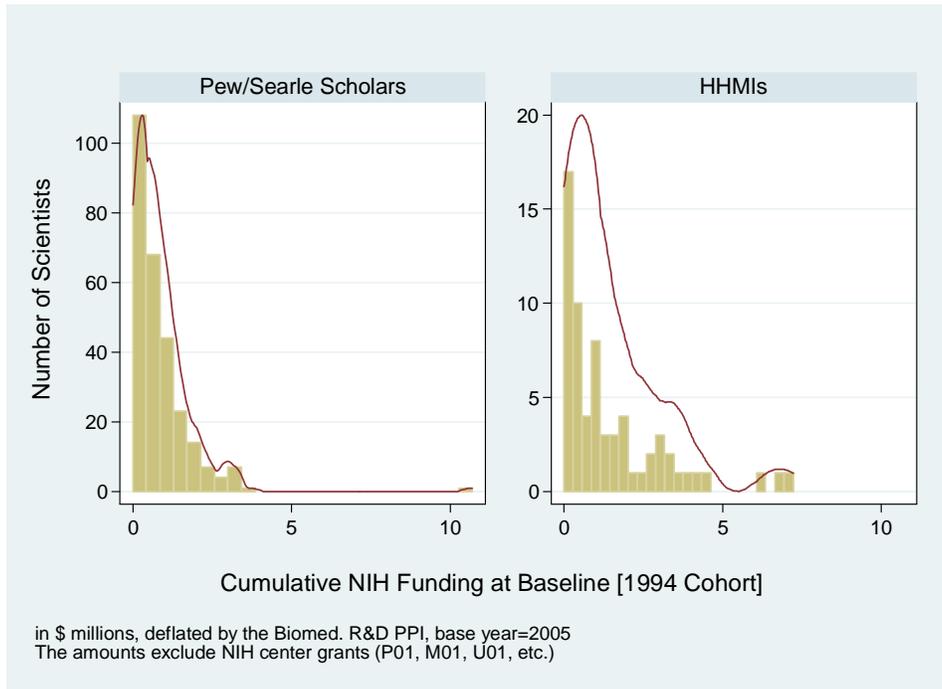
**Figure 3A: Baseline Number of “Hits” as PI — 1994 Cohort
Pew/Searle Scholars vs. HHMI**



**Figure 3B: Baseline Number of “Hits” as PI— 1994 Cohort
MERIT Awardees vs. HHMI**



**Figure 4A: Baseline NIH Funding — 1994 Cohort
Pew/Searle Scholars vs. HHMI**



**Figure 4B: Baseline NIH Funding — 1994 Cohort
NIH MERIT Awardees vs. HHMI**

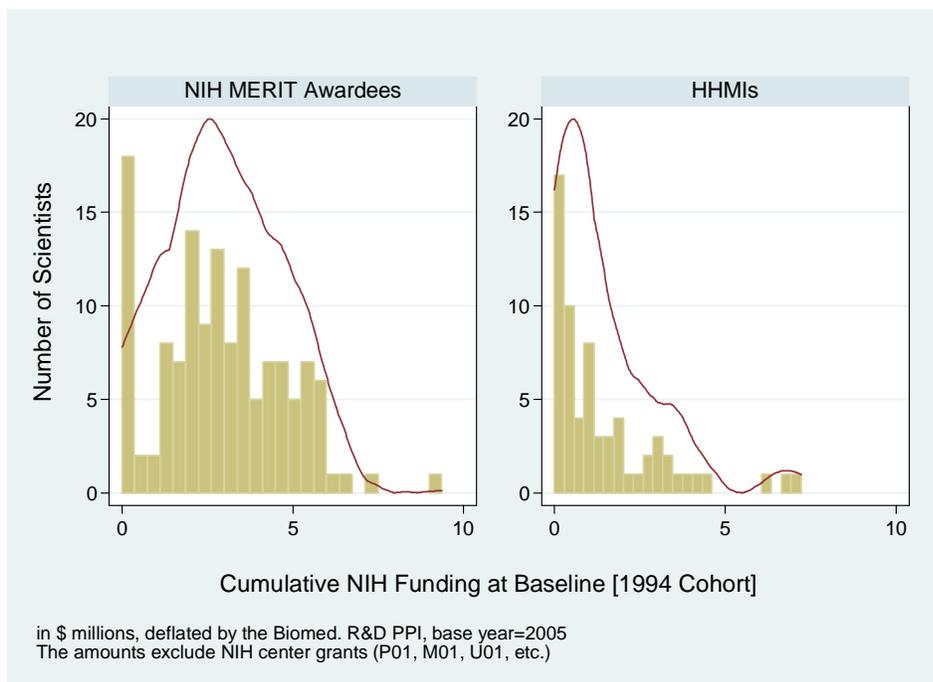


Figure 5: Pubs. in the Top Ventile of the Distribution of Citations

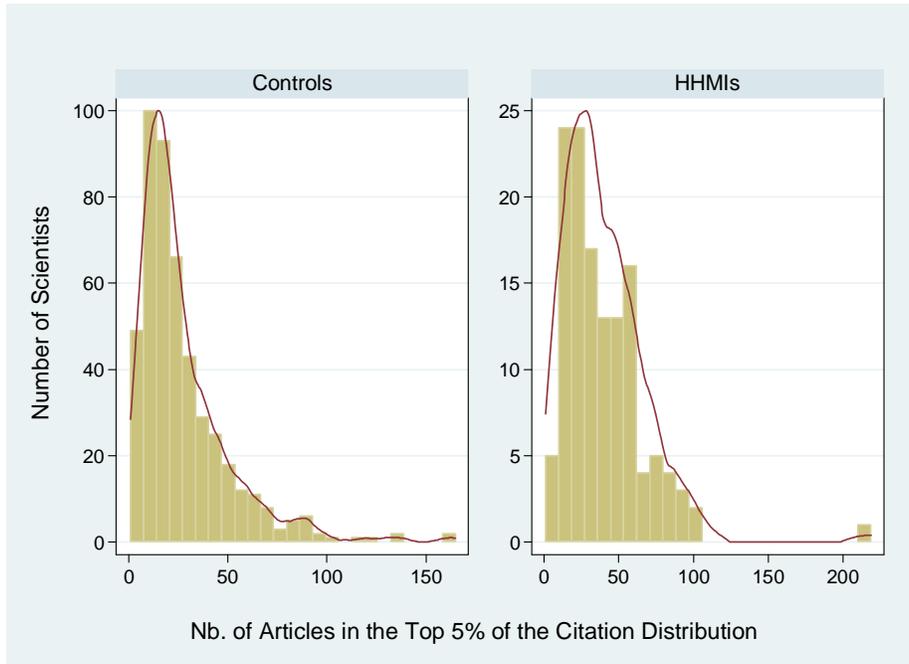


Figure 6: Pubs. in the Top Percentile of the Distribution of Citations

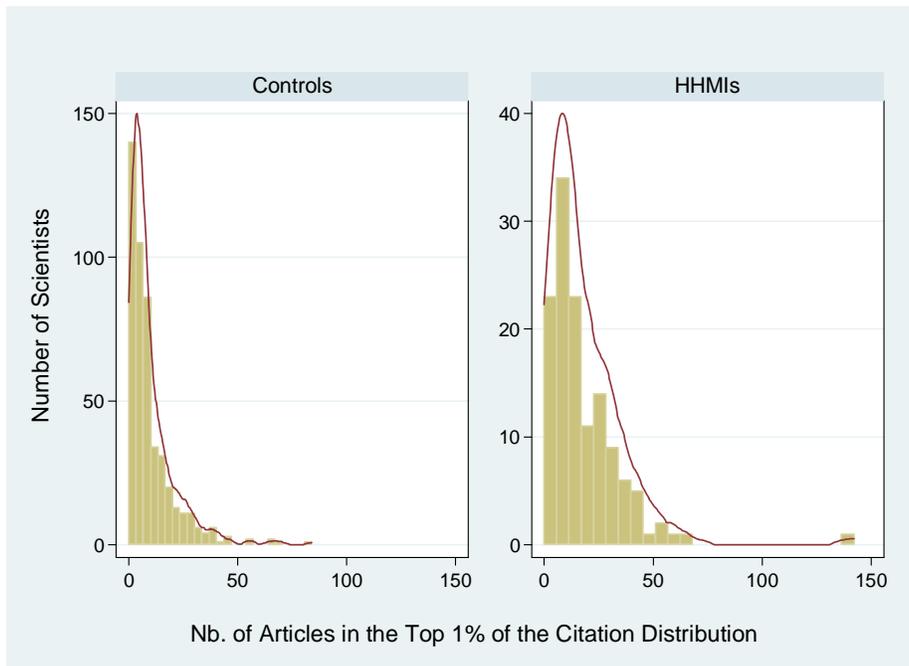


Figure 7: Career Funding, All Sources

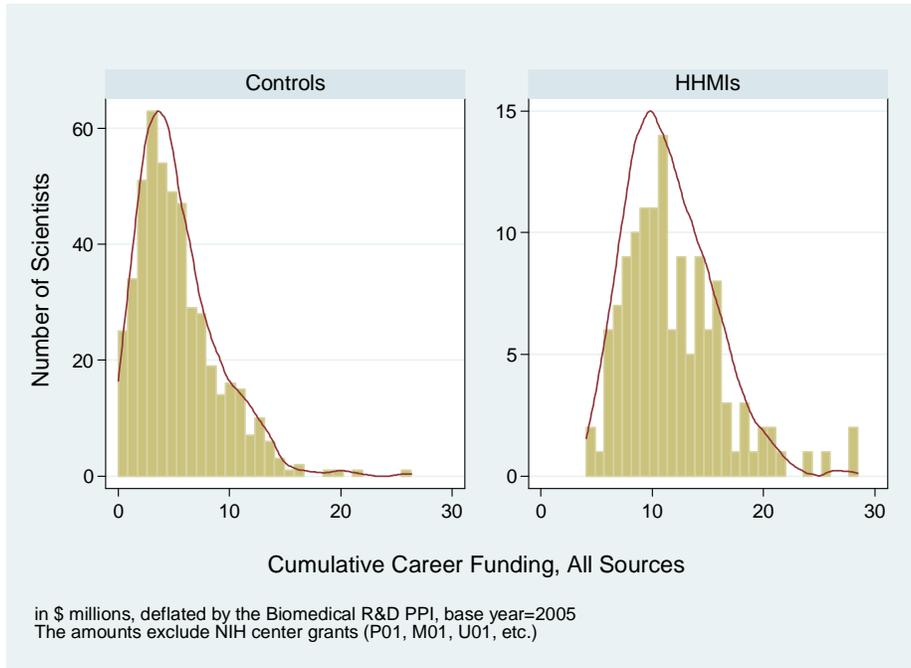
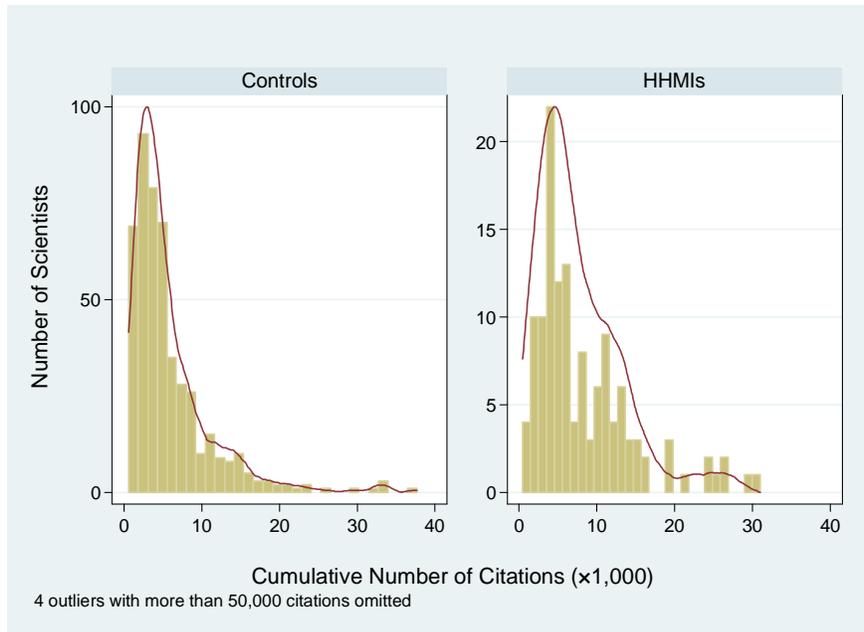
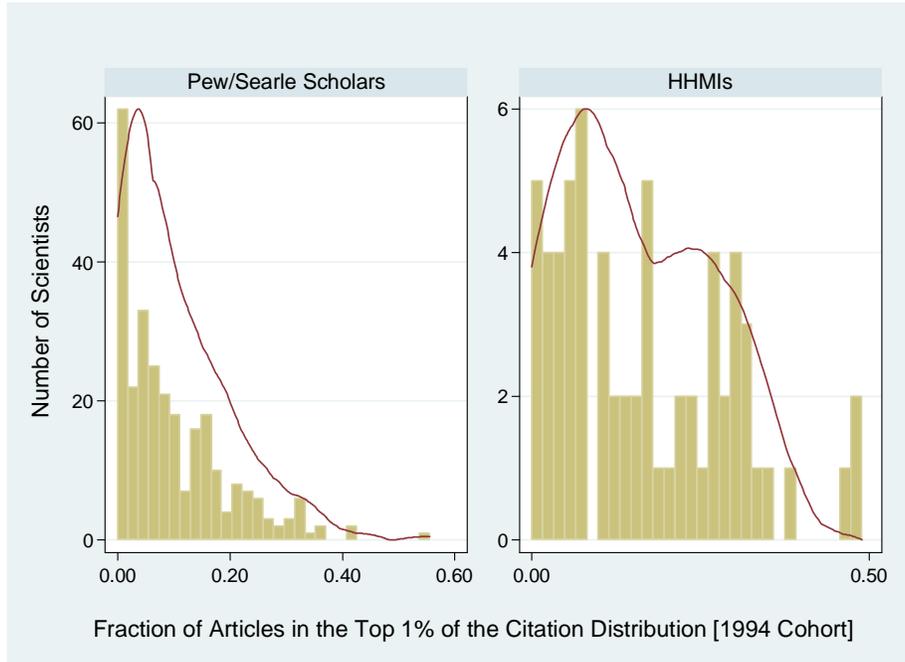


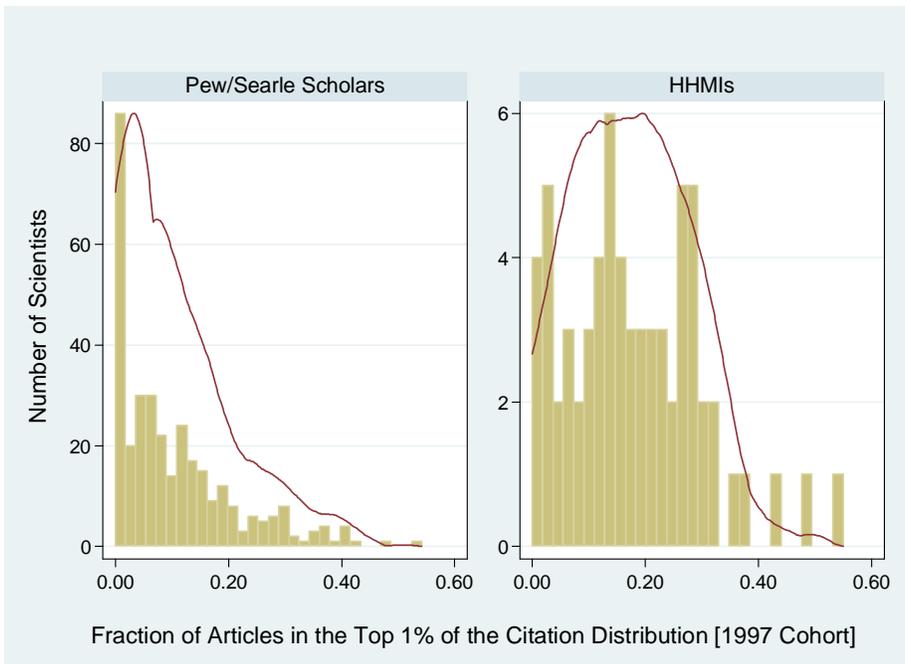
Figure 8: Career Number of Citations



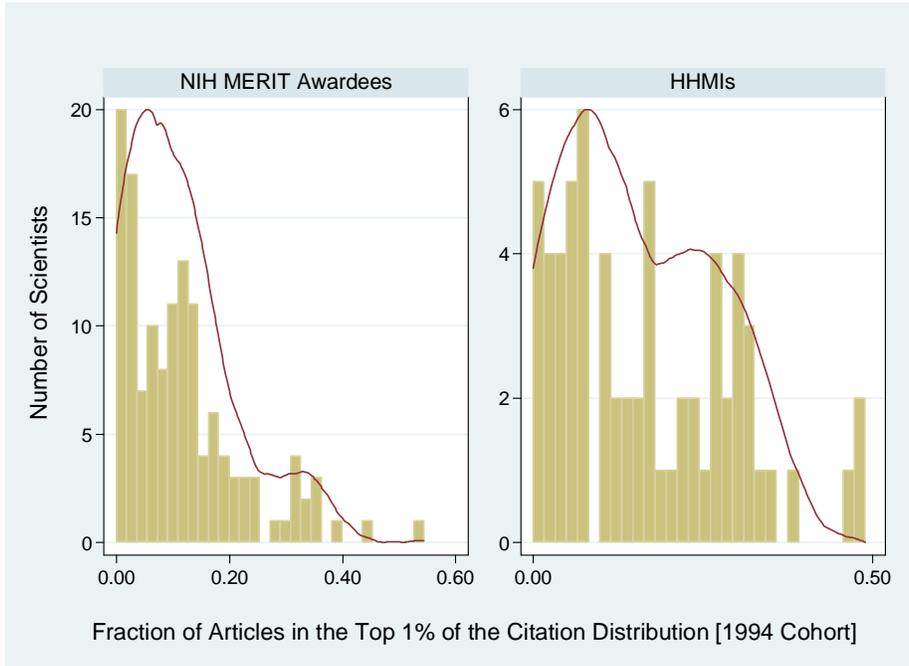
**Figure 9A: Treatment Period “Hit Rate” — 1994 Cohort
Pew/Searle Scholars vs. HHMI**



**Figure 9B: Treatment Period “Hit Rate”— 1997 Cohort
Pew/Searle Scholars vs. HHMI**



**Figure 9C: Treatment Period “Hit Rate” — 1994 Cohort
NIH MERIT Awardees vs. HHMI**



**Figure 9D: Treatment Period “Hit Rate”— 1997 Cohort
NIH MERIT Awardees vs. HHMI**

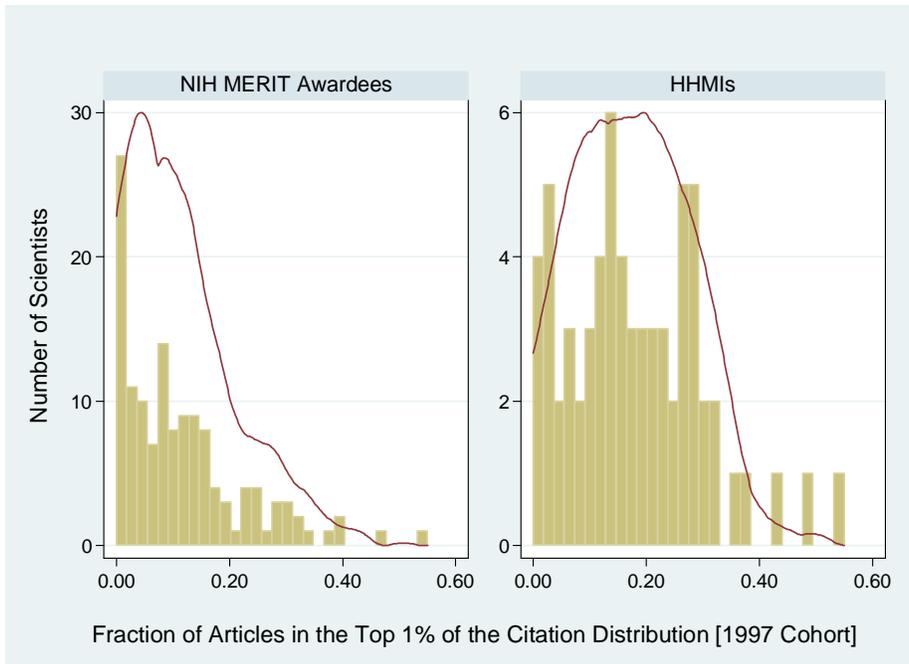


Table 2: Demographic Characteristics

		Female	MD	PhD	MD/PhD	US Born
Controls	477	104 (21.80%)	50 (10.50%)	362 (75.90%)	65 (13.60%)	380 (79.70%)
HHMIs	131	40 (30.50%)	12 (9.20%)	94 (71.80%)	25 (19.10%)	108 (82.40%)
Total	608	144 (23.70%)	62 (10.20%)	456 (75.00%)	90 (14.80%)	488 (80.30%)

Table 3: Lab Type

		Macromolecular	Cellular	Organismal	Translational
Controls	477	149 (31.20%)	114 (23.90%)	155 (32.50%)	59 (12.40%)
HHMIs	131	46 (35.10%)	22 (16.80%)	42 (32.10%)	21 (16.00%)
Total	608	195 (32.10%)	136 (22.40%)	197 (32.40%)	80 (13.20%)

Table 4A: Descriptive Statistics at Baseline — 1994 Cohort

	Mean	Std. Dev	Min.	Max.
Pew/Searle Scholars (n=277)				
Degree Year	1984.285	3.39	1974	1991
Female	0.22	0.415	0	1
Cum. NIH Funding (excl. P01, etc.)	\$827,135	\$963,611	\$0	\$10,698,334
Cum. Nb. of Pubs.	22.278	17.563	2	200
Cum. Nb. of Pubs. in the Top 5%	8.924	6.28	0	42
Cum. Nb. of Pubs. in the Top 1%	3.412	3.174	0	16
Nb. Keywords Overlap before/after	42.516	22.724	2	154
NIH MERIT Awardees (n=134)				
Degree Year	1978.112	2.988	1974	1988
Female	0.194	0.397	0	1
Cum. NIH Funding (excl. P01, etc.)	\$2,930,903	\$1,844,661	\$0	\$9,366,733
Cum. Nb. of Pubs.	60.806	33.448	14	227
Cum. Nb. of Pubs. in the Top 5%	20.336	15.408	0	89
Cum. Nb. of Pubs. in the Top 1%	7.015	7.71	0	41
Nb. Keywords Overlap before/after	83.007	37.025	21	221
HHMIs (n=65)				
Degree Year	1983.723	4.002	1974	1991
Female	0.369	0.486	0	1
Cum. NIH Funding (excl. P01, etc.)	\$1,455,499	\$1,684,935	\$0	\$7,220,260
Cum. Nb. of Pubs.	30.169	26.117	1	172
Cum. Nb. of Pubs. in the Top 5%	15.862	16.834	0	119
Cum. Nb. of Pubs. in the Top 1%	8.092	10.464	0	73
Nb. Keywords Overlap before/after	60.292	35.264	3	216

Table 4B: Descriptive Statistics at Baseline — 1997 Cohort

	Mean	Std. Dev	Min.	Max.
Pew/Searle Scholars (n=333)				
Degree Year	1985.393	4.019	1974	1994
Female	0.234	0.424	0	1
Cum. NIH Funding (excl. P01, etc.)	\$1,225,889	\$1,228,677	\$0	\$12,826,429
Cum. Nb. of Pubs.	28.502	22.265	4	250
Cum. Nb. of Pubs. in the Top 5%	11.132	7.847	0	56
Cum. Nb. of Pubs. in the Top 1%	4.219	3.932	0	23
Nb. Keywords Overlap before/after	43.742	25.151	0	164
NIH MERIT Awardees (n=134)				
Degree Year	1978.112	2.988	1974	1988
Female	0.194	0.397	0	1
Cum. NIH Funding (excl. P01, etc.)	\$4,189,175	\$2,337,379	\$0	\$11,931,537
Cum. Nb. of Pubs.	79.53	43.161	25	317
Cum. Nb. of Pubs. in the Top 5%	27.284	20.539	0	129
Cum. Nb. of Pubs. in the Top 1%	9.455	9.888	0	50
Nb. Keywords Overlap before/after	84.164	39.758	31	225
HHMIs (n=66)				
Degree Year	1984.439	4.811	1974	1994
Female	0.242	0.432	0	1
Cum. NIH Funding (excl. P01, etc.)	\$1,905,030	\$1,670,714	0	\$7,341,554
Cum. Nb. of Pubs.	40.076	25.93	6	141
Cum. Nb. of Pubs. in the Top 5%	19.561	12.054	3	68
Cum. Nb. of Pubs. in the Top 1%	8.652	6.583	0	24
Nb. Keywords Overlap before/after	64.409	28.154	14	147

Table 5: Descriptive Statistics — Career Achievement

	Mean	Std. Dev	Min.	Max.
Controls (n=477)				
Career NIH Funding	\$6,448,440	\$6,393,926	\$0	\$52,104,756
Career NIH Funding (excl. P01, etc.)	\$5,116,235	\$3,739,049	\$0	\$26,333,260
Career HHMI Funding	\$0	\$0	\$0	\$0
Career Total Funding	\$5,351,768	\$3,680,351	\$0	\$26,333,260
Cum. Nb. of Cites	5,853	5,332	626	37,672
h Index	34.954	15.917	8	104
Cum. Nb. of Pubs	76.472	56.722	11	450
Cum. Nb. of Pubs in the Bottom 25%	2.002	2.475	0	18
Cum. Nb. of Pubs. in the Top 25%	57.48	42.582	5	339
Cum. Nb. of Pubs. in the Top 5%	27.528	23.385	1	165
Cum. Nb. of Pubs. in the Top 1%	9.897	11.001	0	84
Cum. Nb. of Pubs. in the Top 5%	5.998	7.552	0	59
Cum. Nb. of Pubs. in the Top 1%	1.543	2.695	0	21
Cum. Nb. of Patents	1.507	3.661	0	36
HHMIs (n=131)				
Career NIH Funding	\$4,774,888	\$4,883,070	\$0	\$32,025,670
Career NIH Funding (excl. P01, etc.)	\$3,626,550	\$2,691,153	\$0	\$14,973,205
Career HHMI Funding	\$8,223,618	\$2,625,364	\$2,410,360	\$17,205,788
Career Total Funding	\$11,957,585	\$4,566,230	\$4,048,120	\$28,469,912
Cum. Nb. of Cites	9,125	11,415	473	117,316
h Index	41.878	17.77	5	141
Cum. Nb. of Pubs	83.351	46.764	7	313
Cum. Nb. of Pubs in the Bottom 25%	1.481	1.516	0	7
Cum. Nb. of Pubs. in the Top 25%	68.45	40.043	2	277
Cum. Nb. of Pubs. in the Top 5%	40.015	27.868	1	219
Cum. Nb. of Pubs. in the Top 1%	18.092	17.669	0	142
Cum. Nb. of Pubs. in the Top 5%	11.87	13.841	0	119
Cum. Nb. of Pubs. in the Top 1%	3.74	6.528	0	60
Cum. Nb. of Patents	3.366	10.51	0	103
Total (n=608)				
Career NIH Funding	\$6,087,855	\$6,135,159	\$0	\$52,104,756
Career NIH Funding (excl. P01, etc.)	\$4,795,267	\$3,590,277	\$0	\$26,333,260
Career HHMI Funding	\$1,771,865	\$3,595,362	\$0	\$17,205,788
Career Total Funding	\$6,775,061	\$4,740,851	\$0	\$28,469,912
Cum. Nb. of Cites	6,558	7,212	473	117,316
h Index	36.446	16.565	5	141
Cum. Nb. of Pubs	77.954	54.766	7	450
Cum. Nb. of Pubs in the Bottom 25%	1.89	2.311	0	18
Cum. Nb. of Pubs. in the Top 25%	59.844	42.257	2	339
Cum. Nb. of Pubs. in the Top 5%	30.219	24.931	1	219
Cum. Nb. of Pubs. in the Top 1%	11.663	13.158	0	142
Cum. Nb. of Pubs. in the Top 5%	7.263	9.57	0	119
Cum. Nb. of Pubs. in the Top 1%	2.016	3.955	0	60
Cum. Nb. of Patents	1.908	5.895	0	103

Table 6: Accolades

		Pew/Searle Scholars Trained	Nobel Prize Winners	Lasker Awardees	Members of the NAS	Members of the IoM
Controls	477	73	1 (0.20%)	1 (0.20%)	21 (4.40%)	16 (3.40%)
HHMIs	131	72	3 (2.30%)	2 (1.50%)	38 (29.00%)	15 (11.50%)
Total	608	145	4 (0.70%)	3 (0.50%)	59 (9.70%)	31 (5.10%)

Table 7A: Aggregate Outcomes — 1994 Cohort

		Total Number of Pubs in Top 5%		Total Number of Pubs in Top 1%		Total Patents		Total Funding	
		Before Selection	After Selection	Before Selection	After Selection	Before Selection	After Selection	Before Selection	After Selection
Pew/Searle (N=246)	Scientist Avg.	5.224	11.747	1.791	3.877	0.332	0.892	\$1,091,028	\$3,282,611
	Total	1,447.00	3,254.00	496	1,074.00	92	247	\$302,214,685	\$909,283,265
MERIT (N=122)	Scientist Avg.	18.284	24.007	6.299	8.478	0.94	1.56	\$2,944,875	\$5,361,533
	Total	2,450.00	3,217.00	844	1,136.00	126	209	\$394,613,198	\$718,445,469
HHMIs (N=65)	Scientist Avg.	12.431	25.877	6.354	11.323	1.692	2.6	\$1,930,957	\$10,685,114
	Total	808	1,682	413	736	110	169	\$125,512,181	\$694,532,396

Table 7B: Aggregate Outcomes — 1997 Cohort

		Total Number of Pubs in Top 5%		Total Number of Pubs in Top 1%		Total Patents		Total Funding	
		Before Selection	After Selection	Before Selection	After Selection	Before Selection	After Selection	Before Selection	After Selection
Pew/Searle (N=302)	Scientist Avg.	7.237	8.892	2.483	2.931	0.646	0.402	\$1,506,266	\$2,564,071
	Total	2,410.00	2,961.00	827	976	215	134	\$501,586,602	\$853,835,644
MERIT (N=122)	Scientist Avg.	25.231	17.06	8.739	6.037	1.716	0.784	\$4,203,147	\$4,103,261
	Total	3,381.00	2,286.00	1,171.00	809	230	105	\$563,221,698	\$549,836,970
HHMIs (N=66)	Scientist Avg.	16.00	18.758	7.015	8.136	1.348	0.955	\$2,461,152	\$8,847,924
	Total	1,056	1,238	463	537	89	63	\$162,436,035	\$583,962,998

Table 8:
Determinants of HHMI First Appointment Length (OLS)

	(1)	(2)	(3)	(4)	(5)
Cum. Nb. of Pubs.	1.001 [0.13]				
Cum. Nb. of Pubs (JIF-weighted)		1.000 [0.69]			
Cum. Nb. of Cites			1.000 [0.87]		
Cum. Nb. of Pubs. in the Top 5%				0.995 [0.54]	
Cum. Nb. of Pubs. in the Top 1%					0.991 [0.64]
Nb. of Years since Start as a PI	1.135** [3.33]	1.156** [3.88]	1.145** [4.75]	1.149** [4.04]	1.145** [4.67]
Female	0.727 [1.30]	0.703 [1.42]	0.708 [1.42]	0.704 [1.41]	0.703 [1.42]
PhD	1.085 [0.15]	1.033 [0.06]	1.046 [0.08]	1.045 [0.08]	1.046 [0.08]
MD/PhD	0.948 [0.10]	0.918 [0.16]	0.913 [0.17]	0.922 [0.16]	0.923 [0.15]
Translational Lab	1.106 [0.18]	1.150 [0.25]	1.165 [0.28]	1.159 [0.26]	1.160 [0.27]
Organismal Lab	1.028 [0.09]	1.021 [0.07]	1.022 [0.07]	1.024 [0.08]	1.029 [0.09]
Macromolecular Lab	1.433 [1.01]	1.471 [1.07]	1.438 [1.03]	1.451 [1.05]	1.444 [1.04]
Out-of-Competition Appointment	0.595 [1.60]	0.592 [1.61]	0.590 [1.61]	0.587 [1.62]	0.587 [1.63]
Nb. Scientists	127	127	127	127	127
Adjusted R ²	0.26	0.26	0.26	0.26	0.26

Note: The dependent variable is the log of the length of the first appointment period, measured in days, among 127 HHMI investigators who did not terminate their appointment voluntarily.

Robust *t*-statistics in parantheses. †significant at 10%; *significant at 5%; **significant at 1%.

Table 9: Univariate Termination Data

	End of First Appointment	End of Second Appointment
Reappointed	106 (83.46%)	54 (77.14%)
Terminated	21 (16.54%)	16 (22.86%)

Note: Only a small proportion of investigators appointed in 1997 had come up for second review at the time of the analysis.

Table 10: Sensitivity of Reappointment to Performance [Logit Models]

	Pubs		JIF-weighted Pubs		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(3c)	(3d)
	First	Second	First	Second	First	Second	First	Second
	Reappt.	Reappt.	Reappt.	Reappt.	Reappt.	Reappt.	Reappt.	Reappt.
Cum. Nb. of Pubs. during relevant period	1.007 [0.42]	0.804** [2.85]						
Cum. Nb. of JIF-weighted Pubs. during relevant period			0.998 [0.96]	0.980* [2.54]				
Cum. Nb. of Pubs. in the Top 5% during relevant Period					0.974 [0.88]	0.816† [1.94]		
Cum. Nb. of Pubs. in the Top 1% during relevant period							0.960 [0.79]	0.645† [1.82]
Female	0.323 [1.64]	0.985 [0.02]	0.302† [1.71]	1.139 [0.13]	0.296† [1.75]	0.664 [0.48]	0.300† [1.73]	0.522 [0.77]
Associate	0.325 [1.57]	0.650 [0.43]	0.329 [1.52]	0.468 [0.73]	0.307 [1.61]	0.756 [0.30]	0.310 [1.60]	0.626 [0.51]
Full	0.351 [1.46]	1.174 [0.08]	0.406 [1.24]	0.813 [0.11]	0.392 [1.30]	0.875 [0.08]	0.390 [1.31]	1.073 [0.04]
Nb. Scientists	127	70	127	70	127	70	127	70
Log Quasi-Likl.	-48.25	-18.73	-47.84	-19.28	-47.92	-23.71	-47.99	-23.61
Pseudo-R ²	0.15	0.50	0.16	0.49	0.16	0.37	0.16	0.37

Note: The dependent variable is the probability of being reappointed, whether at the end of the first term, or at the end of the second term, among 127 HHMI investigators who did not terminate their appointment voluntarily.

Robust z-statistics in parantheses. †significant at 10%; * significant at 5%; ** significant at 1%.

Table 11:
Determinants of Selection into the HHMI Program [Logit Models]

	Pew/Searle Control Group				MERIT Control Group			
	1994 Cohort		1997 Cohort		1994 Cohort		1997 Cohort	
	(1a)	(1b)	(1c)	(1d)	(2a)	(2b)	(2c)	(2d)
Nb. of “Hit Papers” as PI		1.602** [5.07]		1.493** [6.02]		1.183** [3.34]		1.062 [1.59]
NIH Funding		1.081 [0.49]		1.229 [1.04]		0.725† [1.76]		0.765** [2.58]
Female	2.314** [2.74]	3.683** [3.78]	1.281 [0.77]	2.117* [2.09]	3.378* [2.52]	4.623** [3.12]	1.168 [0.27]	1.327 [0.48]
PhD	1.395 [0.42]	1.110 [0.12]	0.467 [1.18]	0.435 [1.25]	0.341† [1.66]	0.419 [1.26]	1.159 [0.26]	1.373 [0.54]
MD/PhD	1.923 [0.97]	1.321 [0.36]	1.237 [0.34]	1.318 [0.40]	0.814 [0.23]	0.829 [0.20]	0.985 [0.03]	1.185 [0.26]
Macromolecular Lab	1.115 [0.28]	1.331 [0.67]	1.845 [1.40]	1.661 [1.07]	2.348 [1.39]	4.339* [2.30]	3.465† [1.95]	4.725* [2.22]
Organismal Lab	0.930 [0.18]	0.993 [0.02]	2.284† [1.91]	2.666* [2.15]	1.413 [0.54]	2.093 [1.13]	1.564 [0.66]	1.770 [0.76]
Translational Lab	1.432 [0.49]	3.003 [1.31]	1.400 [0.53]	1.586 [0.66]	0.536 [0.68]	0.811 [0.22]	2.513 [1.23]	3.019 [1.33]
Nb. Scientists	342	342	399	399	199	199	200	200
Log likelihood	-159.19	-136.85	-169.70	-140.72	-75.97	-70.09	-80.76	-77.65
Pseudo R ²	0.04	0.18	0.05	0.21	0.40	0.44	0.36	0.39

Notes:

- (1) The “cellular lab” indicator variable is omitted. All models include a cubic spline in career age.
(2) Robust z-statistics in parentheses.
(3) †significant at 10%; * significant at 5%; ** significant at 1%.

Table 12:
Effects of HHMI Appointment on Changes in Research Topic [QML Poisson Models]

	Pew/Searle Control Group				MERIT Control Group			
	1994 Cohort		1997 Cohort		1994 Cohort		1997 Cohort	
	(1a)	(1b)	(1c)	(1d)	(2a)	(2b)	(2c)	(2d)
HHMI	1.236** [3.65]	1.114* [1.97]	1.237** [4.34]	1.202** [4.27]	0.902* [2.23]	0.907* [1.98]	0.907* [2.53]	0.904** [2.62]
Female	0.898* [2.23]	0.917† [1.76]	0.923† [1.91]	0.919† [1.75]	0.941 [1.23]	0.946 [1.11]	1.054 [1.19]	1.055 [1.22]
PhD	1.088 [0.81]	1.081 [0.72]	1.058 [0.76]	1.053 [0.61]	0.921 [1.36]	0.911 [1.53]	0.867** [2.69]	0.865** [2.73]
MD/PhD	1.063 [0.63]	1.093 [0.88]	1.072 [0.95]	1.073 [0.85]	0.990 [0.16]	1.005 [0.08]	1.008 [0.15]	1.017 [0.30]
Macromolecular Lab	0.951 [0.99]	0.944 [1.08]	0.969 [0.60]	1.021 [0.33]	0.915† [1.74]	0.907† [1.88]	0.945 [1.11]	0.943 [1.15]
Organismal Lab	0.914 [1.54]	0.871* [2.28]	0.926 [1.39]	0.908† [1.66]	0.862* [2.40]	0.850* [2.55]	0.824** [3.31]	0.816** [3.46]
Translational Lab	1.114 [1.13]	1.121 [1.06]	1.172 [2.11]	1.253** [2.92]	0.990 [0.15]	0.969 [0.45]	1.035 [0.53]	1.024 [0.36]
Nb. Scientists	342	342	399	399	199	199	200	200
Log quasi-likelihood	-1,999.67	-1,964.55	-2,314.24	-2,399.53	-1,213.68	-1,209.01	-1,087.77	-1,084.23

Notes:

- (1) The “cellular lab” indicator variable is omitted. All models include a cubic spline in career age, and an offset for the number of publications in the 1989-1994 period (models 1a, 1b, 2a, and 2b) or 1992-1997 years (models 1c, 1d, 2c, and 2d).
- (2) Robust z-statistics in parentheses.
- (3) †significant at 10%; * significant at 5%; ** significant at 1%.

**Table 13A: Effect of HHMI Appointment on Scientific Output — Pooled Cohorts
Pew/Searle Scholars Control Group [QML Poisson Models]**

	Publications		Pubs. in Bttm. Quartile		Pubs. in Top Quartile		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	X- Section	Fixed Effects	X- Section	Fixed Effects	X- Section	Fixed Effects	X- Section	Fixed Effects	X- Section	Fixed Effects
HHMI	1.495** [6.75]	1.115* [2.07]	0.885 [0.87]	0.768 [1.20]	1.625** [7.80]	1.115* [2.08]	2.019** [9.25]	1.154* [2.05]	2.606** [8.87]	1.146 [1.39]
Female	0.747** [5.38]		0.766† [1.85]		0.734** [5.41]		0.677** [5.49]		0.567** [5.44]	
PhD	0.842 [0.98]		1.245 [0.65]		0.793 [1.39]		0.688* [2.22]		0.598* [2.49]	
MD/PhD	0.987 [0.09]		1.126 [0.44]		0.925 [0.52]		0.789 [1.44]		0.706 [1.61]	
Macromolecular Lab	1.143* [2.05]		1.253 [1.54]		1.109 [1.48]		0.977 [0.26]		0.943 [0.44]	
Organismal Lab	1.031 [0.39]		1.099 [0.51]		1.020 [0.26]		1.002 [0.02]		0.975 [0.19]	
Translational Lab	1.318* [2.01]		2.264** [3.04]		1.208 [1.40]		1.079 [0.51]		1.133 [0.66]	
Nb. Scientists	465	465	465	273	465	465	465	463	465	428
Nb. of Observations	7,656	7,656	7,656	4,749	7,656	7,656	7,656	7,634	7,656	7,048
Log Quasi-Likl.	-18,437	-13,538	-2,114	-1,349	-16,550	-12,310	-12,499	-9,121	-7,520	-5,045

Notes:

- (1) All models control for calendar year indicator variables, and 5 career age indicator variables (career age < 5 is the omitted category).
- (2) The “cellular lab” indicator variable is omitted, as is the indicator variable for investigators holding a MD degree.
- (3) Incidence Rate ratios are displayed, with robust z-statistics in parentheses, clustered around scientists for the cross-sectional models.
- (4) †significant at 10%; * significant at 5%; ** significant at 1%.

**Table 13B: Effect of HHMI Appointment on Scientific Output — Pooled Cohorts
MERIT Awardees Control Group [QML Poisson Models]**

	Publications		Pubs. in Bttm. Quartile		Pubs. in Top Quartile		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	X- Section	Fixed Effects	X- Section	Fixed Effects	X- Section	Fixed Effects	X- Section	Fixed Effects	X- Section	Fixed Effects
HHMI	0.911 [1.33]	1.123* [2.29]	0.631** [2.90]	0.911 [0.42]	0.991 [0.12]	1.097† [1.70]	1.179 [1.62]	1.060 [0.78]	1.492* [2.48]	1.032 [0.28]
Female	0.838* [2.30]		1.233 [1.46]		0.806** [2.75]		0.709** [3.69]		0.592** [4.07]	
PhD	0.848 [1.51]		0.708* [2.00]		0.833 [1.62]		0.795† [1.68]		0.723† [1.77]	
MD/PhD	1.015 [0.15]		0.868 [0.89]		0.983 [0.16]		0.880 [1.02]		0.777 [1.48]	
Macromolecular Lab	1.040 [0.52]		1.080 [0.48]		1.028 [0.34]		0.936 [0.59]		0.855 [0.94]	
Organismal Lab	0.860† [1.70]		0.931 [0.42]		0.795* [2.50]		0.713** [3.01]		0.600** [3.26]	
Translational Lab	1.128 [1.13]		1.193 [1.07]		1.078 [0.68]		1.029 [0.20]		1.024 [0.12]	
Nb. Scientists	265	265	265	203	265	265	265	264	265	251
Nb. of Observations	5,490	5,490	5,490	4,355	5,490	5,490	5,490	5,478	5,490	5,183
Log Quasi-Likl.	-14,805	-11,277	-1,899	-1,378	-13,447	-10,341	-10,617	-7,939	-6,771	-4,714

Notes:

- (1) All models control for calendar year indicator variables, and 5 career age indicator variables (career age < 5 is the omitted category).
- (2) The “cellular lab” indicator variable is omitted, as is the indicator variable for investigators holding a MD degree.
- (3) Incidence Rate ratios are displayed, with robust z -statistics in parentheses, clustered around scientists for the cross-sectional models.
- (4) †significant at 10%; * significant at 5%; ** significant at 1%.

**Table 14A: Effect of HHMI Appointment on Scientific Output — 1994 Cohort
Pew/Searle Scholars Control Group [QML Poisson Models]**

	Publications		Pubs. in Bttm. Quartile		Pubs. in Top Quartile		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW
HHMI	1.558** [5.49]	1.264* [2.44]	0.991 [0.05]	0.942 [0.32]	1.738** [6.53]	1.386** [3.25]	2.250** [7.84]	1.704** [4.30]	3.047** [7.49]	2.106** [4.32]
Female	0.700** [5.64]	0.722** [4.98]	0.650* [2.54]	0.662* [2.31]	0.685** [5.53]	0.710** [4.98]	0.623** [5.40]	0.658** [4.80]	0.495** [5.40]	0.549** [4.61]
PhD	0.782 [1.14]	0.681 [1.28]	1.423 [0.85]	1.896 [1.41]	0.732 [1.53]	0.629 [1.60]	0.642 [2.18]	0.557* [2.12]	0.582* [2.32]	0.516* [2.35]
MD/PhD	0.922 [0.42]	0.902 [0.42]	1.268 [0.72]	1.430 [1.06]	0.839 [0.90]	0.810 [0.88]	0.673† [1.84]	0.652† [1.76]	0.578* [1.98]	0.567* [2.07]
Macromolecular Lab	1.125 [1.55]	1.101 [1.27]	1.228 [1.22]	1.355† [1.76]	1.088 [1.03]	1.046 [0.55]	0.942 [0.55]	0.872 [1.28]	0.886 [0.77]	0.763† [1.72]
Organismal Lab	1.037 [0.40]	1.021 [0.18]	1.216 [0.93]	1.284 [1.09]	1.016 [0.18]	0.982 [0.17]	0.971 [0.27]	0.898 [0.86]	0.938 [0.40]	0.787 [1.34]
Translational Lab	1.196 [1.02]	1.123 [0.49]	2.469** [2.82]	3.199** [3.40]	1.074 [0.41]	0.987 [0.06]	0.901 [0.52]	0.813 [0.84]	0.896 [0.43]	0.773 [0.90]
Nb. Scientists	343	343	343	343	343	343	343	343	343	343
Nb. of Observations	5,951	5,951	5,951	5,951	5,951	5,951	5,951	5,951	5,951	5,951
Log Quasi-Likl.	-14,346	-14,115	-1,678	-1,643	-12,748	-12,435	-9,369	-9,063	-5,467	-5,184

Notes:

- (1) All models control for calendar year indicator variables, and 5 career age indicator variables (career age < 5 is the omitted category).
- (2) The “cellular lab” indicator variable is omitted, as is the indicator variable for investigators holding a MD degree.
- (3) Incidence Rate ratios are displayed, with robust z-statistics in parentheses, clustered around scientists for the cross-sectional models.
- (4) †significant at 10%; * significant at 5%; ** significant at 1%.

**Table 14B: Effect of HHMI Appointment on Scientific Output — 1997 Cohort
Pew/Searle Scholars Control Group [QML Poisson Models]**

	Publications		Pubs. in Bttm. Quartile		Pubs. in Top Quartile		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW
HHMI	1.470** [5.43]	1.119 [1.08]	0.777 [1.31]	0.941 [0.25]	1.582** [6.30]	1.158 [1.51]	2.005** [7.72]	1.334** [2.65]	2.580** [7.81]	1.622** [3.29]
Female	0.766** [4.39]	0.693** [4.83]	0.692* [2.08]	0.589* [2.42]	0.757** [4.45]	0.696** [4.92]	0.728** [4.03]	0.681** [4.38]	0.661** [3.68]	0.634** [3.74]
PhD	0.883 [0.54]	0.427** [4.75]	1.279 [0.61]	4.302† [1.96]	0.851 [0.73]	0.395** [5.98]	0.782 [1.10]	0.363** [6.88]	0.708 [1.26]	0.350** [6.16]
MD/PhD	1.069 [0.38]	0.692* [2.07]	1.026 [0.09]	1.652 [1.45]	1.020 [0.12]	0.633** [2.74]	0.903 [0.57]	0.561** [3.40]	0.835 [0.79]	0.564** [2.91]
Macromolecular Lab	1.174* [2.25]	1.223* [2.46]	1.203 [1.12]	1.185 [0.99]	1.149† [1.86]	1.186† [1.95]	1.020 [0.21]	1.066 [0.59]	1.024 [0.17]	1.087 [0.54]
Organismal Lab	1.061 [0.67]	1.372** [2.58]	1.158 [0.71]	1.294 [0.75]	1.049 [0.56]	1.312* [2.57]	1.023 [0.23]	1.262* [2.10]	0.994 [0.04]	1.296† [1.71]
Translational Lab	1.365† [1.78]	0.920 [0.44]	2.551** [2.81]	5.892** [2.88]	1.265 [1.35]	0.816 [1.11]	1.169 [0.82]	0.740 [1.59]	1.349 [1.27]	0.838 [0.80]
Nb. Scientists	400	400	400	400	400	400	400	400	400	400
Nb. of Observations	6,487	6,487	6,487	6,487	6,487	6,487	6,487	6,487	6,487	6,487
Log Quasi-Likl.	-15,289	-17,727	-1,791	-1,909	-13,560	-15,338	-9,979	-11,277	-5,733	-6,465

Notes:

- (1) All models control for calendar year indicator variables, and 5 career age indicator variables (career age < 5 is the omitted category).
- (2) The “cellular lab” indicator variable is omitted, as is the indicator variable for investigators holding a MD degree.
- (3) Incidence Rate ratios are displayed, with robust z-statistics in parentheses, clustered around scientists for the cross-sectional models.
- (4) †significant at 10%; * significant at 5%; ** significant at 1%.

**Table 14C: Effect of HHMI Appointment on Scientific Output — 1994 Cohort
MERIT Awardees Scholars Control Group [QML Poisson Models]**

	Publications		Pubs. in Bttm. Quartile		Pubs. in Top Quartile		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW
HHMI	0.917 [0.91]	0.913 [1.01]	0.615** [2.60]	0.606** [2.65]	1.023 [0.22]	1.004 [0.04]	1.267† [1.70]	1.208 [1.51]	1.689* [2.34]	1.516* [2.09]
Female	0.810* [2.30]	0.815* [2.17]	1.183 [1.04]	1.195 [1.09]	0.774** [2.70]	0.775** [2.64]	0.666** [3.55]	0.663** [3.60]	0.531** [3.95]	0.521** [4.09]
PhD	0.847 [1.37]	0.845 [1.37]	0.732 [1.63]	0.740 [1.60]	0.829 [1.47]	0.825 [1.49]	0.805 [1.38]	0.802 [1.39]	0.751 [1.35]	0.762 [1.26]
MD/PhD	1.012 [0.10]	1.011 [0.10]	0.902 [0.60]	0.866 [0.82]	0.971 [0.24]	0.978 [0.19]	0.853 [1.05]	0.885 [0.85]	0.728 [1.54]	0.782 [1.28]
Macromolecular Lab	1.034 [0.38]	1.011 [0.12]	1.086 [0.45]	1.099 [0.53]	1.018 [0.18]	0.989 [0.11]	0.919 [0.62]	0.870 [1.01]	0.823 [0.95]	0.752 [1.36]
Organismal Lab	0.874 [1.33]	0.859 [1.44]	1.018 [0.10]	1.056 [0.29]	0.796* [2.13]	0.773* [2.31]	0.692** [2.79]	0.652** [3.15]	0.583** [2.95]	0.522** [3.46]
Translational Lab	1.107 [0.84]	1.051 [0.41]	1.231 [1.14]	1.276 [1.34]	1.050 [0.38]	0.981 [0.15]	0.986 [0.08]	0.879 [0.81]	0.932 [0.29]	0.785 [1.09]
Nb. Scientists	199	199	199	199	199	199	199	199	199	199
Nb. of Observations	4,358	4,358	4,358	4,358	4,358	4,358	4,358	4,358	4,358	4,358
Log Quasi-Likl.	-12,011	-1,1945	-1,594	-1,571	-10,898	-10,837	-8,490	-8,446	-5,290	-5,285

Notes:

- (1) All models control for calendar year indicator variables, and 5 career age indicator variables (career age < 5 is the omitted category).
- (2) The “cellular lab” indicator variable is omitted, as is the indicator variable for investigators holding a MD degree.
- (3) Incidence Rate ratios are displayed, with robust z-statistics in parentheses, clustered around scientists for the cross-sectional models.
- (4) †significant at 10%; * significant at 5%; ** significant at 1%.

**Table 14D: Effect of HHMI Appointment on Scientific Output — 1997 Cohort
MERIT Awardees Control Group [QML Poisson Models]**

	Publications		Pubs. in Bttm. Quartile		Pubs. in Top Quartile		Pubs. in Top Ventile		Pubs. in Top Percentile	
	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)	(5a)	(5b)
	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW	X-Section	IPTW
HHMI	0.867 [†] [1.88]	0.880 [1.61]	0.611* [2.27]	0.646* [2.04]	0.913 [1.13]	0.920 [1.02]	1.055 [0.51]	1.051 [0.48]	1.247 [1.47]	1.246 [1.46]
Female	0.918 [0.94]	0.917 [0.90]	1.268 [1.42]	1.267 [1.39]	0.889 [1.26]	0.886 [1.25]	0.815 [†] [1.83]	0.812 [†] [1.87]	0.746* [2.00]	0.743* [2.01]
PhD	0.829 [1.60]	0.842 [1.48]	0.696* [2.04]	0.689* [2.06]	0.816 [†] [1.66]	0.831 [1.54]	0.786 [1.60]	0.802 [1.50]	0.711 [†] [1.66]	0.731 [1.55]
MD/PhD	1.074 [0.62]	1.095 [0.78]	0.848 [1.00]	0.855 [0.94]	1.045 [0.38]	1.065 [0.55]	0.944 [0.44]	0.962 [0.31]	0.844 [1.00]	0.870 [0.85]
Macromolecular Lab	1.052 [0.58]	1.043 [0.48]	1.010 [0.05]	0.996 [0.02]	1.055 [0.56]	1.045 [0.46]	0.974 [0.20]	0.968 [0.24]	0.894 [0.56]	0.886 [0.58]
Organismal Lab	0.845 [†] [1.66]	0.836 [†] [1.73]	0.975 [0.14]	0.959 [0.23]	0.764* [2.52]	0.753** [2.60]	0.652** [3.24]	0.639** [3.35]	0.494** [4.06]	0.477** [4.20]
Translational Lab	1.074 [0.60]	1.078 [0.62]	1.228 [1.19]	1.241 [1.24]	1.020 [0.16]	1.021 [0.17]	0.963 [0.25]	0.956 [0.29]	0.940 [0.31]	0.920 [0.41]
Nb. Scientists	200	200	200	200	200	200	200	200	200	200
Nb. of Observations	4,321	4,321	4,321	4,321	4,321	4,321	4,321	4,321	4,321	4,321
Log Quasi-Likl.	-11,694	-11,685	-1,578	-1,568	-10,519	-10,496	-8,166	-8,124	-5,005	-4,955

Notes:

- (1) All models control for calendar year indicator variables, and 5 career age indicator variables (career age < 5 is the omitted category).
- (2) The “cellular lab” indicator variable is omitted, as is the indicator variable for investigators holding a MD degree.
- (3) Incidence Rate ratios are displayed, with robust z-statistics in parentheses, clustered around scientists for the cross-sectional models.
- (4) [†]significant at 10%; * significant at 5%; ** significant at 1%.

Appendix

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

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

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:

$$E[p_2] < p_1 < E[p_2|S, 2]. \tag{4}$$

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 $\langle i_k^j \rangle$ to maximize her total expected payoff

$$\begin{aligned} R(\langle i_k^j \rangle) = & \{E[p_i]S + (1 - E[p_i])F - c_i\} \\ & + E[p_i] \{E[p_j|S, i]S + (1 - E[p_j|S, i])F - c_j\} \\ & + (1 - E[p_i]) \{E[p_k|F, i]S + (1 - E[p_k|F, i])F - c_k\} \end{aligned} \tag{5}$$

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

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

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

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

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

$$\alpha(E[p_2] - p_1)(S - F) \geq (c_2 - c_1). \quad (8)$$

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 $\langle 2_2^2 \rangle$. If the researcher chooses action plan $\langle 1_1^1 \rangle$, her total expected payoff is

$$\begin{aligned} R(\langle 1_1^1 \rangle) = & \{E[p_1]S + (1 - E[p_1])F - c_1\} \\ & + E[p_1] \{E[p_1]S + (1 - E[p_1])F - c_1\} \\ & + (1 - E[p_1]) \{E[p_1]S + (1 - E[p_1])F - c_1\} \end{aligned} \quad (9)$$

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

$$\begin{aligned} R(\langle 2_2^2 \rangle) = & \{E[p_2]S + (1 - E[p_2])F - c_2\} \\ & + E[p_2] \{E[p_2|S, 2]S + (1 - E[p_2|S, 2])F - c_2\} \\ & + (1 - E[p_2]) \{E[p_1]S + (1 - E[p_1])F - c_1\} \end{aligned} \quad (10)$$

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

$$\alpha\{(E[p_2](E[p_2|S, 2] - p_1) + (E[p_2] - p_1))(S - F) \geq (1 + E[p_2])(c_2 - c_1) \quad (11)$$

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 (8) implies (11). For the second statement, we construct the following example. If $c_2 > c_1$, (8) implies that the agent never explores under the “specific project” approach. However, from (11), if the payoff from exploration is sufficiently high, the agent will explore under the “people, not projects” approach. ■