RAND Journal of Economics
Vol. 42, No. 3, Fall 2011
pp. 527–554

Incentives and creativity: evidence from the academic life sciences

Pierre Azoulay*

Joshua S. Graff Zivin**

and

Gustavo Manso***

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

1. Introduction

■ 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

*Massachusetts Institute of Technology and NBER; pazoulay@mit.edu.
**University of California, San Diego and NBER; jgraffzivin@ucsd.edu.
***Massachusetts Institute of Technology; manso@mit.edu.
We gratefully acknowledge the financial support of the Kauffman Foundation and the National Science Foundation through its SciSIP Program (award no. SBE-0738142). We thank Peter Cebon, Thomas Cech, Purnell Choppin, David Clayton, Nico Lacetera, Antoinette Schoar, Scott Stern, Jerry Thursby, Heidi Williams, and two anonymous referees for useful comments and suggestions. The usual disclaimer applies.

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 tradeoff 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.

In the anecdote reported above, the scientist was undeterred by his peers' advice to "play it safe," and eventually saw his bold ideas prevail. If incentives play an important role in the production of novel ideas, this heroic story might be atypical. In this article, we provide empirical evidence that nuanced features of incentive schemes embodied in the design of research contracts exert a profound influence on the subsequent development of breakthrough ideas. The challenge is to find a setting in which (i) radical innovation is a key concern; (ii) agents are at risk of receiving different incentive schemes; and (iii) it is possible to measure innovative output and to distinguish between incremental and radical ideas. We argue that the academic life sciences in the United States provides an excellent testing ground.

Specifically, we study the careers of researchers who can be funded through two very distinct mechanisms: investigator-initiated R01 grants from the NIH, or support from the Howard Hughes Medical Institute (HHMI) through its investigator program. HHMI, a 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 National Science Foundation 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 (2011) should provide strong incentives for breakthrough scientific discoveries: the award cycles are long (five years, and typically renewed at least once); the review process provides detailed, high-quality feedback to the researcher; and the program selects "people, not projects," which allows (and in fact encourages) the quick reallocation of resources to new approaches when the initial ones are not fruitful.[1] This stands in sharp contrast with the incentives faced by life scientists funded by the NIH. The typical R01 grant cycle lasts only three years, and renewal is not very forgiving of failure. Feedback on performance is limited in its depth. Importantly, the NIH funds projects

[1] Though not part of Manso's (2011) initial analysis, we extend his model in Appendix A to show that providing the researcher greater latitude in her search activities encourages exploration.

with clearly defined deliverables, not individual scientists, which could increase the costs of experimentation.

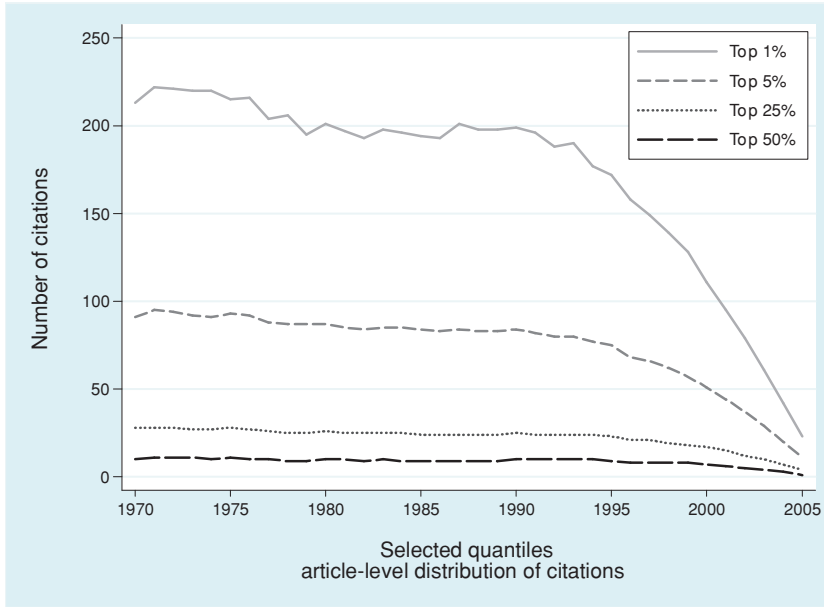
The contrast between the HHMI and NIH grant mechanisms naturally leads to the question of whether HHMI-style incentives result in a higher rate of production of particularly valuable ideas. Three significant hurdles must be overcome to answer this question.

First, we need to identify a group of NIH-funded scientists who are appropriate controls for the researchers selected into the HHMI program. Given the high degree of accomplishment exhibited by HHMI investigators at the time of their appointment, a random sample of scientists of the same age, working in the same fields, would not be appropriate. In the absence of a plausible source of exogenous variation for HHMI appointment, we estimate the treatment effect of the program by contrasting HHMI-funded scientists' output with that of a group of NIH-funded scientists who focus their research on the same subfields of the life sciences as HHMI investigators and received prestigious early career prizes. Furthermore, using an in-depth understanding of the HHMI appointment process, we cull from this control group scientists who look similar to the HHMI investigators on the observable factors that we know to be relevant for selection into the HHMI program.

Second, we must be able to distinguish particularly creative contributions from incremental advances. Although 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 five percentiles, and top percentile. We also use these scientists' *own* citation impact in the pre-appointment period to ask whether they often outperform their most heavily cited article, and conversely, whether they often publish articles that garner less citations than their least-cited article. Another prong in our attempt to measure creativity is to measure explorative behavior directly. Specifically, we examine whether the research agenda of

FIGURE 1

MEASURING THE TAIL OF THE DISTRIBUTION OF CITATIONS



Selected quantiles (0.50, 0.25, 0.05, and 0.01) for the vintage-specific empirical distribution of the number of citations at the article level. These quantiles were computed in early 2008 using the universe of all articles indexed by ISI/Web of Science that appeared in life science journals.

HHMI investigators changes after their appointment; we measure the novelty (both relative to the universe of published research and to the scientists themselves) of the keywords tagging their publications; and we also assess whether the impact of their research broadens, as inferred by the range of journals that cite it.

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

Our results provide support for the hypothesis that appropriately designed incentives stimulate exploration. In particular, we find that the effect of selection into the HHMI program increases as we examine higher quantiles of the distribution of citations. Relative to early career prize winners (ECPWs), our preferred econometric estimates imply that the program increases overall publication output by 39%; the magnitude jumps to 96% when focusing on the number of publications in the top percentile of the citation distribution. Success is also more frequent among HHMI investigators when assessed with respect to scientists' own citation impact prior to appointment, rather than relative to a universal citation benchmark. Symmetrically, we also uncover robust evidence that HHMI-supported scientists "flop" more often than ECPWs: they publish 35% more articles that fail to clear the (vintage-adjusted) citation bar of their least well cited pre-appointment work. This provides suggestive evidence that HHMI investigators are not simply rising stars annointed by the program. Rather, they appear to place more risky scientific bets after their appointment, as theory would suggest.

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

The rest of the article proceeds as follows. In the next section, we present the theoretical motivation for our hypothesis. Section 3 describes the construction of the sample and presents descriptive statistics. Section 4 lays out our econometric methodology. Section 5 reports and discusses the results of the analysis. Section 6 concludes.

2. 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 plague 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 an 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 a recent article, Manso (2011) explicitly models the innovation process as the result of learning through experimentation. In this setting, the tradeoff between the exploitation of well-known approaches and the exploration of new untested approaches first emphasized by March (1991) arises naturally. The main insight of his contribution is that the optimal incentive scheme to motivate exploration exhibits substantial tolerance for early failure and rewards for long-term

success. Tolerance for early failure allows the agent to explore in the early stages of the contractual relationship without incurring the usual negative consequences of lower pay or termination. At the same time, reward for long-term success prevents the agent from shirking early on and induces the agent to explore new ideas that will allow him to perform well in the longrun. The principal can more effectively motivate exploration if he can commit not to terminate an agent after poor short-term performance, even if it is *ex post* efficient for the principal to do so. Another important ingredient of Manso's model is timely feedback on performance. Providing information to the agent about how well he is doing allows the agent to explore more efficiently, reducing the costs of experimentation. An agent who does not get feedback on performance may waste more time on unfruitful ideas.

Empirical evidence on the effects of long-term incentives is scant. Most relevant to the findings presented below is Lerner and Wulf's (2007) study of corporate R&D lab heads. They show that higher levels of deferred compensation are associated with the production of more heavily cited patents, whereas short-term incentives bear no relationship to firm innovative performance. In a similar vein, Tian and Wang (2010) show that start-up firms backed by more failure-tolerant venture capitalists are more innovative *ex post*. The present article presents the first systematic attempt to isolate, in a field setting, the effect of the *bundle* of incentive practices identified by Manso (2011) on exploration and creativity at the individual level (see Ederer and Manso, 2010 for experimental evidence with a similar flavor). We believe that the academic life sciences in the United States provide an appropriate setting, first and foremost because it provides naturally occurring variation in incentives that closely matches the contrast between pay-for-performance and exploration-type schemes emphasized by Manso (2011).

Most academic life scientists must rely on grants from the NIH, the largest public funder of biomedical research in the United States. With an annual budget of $28.4 billion in 2007, support from the NIH dwarfs that available from other public or private funders, including the National Science Foundation ($6 billion in 2007) or the American Cancer Society ($147 million in 2007). The most common type of NIH grant for investigator-initiated projects is the R01 grant. In 2007, their average amount was $225,000 in annual direct costs, and the awards last for a typical three to five years before coming up for renewal (see Figure 2). The NIH "study sections," or peer-review panels in charge of allocating awards, are notoriously risk averse and often insist on a great deal of preliminary evidence before deciding to fund a project. This often leads researchers to resubmit their applications several times and to multiply the number of applications, taking time away from productive research activities. It is an often-heard complaint among academic biomedical researchers that study sections' prickliness encourages them to pursue relatively safe avenues that build directly on previous results, at the expense of truly exploratory research (Kaplan, 2005; Kolata, 2009; McKnight, 2009).

An alternative funding mechanism is provided by the investigator program of the 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 based on nominations from research institutions; once selected, researchers continue to be based at their institutions, typically leading a research group of 10–25 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 (2011). HHMI investigators are initially appointed for five years,[3] and in the case of termination, there is a two-year phase-down period during which the researcher continues to be funded, allowing her to search for other sources of funding without having to close down her lab.

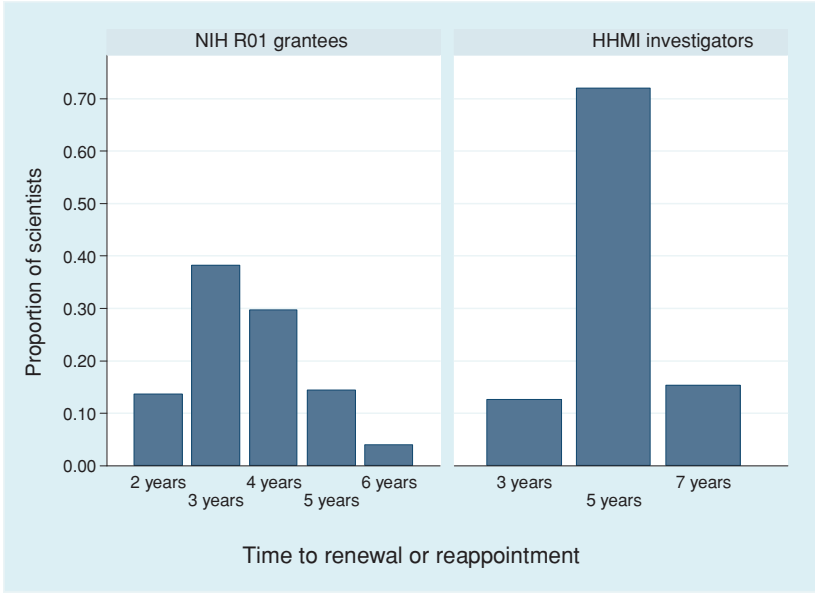
Moreover, HHMI investigators appear to share the perception that their first appointment review is rather lax, with reviewers more interested in making sure that they have taken on new

[2] See www.hhmi.org/research/investigators/

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

FIGURE 2

LENGTH OF NIH R01 GRANTS VERSUS HHMI APPOINTMENTS



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

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 articles 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.

Because 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 articles 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. Although 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.

TABLE 1 Comparison between the Two Sources of Funding

NIH R01 Grants	HHMI Investigator Program
Three- to five-year funding	Five-year funding
First review is similar to any other review	First review is rather lax
Funds dry up upon nonrenewal	Two-year phase-down upon nonrenewal
Some feedback in the renewal process	Feedback from renowned scientists
Funding is for a particular project	"People, not projects"

The economics literature (e.g., Aghion, Dewatripont, and Stein, 2008) views unfettered control over one's research agenda as the key distinguishing feature of innovative activities performed in academia (relative to the private sector). Variation in the unit of selection reminds us that the degree of *effective* control experienced by academic researchers often depends on the arcane details of funding mechanisms. Although not part of Manso's (2011) initial analysis, we extend his model in Appendix A 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.

3. 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 (i) the Howard Hughes Medical investigator sample; (ii) the set of control investigators against which the HHMI scientists will be compared; and (iii) our metrics of scientific creativity. We also present relevant descriptive statistics.

□ **HHMI sample.** We begin with a basic description of the criteria necessary for nomination and appointment as an HHMI investigator. To be eligible, a scientist must be tenured or on the tenure track at a major research university, academic medical center, or research institute. The subfields of the life sciences of interest to HHMI investigators are quite broad, but have tended to concentrate on cell and molecular biology, neurobiology, immunology, and biochemistry. Career-stage considerations have varied over time, although HHMI typically has not appointed scientists until they have had enough independent experience so that their work can be distinguished from that of their postdoctoral or graduate school adviser.

Upon receipt of nominations from participating institutions, HHMI empanels a jury that reviews these nominations in two sequential steps. In a first step, the number of nominees is whittled down to a manageable number, mostly based on observable characteristics. For example, NIH-funded investigators have an advantage because the panel of judges interprets receipt of federal grants as a signal of management ability. The jury also looks for evidence that the nominee has stepped out of the shadow cast by his/her mentors: an independent research agenda, and a "big hit," that is, 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.[4] 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.[5]

[4] Although 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.

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

Our analysis focuses on HHMI investigators appointed in 1993, 1994, and 1995. We exclude the three researchers that withdrew from the program voluntarily, leaving us with a sample of 73 scientists.[6]

□ **Control sample: early career prize winners.** In the absence of information on the runners-up of the HHMI competitions, we must rely on observable characteristics to create viable control groups. The main challenge is that HHMI investigators are extremely accomplished at the time of their appointment. Controls should not only be well matched with HHMI investigators in terms of fields, age, gender, and host institutions; their accomplishments should also be comparable at baseline. Our control group is drawn from early career prize winners in the life sciences.

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

Before presenting descriptive statistics, it is useful to discuss broad features of the control group that will influence the interpretation of the treatment effect. The process that results in the selection of HHMIs and controls is very similar. In both cases, an elite jury of senior scientists is given the mission to identify individuals with an impressive track record as well as exceptional promise; in particular, they are not asked to evaluate the merits of an individual project. The main difference between these programs is that ECPWs are selected at the very start of their independent career, when it is difficult to distinguish their output from that of their postdoctoral mentor. In contrast, the modal HHMI investigator stands at the cusp of the tenure decision when s/he is appointed. As a result, there is more variability in the expected performance of ECPW scholars than is the case among HHMI investigators but, as we will show, it is possible to cull from this group a subsample of scientists whose characteristics match well those of HHMI scientists at baseline.

□ **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 metrics or techniques to measure creative outputs.

The perspective adopted in this article is very pragmatic, and guided by the constraints put on us by the availability of data. Amabile (1996) suggests that whereas 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."* Although 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

[6] One accepted a top administrative position in his/her university (HHMI rules prevent investigators from holding major administrative posts), and one moved to an institution that had no relationship with HHMI. Yet another wished to move to a different institution during his/her first appointment. To prevent the eruption of bidding wars over HHMI investigators, the institute forces such investigators to resign their appointment.

generate publications and citations. Moreover, some types of publications, like review articles, tend to generate a number of citations not commensurate with their degree of originality. It has long been noted that the distributions of publications and citations at the individual level is extremely skewed, and typically follows a power law (Lotka, 1926). The distribution of citations *at the article level* exhibits even more skewness. In this article, we make use of the wide variation in impact across the publications of a given scientist to compute measures of creative output. Specifically, we sum the number of distinct contributions that fall into the higher quantiles (top quartile, top five percentiles, or top percentile) of the article-level distribution of citations for an individual scientist in a given time period.

One practical hurdle is truncation: older articles have had more time to be cited, and hence are more likely to reach the tail of the citation distribution. To overcome this issue, we compute a different empirical cumulative distribution function in each year.[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 five percentiles of the distribution; an article published in 1990, 94 citations; and an article published in 2000, only 57 citations (this is illustrated in Figure 1). With these empirical distributions in hand, it becomes meaningful to count the number of articles that fall, for example, in the top percentile over a scientist's career. Counting the number of contributions that fall "in the tail" is predicated on the idea that exploration is more likely to result in high-impact publications, relative to exploitation.[8] We also assess impact relative to each scientist's own pre-appointment citation performance. Because there are not enough data to estimate individual, vintage-specific citation distributions, we use the entire corpus of work published up until the year of appointment (1993, 1994, or 1995) to compute the citation quantile corresponding to each scientist's most heavily cited article. We then count the number of times a scientist exceeds this level after appointment.

We rely on two additional metrics of scientific excellence. We tabulate elections to the National Academy of Sciences. We also measure the number of students and fellows trained in a scientist's lab that go on to win a Pew, Searle, Beckman, Packard, or Rita Allen scholarship.[9]

HHMI appointments might also fatten the left-hand tail of the outcome distribution, because pushing the boundaries of one's field is a riskier endeavor than cruising along an already-established scientific trajectory. To test this prediction, we compute the number of contributions that fall in the bottom quartile of the vintage-specific, article-level distribution of citations (about three citations or fewer).[10] We also count the number of times each scientist underperforms, relative to the pre-appointment article corresponding to his/her lowest citation quantile. Because HHMI investigators remain eligible for NIH grants, we also examine how funding outcomes change following appointment, relative to ECPW controls. In particular, our data enable us to separate whether funding levels differ because of a change in application behavior or because HHMI investigators' grant applications are scored differently by NIH's review panels in the post-appointment regime.

Finally, explorative behavior should have implications for the *direction* of research endeavors, independently of the success or failure of the associated projects. To investigate this issue, we construct a battery of measures designed to capture potential changes in the scientists' research trajectories. Most of these measures use MeSH keywords as an essential input.[11] First, we calculate the average age of MeSH keywords for the published research of every scientist in the sample,

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

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

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

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

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

separately for each year of their independent career. A keyword is said to be born the first year it appears in any article indexed by PubMed. This measure captures the extent to which a scientist's research is novel relative to the world's research frontier. Equally important is to document the extent to which scientists place new scientific bets in the post-appointment period (1995–2006) relative to the pre-appointment period (1986–1994).[12] We do so by (i) computing the degree of overlap in MeSH keywords corresponding to articles published in both periods; (ii) computing the Herfindahl index of MeSH keywords in both periods (a proxy for variety in topic choice); and (iii) computing one minus the Herfindahl index of citing journal diversity in both periods (a measure of impact breadth, rather than impact depth as with the citation quantiles). If HHMI investigators are induced to explore novel approaches following their appointment, we would expect this behavior to be reflected in these measures.

□ **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 postdoctoral fellowship; gender; and department(s).

We obtain publication and citation data from PubMed and Thomson Scientific's *Web of Science*, respectively. Funding information stems from NIH's Compound Applicant Grant File, and is available for the entire length of these scientists' careers. In contrast, grant applications and their associated priority scores (the "grades" awarded to applications by NIH review panels) are available solely for years 2003–2008.

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 signalling. Yet others work with model organisms (mice, flies, monkeys, worms, etc.), asking questions that require, if not a whole organism, at least the interaction of multiple cells. The translational label is given to labs run by physician-scientists whose research has both a laboratory and a clinical component.

HHMI and control samples at baseline. Table 2 presents baseline descriptive statistics. Approximately 37% of the HHMI sample is female, versus 20% of the ECPW sample. They are of the same career age on average, but better funded than ECPW scholars at baseline ($1.45 million vs. $1.10 million on average). In terms of raw publication output, the pattern is very similar, with HHMI investigators leading ECPW scholars. The breadth of impact and diversity of topics studied by these scientists appears similar for both groups of scientists. ECPWs and HHMI investigators appear to be drawn from a similar set of academic employers in a dimension relevant for HHMI appointment: the number of slots allocated to their institution at the nomination stage.

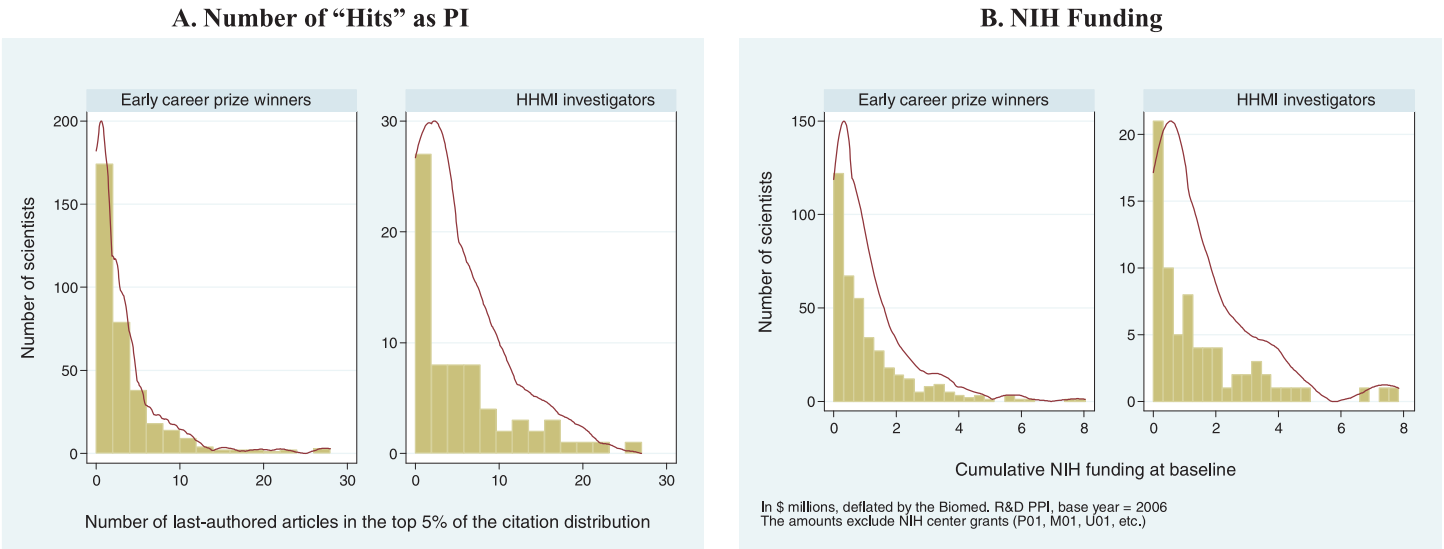
Of course, these averages tell only part of the story. Figure 3 A plots the distribution of baseline publications in the top 5% of the citation distribution. Note that we are only including here publications for which the scientist is the senior author, that is, where s/he appears in last position on the authorship list. The distribution for ECPW scholars appears significantly more skewed than that for HHMI investigators. Similarly, Figure 3B plots the distribution of NIH funding at baseline for treatment and control scientists; the shapes of these distributions 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 who will be good matches for HHMI investigators on these important dimensions.

[12] For investigators appointed in 1993 (resp. 1995), the "after" period begins in 1994 (resp. 1996).

FIGURE 3

BASELINE COMPARISONS BETWEEN HHMI AND CONTROL SCIENTISTS



To compute baseline article counts, we focus on articles in which the scientist appears in last position on the authorship list, because this is the set that clearly identifies both treated and controls as principal investigators in the pre-appointment period. To compute NIH funding totals, we exclude research center grants because these grants are less likely to correspond to individual effort; in some cases, deans or department chairs serve as pro forma PIs on such grants, making it a less useful measure for our purposes.

TABLE 2 Descriptive Statistics: Baseline

	Mean	Median	Standard Deviation	Minimum	Maximum
Controls (N = 393)					
Degree year	1983.689	1984	3.738	1974	1991
Female	0.199	0	0.400	0	1
MD	0.076	0	0.265	0	1
PhD	0.799	1	0.401	0	1
MD/PhD	0.125	0	0.331	0	1
Macromolecular	0.232	0	0.422	0	1
Cellular	0.394	0	0.489	0	1
Organismal	0.265	0	0.441	0	1
Translational	0.104	0	0.305	0	1
No. of nomination slots	2.179	2	1.296	0	8
Cum. NIH funding ($)	1,106,790	676,249	1,375,588	0	11,634,552
Highest citation quantile	40.001	36	24.352	1	100
Lowest citation quantile	99.202	100	2.748	62	100
Cum. no. of pubs.	24.775	20	20.764	2	200
Cum. no. of pubs. in the bottom 25%	0.647	0	1.410	0	15
Cum. no. of pubs. in the top 25%	18.718	15	14.146	0	123
Cum. no. of pubs. in the top 5%	9.647	8	7.822	0	51
Cum. no. of pubs. in the top 1%	3.712	3	3.875	0	27
Average MeSH age	23.376	23	2.808	18	35
Citing journal diversity, 1986–1994	0.963	1	0.020	0.837	0.992
HHMIs (N = 73)					
Degree year	1983.723	1984	4.002	1974	1991
Female	0.369	0	0.486	0	1
MD	0.082	0	0.274	0	1
PhD	0.753	1	0.431	0	1
MD/PhD	0.164	0	0.370	0	1
Macromolecular	0.288	0	0.453	0	1
Cellular	0.329	0	0.470	0	1
Organismal	0.274	0	0.446	0	1
Translational	0.110	0	0.313	0	1
Nb. of nomination slots	2.194	2	1.222	0	8
Cum. NIH funding ($)	1,502,810	1,005,176	1,768,341	0	7,852,110
Highest citation quantile	33.626	28	23.197	1	89
Lowest citation quantile	99.762	100	0.847	93	100
Cum. no. of pubs.	32.657	23	27.399	3	172
Cum. no. of pubs. in the bottom 25%	0.627	0	0.902	0	4
Cum. no. of pubs. in the top 25%	26.866	19	23.398	3	148
Cum. no. of pubs. in the top 5%	16.910	13	16.889	1	119
Cum. no. of pubs. in the top 1%	8.478	5	10.224	0	73
Average MeSH age	22.824	23	2.253	17	29
Citing journal diversity, 1986–1994	0.965	1	0.018	0.921	0.992

Career achievement. Although the differences between treatment and control samples are relatively modest at baseline, their magnitude increases when we examine achievements over the entire career. In Table 3, we see that HHMI scientists publish many more articles than ECPW scientists, with this output of higher quality, regardless of the quantile threshold one chooses to focus on. Of course, these accomplishments should be viewed in light of HHMI investigators' funding advantage: although they have garnered fewer resources from NIH by the end of the sample period than ECPW scholars, they also benefit from HHMI's relatively lavish research budgets. In fact, HHMI scientists apply less for R01 grants than controls who have no alternative sources of funding: 3.2 versus 5.1 applications on average between 2003 and 2008. On the other hand, conditional on applying, these same applications are judged more harshly by NIH study

TABLE 3 Descriptive Statistics: Career Achievement

	Mean	Median	Standard Devision	Minimum	Maximum
Early career prize winners (N = 393)					
Early career prize winners trained	0.229	0	0.630	0	1
Nobel Prize winner	0.003	0	0.050	0	1
Elected NAS member	0.041	0	0.198	0	1
Career no. of articles	65.003	53	43.444	11	314
Career no. of citations	4,489	3,504	3,489	242	21,448
Career no. of articles in the top 25%	47.952	40	30.829	7	212
Career no. of articles in the top 5%	22.214	18	15.760	0	96
Career no. of articles in the top 1%	7.926	6	7.410	0	38
Number of post-appointment hits	4.087	2	6.150	1	69
Number of post-appointment flops	3.448	2	5.287	0	41
Career NIH funding ($)	5,229,193	4,805,193	3,458,834	160,249	23,350,194
Avg. length (in years) for R01 grants	3.680	3.500	1.151	2	6
No. of R01 grant apps., 2003–2008	5.119	4	3.339	1.000	23.000
Avg. priority score, 2003–2008	161.842	158	36.637	100.000	283.000
Citing journal diversity, 1995–2006	0.968	1	0.025	0.667	0.992
Normalized MeSH keyword overlap	0.104	0	0.062	0	0.462
HHMI investigators (N = 73)					
Early career prize winners trained	1.137	0	2.388	0	1
Nobel Prize winner	0.014	0	0.117	0	1
Elected NAS member	0.329	0	0.473	0	1
Career no. of articles	95.521	83	56.126	17	321
Career no. of citations	10,550	6,672	14,542	798	117,401
Career no. of articles in the top 25%	78.219	69	48.843	10	284
Career no. of articles in the top 5%	45.562	38	33.863	4	224
Career no. of articles in the top 1%	21.014	16	21.270	0	144
Number of post-appointment hits	5.967	4	8.663	1	62
Number of post-appointment flops	3.483	1	5.890	0	32
Career NIH funding ($)	4,331,909	3,587,172	3,368,619	0	15,917,327
Avg. length (in years) for R01 grants	3.013	2.500	1.414	2	5
No. of R01 grant apps., 2003–2008	3.217	2	2.358	1.000	10.000
Avg. priority score, 2003–2008	178.289	173	33.405	111.500	326.000
Citing journal diversity, 1995–2006	0.975	1	0.013	0.921	0.993
Normalized MeSH keyword overlap	0.085	0	0.037	0	0.188

sections, because they are associated with higher priority scores.[13] Among our "direct" measures of explorative behavior, only the average level of normalized keyword overlap appears to be lower for HHMI investigators, compared with ECPW controls in these univariate comparisons.

When we focus on discrete career accolades, we observe an even greater contrast between HHMI and control scientists. Approximately a third of the HHMI investigators are elected members of the National Academy of Sciences, versus 4.1% for the control sample. Our 73 HHMI investigators collectively train 83 future early career prize winners (an average of 1.13 per scientist), whereas the control investigators are mentors to 90 such "young superstars" (an average of 0.23 per scientist).

4. Econometric considerations

■ In order to estimate the treatment effect of the HHMI investigator program, we must confront a basic identification problem: appointments are driven by expectations about the creative potential of scientists, and selected investigators might have experienced very similar outcomes had they

[13] Priority scores vary between 100 and 500, with lower scores indicating applications with a higher chance of funding.

not been appointed. As a result, traditional econometric techniques, which assume that assignment into the program is random, cannot recover causal effects.

□ **Propensity-score weighting.** As an attempt to overcome this challenge, we estimate the effects of the program using inverse probability of treatment-weighted estimation (Robins and Rotnitzky, 1995; Hirano and Imbens, 2001; Busso, DiNardo, and McCrary, 2008). Suppose we have a random sample of size N. For each individual i in this sample, let $TREAT_i$ indicate whether s/he received treatment. Using the counterfactual outcome notation (e.g., Rubin 1974), let y_i^1 be the value of the outcome y that would have been observed had i received treatment, and y_i^0 the value of the outcome had i been assigned to the control arm of the experiment. In addition, we will assume that we observe a vector of covariates denoted by $X = (W, Z)$. The variables included in W are assumed to be strictly exogenous; in contrast, the vector Z includes pretreatment variables such as lagged outcomes.

For each individual i, the treatment effect is $y_i^1 - y_i{}^0$. For the population as a whole, we are interested in two distinct estimands, the average treatment effect (ATE) and the average treatment effect on the treated (ATT). Formally,

$$\beta^{ATE} = E\left[y_i^1 - y_i^0\right]$$
$$\beta^{ATT} = E\left[y_i^1 - y_i^0 \mid TREAT_i = 1\right].$$

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

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

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

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

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

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

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

In order to develop the intuition for this weighting strategy, we examine the formula corresponding to w^{ATE} a bit more closely. Each factor in the denominator is the probability that an individual received her own observed treatment, conditional on her past history of "prognosis factors" for treatment. Suppose that all relevant variables are observed and included in X. Then, weighting effectively creates a pseudopopulation in which X no longer predicts selection into treatment and the causal association between treatment and outcome is the same as in the original population.[14]

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

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

Through interviews with HHMI senior administrators, we have sought to identify the criteria that increase the odds of appointments, conditional on being nominated. As described earlier, the institute appears focused on making sure that its new investigators have stepped out of the shadow cast by their graduate school or postdoctoral mentors. They also 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 proxy PI leadership skills with a measure of cumulative R01 NIH funding at baseline. Of course, our selection equation also includes important demographic characteristics, such as gender, laboratory type, degree, and career age.

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

In such a case, Abadie (2005) proposes a semiparametric difference-in-differences (SDD) estimator that combines the advantages of adjustment for observed heterogeneity with differencing. The idea is to apply propensity-score reweighting not to the *levels* of outcome y as above, but to the *differences* in outcome between the post- and pretreatment periods. Under some additional regularity conditions, the ATT is identified and can be recovered by weighting $y^{post} - y^{pre}$ using

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

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

The SDD estimates are still vulnerable to the critique that time-varying sources of unobserved heterogeneity could bias the effects, but they greatly narrow the scope of selection concerns.

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

Because they rely on within-scientist variation, fixed personality differences that impact the creative potential of individual scientists (such as conscientiousness [Charlton, 2009] or desire for intellectual challenge [Sauermann and Cohen, 2010]) do not jeopardize a causal interpretation of the effect of HHMI appointment. Rather, one might worry that the appointment committee is able to recognize and select for "exploratory tendencies" *before* they manifest themselves in the researcher's published work. If this were the case, these latent explorers might have branched out in new directions even in the absence of their HHMI appointment. Although we cannot rule out this possibility, we take solace in the fact that ECPW scholars and HHMI investigators are very well matched at baseline along the dimensions of topic novelty and citation breadth, dimensions that we argue are good proxies for exploration. Furthermore, ECPW scholars are selected through a very similar process at an earlier career stage; given that the same individuals, or at least the same type of individuals, often serve on these panels, it is unlikely that the HHMI committee is more skilled at identifying those scientists that are "itching to branch out."

5. Results

■ Our presentation of results is organized in three sets of tables. Table 4 pertains solely to HHMI investigators, and validates empirically some of the purported distinctive features of the program. Table 5 presents evidence on the determinants of HHMI appointment. Finally, Tables 6–8 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. The unconditional probability of termination at the end of the first appointment term is 15.5%, versus 28.33% at the end of the second appointment term (conditional on being renewed once). However, our contention that the first review is laxer than the second has implications for the *conditional* probability of first and second reappointment.

TABLE 4 Sensitivity of HHMI Reappointment to Scientific Output

	First Reappt. (1a)	Second Reappt. (1b)	First Reappt. (2a)	Second Reappt. (2b)	First Reappt. (3a)	Second Reappt. (3b)	First Reappt. (4a)	Second Reappt. (4b)
Pubs	−0.001	0.024**						
	(0.001)	(0.005)						
Pubs in the top 25%			−0.002	0.027**				
			(0.002)	(0.007)				
Pubs in the top 5%					−0.003	0.027**		
					(0.003)	(0.010)		
Pubs in the Top 1%							−0.003	0.053**
							(0.006)	(0.020)
Female	0.039	0.022	0.040	0.035	0.036	0.053	0.045	0.086
	(0.100)	(0.114)	(0.102)	(0.115)	(0.105)	(0.121)	(0.107)	(0.119)
Associate	0.028	0.096	0.029	0.076	0.023	0.128	0.027	0.153
	(0.100)	(0.104)	(0.100)	(0.117)	(0.097)	(0.121)	(0.099)	(0.119)
Full	0.070	0.001	0.066	−0.026	0.059	0.074	0.057	0.098
	(0.114)	(0.146)	(0.112)	(0.192)	(0.110)	(0.206)	(0.110)	(0.213)
No. scientists	71	60	71	60	71	60	71	60
Log quasi-likelihood	−27.497	−19.841	−27.653	−21.251	−27.674	−24.150	−27.895	−24.176
Pseudo-R^2	0.102	0.338	0.097	0.291	0.096	0.194	0.089	0.193

Note: The dependent variable is the probability of being reappointed, whether at the end of the first term (models 1a, 2a, 3a, and 4a) or at the end of the second term (models 1b, 2b, 3b, and 4b) among 71 HHMI investigators who did not terminate their appointment voluntarily. The sample relevant to specifications 1b, 2b, 3b, and 4b comprises only 60 observations because 11 investigators were either not renewed at the end of the first appointment period or resigned their posts voluntarily. Estimates correspond to marginal effects from logit specifications, with robust standard errors in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

TABLE 5 Determinants of Selection into the HHMI Program

	(1)	(2)	(3)	(4)	(5)
Cum. no. pubs as PI	0.006**				0.002
	(0.002)				(0.014)
Cum. no. pubs in top 25% as PI		0.013**			0.003
		(0.002)			(0.006)
Cum. no. pubs in top 5% as PI			0.023**		0.015†
			(0.004)		(0.008)
Cum. no. pubs in top 1% as PI				0.039**	0.033**
				(0.008)	(0.009)
NIH funding	0.004	−0.018	−0.015	−0.001	−0.021
	(0.024)	(0.019)	(0.018)	(0.015)	(0.022)
Female	0.121**	0.123**	0.119**	0.122**	0.125**
	(0.036)	(0.035)	(0.034)	(0.034)	(0.034)
PhD	−0.082	−0.078	−0.058	−0.032	−0.049
	(0.087)	(0.096)	(0.104)	(0.100)	(0.110)
MD/PhD	−0.048	−0.053	−0.022	0.007	−0.017
	(0.082)	(0.087)	(0.092)	(0.089)	(0.097)
No. of nomination slots	−0.010	−0.011	−0.008	−0.006	−0.007
	(0.014)	(0.013)	(0.012)	(0.012)	(0.012)
Macromolecular lab	−0.039	−0.041	−0.024	−0.030	−0.028
	(0.043)	(0.042)	(0.041)	(0.042)	(0.042)
Organismal lab	0.002	0.004	0.002	−0.004	0.001
	(0.046)	(0.045)	(0.044)	(0.044)	(0.044)
Translational lab	−0.014	−0.005	0.008	0.013	0.010
	(0.085)	(0.087)	(0.090)	(0.083)	(0.090)
Pseudo-R^2	0.074	0.111	0.143	0.133	0.160
No. of scientists	466	466	466	466	466

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

$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

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 4 provides evidence consistent with this hypothesis. It reports estimates from logit models of reappointment as explained by various indicators of achievement during the preceding term. We find a consistent pattern, regardless of the achievement variable on the right-hand side: higher achievement significantly increases the likelihood of renewal at the end of the second term, but not at the end of the first term. Moreover, the marginal effect for blockbuster articles produced in the previous period is twice as large as the marginal effect for total publication output. This is consistent with the idea that HHMI review panels care more about whether investigators "transform their fields" than they care about counting lines on their CVs.

From these results, we conclude that the HHMI program conforms both in its *stated* and *actual* practices with the features that Manso (2011) predicts should encourage exploration.

□ **Determinants of HHMI appointment.** We now turn to the observable determinants of selection into the HHMI program (Table 5). We present the results from logit specifications that include demographic characteristics as controls, as well as cumulative NIH funding at baseline, and achievements as PIs in the pre-appointment period. Among the demographic characteristics, the only consistent pattern is the higher appointment probability of female scientists. Consistent with the qualitative evidence on the selection process, we find that the number of "hit articles"

TABLE 6 Effects of HHMI Appointment on Citation Impact (N = 417 Scientists)

Benchmark	Achievement Metric	"Naive" X-Section	ATE	ATT	DD	SDD
Universal article-level citation distribution	All pubs	0.419**	0.235**	0.227*	0.178*	0.333**
		(0.076)	(0.078)	(0.088)	(0.072)	(0.109)
	Top 25%	0.514**	0.297**	0.305**	0.212**	0.268*
		(0.079)	(0.085)	(0.087)	(0.074)	(0.114)
	Top 5%	0.733**	0.482**	0.510**	0.293**	0.439**
		(0.093)	(0.111)	(0.102)	(0.108)	(0.161)
	Top 1%	0.964**	0.663**	0.817**	0.363*	0.678**
		(0.133)	(0.138)	(0.133)	(0.148)	(0.240)
	Bottom 25%	0.181	0.094	0.154	0.187	0.155
		(0.128)	(0.131)	(0.135)	(0.292)	(0.887)
Relative to the scientist's own citation impact pre-appointment	Number of Hits	0.401**	0.299*	0.356**		
		(0.125)	(0.128)	(0.128)		
	Number of Flops	0.341*	0.272*	0.317†		
		(0.146)	(0.121)	(0.162)		

Note: The first five lines pertain to the analysis of citation impact using the total number of citations for the universe of all articles in the life sciences field, as coded by ISI/Web of Science. Each coefficient corresponds to the treatment effect of HHMI appointment in a specification that regresses output on treatment status, five age indicator variables (5–10 years of career age, 10–15 years, 15–20 years, 20–25 years, and 25 years and more of career age), and year indicator variables in all models. The cross-sectional models (corresponding to the first three columns) also include three lab indicator variables, a gender indicator variable, and two degree-type indicator variables (coefficients not reported). Estimates derive from quasi-maximum likelihood (QML) Poisson estimation, with robust standard errors in parentheses, clustered around scientist (X-section, ATE, ATT, and DD columns); bootstrapped standard errors are reported for the semi-parametric difference-in-differences estimates. All specifications except the naive cross-sections and the plain difference-in-differences include regression weights computed using fitted values for the probability of HHMI appointment estimated in Table 5. The weights differ depending on whether ATT or ATE is the effect of interest, and whether the focus is on generating a between-scientist comparison (ATE and ATT columns) or a within-scientist comparison (SDD column). See Section 4 for more details.

The last two lines use each scientist's own citation impact in the pre-appointment period as a benchmark. We code the highest (resp. lowest) quantile of the article-level citation distribution for any article published by each scientist prior to appointment. We then compute the number of hits (resp. flops) for each scientist by counting the number of articles whose citation quantile places them above (resp. below) this level in the post-appointment period. The corresponding specifications also include year of highest degree indicator variables, three lab-type indicator variables, a gender indicator variable, two degree-type indicator variables, as well as the pre-appointment highest or lowest pre-appointment quantile mentioned above. Because we use the whole pre-appointment citation data to calculate the benchmark, there are no DD or SDD specifications when assessing citation impact relative to the scientists' own prior performance.

$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

at baseline is highly predictive of appointment. In contrast, the level of funding appears to play no role in the odds of selection. Using the most saturated model of selection (column 5), we find that the region of common support excludes 4 HHMI investigators whose superlative record of achievement prior to appointment makes them difficult to compare to any member of the control group. Conversely, 45 early career prize winners have a very low predicted probability of appointment, mostly because they do not produce an impactful article after they set up their lab. In all that follows, we have excluded from the estimation sample these 49 scientists. This ensures the validity of the common support assumption, which is necessary to identify the ATE or the ATT using inverse probability of treatment-weighted estimation. The final sample contains information on 417 scientists (69 HHMIs and 348 controls).

□ **Effects of HHMI appointment on citation impact.** The first four lines of Table 6 report the effect of the program on the rate of publication output falling in distinct citation quantile bins: all publications, publications in the top quartile, in the top five percentiles, and in the top percentile. For each outcome variable, we present five coefficients corresponding to different ways of assessing the program's effects. The first column reports naive cross-sectional results, which ignore the selection process. The second and third columns weight the outcome

equations by the inverse probability of treatment so as to recover the ATE, and the ATT under unconfoundedness. The fourth column reports simple conditional fixed-effects estimates, a naive DD. Finally, the fifth column reports results corresponding to SDD estimates as in Abadie (2005). Because the SDD estimator adjusts the treatment effect for selection on observables while purging the estimates of time-invariant unobserved heterogeneity, it is our preferred specification.

Following a long-standing tradition in the study of scientific and technical change, the cross-sectional, ATE, ATT, and DD effects are estimated on the full panel using quasi-maximum likelihood Poisson.[16] In contrast, the SDD effects stem from a two-step procedure detailed in Appendix C.

The naive cross-sectional estimate is always the largest in magnitude, and using propensity-score weighting reduces the magnitude of the effect by approximately a third. In contrast, the DD estimate is systematically lower than the SDD estimate, as is possible if HHMI investigators and controls are on different output trends even before appointment. The magnitudes of the effects are large. For instance, the SDD estimates imply that the HHMI program increases the rate at which appointed scientists produce publications by $e^{.333} - 1 = 39\%$; the figure for articles in the top 5% of the citation distribution is 55%; and for articles in the top 1%, a 97% increase. The observed pattern is that the program has a bigger effect on the upper tail of the distribution of accomplishments, regardless of the estimation method used.

Figure 4 display the time path of the average publication count and top 5% outcome for HHMIs and ECPWs separately. While computing the averages, we weight each control scientist's outcome by his/her inverse probability of being selected into the program, while leaving the treated scientists' outcomes unchanged. Loosely, Figures 4A and 4B provide a graphical intuition for the SDD estimates: they correspond to the difference between the change in outcomes for the HHMI investigators and for a pseudopopulation of control scientists matched on observables. A necessary condition for the plausibility of this exercise is that the treated and control groups display parallel output trends prior to the appointment event. This appears to be the case here.

Interestingly, for three years after appointment, the outcomes for treated and control scientists continue to track each other closely. Figure 4B even suggests that the control group (appropriately selected on observables) briefly outpaces the treatment group following the appointment, consistent with Manso's (2011) theory which predicts both slower and more variable returns under an exploration incentive scheme. This difference is not statistically significant, however, which is perhaps unsurprising given our sample's relatively small size. HHMI investigators' output begins to diverge from that of ECPWs only four to five years after appointment, and this divergence is more marked in Figure 4B.

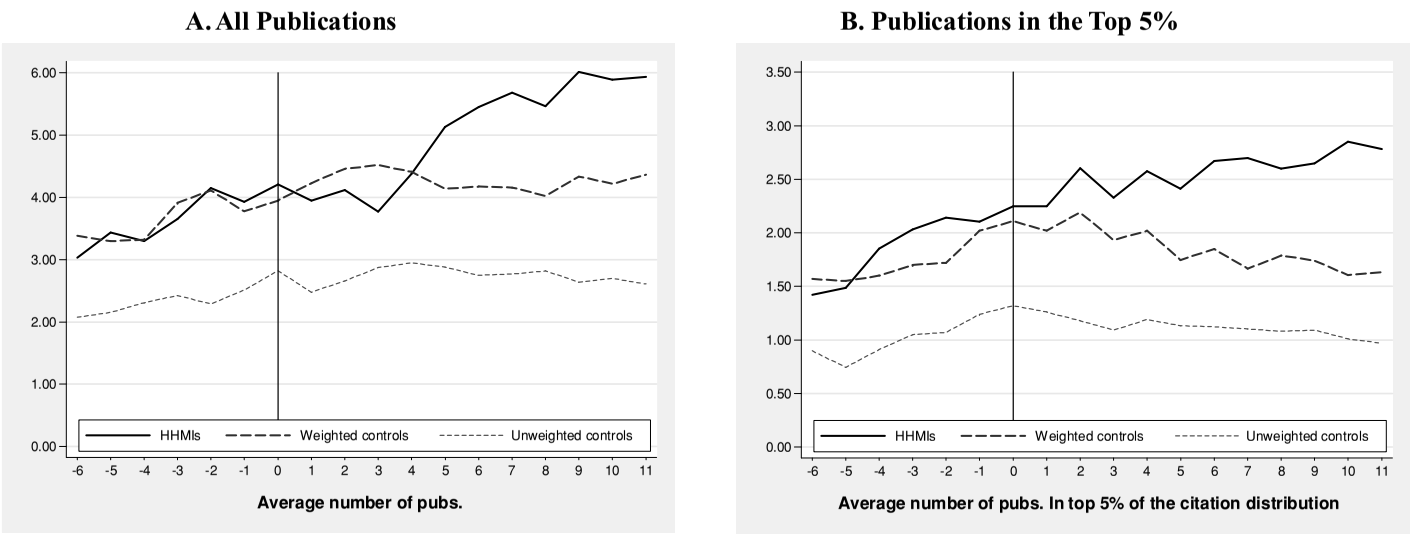
We have explored the hypothesis of a temporary, post-appointment slowdown qualitatively by asking eight current and former HHMI investigators about the "retooling" necessary to take advantage of the freedom afforded by the program. This idea resonated with these scientists, but it also seems clear that these lags are very heterogeneous across labs. Some of them mentioned waiting until their first renewal before branching out; others were clearly itching to begin new projects, for example by focusing on a new disease (autism vs. Huntington's), model organism (mice vs. yeast), or discipline (chemistry vs. cell biology). Still others described a less deliberate exploration process whereby the logic of their traditional research projects opened up novel opportunities, which they could more easily take advantage of as HHMI investigators.

□ **Effects of HHMI appointment on failure.** It seems intuitive that exploration would lead scientists to "strike out" more often. Measuring failure is difficult, because it might lead researchers to abort projects altogether. Here we ask whether HHMI investigators produce more

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

FIGURE 4

DYNAMIC EFFECTS OF HHMI APPOINTMENT ON PUBLICATION RATES



The dashed and solid black lines correspond to the average yearly number of articles for early career prize winners and HHMI investigators, respectively. The averages for the control scientists are weighted by each researcher's inverse probability of treatment, where the weights are computed using fitted values of the logit specification. The dashed light gray line corresponds to the unweighted average yearly number of articles for the control scientists. Panel A displays our results for total publications (regardless of impact), whereas panel B restricts the outcome data to "hits" (publications that fall in the top five percentiles of the vintage-specific article-level distribution of citations).

TABLE 7 Effects of HHMI Appointment on NIH Funding

Dependent Variable	"Naive" X-Section	ATE	ATT	DD	SDD
NIH funding ($)	−0.404**	−0.549**	−0.497**	−0.546**	−0.426**
	(0.095)	(0.099)	(0.094)	(0.105)	(0.115)
No. of R01 Apps.	−0.521**	−0.603**	−0.486**		
	(0.122)	(0.126)	(0.122)		
Avg. priority score for R01	−0.077**	−0.075**	−0.057†		
	(0.029)	(0.025)	(0.032)		
No. of scientists	417	417	417	417	417

Note: Each coefficient corresponds to the treatment effect of HHMI appointment in a specification that regresses the dependent variable on treatment status, five age indicator variables (5–10 years of career age, 10–15 years, 15–20 years, 20–25 years, and 25 years and more of career age), and year indicator variables. The cross-sectional models (corresponding to the first three columns) also include three lab indicator variables, a gender indicator variable, and two degree-type indicator variables (coefficients not reported). Estimates derive from QML Poisson estimation, with robust standard errors in parentheses, clustered around scientist (X-section, ATE, ATT, and DD columns); bootstrapped standard errors are reported for the semiparametric difference-in-differences estimates. All specifications except the naive cross-sections and the plain difference-in-differences include regression weights computed using fitted values for the probability of HHMI appointment estimated in Table 5. The weights differ depending on whether ATT or ATE is the effect of interest, and whether the focus is on generating a between-scientist comparison (ATE & ATT columns) or a within-scientist comparison (SDD column). Because grant application data are only available for the period 2003–2008, the determinants of application rates and priority scores are estimated using a single cross-section that pools together all of the data for the corresponding period. See Section 4 for more details.

†$p < 0.10$, *$p < 0.05$, $p < 0.01$.

articles of little import, relative to controls. To answer this question, we examine whether HHMI appointment increases the rate of publications that fall in the bottom quartile of citations. Relative to ECPW scholars, HHMI investigators indeed fail more often, regardless of estimation method; some of these estimates are large in magnitude, but they are also imprecisely estimated. The lack of statistical significance is not terribly surprising, because relatively few of the articles produced by these elite scientists will fail to garner the three citations that correspond to the 25th percentile of the citation distribution in most years.

An alternative approach is to use these scientists' own citation impact prior to appointment to assess their performance in the post-appointment regime, rather than a universal citation benchmark as above. The corresponding results are displayed in the last two lines of Table 6. The coefficient estimates pertain to the HHMI treatment effect in cross-sectional comparisons where the rates of "hit" and "flop" publications are modelled using quasi-maximum likelihood Poisson. To compute the number of hits, we count the number of times each scientist publishes an article whose citation quantile places it above the highest citation quantile of any article published prior to appointment. Symmetrically, the number of flops is computed by counting the number of times each scientist publishes an article whose citation quantile places it below the lowest citation quantile of any article published prior to appointment (further details are provided in Appendix D). Because we use citation data for the entire pre-appointment period to compute individual citation benchmarks, there are no DD and SDD specifications for these two outcomes.

We find robust statistical evidence that HHMI appointment increases the frequency of both hits and flops. We focus on the latter result, because an increased rate of failure under an exploration incentive scheme is a strong prediction of the theory. It is also more challenging to reconcile with the view that HHMI simply picks extraordinarily talented scientists and then takes credit for their accomplishments. Of course, we cannot rule out a more nuanced selection story whereby the elite scientists who serve as judges in HHMI competitions are skilled at identifying scientists destined to push the scientific frontier outward.

□ **Effects of HHMI appointment on NIH grant outcomes.** In Table 7, we document a negative association between HHMI appointment and NIH funding that holds both in the cross-sectional and the within-scientist dimensions of the data. This effect corresponds in large part to

TABLE 8 Effects of HHMI Appointment on the Direction of Research

Impact of Treatment on:	Dependent Variable	X-Section	ATE	ATT	DD	SDD
Topic novelty	Avg. MeSH keyword age	−0.028**	−0.014	− 0.016†	−0.020	−0.027*
		(0.009)	(0.009)	(0.009)	(0.013)	(0.013)
Change in research direction	Normalized MeSH keyword overlap	−0.258**	−0.206**	−0.259**		
		(0.060)	(0.058)	(0.059)		
Breadth of impact	Citing journal diversity index	0.223**	0.192**	0.231**		
		(0.071)	(0.060)	(0.073)		
No. of scientists		417	417	417	417	417

Note: For the analysis of the determinants of topic novelty, each coefficient corresponds to the treatment effect of HHMI appointment in a specification that regresses measures of scientific novelty on treatment status, five age indicator variables (5–10 years of career age, 10–15 years, 15–20 years, 20–25 years, and 25 years and more of career age), and year indicator variables in all models. The cross-sectional models (corresponding to the first three columns) also include three lab indicator variables, a gender indicator variable, and two degree-type indicator variables (coefficients not reported). Estimates derive from QML Poisson estimation, with robust standard errors in parentheses, clustered around scientist (X-section, ATE, ATT, and DD columns); bootstrapped standard errors are reported for the semiparametric difference-in-differences estimates.

For the other two outcomes (keyword overlap and diversity of citing journals), each coefficient corresponds to the treatment effect of HHMI appointment on various measures of an investigator's scientific direction in the "after" period (1995–2006). All models include as independent variables year of highest degree indicator variables, three lab type indicator variables, a gender indicator variable, and two degree-type indicator variables (coefficients not reported). Also included is an offset for the dependent variable in the "before" period (1986–1994). Because all of the dependent variable are bounded inclusively by 0 and 1, estimates stem from a QML fractional logit procedure (Papke and Wooldridge, 1996), with robust standard errors in parentheses. All specifications except the naive cross-sections and the plain difference-in-differences include regression weights computed using fitted values for the probability of HHMI appointment estimated in Table 5. The weights differ depending on whether ATT or ATE is the effect of interest; and whether the focus is on generating a between-scientist comparison (ATE and ATT columns), or a within-scientist comparison (SDD column). See Section 4 for more details.

$^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

a mechanical substitution of HHMI resources for traditional NIH funding: the number of R01 grant applications by HHMI investigators is, on average, only about 60% of the corresponding number for ECPW scholars. But application behavior does not on its own explain the decline in NIH funding; conditional on applying, applications from HHMI scientists receive higher priority scores (i.e., are judged more harshly by study sections). When combined with the results pertaining to citation impact, this evidence supports the view that the punctiliousness of the NIH peer-review process crowds out scientific exploration.[17]

□ **Effects of HHMI appointment on the direction of research.** So far, our presentation of results has conflated intensity of exploration with the rate at which tail outcomes are produced. But taken literally, the Manso (2011) 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. This is the hypothesis we examine in Table 8. Choosing less-traveled scientific avenues could also leave trails in the content of what scientists publish, and in particular affect the keywords that tag their publications. We first focus on whether HHMI investigators are prone to define the scientific frontier, by examining the vintage of the MeSH keywords in their output. In our analysis, a keyword is born the earliest year in which it appears in any publication indexed by PubMed. We then compute the average age of all keywords or all keyword pairs in each scientist's yearly output. Table 8 shows that HHMI investigators indeed tackle more novel topics; the coefficient estimates are negative regardless of estimation method.

Next, we ask whether evidence exists that HHMI investigators alter the direction of their scientific trajectory following their appointment. We first examine the program's effect on the

[17] More prosaically, they might put less effort into preparing these proposals, because they benefit from the safety of Hughes funding.

number of unique publication keywords that overlap between the set of articles published in the "before" period (1986–1994) and the "after" period (1998–2006). This measure is then normalized by the number of unique keywords used in the after period. For each control group, we report both the results of a "naive" specification and the results of two specifications which incorporate inverse probability of treatment weights corresponding to ATE and ATT. Because the dependent variable is a proportion, we estimate these models using the quasi-maximum likelihood fractional logit estimator of Papke and Wooldridge (1996). Relative to ECPW scholars, HHMI investigators exhibit unambiguously lower overlap in keyword use. The effects are statistically significant, and imply that HHMI appointment is associated with about a 10% lower rate of overlap.

Our last test focuses on the *breadth*—rather than the *depth*—of impact for these scientists' publications. To do so, we examine the journals in which citing articles appear, and compute the Herfindahl of journal concentration H. We find that HHMI investigators exhibit higher levels of $(1 - H)$ in the post-appointment period, that is, their work is cited by a more diverse set of journals than the articles published by ECPWs. Overall, the results in Table 8 are consistent with the idea that HHMI investigators broaden their research agenda in the post-appointment period, a necessary condition for exploration.

□ **Incentives versus alternative mechanisms.** Even if our estimates of HHMI appointments' treatment effects can be given a causal interpretation, ascribing them to the program's *incentive* features requires an interpretive leap.

First, we are unable to ascertain the extent to which the program increases productivity rather than output. It is hard to compare directly HHMI and NIH levels of funding, because the two programs are structured in a different fashion.[18] Yet, it appears likely that, per dollar of funding, HHMI investigators do not publish more articles 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.[19] We also note that the results pertaining to the diversity of experimentation are less vulnerable to this critique, because they essentially hold output constant.

Second, the prestige conferred by HHMI appointment might have independent effects on scientists' achievements, either by increasing exposure to their research or through a dynamic of ascription that has long been the focus of sociologists of science (Merton, 1968). Azoulay, Stuart, and Wang (2010) provide estimates of the HHMI investigator program's "anointment effects" by examining whether appointment shifts the citation rate of articles written in the pre-appointment period; their evidence points to effects of very modest magnitude. As such, an interpretation of our results that emphasizes the status benefits of HHMI appointment appears unwarranted.

Third, collaboration between scientists in the treatment and control groups might threaten the validity of the comparisons drawn in the analysis. Fifty-nine out of the 73 HHMI investigators have at least one control collaborator; 10 have five or more. However, peer effects from coauthorship (e.g., Azoulay, Graff Zivin, and Wang, 2010), which enhance the accomplishments of the control group, would tend to dampen the magnitudes of the effects estimated above.

A more subtle reinterpretation of our main results is that "explorer types" are more likely to seek HHMI appointment. A sorting process in this vein would imply that the freedom and long-term funding bundled with HHMI appointment have real effects, even if they do not really induce behavioral changes in the treated group. However, a peculiar feature of the HHMI appointment process is that candidates do not apply but rather are nominated by their universities, which are endowed by HHMI with a very limited number of nomination slots. This casts doubt on the

[18] For example, HHMI does not pay host institutions standard overhead rates, but does make a contribution toward rent and occupancy.

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

sorting interpretation at the principal investigator level. It is altogether more likely (and also more feasible) for talented postdoctoral researchers and graduate students with idiosyncratic tastes for exploration to sort themselves into HHMI-funded labs. The fact that HHMI labs train a much higher number of young scientists who go on to win early career prizes is consistent with a sorting process at the level of trainees.

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

6. Conclusion

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

Our findings are important for at least two reasons. First, they demonstrate the impact of nuanced features of research contracts for the rate and direction of scientific progress. Given the prominent role that scientific change is presumed to play in the process of economic growth (e.g., Mokyr, 2002), this has important implications for the organization of public and private research institutions. Second, they offer empirical support for the theoretical model developed by Manso (2011), and as such may provide insights relevant to a wider set of industries that rely on creative professionals, ranging from advertising and computer programming to leadership roles at the upper echelons of the corporate world.

Finally, our results should not be interpreted as a critique of NIH and its funding policies. Although "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 includes gifted individuals as well as those with more modest talent. Moreover, HHMI provides detailed evaluation and feedback to its investigators. The richness of this feedback consumes a great deal of resources, particularly the time of the elite scientists who serve on review panels, and its quality might degrade if the program were expanded drastically.

It is also vital to recognize that NIH operates under political constraints that a private foundation like HHMI can safely ignore. For instance, all public research agencies need to spread their support across many institutions, including those of lesser renown. Similarly, supporting individual projects, rather than individual scientists, introduces a level of impersonality in the funding decisions that may make them easier to defend vis-à-vis congressional appropriators.

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

Appendix A
Funding people versus funding specific projects

We develop a simple model to contrast the "specific project" and the "people, not projects" approaches to scientific funding. The researcher lives for two periods. In each period, she chooses a project $i \in \mathcal{I}$, producing output S ("success") with 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, $E[p_i \mid S, j]$ denote the conditional expectation of p_i given a success on project j, and $E[p_i \mid F, j]$ denote the conditional expectation of p_i given a failure on project j.

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

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

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

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

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

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

We assume that the innovative research project is of an exploratory nature. This means that when the researcher experiments with the innovative research project, she is initially not as likely to succeed as when she takes 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

$$E[p_2] < p_1 < E[p_2|S, 2]. \tag{A1}$$

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 this research as well as some weight to her private preferences between the two projects. These private preferences are represented with a cost c_i that is incurred by the researcher whenever she pursues project i. The researcher thus chooses an action plan $\langle {}_i{}^j_k \rangle$ to maximize her total expected payoff:

$$\begin{aligned} R(\langle {}_i{}^j_k \rangle) = {} & \{E[p_i]S + (1 - E[p_i])F - c_i\} + E[p_i]\{E[p_j|S, i]S + (1 - E[p_j|S, i])F - c_j\} \\ & + (1 - E[p_i])(E[p_k|F, i]S + (1 - E[p_k|F, i])F - c_k), \end{aligned} \tag{A2}$$

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

$$\begin{aligned} R(\langle {}_1{}^1_1 \rangle) = {} & \{E[p_1]S + (1 - E[p_1])F - c_1\} + E[p_1]\{E[p_1]S + (1 - E[p_1])F - c_1\} \\ & + (1 - E[p_1])\{E[p_1]S + (1 - E[p_1])F - c_1\}. \end{aligned} \tag{A3}$$

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

$$\begin{aligned} R(\langle {}_2{}^2_2 \rangle) = {} & \{E[p_2]S + (1 - E[p_2])F - c_2\} + E[p_2]\{E[p_2|S, 2]S + (1 - E[p_2|S, 2])F - c_2\} \\ & + (1 - E[p_2])\{E[p_2|F, 2]S + (1 - E[p_2|F, 2])F - c_2\}. \end{aligned} \tag{A4}$$

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

$$\alpha(E[p_2] - p_1)(S - F) \geq (c_2 - c_1). \tag{A5}$$

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

$$\begin{aligned} R(\langle {}_1{}^1_1 \rangle) = {} & \{E[p_1]S + (1 - E[p_1])F - c_1\} + E[p_1]\{E[p_1]S + (1 - E[p_1])F - c_1\} \\ & + (1 - E[p_1])\{E[p_1]S + (1 - E[p_1])F - c_1\}. \end{aligned} \tag{A6}$$

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

$$\begin{aligned} R(\langle {}_2{}^2_1 \rangle) = {} & \{E[p_2]S + (1 - E[p_2])F - c_2\} + E[p_2]\{E[p_2|S, 2]S + (1 - E[p_2|S, 2])F - c_2\} \\ & + (1 - E[p_2])\{E[p_1]S + (1 - E[p_1])F - c_1\}. \end{aligned} \tag{A7}$$

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

$$\alpha\{E[p_2](E[p_2|S,2]-p_1)+(E[p_2]-p_1)\}(S-F)\geq(1+E[p_2])(c_2-c_1). \tag{A8}$$

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 (A5) implies (A8). For the second statement, we construct the following example. If $c_2 > c_1$, (A5) implies that the agent never explores under the "specific project" approach. However, from (A8), if the payoff from exploration is sufficiently high, the agent will explore under the "people, not projects" approach.

Appendix B
Career and output data

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

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

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

Appendix C
Estimation procedure for the semiparametric DD estimates

The ATE, ATT, and DD effects stem from panel specifications; the sample size is equal to the total number of independent career years for each scientist ($N \times T = 8{,}767$). The procedure followed to estimate the SDD effects is slightly different. We first regress the various measures of output on calendar year and age indicator variables using the full panel, and compute the residuals ε_{it}. In a second step, we sum the residuals corresponding to the pre-appointment (1986–1994) and post-appointment (1998–2006) periods separately for each scientist. In the final step, the SDD effects are obtained by regressing $\sum_{t=1986}^{1994}\varepsilon_{it}-\sum_{t=1998}^{2006}\varepsilon_{it}$ on treatment status, weighting these differences as described in Section 4. Note that the sample size corresponds in this case to the number of scientists ($N = 417$), not the number of scientist-year observations.

Appendix D
Scientist-specific citation benchmarks

We illustrate the computation of the number of hits and flops using the example of Iva Greenwald, an HHMI investigator from Columbia University. Prior to 1994 (the year of her appointment), her publication with the highest citation quantile is an article which appeared in the journal *Cell* in 1993 (341 citations as of the end of 2008, which places it in the top percentile of the article-level distribution). Conversely, her publication with the lowest citation quantile is an article which appeared in the journal *Molecular and Cellular Biology*, also in 1993. It garnered only 11 citations, which places it at the 52nd percentile of the distribution. Between 1995 and 2006, Greenwald published three more publications in the top percentile of the citation distribution, given their vintage. And she published three more publications which fell in the 32nd, 44th, and 50th percentiles of the distribution in the years they were published. As a result, both the number of hits, and of flops, are equal to three for this investigator.

□ **Estimation.** Because we use the entire pre-appointment data to compute citation benchmarks specific to each individual scientist, we can only analyze the effect of HHMI appointment on these measures of impact using the cross-sectional dimension of the data, collapsing all post-appointment years into a single observation for each scientist. For the number of flops, the equation we estimate can be written

$$E[FLOPS_i \mid X_i] = exp(\beta_0 + \beta_1 HHMI_i + \beta_2' Z_i + \beta_3 MIN\_QNTL_i + \phi(SCIENTIST\_AGE_i)), \quad \text{(D1)}$$

where *HHMI* denotes the treatment effect, the variables in Z include degree type, lab type, and gender indicator variables, *MIN_QNTL* is the citation quantile corresponding to scientist *i*'s least impactful article published prior to appointment,[20] and $\phi(SCIENTIST\_AGE_i)$ is a flexible function of scientist *i*'s career age—in practice a full set of indicator variables for the different years in which our scientists received their highest degree. Estimation proceeds by quasi-maximum likelihood. In some of the specifications, the data are weighted to reflect each scientist's inverse probability of being appointed to the program, as explained in Section 4.

References

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

Aghion, P., Dewatripont, M., and Stein, J.C. "Academic Freedom, Private Sector Focus, and the Process of Innovation." *RAND Journal of Economics,* Vol. 39 (2008), pp. 617–635.

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

Azoulay, P., Stellman, A., and Graff Zivin, J. "PublicationHarvester: An Open-Source Software Tool for Science Policy Research." *Research Policy,* Vol. 35 (2006), pp. 970–974.

———, Graff Zivin, J. and Wang, J. "Superstar Extinction." *Quarterly Journal of Economics,* Vol. 125 (2010), pp. 549–589.

———, Stuart, T., and Wang, Y. "Matthew: Effect or Fable?" Working Paper, MIT Sloan School of Management, 2010. Available at http://pazoulay.scripts.mit.edu/

Burt, R.S. "Structural Holes and Good Ideas." *American Journal of Sociology,* Vol. 110 (2004), pp. 349–399.

Busso, M., DiNardo, J., and McCrary, J. "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects." Working Paper, University of Michigan, 2008.

Capecchi, M.R. "Response." *Science,* Vol. 319 (2008), pp. 900–901.

Charlton, B.G. "Why Are Modern Scientists So Dull? How Science Selects for Perseverance and Sociability at the Expense of Intelligence and Creativity." *Medical Hypotheses,* Vol. 72 (2009), pp. 237–243.

Dehejia, R.H. and Wahba, S. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics,* Vol. 84 (2002), pp. 151–161.

Ederer, F. and Manso, G. "Is Pay-for-Performance Detrimental to Innovation?" Working Paper, MIT Sloan School of Management, 2010.

Hirano, K. and Imbens, G.W. "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catherization." *Health Services & Outcomes Research Methodology,* Vol. 2 (2001), pp. 259–278.

Holmström, B. "Agency Costs and Innovation." *Journal of Economic Behavior and Organization,* Vol. 12 (1989), pp. 305–327.

Jacob, B. and Lefgren, L. "The Impact of Research Grant Funding on Research Productivity." Working Paper no. 13519, NBER, 2007.

Kaplan, D. "How to Improve Peer Review at NIH." *Scientist,* Vol. 19 (2005), p. 10.

Kolata, G. "Grant System Leads Cancer Researchers to Play It Safe." *New York Times*, June 28, 2009.

Lazear, E. "Performance Pay and Productivity." *American Economic Review,* Vol. 90 (2000), pp. 1346–1361.

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

Lotka, A.J. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences,* Vol. 16 (1926), pp. 317–323.

Manso, G. "Motivating Innovation." *Journal of Finance*, Vol. 66 (2011).

March, J.G. "Exploration and Exploitation in Organizational Learning." *Organization Science,* Vol. 2 (1991), pp. 71–87.

McKnight, S.L. "Unconventional Wisdom." *Cell,* Vol. 138 (2009), pp. 817–819.

Merton, R.K. "The Matthew Effect in Science." *Science,* Vol. 159 (1968), pp. 56–63.

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

Papke, L.E. and Wooldridge, J.M. "Econometric Methods for Fractional Responses with an Application to 401(k) Plan Participation Rates." *Journal of Applied Econometrics,* Vol. 11 (1996), pp. 619–632.

Robins, J.M. and Rotnitzky, A. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association,* Vol. 90 (1995), pp. 122–129.

[20] In the case of hits, the corresponding specification includes MAX_QNTL—the citation quantile corresponding to scientist *i*'s most impactful article published prior to appointment—as an independent variable.

ROSENBAUM, P.R. AND RUBIN, D.B. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika,* Vol. 70 (1983), pp. 41–55.

RUBIN, D.B. "Characterizing the Estimation of Parameters in Incomplete-Data Problems." *Journal of the American Statistical Association,* Vol. 69 (1974), pp. 467–474.

SANTOS SILVA, J.M.C. AND TENREYRO, S. "The Log of Gravity." *Review of Economics and Statistics,* Vol. 88 (2006), pp. 641–658.

SAUERMANN, H. AND COHEN, W.M. "What Makes Them Tick? Employee Motives and Firm Innovation." *Management Science,* Vol. 56 (2010), pp. 2134–2153.

SIMONTON, D.K. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist.* New York: Cambridge University Press, 2004.

SMITH, J.A. AND TODD, P.E. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics,* Vol. 125 (2005), pp. 305–353.

TIAN, X. AND WANG, T.Y. "Tolerance for Failure and Corporate Innovation." Working Paper, Indiana University, 2010.

WEITZMAN, M. Recombinant growth, *Quarterly Journal of Economics*, Vol. 13(1998), pp. 331–360.

WOOLDRIDGE, J.M. "Quasi-Likelihood Methods for Count Data." In M.H. Pesaran and P. Schmidt, eds., *Handbook of Applied Econometrics,* Vol 2. Oxford: Blackwell, 1996.

———, "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification." *Portuguese Economic Journal,* Vol. 1 (2002), pp. 117–139.