Project Selection in NIH: A Natural Experiment from ARRA[☆]

Hyunwoo Park[a], Jeongsik (Jay) Lee[b,*], Byung–Cheol Kim[c]

[a] *Georgia Institute of Technology, School of Industrial & Systems Engineering and Tennenbaum Institute, 755 Ferst Dr. NW, Atlanta, GA 30332-0205, hwpark@gatech.edu*

[b] *Drexel University, LeBow College of Business, 3141 Chestnut Street, Philadelphia, PA 19104, jaylee@drexel.edu*

[c] *Georgia Institute of Technology, School of Economics, 221 Bobby Dodd Way, Atlanta, GA 30332–0615, byung-cheol.kim@econ.gatech.edu*

Abstract

Using a natural experiment in research funding by the National Institutes of Health (NIH) following the American Recovery and Reinvestment Act (ARRA) of 2009, we study the NIH's revealed preference in project selection. We do so by comparing the characteristics of the projects additionally selected for funding due to an unexpected increase in resources under the ARRA with those supported through regular NIH budget. We find that the regular-funded projects are on average of higher quality, as measured by the number of publications per project and the impact of these publications, than ARRA-funded projects. Moreover, compared to ARRA projects, regular projects are more likely to produce highest-impact articles and exhibit greater variance in research output. The output from regular projects also seems more closely fits the intended purpose of funding. The differences in project quality are largely explained by observable attributes of the projects and research teams, suggesting that the NIH may use these attributes as cues for discerning underlying project quality. In addition, ARRA projects are more likely than regular projects to involve investigators with past grant experience. Many of these inter-group differences are specific to R01 grants, the largest funding category in the NIH. Overall, these results suggest that the NIH's project selection appears generally in line with its purported mission. In particular, our findings do not support the frequent criticism that the NIH is risk-averse and favors experienced investigators.

Keywords: public research funding, project selection, natural experiment, National Institutes of Health, American Recovery and Reinvestment Act, revealed preference
JEL: D2, H0, H5, O3

[☆]The authors thank the Editor and the two anonymous referees for their insightful comments and suggestions. All errors remain ours.

[*]Corresponding author

1. Introduction

The National Institutes of Health (NIH) as a part of the U.S. Department of Health and Human Services is the largest public source of funding for biomedical and health-related research in the world. Few disagree on the institution's crucial role in improving human health, economic growth, and job creation. However, the adequacy of project selection in the NIH has been more controversial. For example, Azoulay et al. (2011) find that the scientists supported by the Howard Hughes Medical Institute (HHMI), a U.S. non-profit private biomedical research organization, produce high-impact articles at a much higher rate than a control group of similarly accomplished NIH-funded scientists. It has been argued that the NIH tends to be risk-averse and peer reviews are too conservative, putting greater weights on the likelihood of success rather than the potential impact of the projects (Zerhouni, 2003; Nurse, 2006). These selection criteria thus tend to put young and less experienced scientists at a disadvantage in securing grants from the NIH (Weinberg, 2006). They also generate perverse incentives for scientists to strategically submit proposals that are already close to completion, rather than their most innovative applications (Zerhouni, 2003; Nurse, 2006; Stephan, 2012). To its credit, the NIH has actively sought to address these concerns in various ways. For instance, it recently began offering "High-Risk High-Reward (HRHR)" Funding Opportunity Announcements (FOAs) and has initiated special programs such as the Director's Pioneer Awards and New Innovator Awards (Zerhouni, 2003; Austin, 2008). As their titles suggest, these initiatives are specifically designed to promote highly innovative yet risky research ideas. Nonetheless, concerns seem to remain unabated.

Given the NIH's enormous influence on individual scientists' career as well as on the national-level innovation in the biomedical field, it is imperative to ensure that the NIH supports the "right" projects. If the above-mentioned allegations were true, the NIH may not be doing their job and public resources may be used inefficiently. It is thus important to accumulate evidence on the effectiveness of the NIH as a public institution, particularly given their crucial role in biomedical research. However, despite the frequent criticism on the NIH's review criteria and selection process, we know surprisingly little about the NIH's

"preference" in its project selection. Because the NIH does not disclose how it actually selects projects, we simply have little way of confirming or disproving these allegations, let alone assessing the institution's overall effectiveness. Hence, what seems still in order is to find a way to systematically investigate the nature of projects or scientists that the NIH chooses to support.

This paper is our attempt to do just that. Specifically, we take a revealed preference approach (Samuelson, 1938) to deduce the NIH's preference from the observed funding decisions. Our logic behind this approach is as follows. If choices $\mathbf{x}$ and $\mathbf{y}$ are both eligible for selection and $\mathbf{x}$ is chosen over $\mathbf{y}$, then $\mathbf{x}$ is revealed to be deemed at least as good as $\mathbf{y}$. Thus, by comparing the characteristics associated with choice $\mathbf{x}$ relative to those with choice $\mathbf{y}$, one can infer the decision-maker's preference. Several challenges arise in applying this approach to our context because the following conditions have to be satisfied: (i) projects $\mathbf{x}$ and $\mathbf{y}$ are both identified as eligible for funding when $\mathbf{x}$ is revealed preferred to $\mathbf{y}$ (i.e., both projects belong to a feasible set); (ii) $\mathbf{y}$ is also funded eventually (so that the research output from the projects are comparable); and (iii) underlying characteristics of $\mathbf{x}$ and $\mathbf{y}$ are independent of the nature of the funding resource for $\mathbf{y}$ (i.e., the changes in funding resource are exogenous to the attributes of the projects). The American Recovery and Reinvestment Act (ARRA) in 2009 offers a natural experiment that successfully meets these requirements.

In February 2009, the 111th United States Congress signed on the ARRA that stipulated outlays of later-revised $831 billion as the economic stimulus package, which is the largest single economic stimulus program in the U.S. history. $10.4 billion of that package was allocated to the NIH to be spent within two years from the enactment. Considering the NIH's annual budget of around $25-30 billion, this additional fund was substantial (Steinbrook, 2009). The NIH accordingly disbursed most of the ARRA fund to extramural scientific research in the form of grants ($8.97 billion to 21,581 projects). In disbursing the ARRA fund, the NIH used two distinct categories, on top of its regular NIH grants. The first category is "ARRA Solicited," under which the NIH selected and funded projects from competing applications that were newly submitted in response to the ARRA FOA. Under the second category, "Not ARRA Solicited," the NIH selected and funded projects from

a pool of *past* applications that "received meritorious priority scores from the initial peer review process" and "received priority scores that could not otherwise be paid in FY 2008 or 2009."[1] In fact, the NIH explicitly acknowledges that it "extended beyond payline to pick them up."[2] In other words, these second-category projects were evaluated as having sufficient scientific merits but, due to budget constraints, could not be selected for funding initially; absent the ARRA enactment, these would never have been awarded NIH grants. This second category of projects forms the core of our study. Under this category, 3,869 projects ($1.41 billion) were awarded NIH grants in FY2009 and 628 projects ($0.32 billion) in FY2010. The ARRA thus disturbed the NIH funding mechanism temporarily but substantially.

Given the nature of the event, the ARRA experiment provides several important merits for our approach. First, the infusion of the ARRA fund was exogenous to the NIH's agenda (Chodorow-Reich et al., 2012; Wilson, 2012). Thus, the project selection is unlikely to have been influenced by some NIH-specific policy initiatives. Second, the ARRA event has accidentally revealed the projects that were deemed worthy of support but were not funded due to budget constraints (for convenience, we label them as "ARRA projects"). This set of projects thus belonged to the same risk set as the projects that initially cleared the hurdle and were selected for funding from the same pool of proposals (we label the latter as "regular projects"). Third, ARRA projects were also funded later by the NIH. Hence, both groups of projects are subject to a fair comparison. Lastly, because the NIH requires all funded projects to acknowledge their funding in all publications resulting from the projects, we can precisely identify the research output and link it to individual grants.

Exploiting this unusual setting to study the NIH's preference in project selection, we examine the following questions: (i) Does the NIH select higher quality projects?; (ii) Does the NIH prefer riskier projects?; (iii) What are the cues that the NIH uses to identify fundable projects?; (iv) How close are the selected projects to the intended purpose of funding?; and (v) Does the NIH favor experienced investigators? In addressing these questions, we also look at variations across different types of grants.

[1] http://grants.nih.gov/grants/guide/notice-files/not-od-09-078.html
[2] http://report.nih.gov/recovery/NIH_ARRA_Funding.pdf

We test the first question by examining the differences in research output between the two groups, in terms of their impact (measured by the dollar-adjusted number of citations of journal publications per project) and productivity (measured by the dollar-adjusted number of journal publications per project). We find that ARRA projects on average produce per-dollar publications of significantly lower impacts (13.8%) than regular projects and that this difference is primarily driven by R01 grants, the largest funding category among NIH grants. We find no difference in the (dollar-adjusted) number of publications overall, though for R01 the difference (19%) is again considerably in favor of the regular group. Moreover, regular projects are significantly more likely (by 2.2%p) than ARRA projects to produce articles of the highest impact, as defined by the probability of belonging to the top 5% of the citation distribution. This pattern holds across different grant types. In contrast, ARRA projects are no more likely than regular projects to "fail," as defined by producing zero publications. Overall, regular projects appear to be of higher quality than ARRA projects.

We next explore which group of projects exhibits higher "risk." In fact, the NIH claims that it considers risk as an important selection criterion along with impact (Austin, 2008). Our analysis above indicates that the regular group is more prone than the ARRA group to produce right-tail outcomes though the likelihood of left-tail outcomes is similar between the two groups. We further conduct an inter-group comparison of the variance in research output in terms of the citation-based impact and the quantity-based productivity. We find that the distribution of impact (for a given FOA) exhibits a greater dispersion for regular projects than for ARRA projects and this difference is not specific to a particular grant type. No difference in dispersion is found for the number of publications either at the aggregate level or between grant types. Thus, on the whole, regular projects seem to exhibit greater variations in research output relative to ARRA projects. Taken together, these findings do not render support to the criticism that the NIH is too risk-averse (e.g., Zerhouni, 2003). We confirm these results on a subset of projects that started in the same time period and hence are less subject to differences in time window.

The next question concerns potential cues that the NIH might use to determine which projects are more promising and hence warrant funding among other projects. Obviously,

we do not have data on the actual criteria or check points that the NIH uses to evaluate each project. Instead, we employ a set of observable attributes of projects and investigators and relate them to research output to see which factor better explains the characteristics of the output. We find that, among other factors, team size, recent grant money and the institution-level grant award history positively explain the impact of research output. In contrast, team size is the only significant correlate of the quantity-based productivity. Interestingly, when these observable attributes are controlled for, the differences in research output between regular projects and ARRA projects almost disappear. This suggests that these project- and investigator-level attributes help predict the outcomes of the projects reasonably well and hence the NIH might be using them as useful cues for identifying promising projects.

Each FOA has its own objectives of funding. We thus compare between the groups the extent to which the project produces an output that is close to that funding purpose. For this, we devise a metric ("research fit") that quantifies the proximity between the objectives of an FOA and the content of the resulting publications from the project funded under the FOA. On this metric, the regular group of projects exhibits a significantly greater research fit than the ARRA group. This implies that, in choosing projects of higher quality and greater risk, the NIH does not make a trade-off with their fit with the funding objectives.

Lastly, we examine if experienced investigators, defined as those with a record of past NIH grants, are favorably treated in receiving NIH grants. We do this by looking at the probability that the applicant team with at least one principal investigator (PI) who previously received an NIH grant is awarded an ARRA grant, as opposed to receiving a regular grant. We find that, controlling for other project attributes, ARRA funding is significantly more likely given to experienced PIs. This pattern is particularly observed for R01 grants. The flip-side of this result is that, in the regular funding cycle, these experienced PIs are *less* likely to be selected for funding, all else equal. This result runs counter to the frequent allegation that the NIH favors PIs with proven records (e.g., Weinberg, 2006). In addition, the grant history of the PIs' institutions has no influence on the probability of an ARRA award, suggesting that the so-called Matthew effect (Merton, 1968) does not apply to NIH grants.

Reflecting the importance of NIH funding to scientific research, the selection process at the agency has been under academic scrutiny on various aspects such as researcher ethnicity (Ginther et al., 2011) and expertise (Li, 2012), political influences (Hegde and Mowery, 2008; Hegde, 2009), and peer review (Hegde, 2009; Azoulay et al., 2012; Nicholson and Ioannidis, 2012). We add to this literature by documenting the agency's revealed preference in selecting projects; our unique research setting provides a natural variation that allows for such an attempt. In spirit, our work is close to Bisias et al. (2012) who assess the efficiency of the NIH's funding allocation across disease categories by applying the modern portfolio theory to analyze its risk attitude.

More broadly, we join the recent policy debate on scientific research funding (Bourne and Lively, 2012; Fineberg, 2013; McDonough, 2013). In doing so, our study complements the body of literature on the effect of public funding on research output (Carter et al., 1987; Averch, 1989; Gordin, 1996; Arora and Gambardella, 2005; Jacob and Lefgren, 2011a,b; Benavente et al., 2012).[3] Our paper, however, is distinct from these studies in that our primary focus is not on estimating the effect of funding on research output per se. Tangentially, our study is also related to the literature on the policy evaluation of the ARRA program, which has so far focused almost exclusively on the program's effect on employment (Chodorow-Reich et al., 2012; Wilson, 2012).

2. NIH Grants and the ARRA Program

2.1. NIH Grants

The NIH is the largest single funding source for biomedical research in the world, accounting for 28% of the entire U.S. biomedical research (Moses III et al., 2005). An NIH-funded research led to development of innovative technologies such as the magnetic resonance imaging (MRI), and 138 NIH-supported researchers won the Nobel Prize in chemistry, economics, medicine, and physics.[4] Such evidence of achievements supports the view that the NIH is

[3] See Dorsey et al. (2010) and Moses III et al. (2005) for overall trends of scientific research funding, particularly those of the NIH.

[4] `http://www.nih.gov/about/` accessed on June 26, 2013.

a critical source of scientific development and economic growth by sponsoring academic research in health and igniting private sector innovation. For instance, Toole (2012) provides empirical evidence that NIH-funded basic research helps new drug developments in the pharmaceutical industry.

The NIH is a collective body of 27 institutes and centers (ICs) such as the National Cancer Institute or the Center for Information Technology. Each individual IC is responsible for administrating and disbursing research funding focusing on a specific health problem domain. Of the $25-30 billion annual budget, more than 80% is given to outside research communities (called "extramural" research grants) such as universities, colleges, and private research institutes. Besides these extramural grants, the NIH also operates its own research laboratories and about 10% of the budget goes to supporting these "intramural" research activities.

The overall process of NIH funding, illustrated in Figure 1, is the following. When the need for study in a specific area or domain is identified, one of the ICs issues an FOA. There are two major types of FOAs: the Request for Applications (RFA) and Program Announcements (PA). An RFA calls for research proposals in a narrowly defined area of study, while a PA aims to support projects researching in a broad area. A researcher (or a team of researchers) applies for grants upon noticing an FOA. All proposals must be submitted in response to an FOA. Researcher-initiated proposals are also required to refer to a specific FOA number. Once proposals are received, the NIH first examines through a peer review process the scientific merit that each proposal carries. Reviewers, selected to have no conflict of interest, grade each proposal using a 9-point grading system (in which 1 denotes 'exceptional' and 9 'poor'). The NIH provides reviewers with detailed guidelines for grading proposals. A council meeting then reviews the scores, sets the payline, and prioritizes projects. Proposals whose scores fall beyond the payline are not funded for the term. There are three (occasionally four) council meetings per fiscal year; accordingly, there are three standard due dates for proposals and review cycles. Depending on the timing within a fiscal year, the payline carries different weights in selecting projects. With the final budget unapproved at the beginning of fiscal year, the council sets the payline

conservatively to prepare for potential high-quality proposals in later cycles of the year. Most projects are funded and initiated towards the end of fiscal year when the final budget is determined. When the research output supported by NIH grants is published, the NIH requires the authors to acknowledge the financial support by citing the grant number to their publications. The U.S. Congress mandates the average length of NIH projects to be four years. Project end date may be extended only with a prior approval of the NIH, even if the extension request does not ask for additional funding.

Not all proposals deemed to have scientific merits are funded. Although undisclosing to the public, the NIH internally keeps record of the review scores of the proposals that earned meritorious scores (i.e., scores that deserve funding) but fell beyond the payline that is determined by funding availability. Once a proposal is selected and funded, the applicant becomes the PI of the project responsible to carry out the proposed research. Multiple researchers can jointly submit a proposal, in which case all the applicants become PIs with one of them designated as the contact PI.

Funded projects are classified by their activity and application type. The activity type, an alphabet followed by two-digit number or two alphabets followed by a single-digit number, characterizes the purpose of fund and how it will be spent. Examples include R01, R03, and P01. R01, the oldest and largest funding mechanism of the NIH, supports normal-size ($\sim$\$500,000) research projects proposed by investigators. R03 provide relatively smaller amounts of fund ($<$\$50,000 per year) to preliminary short-term research projects with an explicit non-renewal term attached. P01, on the other hand, funds the initiation of a program that addresses a broad area of biomedical study. The application type is another dimension of classification. Not all projects are proposed as new or short-term. A funded project that spans more than a year is the norm, not the exception. Thus, every year the PIs of an existing project must submit a renewal application to secure continued funding. Renewal applications may or may not go through a competitive review process, depending on initial terms or other conditions. In some cases, a project can request additional funding as administrative supplements. All these different types of applications are labeled accordingly and recorded as separate projects for that fiscal year. These fine-grained classification systems and detailed

labeling information per project allow us to examine differences across grant types and in some specifications to control for much of the unobserved heterogeneity across types of funding and research activities.

2.2. ARRA and NIH Funding

In February 2009, the U.S. government earmarked $831 billion for the economic stimulus package based on the ARRA enacted by the 111th U.S. Congress. As a result, the US government raised more than $800 billion and have paid out $290.7 billion for tax benefits, $254.5 billion for contracts, grants, and loans, and $250.8 billion for entitlements.[5] One of the five main purposes of the ARRA stated in the Act[6] is to make "investments needed to increase economic efficiency by spurring technological advances in science and health." The law also explicitly directs to "commence expenditures and activities as quickly as possible." This single largest fund flowed into the U.S. economy through many government agencies including the NIH as part of the Department of Health and Human Services. Figure 2 describes the ARRA timeline for NIH-related events, some of which stress the urgency of expending the fund to stimulate the economy.

The NIH was appropriated to allocate $10.4 billion, of which $8.97 billion was spent as extramural research grants. Among these, $2.71 billion (30.2%) were awarded through ARRA-specific funding opportunities such as Challenge Grants and Grand Opportunity Grants. $1.93 billion (21.5%) were granted to existing projects as administrative supplements and $2.31 billion (25.8%) were awarded to ARRA-funded projects taking more than 2 years via noncompeting continuation mechanism. A notable awarding mechanism, which is the focus of our study, that allocated $1.73 billion (19.3%) exclusively targeted the previously reviewed applications that had been submitted to funding opportunities unrelated to the ARRA.

The NIH released the notice on April 3, 2009 stating that it would consider funding proposals that had previously been reviewed and earned meritorious scores, but had not

[5] `http://www.recovery.gov/` as of May 31, 2013.
[6] `http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/pdf/PLAW-111publ5.pdf`.

been funded. All but a few of these proposals had been submitted to an FOA unrelated to the ARRA. Thus, these researchers had submitted proposals without knowing that the NIH would soon obtain substantial amount of additional funding which must be expended "as quickly as possible." The NIH awarded this fund to the proposals that had received meritorious scores from the review but had fallen below the payline in fiscal years of 2008 and 2009. In effect, the NIH temporarily extended beyond the payline to "pick up" projects that would not have been funded without payline extension, and utilized some of the ARRA funding to support these projects. This temporary shift, triggered by an exogenous event, incidentally revealed the proposals around the margin of the payline. Figure 3 illustrates how additional funding from the ARRA triggered payline extension and which group of projects were affected. As detailed in the next section, these additionally selected projects under the ARRA along with the projects selected under regular funding cycles for the same FOAs collectively form the subject of our empirical investigation.

3. Data and Sample

The project-level funding data come directly from the NIH. The NIH makes its research activities publicly available in line with the open government initiatives to ensure transparency in its operation. It provides a web interface for the public to browse the funded projects, as well as a bulk download channel for those who want to conduct a more systematic analysis. The entire project funding data span from 1985 to 2012. Our main dataset focuses on the projects funded in fiscal years 2009 and 2010, though we utilize project records in previous years to construct some of our variables. Each project record contains fiscal year, project number, administrating IC, activity code, application type, indication of funding by the ARRA appropriation, associated FOA number, project start and end dates, list of PIs, affiliation of the contact PI, education institution type, funding mechanism, and award amount. The NIH records the name of each PI and assigns a serial number to each PI for identification. The education institution type is the category of the contact PI's affiliation (e.g., School of Medicine or School of Arts and Sciences). The funding mechanism indicates the general purpose of the fund such as Research Projects, Training, or Construction.

Not only does the NIH compile and disclose detailed project records, but it also keeps track of the list of publications generated from its funded projects and provides a linking table that contains pairs of the publication identifier and the project number. The detailed bibliographic information of the publications listed in this linking table comes from the PubMed database. Each publication record contains authors list, affiliation of the corresponding author, journal title, ISSN, volume, issue, and year of publication. We also use forward citations data from Scopus®. By matching these data with the project-level funding data, we can identify how many publications result from a particular project and where and when these articles are published. The citations data allow us to construct a measure of impact of a research project funded by the NIH.

We collect all NIH projects funded in fiscal years 2009 and 2010. Using application type and funding mechanism, we filter down to only research projects initiated from a *new* application. We then drop grants for non-U.S. institutions and projects funded by non-NIH agencies. Because this sample only contains new application-based projects, administrative supplements to existing projects and continuation funding records are accordingly excluded.

Among ARRA-funded projects, we are left with both ARRA-solicited and non-ARRA-solicited projects. The ARRA-solicited projects are the ones that are selected from the applications associated with ARRA-specific FOAs. We remove these projects using the corresponding FOA numbers.

The purpose of our analysis is to examine the NIH's preference in project selection primarily by comparing between the research output of projects supported by the ARRA fund and that of projects supported by regular NIH grants. To that end, we need to ensure that these two groups of projects are intended to solve the same problem defined by the FOAs. Hence, among the projects supported by regular NIH grants, we exclude all projects that do not share an FOA number with one of ARRA-funded projects in the sample. Our final sample thus contains 12,656 projects (2,790 ARRA projects and 9,866 regular projects).[7] To merge project data with the publication list, we search the link table

[7]We remove two projects whose project start date is in FY2008, which seem apparent coding errors.

with project numbers. A project can have multiple publications. The entire set of matched publications thus contains 33,793 articles.[8]

4. Empirical Analysis

4.1. Project Quality

We first examine if the NIH selects higher quality projects on average. We do this by comparing the impact and the quantity of research output (i.e., journal publications) between regular projects and ARRA projects. Given that both groups of projects ultimately received NIH grants but regular projects have been selected first, any superiority of the regular group on our quality measures would indicate the NIH's preference for (or its ability to select) higher quality projects.

In addition to investigating this on the aggregate level, we also examine if the pattern (if any) varies across grant types. Our sample includes 14 activity categories but the top three categories are R01, R21, and R03 (in a descending order of frequency), which together account for over 90% of the projects. According to the NIH[9], R01 is the classic type of research grant awarded to major research projects, R03 is intended to support pilot or feasibility studies, and R21 is for new and exploratory research that may involve high risk but may lead to a breakthrough. Both R03 and R21 projects are encouraged to be followed up by R01 applications. In sum, R01 is a mechanism to support research that may show some initial results, while R03 and R21 are to seed-fund risky but potentially highly impactful research initiatives. Thus, the nature of projects and hence the characteristics of research output may well vary between R01 and the other two (R03 and R21) grant types.

For the measure of impact, we calculate for each funded project the maximum number of citations of all journal publications that result from the project (e.g., Benavente et al., 2012). For the quantity-based productivity measure, we count the number of publications from a given project. Because projects vary in the resources expensed for the research, we

[8]We remove articles published prior to 2009 as they seem obvious errors in the raw data.

[9]`https://www.nichd.nih.gov/grants-funding/opportunities-mechanisms/mechanisms-types/comparison-mechanisms/Pages/default.aspx`.

also compute the normalized measures of impact and quantity by dividing the raw measures by the total cost of project (in million U.S. dollars). To distinguish between the groups, we define a dummy, $ARRA$, that equals one if a given project is funded under the ARRA and zero if it is supported through regular NIH grants. Table 1 provides the summary statistics of these variables, along with those of all other variables used in our study.

For this part of analysis, we regress the measures of project quality on the ARRA dummy, with FOA-level fixed effects included to minimally control for variations across funding decision units at the NIH.[10] Hence, our regression takes the following form:

$$y_{ij} = \alpha + \beta \cdot ARRA_i + \vartheta_j + \varepsilon_{ij} \tag{1}$$

where y_{ij} is our measure of project i's quality, $ARRA_i$ indicates the ARRA status of project i, ϑ_j is a dummy for FOA j that project i belongs to, and ε_{ij} is an idiosyncratic error term. To examine the differences between grant types, we interact the $ARRA$ dummy with $R01_i$, which equals one if project i is supported by an R01 grant and zero otherwise.

Table 2 reports the results. Column 1 shows that the projects funded under the ARRA on average generate publications of 9.9% fewer citations per project, as compared to the projects supported by regular grants. When normalized by project cost, the gap in citations increases to 13.8% (column 3). The inter-group difference in the number of publications (columns 5 and 7) is somewhat weaker but still sizable (about 9% fewer publications per project-dollar for ARRA projects). These results imply that regular-funded projects are not only more productive than ARRA-funded projects but the research output from regular projects also commands significantly higher impacts, compared to that of ARRA projects. Insofar as our measures represent the inherent quality of projects, our results suggest that the NIH has a reasonable capacity to sort and prioritize grant proposals based on the quality of the projects.

[10] By this, we are essentially comparing the raw values of quality measures between the two groups. This is because, given our purpose of examining if the projects funded under the ARRA are fundamentally different from those funded through regular grants, controlling for other observable attributes masks such differences in underlying quality of projects.

A breakdown analysis by grant type indicates that the differences in research impact between the groups, particularly in terms of per-dollar citations, come primarily from the R01 type grant (columns 2 and 4). R01 also seems largely responsible for the difference in the quantity of research output (columns 6 and 8). As indicated by the coefficients on the R01 dummy, projects supported through R01 grants on average generate more and higher-impact research output than those supported through other types of grant. Extending the logic above, the differences between grant types imply that the NIH is relatively good at identifying "better" projects in the R01 category, but may be less so in other categories that involve more exploratory research proposals on nascent opportunities.[11]

4.2. Project Risk

The analysis in the previous section shows that regular projects are on average of higher quality than ARRA projects, suggesting that the NIH gives a priority to higher-quality proposals. As mentioned earlier, the NIH is also interested in promoting high-risk projects, in the hope that such projects will produce major breakthroughs even if that also means higher chances of failure. We thus examine in this section the NIH's risk preference in project selection. We do this in two ways: first, by comparing between the two groups of projects the likelihood of tail outcomes (i.e., extreme successes and complete failures) and second, by contrasting the variances in our measures of quality of research output.

For the first part of analysis, we construct for each project two binary indicators of tail outcomes: the *Top 5%* dummy, which indicates if the project's maximum number of citations (adjusted by the number of months since publication) makes into top 5% of all publications in our sample[12]; and the *No publication* dummy, which indicates if the project

[11]There may be two reasons for this. One is that exploratory research proposals are inherently more difficult to evaluate their potential even though the NIH attempts to carefully prioritize projects based on the assessed quality. The other is that in these exploratory grant categories, the NIH purposefully selects projects that are not guaranteed to succeed, rather than based on quality assessment. We are unable to discern which of the two is more likely.

[12]We would ideally want to construct this measure based on the entire publication pool of articles in biomedical research. However, defining the boundary of biomedical research is challenging and thus collecting all publications in the field is practically infeasible. Instead, we use as the base all publications produced from the projects in our sample. This in fact makes our definition of top 5% quite stringent because the

produces zero publication. Hence, the former represents a right-tail outcome (i.e., extreme success) and the latter a left-tail outcome (i.e., complete failure).

We estimate an analog of Equation (1), with the dependent variable now being one of the two dummies of tail outcomes. Table 3 presents the results. On average, ARRA projects are significantly less (by 2.2%p) likely to produce a highly impactful publication than regular projects (column 1). In contrast, no inter-group difference is found for the likelihood of zero publication (column 3). Thus, regular projects as a group exhibit a greater tendency to generate tail outcomes, particularly those on the right tail of the distribution. By grant type, the R01-supported projects in general are considerably more prone to generate a top-5%-class research output and are much less likely to fail, relative to projects supported through other types of grants. However, these distinctions hold equally for regular projects and ARRA projects, as indicated by the insignificant coefficients on the interaction terms (columns 2 and 4). Taken together, across all grant types, regular projects appear to command a higher risk of producing tail outcomes than ARRA projects.

We further investigate the risk profile of selected projects by looking at the degree of variations in research output between the two groups. We measure variations by the statistical variance of our measures of project quality (i.e., citation-based impact and quantity-based productivity). For this part of analysis, we collapse the data to the FOA level and compute the variances of output measures for each FOA by the ARRA status.[13] Hence, for each FOA and for each measure of quality, we obtain two values of variance, one for the regular project group and the other for the ARRA group. We then estimate an analog of Equation (1), with the dependent variable being one of our measures of quality as in Table 2.

Columns 1 and 3 of Table 4 show that, for the same FOA, the projects supported through

projects in our sample are already among a highly selective group of projects that secured funding from the NIH through a rigorous scientific review process. Therefore, the top 5% in our sample could well indicate an even higher rank in the percentile distribution based on a (hypothetical) pool of full publications.

[13]Note that variances can be calculated only when there are multiple observations. Because a majority of projects in our sample has either zero or a single publication, project-level variances are almost meaningless. Further, since each FOA is designed to address a certain area of problems (in the case of PA) or a specific problem (in the case of RFA), all projects under the same FOA are in principle to provide a solution to the same set of problems. Thus, defining the variance at the FOA level helps reduce idiosyncratic variations across projects.

regular grants exhibit a significantly greater (by 18-25%) variance in citation-based research impact than the projects selected under the ARRA. The inter-group difference in the quantity of research output also indicates the same direction, but is not statistically significant. No difference seems to exist in the pattern across grant types, even for the research impact (columns 2 and 4). On the whole, across all types of grant, regular NIH projects are considerably less predictable than ARRA projects in the quality of research output (at least on the aspect of research impact). Therefore, to the extent that the propensities of tail outcomes and the variances of research output represent the relative "risk" of the projects, we can interpret this section's results as suggesting that the NIH gives a priority to higher-risk proposals.

4.3. Robustness Checks on a Cohort Sample

Our analysis in the previous sections uses the full sample of projects. However, due to differences in the timing of cost disbursement, the projects in our sample have different time windows for producing research output. In particular, for the institutional reasons explained in Section 2, even for the same fiscal year ARRA projects systematically started later than non-ARRA projects (Figure 4). This difference in time window could affect our measurement of project output, particularly the quantity side of it. Thus, as a robustness check, we restricted the sample to the projects that started in the same time period (May 1-Sep. 30, 2009) and repeated the same analyses as we did on the full sample.

On project quality, presented in Table 5(a), we find results that are qualitatively similar to those obtained from the full sample analysis (Table 2). In fact, the differences between the groups become starker on this cohort sample. Compared to ARRA projects, regular NIH projects on average generate research output with significantly higher (14-19%) impacts and result in a greater (8-13%) number of publications per project. Notably, the significance level of the difference in the output volume has increased, underscoring the importance of harmonizing the project time between the two groups. There is even stronger evidence that these differences in project quality are largely specific to the R01 category.

We also find similar patterns on project risk (Table 5(b)): regular NIH projects exhibit

greater variances in research output than ARRA projects, at least in terms of citation-based impact. The overall differences between the two groups are statistically weaker, though the economic significances remain generally comparable to those in Table 4. Once again, we find no inter-group difference across grant types, as the coefficients on the interaction terms stay insignificant throughout. We suspect that the overall weakening of the statistical significance on this cohort sample may be due to a considerable reduction in the sample size (by about 28%). Nonetheless, the previous sections' results on project quality and risk seem generally robust to differential timing of project start.

4.4. Correlates of Project Quality

In this section we investigate the factors that help explain the differences in research output between regular projects and ARRA projects. This will allow us to speculate on the potential cues that the NIH might be using to identify and select promising proposals. Specifically, we re-estimate Tables 2 and 3 with a full set of controls for available observable attributes of the project and the investigator(s). The regression model thus takes the following form:

$$y_i = \alpha + \beta \cdot ARRA_i + \mathbf{x}_i\boldsymbol{\gamma} + \varepsilon_i \tag{2}$$

where y_i is one of our measures of project i's output characteristics; $ARRA_i$ is a dummy indicating the ARRA status of project i; $\mathbf{x}_i$ is a vector of observable attributes of project i including (log) grant size (in U.S. dollar), (log) number of unique authors, a dummy indicating whether project i ends within two years, a dummy indicating whether project i is funded in FY2010, the number of PIs associated with project i, a dummy indicating whether any of the PIs has an experience of any NIH grant in the past, (log) mean grant amount that the PIs have received in the preceding five years, a dummy indicating no grant in the preceding five years, the number of grants awarded to the organization in the preceding five years, the number of months since publication, and the activity-FOA-IC-education institution type-fiscal year dummies; and ε_i is an idiosyncratic error term. The

number of unique authors, identified from reported publications, captures the size of lab that the PI(s) operate. To identify a PI's experience in NIH grants, we search the entire grant data between 1985 and 2008 (or 2009). If any of the PIs of projects funded in FY 2009 (or 2010) appears in project records between 1985 and 2008 (or 2009), then we mark the project as including an experienced PI. The amount of NIH grants of PIs in the preceding five years (2004-2008) measures the PIs' recent funding track record. For projects with multiple PIs, we take the mean of all PIs' recent grant amounts. Since this variable is left-censored at zero, we include an indicator of whether the value is zero (i.e., no grant in 2004-2008). The number of grants that the contact PI's institution received in the preceding five years (2004-2008) represents the institution-level quality. The time since publication controls for differences in citation window. Lastly, the joint fixed effects between activity type (R01, R03, etc.), FOA number (RFA or PA's serial number), IC code (National Institute of Allergy and Infectious Disease, National Cancer Institute, etc.), education institution type (School of Medicine, School of Arts and Sciences, etc.), and fiscal year of application help control for other potential unobserved heterogeneity across grants.

Table 6 reports the results from this analysis. Among the explanatory variables, team size, measured by the number of unique authors, is estimated to be the strongest correlate of project outcomes: projects with a greater team size on average generate more and higher-impact publications (columns 1 and 2), are more likely to produce the highest-impact research (column 3), and are less likely to fail (column 4). In contrast, project cost, once team size is controlled for, performs poorly in explaining research output, and in fact is negatively related to quantity-based output measures (columns 2 and 4). Not surprisingly, having more research resources secured by the investigators recently (conditioning on having received some grants) helps produce significantly more impactful research output (column 1), though it does not increase the likelihood of highest-impact output (column 3). It also has no influence on the quantity of publication (column 2) or the likelihood of project failure (column 4). The institution-level quality is a positive and significant correlate of research impact (in terms of both the number of citations and the likelihood of receiving top 5% citation). Interestingly, when none of the PIs associated with the project has a recent grant

record, the project tends to produce more impactful research output (column 1). This implies that, all else equal, new PIs might perform well in generating impactful research, if not more publications. Among other variables, projects with a shorter time window produce fewer publications (columns 2 and 4), while the impact of their output is generally comparable to that of longer-term projects. Projects with earlier publications elicit more citations, validating the importance of controlling for the citation window (column 1).

Notice that when these project- and investigator-level attributes are accounted for, the coefficient on the ARRA dummy largely loses significance (that for the top 5% indicator stays significant but only marginally). In other words, once we account for the observable attributes of the project, regular NIH projects appear qualitatively similar to ARRA projects. This result implies that these project-level attributes are collectively highly correlated with the underlying quality of projects. Hence, they may indeed be useful cues for the NIH in identifying and selecting promising projects among numerous grant proposals. Pushing this a bit further, one can interpret this result as suggesting that the NIH may have limited additional insight beyond these observables in distinguishing good projects from less promising ones.

4.5. Research Fit

Each funding opportunity of the NIH is an attempt to address a specific research problem. We thus explore how closely the selected projects fit the intended purpose of the funding. This question is particularly relevant because selections of high-impact, high-risk projects may be pursued at the expense of the original funding purposes. If such were the case, superiority of projects simply based on our measures of research output may not necessarily indicate a quality work of project selection on the part of the NIH. To examine this question, we measure the closeness of projects to the purpose of a given FOA ("research fit") and compare it between regular projects and ARRA projects. If the NIH did not trade it off with project quality in selecting projects, we should observe a similar difference in research fit between the two groups as found in the previous analyses.

We construct a measure of research fit by textually comparing the FOA research ob-

jectives and the abstracts of publications from the corresponding projects. For the textual analysis, we use the term frequency-inverse document frequency method (Manning et al., 2008; Bird et al., 2009; Rehurek and Sojka, 2010). Note that this measure is defined *ex post* because we use "publication" abstracts instead of grant proposal abstracts. A publication abstract is by definition written after the project is funded, and hence the investigators have much less incentive to intentionally make it close to the FOA's stated objectives. Thus, using publication abstracts, rather than proposal abstracts, to capture the content of projects better serves our purpose. We provide in the Appendix the detailed process of constructing this variable. With this measure as the dependent variable, we estimate an analog of Equation (1).

As shown in Table 7, ARRA projects on average exhibit a significantly lesser fit with the FOA research objectives (column 1). This implies that regular NIH projects as a group produce research output that is more closely aligned with the purpose of funding than the output from ARRA projects. There is no difference in the pattern between grant types, as the coefficient on the interaction term is insignificant (column 2). Therefore, across all grant types, the NIH's selection of high-quality high-risk projects, as evidenced in previous sections, seems achieved within the range of closely meeting the objectives of the funding.

4.6. Preference for Experienced PIs

A final piece of our analysis concerns if the NIH favors experienced investigators in selecting projects. In fact, this is one of the frequent allegations used for questioning the NIH's effectiveness in allocating resources (e.g., Weinberg, 2006). If true, this unwarranted favoritism would indeed stifle young scientists, potentially miss opportunities of great promises, and ultimately lead to an overall decline of the biomedical field. Even our results in the previous sections suggest that a favorable treatment for experienced PIs would not be justifiable because the projects led by experienced PIs do not result in superior research outcomes (Table 6). Given the importance of the question, a systematic investigation of this aspect seems in order. Our approach to that question is to examine the probability that the applicant team involving any PI who has a history of past NIH grants is awarded an ARRA grant, relative

to that of receiving a regular grant.

We estimate a project-level regression model of the following form:

$$y_i = \alpha + \mathbf{z}_i\boldsymbol{\gamma} + \varepsilon_i \tag{3}$$

where y_i is a dummy indicator of project i's ARRA grant status (i.e., one if ARRA-funded and zero otherwise); and $\mathbf{z}_i$ is a vector of explanatory variables including a dummy indicating whether any of the PIs has an experience of any NIH grant in the past ("Existing PI"), the number of grants awarded to the organization in the preceding five years, the number of unique authors, research fit, and the activity-FOA-IC-education institution type-fiscal year dummies. The baseline represents a regular NIH grant, hence a negative coefficient on the *Existing PI* dummy would suggest a greater chance of grant award for experienced PIs in the regular funding cycle.

Table 8 presents the results. Most notably, the coefficient on the *Existing PI* dummy is significantly positive (column 1). Projects involving at least one experienced PI are 4.5%p more likely given an ARRA grant than a regular grant, controlling for other factors. The flip-side of this result is that in the regular funding cycle, these experienced PIs are *less* likely to be selected for funding. This is in stark contrast to the frequent allegation that the NIH gives a favor to PIs with proven track records.[14] Moreover, the institutional "reputation," represented by the institution-level grant record in the preceding five years, seems to have no influence on the probability of an ARRA grant. Thus, at least in our data, we find no evidence of the Matthew effect (Merton, 1968) that is often found in other settings (e.g., Bhattacharjee, 2012).

Looking at other variables, ARRA projects on average have fewer unique authors but have greater project costs than regular projects. This suggests that the NIH may have given a priority to projects that are smaller but could expend more money in research. It is intriguing that ARRA projects are smaller in team size than regular projects. Considering

[14] In fact, our result is consistent with the NIH's own data: the investigators who received the top 20% of funding in 2009 had an average of only 2.2 grants each (Rockey, 2013).

that the ARRA was primarily aimed at boosting employment, granting larger amount of money to smaller-size teams might have helped achieve the original purpose of the ARRA. Recall that, in our previous analysis, team size was a strong indicator of the quality of research output, whereas grant size was not. This interpretation is thus consistent with the results in Table 2 that, without accounting for project-level attributes, ARRA projects are on average associated with lower quality research output.

A split-sample analysis by grant types (columns 2 and 3) indicates that these results are entirely driven by R01 grants (the regression model for the R03 and R21 sample is not even properly specified).

5. Discussion

How does the NIH select projects? Our analysis exploiting a natural experiment setting from the ARRA suggests that in the regular funding cycle, the NIH tends to opt for a high-risk, high-return portfolio with greater likelihoods of breakthrough research outcomes. The selected projects also seem aligned well with the funding objectives. Some project- and investigator-level attributes effectively explain the characteristics of research output from the funded projects, suggesting that these may be useful cues for the NIH to identify high quality proposals. We find no evidence that the grant history of investigators or their affiliated institutions provides an advantage in regular grants awards. There is some heterogeneity across grant types, though R01 is generally the one driving most of the observed patterns.

A natural interpretation of our results is that the NIH may be doing a reasonable job in selecting and funding promising yet uncertain projects. In particular, our findings counter to the frequent allegations that the NIH is risk-averse, preferring surer bets, and gives disproportionately more favors to experienced investigators. Overall, the NIH's selection scheme seems to follow an efficiency trajectory as with the sudden increase in funding resource, they went after lower-risk, lower-return projects (relative to those selected in the regular funding cycle). This in turn implies that more funding resources to the institution may not necessarily lead to the support of higher-risk higher-return projects. We however notice some inconsistency in the selection principle between grant types. For instance, there

is no difference in the risk-return profile between regular grants and ARRA grants in the category of R03 and R21, both of which are designed to promote less proven scientists and projects. If the NIH were faithful to the stated goal, we should observe greater inter-group differences for these grant types than for R01. Nonetheless, given that R01 is the largest funding category at the NIH and is the major source of funding for scientists, our findings appear quite representative and hence help form direct policy implications for the practice.

Our approach in this study is to deduce the NIH's preference from their revealed choices in project selection between two rounds of selection from the same pool of grant proposals. In particular, we take the projects selected under the ARRA as the comparison group to characterize the underlying preference in the regular funding cycle. One might argue that, given the special (and unprecedented) circumstances that provoked the ARRA enactment, the projects selected under that initiative are not appropriate for a comparison group to deduce the NIH's behavior under "normal" circumstances. While this is a legitimate concern, we do not believe our choice of this comparison group to compromise our findings. First of all, the fairly low success rates for new proposals during the period (17.3% in 2009-2010) imply that there must have been many projects with sufficient scientific merits. Thus, the differences in project quality we document are unlikely to be orthogonal to the differences in the underlying distribution of quality, unless the NIH *intentionally* chose lower quality projects under the ARRA; we have no reason to think they would have done so. In terms of project risk, though we see a drop in variance for the ARRA group, there is no reason to expect that the NIH's goal of expeditiously disbursing the money should necessarily lead to a selection of less risky projects (i.e., projects whose expected output is more "centered" than those supported under regular funding). If the policy goal was the main consideration in project selection under the ARRA, the NIH may have given priority to the expendability and employability of the project, as in fact implied by one of our results (ARRA projects had fewer team members; perhaps the NIH thought that these teams had greater room for hiring people). Regardless, that need not be correlated with lower risk. A similar argument can be made for research fit: the ARRA's immediate policy goal need not have forced the NIH to forgo the original funding objectives in selecting projects. The case of the experienced

PI analysis is less clear. One may interpret the positive coefficient of the ARRA dummy (Table 8) to indicate that the NIH has given even more favor to experienced PIs in selecting projects under the ARRA, over and above what they normally give to them. While we cannot entirely rule out this possibility, we also wonder why the NIH would have behaved that way when they certainly knew that the temporary increase in money would soon go away. What better opportunity than this will there be for the NIH to make up for their adverse reputation by purposefully selecting more young and new scientists with the ARRA money? We therefore believe that ARRA projects serve as a reasonable basis for comparison and hence our findings are unlikely to suffer from any idiosyncrasy of the event.

Another potential issue with our analysis is that our sample is conditional on the project being selected. That is, we do not observe the full set of projects at risk of being selected. Given the highly competitive nature of NIH grants, there must be many projects that cleared the (more objective) threshold of scientific merits but were still not selected under the ARRA. Also, many promising but riskier proposals might not have made the cut in scientific merits. Thus, the profile of projects in our sample may not be representative of the underlying distribution of projects. For instance, it is possible that our sample represents the lower-risk group among the population of projects. Then, our finding of the NIH's preference for high-quality high-risk projects might lose some significance. However, we do not claim that the NIH *unconditionally* favors high-quality high-risk projects. Insofar as the peer review process, which is beyond the scope of our study, is performed objectively and that the NIH maintains their consistency in prioritizing the risk-return portfolio among the eligible proposals, our findings reasonably reflect the NIH's preference in project selection.

One might also argue that our analysis based on inter-group mean comparison is unfair to the ARRA group, as at least some of the projects in the regular group should be unambiguously superior ones. Thus, including these exceptional projects in calculating the means can give a natural advantage to the regular group. We do not disagree. In fact, a more interesting and relevant analysis would be to compare between the projects around the cutoff (prior to the ARRA), i.e., projects that are marginally funded in the regular funding cycle and those that are marginally unfunded then but funded later under the ARRA.

Unfortunately, our data do not allow for this approach. Nonetheless, even in this group-based comparison, employing a full set of observable attributes almost entirely eliminates the inter-group differences (see Table 6). Hence, while we acknowledge this as a limitation, we believe that our findings still claim considerable validity.

On a related point, lack of data renders our analysis necessarily crude. Absent data limitations, we would directly use the scores given to each project in the selection process. This would make our analysis much more precise. We are convinced that the question we are addressing in this paper is an extremely important one. Surprisingly, however, this crucial issue has eluded scholars in the field, even those who seem to have access to the full data of NIH proposals and review scores (both funded and unfunded ones) (e.g., Li, 2012). We thus make our best attempt to examine this issue by exploiting what is currently available to us. The findings we report in this study suggest that such an attempt is a worthwhile exercise.

A few caveats to our study seem also in order. First of all, the receipt of a particular grant may have a limited impact on the research productivity of the grant awardees. Prior studies have shown that the productivity of funded applicants near the selection cutoff is often fairly similar to that of unfunded applicants near the cutoff (Carter et al., 1987; Jacob and Lefgren, 2011b). It may be that, given the competitive nature of NIH funding, even researchers who fail to receive an NIH grant can easily find another source of funding to pursue their research (Jacob and Lefgren, 2011b). If so, the research output we observe from funded projects may be only partially influenced by NIH grants. This would limit our inference based on all observed output from the projects. Moreover, in the analysis, we use "realized" outcomes to infer the underlying characteristics of the projects. This approach admittedly ignores the uncertainty and possible idiosyncrasies in the process of scientific research. Some projects may have produced results that far exceeded their initial expectations. Some high quality projects may have failed to realize their full potential because of unexpected disturbances in the process. Thus, it would be naive to regard research output as a precise reflection of the underlying quality of the projects. Nonetheless, to the extent that these idiosyncrasies and uncertainty associated with scientific research equally apply to both groups of projects, our results should reasonably demonstrate the inter-group differences in the underlying project

quality. Lastly, our sample covers only very recent years (2009-2010). Considering the NIH's aggressive push toward high-risk high-return projects reflects fairly recent policy initiatives, our findings from these recent data may not apply equally to the years preceding our sample period. While this may suggest that the NIH's recent initiatives might have been working, it may also limit the generalization of our findings to a broader time span during which the frequent criticisms on the effectiveness of the NIH's selection process have been formed.

6. Conclusion

Important matters tend to elicit greater attention. The NIH's central role in promoting scientific research in the biomedical field has drawn corresponding interests in the process by which the institution selects projects and the effectiveness of the selection mechanism. Selection, by definition, means exclusion for at least some. Complaints can thus arise accordingly. The decade-long stagnation in funding resources at the NIH may have added to such tendency. With the recent effectuation of the U.S. budget sequestration, which called for a spending cut of over $85 billion in the fiscal year of 2013 only, the discordant voices may well be amplified. Allegations abound, but the findings from our study do not render support to such claims. Perhaps the NIH will want to be more transparent in selection criteria and communicate them more clearly to the interested audience. Without a doubt, a richer analysis supported by more fine-grained data will help bring additional insight into this essential process. Some limitations to our study notwithstanding, however, we hope to be able to claim a modest contribution to the policy discussion of this important institutional arrangement.

References

Arora, A., Gambardella, A., 2005. The impact of NSF support for basic research in economics. *Annals of Economics and Statistics* , 91–117.

Austin, F., 2008. High-Risk High-Reward Research Demonstration Project. *NIH Council of Councils* .

Averch, H.A., 1989. Exploring the cost-efficiency of basic research funding in chemistry. *Research Policy* 18, 165–172.

Azoulay, P., Zivin, J.S.G., Manso, G., 2011. Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics* 42, 527–554.

Azoulay, P., Zivin, J.S.G., Manso, G., 2012. NIH peer review: challenges and avenues for reform. *National Bureau of Economic Research Working Paper Series* No. 18116.

Benavente, J.M., Crespi, G., Figal Garone, L., Maffioli, A., 2012. The impact of national research funds: A regression discontinuity approach to the Chilean FONDECYT. *Research Policy* 41, 1461–1475.

Bhattacharjee, Y., 2012. NSF's 'Big Pitch' Tests Anonymized Grant Reviews. *Science* 336, 969–970.

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python. O'Reilly Media.

Bisias, D., Lo, A.W., Watkins, J.F., 2012. Estimating the NIH efficient frontier. *PloS one* 7, e34569.

Bourne, H.R., Lively, M.O., 2012. Iceberg alert for NIH. *Science* 337, 390.

Carter, G.M., Winkler, J.D., Biddle, A.K., 1987. An evaluation of the NIH research career development award. *RAND Corporation Report* , R–3568–NIH.

Chodorow-Reich, G., Feiveson, L., Liscow, Z., Woolston, W.G., 2012. Does State Fiscal Relief During Recessions Increase Employment? Evidence from the American Recovery and Reinvestment Act. *American Economic Journal: Economic Policy* 4, 118–145.

Dorsey, E.R., de Roulet, J., Thompson, J.P., Reminick, J.I., Thai, A., White-Stellato, Z., Beck, C.A., George, B.P., Moses III, H., 2010. Funding of US biomedical research, 2003-2008. *Journal of American Medical Association* 303, 137–143.

Fineberg, H.V., 2013. Toward a new social compact for health research. *JAMA* 310, 1923–4.

Ginther, D.K., Schaffer, W.T., Schnell, J., Masimore, B., Liu, F., Haak, L.L., Kington, R., 2011. Race, ethnicity, and NIH research awards. *Science* 333, 1015–1019.

Gordin, B., 1996. The impact of research grants on the productivity and quality of scientific research. *Working Paper. INRS. Montreal, Quebec.* .

Hegde, D., 2009. Political Influence behind the Veil of Peer Review: An Analysis of Public Biomedical Research Funding in the United States. *Journal of Law & Economics* 52, 665–779.

Hegde, D., Mowery, D.C., 2008. Politics and funding in the US public biomedical R&D system. *Science* 322, 1797–1798.

Jacob, B.A., Lefgren, L., 2011a. The impact of NIH postdoctoral training grants on scientific productivity.

Research Policy 40, 864–874.

Jacob, B.A., Lefgren, L., 2011b. The impact of research grant funding on scientific productivity. *Journal of Public Economics* 95, 1168–1177.

Li, D., 2012. Information vs. Bias in Evaluation: Evidence from the NIH. *Working Paper* .

Manning, C.D., Raghavan, P., Schutze, H., 2008. Introduction to information retrieval. Cambridge University Press Cambridge.

McDonough, J.E., 2013. Budget Sequestration and the U.S. Health Sector. *New England Journal of Medicine* 368, 1269–1271.

Merton, R., 1968. The Matthew effect in science. *Science* .

Moses III, H., Dorsey, E.R., Matheson, D.H.M., Thier, S.O., 2005. Financial anatomy of biomedical research. *Journal of American Medical Association* 294, 1333–1342.

Nicholson, J.M., Ioannidis, J.P.A., 2012. Research grants: Conform and be funded. *Nature* 492, 34–36.

Nurse, P., 2006. US biomedical research under siege. *Cell* 124, 9–12.

Rehurek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* , 46–50.

Rockey, S., 2013. Transparency: Two years of blogging the NIH. *Nature* 493, 298–299.

Samuelson, P., 1938. A note on the pure theory of consumer's behaviour. *Economica* 5, 61–71.

Steinbrook, R., 2009. Health care and the American Recovery and Reinvestment Act. *New England Journal of Medicine* 360, 1057–1060.

Stephan, P., 2012. Research efficiency: Perverse incentives. *Nature* 484, 29–31.

Toole, A.A., 2012. The impact of public basic research on industrial innovation: Evidence from the pharmaceutical industry. *Research Policy* 41, 1–12.

Weinberg, R.A., 2006. A lost generation. *Cell* 126, 9–10.

Wilson, D.J., 2012. Fiscal Spending Jobs Multipliers: Evidence from the 2009 American Recovery and Reinvestment Act. *American Economic Journal: Economic Policy* 4, 251–282.

Zerhouni, E., 2003. The NIH Roadmap. *Science* 302, 63–72.

Tables

Table 1: Summary Statistics

	Regular Projects					ARRA Projects				
	N	Mean	Std. Dev.	Min	Max	N	Mean	Std. Dev.	Min	Max
# of citations	6,818	24.64	39.93	0	649	1,766	19.93	37.71	0	988
# of citations per $M	6,818	963.23	72,663.93	0	6,000,000	1,766	90.17	449.31	0	17,326.98
(Dummy) Top 5%	6,818	0.07	0.25	0	1	1,766	0.04	0.19	0	1
# of publications	9,866	2.79	4.10	0	63	2,790	2.12	3.18	0	36
# of publications per $M	9,866	111.00	10,067.62	0	1,000,000	2,790	9.24	34.23	0	1,509.43
(Dummy) No Publication	9,866	0.31	0.46	0	1	2,790	0.37	0.48	0	1
Research fit	6,740	3.06	6.15	0	82.71	1,754	3.42	5.83	0	50.38
(Dummy) R01	9,866	0.62	0.49	0	1	2,790	0.37	0.48	0	1
(Dummy) R03 or R21	9,866	0.30	0.46	0	1	2,790	0.52	0.50	0	1
Total cost	9,866	364,400.44	308,496.22	1	5,934,572	2,790	325,756.23	294,923.37	1	5,566,450
# of unique authors	9,866	12.35	24.39	0	1,022	2,790	9.23	16.79	0	318
(Dummy) Within 2 years	9,866	0.35	0.48	0	1	2,790	0.97	0.18	0	1
Fiscal year	9,866	2,009.50	0.50	2009	2010	2,790	2,009.07	0.25	2009	2010
# of PIs	9,866	1.11	0.36	1	6	2,790	1.09	0.33	1	6
(Dummy) Existing PI	9,866	0.75	0.43	0	1	2,790	0.76	0.43	0	1
Mean cumulative $ grants for PIs (2004-2008)	9,866	202,173.73	237,561.81	0	3,502,544	2,790	217,905.72	261,220.79	0	4,620,253
(Dummy) No PI has a grant (2004-2008)	9,866	0.31	0.46	0	1	2,790	0.30	0.46	0	1
# of grants for organization (2004-2008, thousands)	9,866	1.75	1.84	0	6.76	2,790	1.67	1.79	0	6.76
# of months since published	6,818	30.08	9.44	11	56	1,766	30.77	9.23	12	55

Table 2: Analysis of Project Quality

	1	2	3	4	5	6	7	8
	(Log) Citation		(Log) Citation per $M		(Log) N_{pub}		(Log) N_{pub} per $M	
(Dummy) ARRA	-0.099**	-0.052	-0.138**	-0.005	-0.047†	-0.002	-0.090†	0.033
	(0.027)	(0.040)	(0.052)	(0.065)	(0.027)	(0.015)	(0.049)	(0.028)
(Dummy) R01		0.379**		0.540**		0.382**		0.512**
		(0.044)		(0.053)		(0.017)		(0.040)
ARRA × R01		-0.062		-0.211**		-0.058**		-0.193**
		(0.044)		(0.068)		(0.019)		(0.039)
Constant	2.504**	2.252**	3.497**	3.136**	1.369**	1.115**	2.333**	1.990**
	(0.006)	(0.031)	(0.011)	(0.039)	(0.006)	(0.011)	(0.010)	(0.027)
N	8,499	8,499	8,499	8,499	8,499	8,499	8,499	8,499
F-stat	13.41	42.42	6.96	72.68	2.91	189.82	3.37	92.61
Adj. R^2	0.06	0.06	0.04	0.04	0.06	0.07	0.12	0.13

Note: FOA-fixed effects are included in all models. Robust standard errors, clustered by FOA, are in parentheses. †, ** denotes statistical significance at 10%, and 1%, respectively. All models are conditioned on the project having at least one publication.

Table 3: Analysis of Tail Outcomes: Top 5% Probability and No Publication Probability

	1	2	3	4
	(Dummy) Top 5%		(Dummy) No Publication	
(Dummy) ARRA	-0.022**	-0.025**	0.027	0.010
	(0.005)	(0.007)	(0.018)	(0.020)
(Dummy) R01		0.030**		-0.343**
		(0.009)		(0.017)
ARRA × R01		0.008		-0.005
		(0.009)		(0.021)
Constant	0.062**	0.043**	0.317**	0.516**
	(0.001)	(0.006)	(0.004)	(0.011)
N	8,499	8,499	12,558	12,558
F-stat	22.33	13.12	2.32	200.96
Adj. R^2	0.02	0.02	0.09	0.11

Note: FOA-fixed effects are included in all models. Robust standard errors, clustered by FOA, are in parentheses. ** denotes statistical significance at 1%.

Table 4: Analysis of Variance in Project Quality (FOA-Level)

	1	2	3	4	5	6	7	8
	(Log) Citation		(Log) Citation per $M		(Log) N_{pub}		(Log) N_{pub} per $M	
(Dummy) ARRA	-0.183*	-0.197*	-0.252*	-0.264*	-0.018	-0.021	-0.052	-0.058
	(0.075)	(0.092)	(0.108)	(0.131)	(0.036)	(0.044)	(0.058)	(0.072)
(Dummy) R01		-0.210		-0.326		-0.035		-0.069
		(0.238)		(0.341)		(0.116)		(0.189)
ARRA × R01		0.019		-0.002		0.010		0.019
		(0.181)		(0.259)		(0.085)		(0.138)
Constant	1.193**	1.272**	1.553**	1.675**	0.648**	0.660**	1.103**	1.126**
	(0.039)	(0.097)	(0.056)	(0.139)	(0.021)	(0.045)	(0.034)	(0.073)
N	271	271	271	271	376	376	376	376
F-stat	5.97	2.23	5.48	2.15	0.24	0.11	0.80	0.31
Adj. R^2	0.24	0.23	0.27	0.26	0.21	0.20	0.26	0.25

Note: FOA-fixed effects are included in all models. Standard errors are in parentheses. *, ** denotes statistical significance at 5%, and 1%, respectively.

Table 5: Robustness Checks: Analysis on Cohort Sample

(a) Mean Values

	1	2	3	4	5	6	7	8
	(Log) Citation		(Log) Citation per $M		(Log) N_{pub}		(Log) N_{pub} per $M	
(Dummy) ARRA	-0.144**	-0.023	-0.188*	0.038	-0.083*	-0.025	-0.132*	-0.001
	(0.051)	(0.053)	(0.087)	(0.076)	(0.032)	(0.020)	(0.053)	(0.033)
(Dummy) R01		0.211**		0.334**		0.392**		0.517**
		(0.062)		(0.091)		(0.021)		(0.061)
ARRA × R01		-0.197**		-0.371**		-0.088**		-0.208**
		(0.056)		(0.080)		(0.026)		(0.043)
Constant	2.573**	2.437**	3.600**	3.380**	1.408**	1.169**	2.407**	2.088**
	(0.022)	(0.043)	(0.038)	(0.063)	(0.014)	(0.013)	(0.023)	(0.038)
N	3,912	3,912	3,912	3,912	3,912	3,912	3,912	3,912
F-stat	8.02	51.82	4.66	66.52	6.72	279.89	6.21	44.02
Adj. R^2	0.08	0.08	0.06	0.06	0.08	0.09	0.18	0.19

(b) Variances

	1	2	3	4	5	6	7	8
	(Log) Citation		(Log) Citation per $M		(Log) N_{pub}		(Log) N_{pub} per $M	
(Dummy) ARRA	-0.179†	-0.237†	-0.245†	-0.296	0.038	-0.027	0.067	0.001
	(0.104)	(0.130)	(0.142)	(0.180)	(0.049)	(0.061)	(0.077)	(0.098)
(Dummy) R01		0.778		0.796		-0.287		-0.504
		(0.558)		(0.774)		(0.204)		(0.326)
ARRA × R01		0.256		0.234		0.169		0.166
		(0.219)		(0.304)		(0.103)		(0.164)
Constant	1.196**	0.889**	1.532**	1.218**	0.613**	0.723**	1.002**	1.193**
	(0.063)	(0.229)	(0.086)	(0.318)	(0.031)	(0.082)	(0.049)	(0.132)
N	193	193	193	193	280	280	280	280
F-stat	2.93	2.29	2.98	1.63	0.61	1.47	0.76	1.20
Adj. R^2	0.12	0.15	0.20	0.20	0.27	0.28	0.36	0.36

Note: FOA-fixed effects are included in all models. Robust standard errors, clustered by FOA, are in parentheses. †, *, ** denotes statistical significance at 10%, 5%, and 1%, respectively. All models are conditioned on the project having at least one publication.

Table 6: Correlates of Project Characteristics

	1 (Log) Max Citation	2 Max IF	3 (Log) N_{pub}	4 No Pub
(Dummy) ARRA	0.011	-0.466	-0.008	-0.012
	(0.058)	(0.347)	(0.021)	(0.010)
(Log) Total cost	-0.054	0.090	-0.046†	0.016*
	(0.043)	(0.324)	(0.026)	(0.008)
(Log) # of unique authors	0.687**	4.205**	0.562**	-0.296**
	(0.020)	(0.242)	(0.011)	(0.007)
(Dummy) Within 2 years	-0.081	0.082	-0.050†	-0.019†
	(0.060)	(0.451)	(0.026)	(0.011)
# of PIs	-0.042	-0.196	-0.014	0.018*
	(0.054)	(0.310)	(0.019)	(0.008)
(Dummy) Existing PI	-0.074	-0.447	-0.028	-0.011
	(0.048)	(0.380)	(0.023)	(0.010)
(Log) Mean cumulative $ grants for PIs (04-08)	0.076**	0.780**	-0.003	0.007
	(0.029)	(0.241)	(0.010)	(0.005)
(Dummy) No PI has a grant (04-08)	0.910**	9.256**	-0.104	0.086
	(0.348)	(2.944)	(0.123)	(0.062)
# grants for organization (04-08, thousands)	0.022**	0.280**	-0.001	0.001
	(0.008)	(0.066)	(0.003)	(0.002)
# of months since published	0.041**			
	(0.002)			
Constant	-0.622	-11.859*	0.655*	0.528**
	(0.576)	(5.084)	(0.330)	(0.114)
N	8,499	8,499	8,499	12,558
F-stat	252.59	44.29	337.56	256.85
Adj. R^2	0.42	0.22	0.62	0.76

Note: Activity-FOA-IC-institution type-year fixed effects are included in all models. Robust standard errors, clustered by activity-FOA-IC-institution type-year, are in parentheses. †, *, ** denotes statistical significance at 10%, 5%, and 1%, respectively. Models 1, 2, and 3 are conditioned on the project having at least one publication.

Table 7: Analysis of Research Fit

	1	2
(Dummy) ARRA	-0.334*	-0.446†
	(0.145)	(0.241)
(Dummy) R01		0.359
		(0.228)
ARRA × R01		0.228
		(0.290)
Constant	3.204**	2.970**
	(0.030)	(0.158)
N	8,494	8,494
F-stat	5.34	3.33
Adj. R^2	0.63	0.63

Note: FOA-fixed effects are included in all models. Robust standard errors, clustered by FOA, are in parentheses. †, *, ** denotes statistical significance at 10%, 5%, and 1%, respectively.

Table 8: Correlates of ARRA Project Selection

	1 Full	2 R01	3 R03+R21
(Dummy) Existing PI	0.045**	0.066**	-0.006
	(0.015)	(0.017)	(0.030)
# grants for organization	-0.002	-0.002	-0.003
(04-08, thousands)	(0.003)	(0.003)	(0.008)
(Log) Total cost	0.044*	0.061*	-0.037
	(0.022)	(0.027)	(0.047)
(Log) # of unique authors	-0.029**	-0.034**	0.001
	(0.008)	(0.008)	(0.017)
Research Fit	-0.002	-0.004†	0.002
	(0.002)	(0.002)	(0.003)
Constant	-0.305	-0.590†	0.784
	(0.279)	(0.338)	(0.566)
N	8,494	5,557	2,436
F-stat	4.42	6.58	0.20
Adj. R^2	0.28	0.17	0.29

Note: All models are conditioned on the project having at least one publication. Activity-FOA-IC-institution type-year fixed effects are included in all models. Robust standard errors, clustered by activity-FOA-IC-institution type-year, are in parentheses. †, *, ** denotes statistical significance at 10%, 5%, and 1%, respectively.

Figures

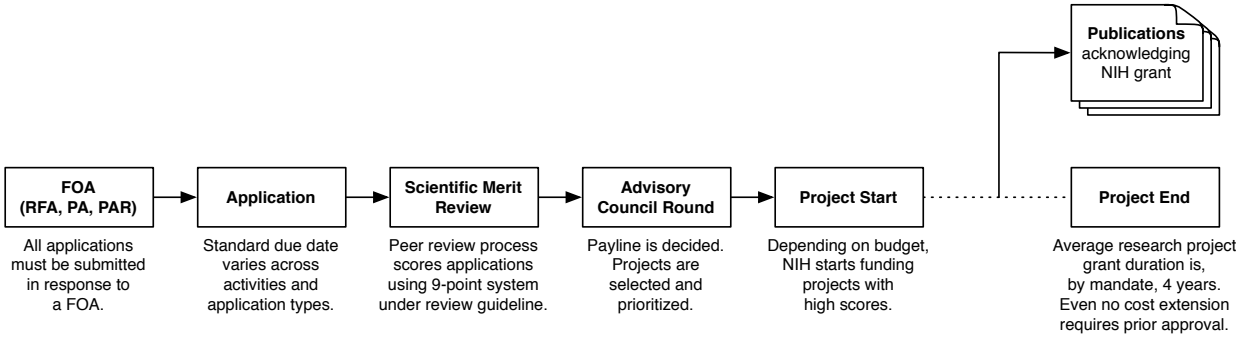

Figure 1: Schematic Illustration of NIH Funding Process

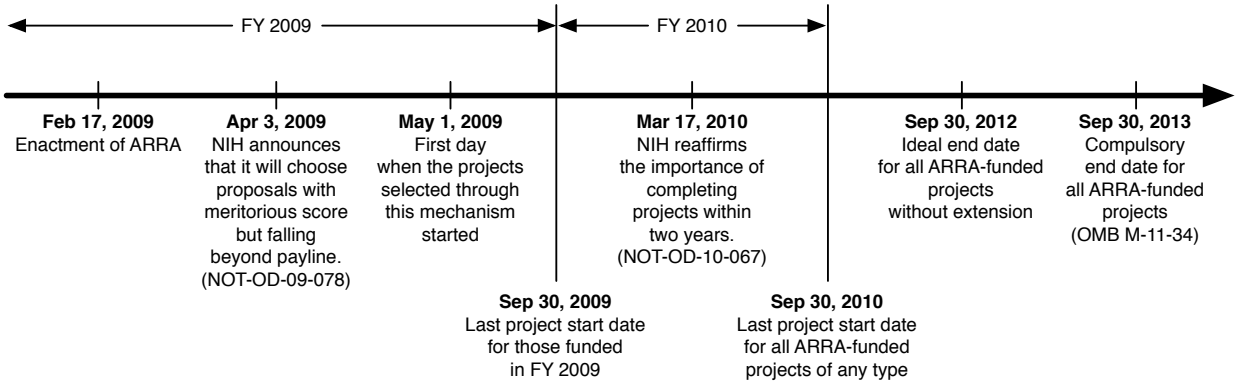

Figure 2: NIH-specific ARRA Timeline

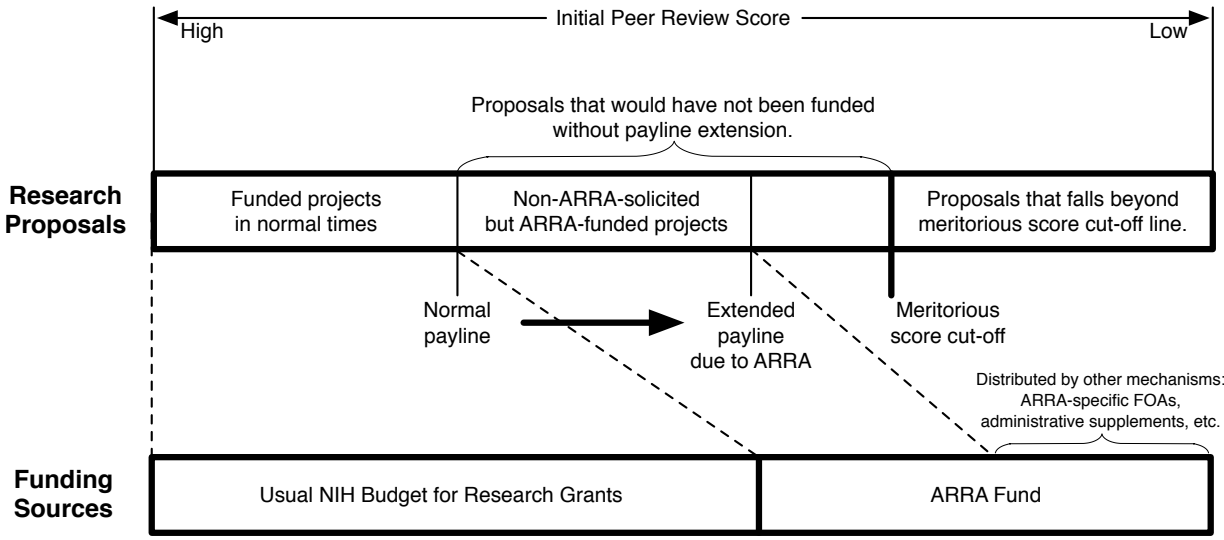

Figure 3: Graphic Illustration of NIH-ARRA and Payline Extension (not drawn to scale)

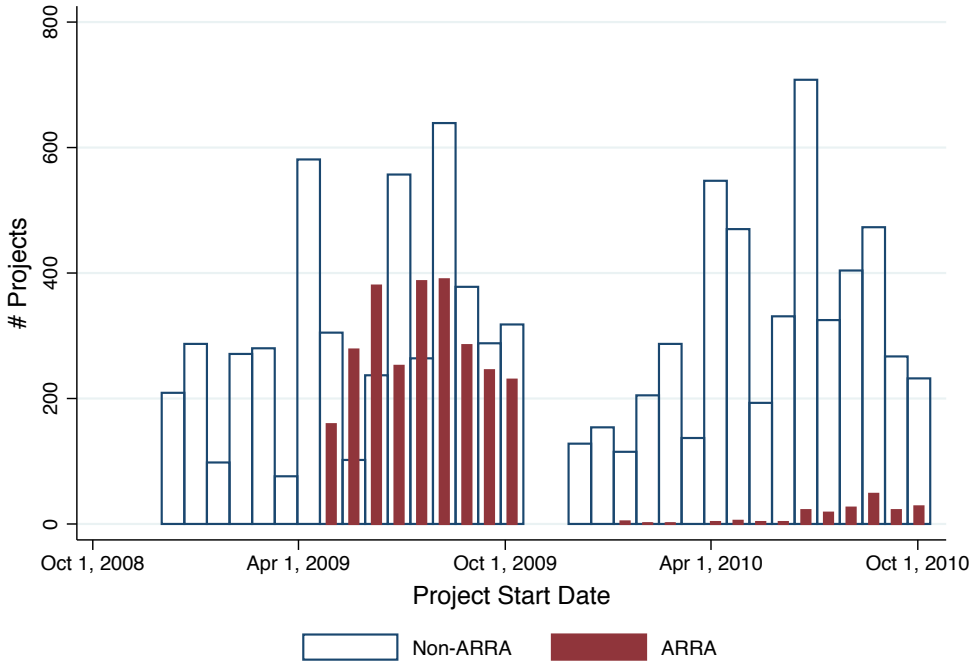

Figure 4: Histogram of Project Start Date

Appendix: Construction of Research Fit

Each FOA document starts with "Research Objectives," which describe the purpose of the funding opportunity. Two possible inputs for comparison from the project side are the abstract of the project proposal and abstracts of the publications resulted from the project. Abstracts of the project proposal are subject to investigators' deliberate efforts to make them as close as possible to the FOA objectives in order to increase the chance of funding. In contrast, because the abstracts of the publications are in general finalized after the funding decision, investigators have less incentive to make these abstracts close to the FOA objectives. Therefore, we choose to use publication abstracts for the comparison with FOA Research Objectives.

We employ the natural language processing approach to compute the similarity between FOA objectives and abstracts (Rehurek and Sojka, 2010; Bird et al., 2009). The term frequency-inverse document frequency (tf-idf) model is one of the classical vector space models in natural language processing (Manning et al., 2008). The key idea behind this model is that the more frequently a particular term appears in a document (i.e., local property), the more representative the term is of the document. However, if the term appears in all documents (i.e., global property), the discerning power of that term should be lower. In our context, suppose the term "medical" appears many times in an abstract. Then, we know that the content of the paper is about some medical topics. However, if "medical" appears in all other FOA objectives and abstracts, this term offers little help in uniquely identifying the content relative to other abstracts. By balancing between these local and global perspectives, we can identify nontrivial characterizing terms from the collection of FOA objectives and abstracts.

We start by collecting 369 FOA objectives and 29,995 abstracts of publications from the projects in our sample. After removing English stop words (e.g., a, an, the), we tokenize each document and create a corpus (i.e., a formatted collection of documents). This corpus provides us with the global perspective on determining which terms have distinguishing power. Based on the tf-idf model, we then compute for each project the similarity between the publication abstracts and the corresponding FOA objectives. When a project has multiple publications, we take the maximum value of the computed similarities for the project research fit.