# Engineering science: The impact of public investment in PhD training on the size and shape of the U.S. scientific workforce*

Dror Shvadron[†] Duke University
Hansen Zhang[‡] Duke University
Lee Fleming[§] UC Berkeley
Daniel P. Gross[¶] Duke University and NBER

January 2025

**Abstract:**

Using data on nearly one million U.S. doctoral dissertations in the natural sciences and engineering since 1950 and new measures of PhD graduates' sources of financial support, we examine how public investment in scientific training affects the scientific workforce. We show that government funding has large effects on the annual number and composition of new scientists the U.S. higher education system produces: supporting 10% more PhD graduates in a given field on average increases PhD production in that field by 8-9% in the short-run and working scientists by 3-4% in the long-run. However, increases in public support for doctoral training on the margin increase contingent faculty more than tenure-track faculty or industrial scientists, suggesting a possibility of over-saturation and/or distortion. Ongoing analysis examines how many government-funded U.S. PhDs subsequently move abroad, and where—and in turn, whether the fruits of these investments are appropriated by other countries. The results provide insight on how government investment in PhD training shapes the scientific expertise available to the U.S. innovation system.

**JEL Classification:** I23, I28, O31, O32, O33, O38

**Keywords:** Public R&D, Doctoral training, Scientific careers, Scientific workforce

*We thank Michelle Qiu and Max Murakami-Moses for valuable research assistance and the Duke University Fuqua School of Business and Alfred P. Sloan Foundation for financial support. All errors are our own.
[†]Address: University of Toronto, Rotman School of Management, Torono, Canada; email: dror.shvadron@utoronto.edu.
[‡]Address: Duke University, Fuqua School of Business, Durham, NC, USA; email: hansen.zhang@duke.edu.
[§]Address: UC Berkeley, Haas School of Business, Berkeley, CA, USA; email: lfleming@berkeley.edu.
[¶]Address: Duke University, Fuqua School of Business, Durham, NC, USA; email: daniel.gross@duke.edu.

# 1 Introduction

High-skill workers are the backbone of the U.S. and global innovation system (Nelson and Rosenberg 1993). Scientifically-trained workers are an especially important input to modern innovation, which is increasingly organized around the discovery and commercial application of new science (e.g., Mansfield 1991, Rosenberg and Nelson 1994, Mowery and Sampat 2006). In 2021, the U.S. produced around 32,000 doctoral graduates in the physical sciences, life sciences, and engineering—a nearly 7-fold increase since 1960, more than tripling the population growth rate over the same period (ProQuest 2023). Despite debate over whether doctoral programs produce too few or too many PhDs (e.g., Cyranoski et al. 2011), there appears to be consensus that the size of the scientific workforce is an important determinant of the rate of scientific and technological advance (e.g., Nelson and Phelps 1966, Romer 1990, or Bloom et al. 2020).

Aspiring scientists require extensive training. This training, however, is a "long and expensive process" (Bush 1945) whose cost appears to be inexorably rising over time: today, it can cost more than (2024 USD) $1.5 million in direct, indirect, and opportunity costs to train a new PhD (Smolka et al. 2015). Though once a domain for the independently wealthy, doctoral training in the U.S. is now heavily subsidized by a diverse set of public, private, and non-profit organizations whose goal is putatively to increase the size of the scientific workforce, overall or in specific subjects. Since World War II, the largest of these has been the U.S. federal government.

In this paper, we use 70 years of data on U.S. PhD production in the natural sciences and engineering—nearly one million graduates in total—to examine how federal investment in scientific training affects how many PhDs the U.S. higher education system produces, in what fields, what jobs they take, and how many stay in the U.S. and in scientific careers post-graduation. Because PhD graduates' sources of financial support have not previously been comprehensively, systematically measured across this period, we develop these measures ourselves from the text of PhD dissertations (primarily, from acknowledgments) between 1950 and 2020. These data enable us to first document who funds scientific education and in what subjects, including for public funders (where the principal sources of support have shifted from the Department of Defense (DoD) to the National Institutes of Health (NIH)) and private ones (which have shifted from the chemical, to oil, to pharmaceutical industry and from firms like Du Pont to IBM, Intel, and Google).

A plurality of PhD graduates throughout this period report U.S. government support. Using a shift-share instrumental variable strategy exploiting variation in individual federal agencies' subject matter priorities and fluctuations in each agency's total level of support, we show that government funding has a large effect on PhD production: funding an additional 100 dissertations increases PhD production by 80-90 graduates, suggesting public funding significantly grows the scientific workforce, and only modestly overlaps with PhD training that would have taken place otherwise. We further show that funding choices at any given point in time have decades-long consequences, as trainees will be the next many decades' scientific workforce.

However, policy efforts to sustain or increase the U.S. scientific workforce may have unintended consequences, especially the possibility of attracting a surfeit of graduates students relative to the labor demand that will exist in five to 50 years (from graduation through the end of a career). The recent (and projected further) decline in U.S. college enrollments, universities' reallocation

from tenure-track to contingent faculty hiring (e.g., Colby 2023), and the significant lengthening of doctoral and post-doctoral training have for several years raised concerns about excess supply in academic labor markets—a condition which could (for example) be brought about by subsidized entry in the form of abundant fellowship funding. Using data from the NSF's Survey of Doctorate Recipients (SDR), we evaluate the link between PhD cohorts' share of graduates later employed in tenure track, non-tenure track, and non-university positions and the number of graduates with federal support at their time of education. Although public investment increases the scientific workforce across all tracks and sectors, the marginal graduate is more likely to be non-tenure track faculty than inframarginal graduates. Given the ex-ante unpredictability of both future demand and trainees' scientific potential, however, adjusting investments in PhD training to discourage the marginal graduate or attempt to balance supply and demand may not be possible: large reductions in public support could also lead to an undersupply of scientifically trained workers due to well-known causes of market failure (Nelson 1959, Arrow 1962, Lazear 2009).

Insofar as the policy goal of supporting doctoral training is to increase the domestic supply of scientists, a third question is whether and how many PhD graduates become practicing (publishing) scientists and remain in the U.S. post-graduation. We evaluate this question using newly-developed dataset linking the PhD graduates in our sample to their scientific publications, and in turn their institutional affiliations and locations. To make this link, we train a classifier based on subject matter (dissertation and publication titles and abstracts), citation patterns, ethnicity, and name frequency, using ORCID-linked authors and publications as a training sample. The construction of this sample is in final stages of refinement, and analysis is pending—but the principal analytical output will be a chart documenting what countries federally-funded U.S. PhD graduates are in 5, 10, and 15 years post-graduation, especially for recent (post-2000) cohorts.

Beyond new evidence, an additional contribution of this paper is data. To our knowledge, no prior effort has produced comprehensive, long-run, systematic data on U.S. PhD graduates' sources of financial support at the level of individual students, funders, and student-funder pairs. Because federal agencies (and other funders) run their own research funding programs, no centralized source for PhD funding data exists, creating challenges in systematically documenting who pays for scientific training in the U.S. and how (and to whom) these investments pay off. The most complete prior measures may be the National Science Foundation's (NSF) Survey of Earned Doctorates (SED), which for many years collected data on graduates' sources of financial support but removed the question in 1997, and even then had only done so for a small number of (mainly public) funders. And although the NSF's Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) currently provide data on enrolled graduate students' sources of financial support, it (i) measures students rather than graduates, and all graduate students rather than doctoral students, (ii) aggregates sources of support into coarse categories, and only individually measures a few specific (government) funders, and (iii) is too recent to study long-term outcomes.

Dissertation text presents an unconventional opportunity for systematic measurement: they are required of every graduating doctoral student, typically include an "Acknowledgments" section thanking research sponsors, and are broadly available. These acknowledgments are primarily what we use to compile information on who supports U.S. scientific training, though we also draw information from elsewhere in a dissertation, such as curriculum vitae. An independent, flexible source like

author-written acknowledgments is useful not only due to its ability to capture the wide variety of potential funders, but also the diverse forms this support can take—from part-time employment or fellowship funding to in-kind support like equipment use. The availability of U.S. PhD dissertations from (i) ProQuest Dissertations & Theses Global (PQDT) and (ii) university repositories enables us to produce these measures for a very large large share of graduates in the natural sciences and engineering over the period we study, which we validate against the SED.

We proceed as follows. The rest of this section connects our work to previous research, particularly on U.S. federal investments in doctoral training. In Section 2 we provide context for this paper. Section 3 describes our data sources for dissertation acknowledgments and describes how we extract information on dissertation research sponsors from them. Section 4 presents preliminary analysis of these data. Section 5 discusses these results and concludes.

## 2 Historical and Institutional Background

University research in the U.S. today is supported by a broad range of funding streams, the largest of which is federal funding—which in 2020 accounted for over half of university R&D expenditures. Alongside nearly $50 billion of federal spending on university research in 2021, state and local governments and firms each sponsored nearly $5 billion of research, and institution-provided funds the remainder (Gibbons 2022). An important share of this funding goes towards supporting scientific trainees—e.g., via fellowships, assistantships, or tuition support.

The current scale and distribution of funding for university research and training is a product of the postwar era. Prior to World War II, academic research was severely resource constrained, and when funded was supported by philanthropic foundations or industry, with limited public funding outside of agriculture. Support for training was even more meager, limiting who could partake in the scientific enterprise. As Vannevar Bush noted in *Science, The Endless Frontier* (Bush 1945), "Higher education in this country [was] largely for those who have the means ... with few exceptions those without the means of buying higher education go without it."

The achievements of science and technology in the war (e.g., Gross and Sampat 2023) drove a major postwar expansion of research policy. After the war ended, several federal agencies picked up the extramural research investments that the wartime Office of Scientific Research and Development (OSRD) left behind when it was decommissioned: Office of Naval Research (and later other military offices) absorbed military contracts, the Atomic Energy Commission (AEC) nuclear work, and the National Institutes of Health (NIH) medical research.[1] These government agencies funded both basic and applied R&D, at firms and universities. The NSF, created in 1950, was explicitly established to support basic research and training at universities. In 1958, the launch of Sputnik triggered the enactment of the National Defense Education Act, which provided a new source of fellowship funding, and the creation of NASA, which funded university research and education as well. In the

[1]Throughout this paper we reference government agencies with their acronyms. In our analysis we group up federal agencies to their ultimate parents, which are: the Departments of Commerce (DOC), Defense (DoD), Energy (DOE), Health and Human Services (HHS), Interior (DOI), and Transportation (DOT); Environmental Protection Agency (EPA); National Aeronautics and Space Administration (NASA); National Science Foundation (NSF); Department of State (STATE); and Department of Agriculture (USDA). We associate defunct historical agencies to their modern descendants, where appropriate (e.g., AEC is mapped to DOE, and HEW to HHS).

meantime, corporate funding for research and graduate training also began to grow, especially by firms in the burgeoning defense and pharmaceutical industries.

The institutional arrangement that emerged after the war has essentially persisted to the present, with changes in distribution across across funders (due to the expansion of the NIH, firm turnover, and growth of new philanthropic foundations). A defining feature of modern funding for graduate training is, in turn, how diffuse its sources are. One casualty is that no single, centralized source for data on PhD funding exists (historically or at present) for understanding how support for graduate trainees affects first-order outcomes like who becomes a scientist, or what science gets done. One goal of this paper is to address this measurement gap.

# 3 Who Pays for Scientific Training?

To evaluate the effects of public (or private) investment in doctoral training on the U.S. scientific workforce and the trajectory of scientific careers, we first need to identify doctoral graduates and develop measures of their financial sponsors that can span universities, fields, funders, and time. Our analytical sample consists of PQDT-listed PhD graduates in the natural sciences and engineering—a sample which, as we will show, approximates the universe of graduates as measured by administrative sources. We measure their sources of support from their dissertations, in a three-step process which requires (i) building a sample of dissertations, (ii) obtaining their full text, and (iii) processing that text to extract, harmonize, and classify research funders.

## 3.1 Building a dissertation sample

Our main data source is ProQuest Dissertations & Theses Global (PQDT). For nearly a century, ProQuest has been acquiring and re-publishing dissertations of U.S. PhD graduates. Its dissertation collection has since been digitally catalogued and includes extensive metadata on each dissertation (e.g., author, institution, degree earned, title, abstract, subject).[2]

At the time of extraction, the PQDT universe contained nearly six million global dissertations, spanning various degree levels (both master's and doctoral) and fields of study including the physical and life sciences, engineering, humanities, and social sciences. We focus our attention on doctoral dissertations at U.S. institutions in natural sciences since 1945. To do so, we crosswalk students' self-reported subjects (as provided in the PQDT data) to 17 SED major fields which are typically considered natural sciences, and restrict our sample to dissertations that map to these fields.[3] Our final sample contains 1,148,292 dissertations from 1945 to 2020.

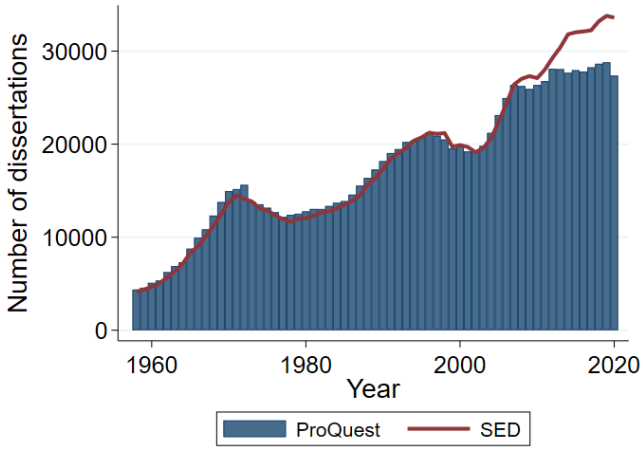
To evaluate the completeness of our data, we compare annual PQDT dissertation counts to SED counts in these fields. Figure 1 shows that PQDT (the blue bars) tracks the SED (red line) closely

[2]PQDT data have been used in a range of recent research (e.g., Toole and Czarnitzki 2010, Bikard et al. 2015, Buffington et al. 2016, Jiang et al. 2023, Antman et al. 2023, Arora et al. 2023, among others).

[3]Our focal major fields are: Agricultural sciences and natural resources, Biological and biomedical sciences, and Health sciences (belonging to the broad field "Life Sciences"); Physics and astronomy, Geosciences, atmospheric sciences, and ocean sciences, and Chemistry (belonging to "Life Sciences"); Computer and information sciences and Mathematics and statistics (belonging to "Mathematical and Computer Sciences"); and Aerospace, aeronautical, and astronautical engineering, Bioengineering and biomedical engineering, Chemical engineering, Civil engineering, Electrical and electronics engineering, Industrial and manufacturing engineering, Materials science engineering, Mechanical engineering, and Other engineering (belonging to "Engineering").

to the late 2000s. Even when they diverge slightly post-2005 (possibly due to changes in PQDT's sample, field labeling, or lags in measurement), PQDT still captures >90% of the SED. Appendix Figure A.1 disaggregates this comparison to individual fields and finds similar patterns at the field level. Together, these comparisons suggest our dissertation sample (from PQDT) closely tracks the underlying population (SED), building confidence in the sample.

Figure 1: ProQuest vs. SED dissertations in the hard sciences, by year



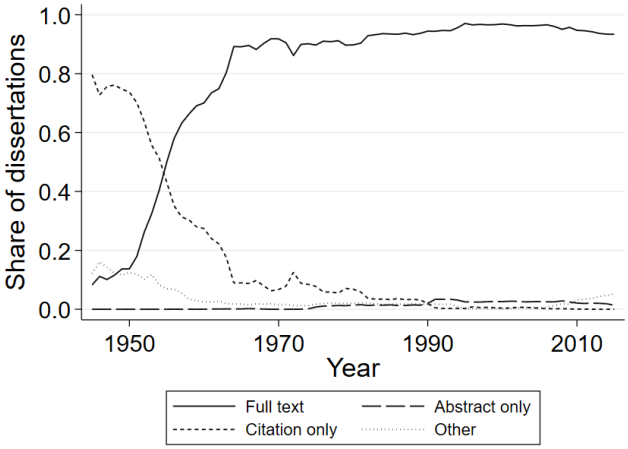

Notes: Figure shows annual PQDT dissertation counts in the physical and life sciences and engineering (blue bars) and SED counts (red line) for comparison, from 1958 (the first year of the SED) to 2020.

### 3.2 Obtaining dissertation text

An important feature of PQDT is that in addition to metadata, it also has dissertations' full text. Dissertation structures have for many decades followed a standard publishing template specified by ProQuest, and most dissertations have an "Acknowledgments" section, where authors thank funders and other supporters. These acknowledgments are a crucial resource for this paper: they provide an opportunity to systematically measure sources of PhD student support. Even when an acknowledgments section is missing, funders tend to be mentioned in the biography section of the dissertation or in footnotes in the text. Access to the full text of dissertations allows us to identify these cases regardless of where they are mentioned in the text.

Though PQDT provides the full text for most modern dissertations, this is not always the case historically, for two reasons. First, some early dissertations were indexed citation-only. Second, until recently, dissertations were stored on microfilm—not all of which has yet been digitized. Using ProQuest document numbering, we can determine how a dissertation was indexed, and whether full text may be available from PQDT. In Figure 2, we plot these frequencies over time. Though early dissertations were primarily indexed citation-only, by the mid-1960s, 90%+ dissertations have full text indexing. In principle, this set should be an upper bound to an obtainable full text sample. In practice, not all such dissertations have yet been digitized. Figure 3 shows this, plotting the annual share of dissertations for which full text is available from PQDT, by field. Between 1950 and 1990, the observable full text share of dissertations is around 30-40% in all fields. This rate discretely jumps in 1990, 1996, and 2000 and holds near 100% thereafter.

Figure 2: ProQuest dissertations by indexing category (according to ProQuest identifiers)

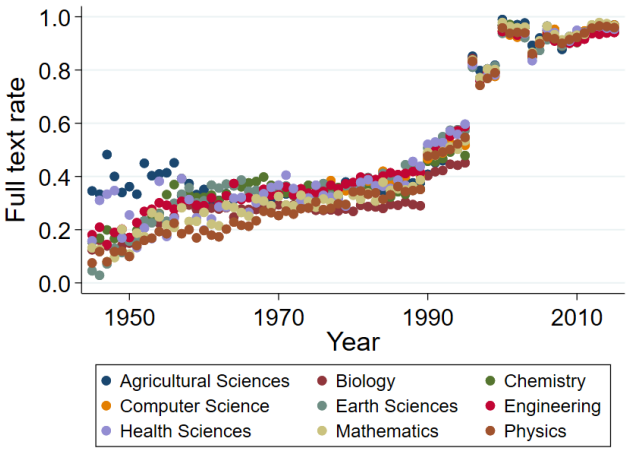



Notes: Figure shows the annual share of PQDT dissertations in our sample that ProQuest identifies has having been indexed full text (i.e., ProQuest received metadata (author, title, subject, etc.) and a full copy of the dissertation), abstract-only (metadata plus abstract), or citation only (metadata only).

Of the 1.15 million dissertations in our sample, we have full text for 846,992 (73.8%). Gaps are not necessarily a problem for our analysis provided there are no systematic differences: as long as the availability of full text is as good as random, the data can be used to produce representative estimates. A key question is then: what explains why some dissertations have full text, and others do not? According to ProQuest, this variation is a result of (i) differences in indexing methods (Figure 2), and (ii) idiosyncratic variation in what dissertations have been retrieved from microfilm. Diagnostic testing is similarly reassuring on several dimensions, one of which is visible in Figure 3: full text availability does not vary differentially across fields.[4]

[4]Probing more deeply, we have found that university-year fixed effects can explain most (nearly 90%) of the variation in indexing category (Figure 2)—reflecting that full-text vs. citation-only indexing may be a result of agreements between ProQuest and these universities—but that they explain a substantially lower fraction (65%) of the variation in the observed availability of full text (vis-à-vis the digitized set, per Figure 3), and only slightly more than university and year fixed effects alone (50%), which indicates residual idiosyncrasy.

Figure 3: ProQuest dissertations with full text available, by field




Notes: Figure shows the annual share of PQDT dissertations for which we have access to full text: the subset of full-text indexed dissertations (Figure 2) which were stored electronically or have been digitized from their original physical storage medium (microfilm).

### 3.3 Identifying dissertation sponsors

With this sample in hand, we then search dissertation text for acknowledged research sponsors. In most cases, these acknowledgments appear in a dedicated "Acknowledgments" section near the beginning of the dissertation. Figure 4 provides an example from the dissertation of Ian Buck, a 2005 Stanford University Computer Science graduate who developed the precursor to Nvidia's CUDA computing platform during his PhD studies and is now a senior executive at Nvidia. This example presents a large number and range of sponsors and illustrates the variation that exists in the data. Though Buck provides a thorough accounting in a formal acknowledgments section, in some cases authors acknowledge funders throughout their dissertation.

Figure 4: Example dissertation acknowledgment (Ian Buck, Stanford University, 2005)

> I have had a variety of funding sources throughout my graduate career including DARPA, DOE, ASC, NVIDIA, ATI. Specifically, I would like to thank Randy Frank, Bob Graybill, and Mike Macedonia for putting money where my mouth was. I also have been fortunate in receiving fellowships from the Stanford School of Engineering as well as NVIDIA.

To flexibly identify research sponsors across this corpus, we develop a pipeline that leverages natural language processing tools and large language models (LLMs).[5] Our procedure takes the following steps. First, we partition each dissertation into sentences and identify "acknowledgment sentences". To isolate potential acknowledgment sentences, we applied a rule-based string matching approach by splitting the text into sentences (using NLTK's Punkt Sentence Tokenizer; Bird et al. 2009) and identifying sentences with tokens such as "fund", "grant", "thank", and so on. This procedure tags roughly 92 million potential acknowledgment sentences.

[5] A more detailed description of our methodology, including the LLM prompts we use, and our approach to consolidation and data linking, is provided in Appendix A.1. Here we provide an abridged summary.

We next identify named entities within these sentences, classify entities into sectors, and identify the type of support and grant identifiers (where provided), using the recent open-source LLMs Solar 10.7B and Smaug 34B (Kim et al. 2023, Pal et al. 2024). We established a procedure that uses the LLM to verify that a sentence acknowledges support from an external entity, identify the supporting entities, categorize them by organization and support type, and record grant and contract identifiers where applicable. We used the VLLM Python package (Kwon et al. 2023) and the LM format enforcer (Gat 2023) for efficient batch inference and to ensure the model produces a well-structured JSON schema. The LLM output indicates that 11 million of our potential acknowledgment sentences (12%) indeed acknowledged support or were part of a biography section. Among all sentences, the LLM extracted 9.3 million mentions of organizations (4 million unique), with 4.8 million of these mentions originating from sentences classified as support or biography.

We then consolidate named entities (accounting for name and spelling variants by grouping them together) and link them to external firm registries. We consolidate extracted entities using the Smaug 34B model and match them with the Research Organization Registry (ROR) and Wikidata to disambiguate institution names and obtain external identifiers.

### 3.4 Example

Table 1 demonstrates the performance of our procedure on the acknowledgment text in Figure 4. Each row is an extracted funder, and the columns report the name, sector, type of support, and the associated ROR organization (where matched to ROR; otherwise blank). Comparing the text to the first three columns of this table, we find that our algorithm correctly extracts all entities by name and classifies them into the correct sectors and type of support.

Table 1: Example extraction of acknowledged funders for Ian Buck (Figure 4)

| Entity Name | Entity Type | Support Type | ROR Organization |
|---|---|---|---|
| DARPA | government | funding | Defense Advanced Research Projects Agency |
| DOE | federal agency | funding | United States Department of Energy |
| ASC | unknown | funding | – |
| NVIDIA | private company | funding | Nvidia |
| ATI | private company | funding | Advanced Micro Devices |
| Stanford School of Engineering | academia | fellowship | – |
| NVIDIA | private company | fellowship | Nvidia |

Notes: Table lists organizations identified in the text in Figure 4 as dissertation funders using our LLM-based procedure and provides their links to ROR. The LLM associates ATI with Advanced Micro Devices (AMD) because ATI Technologies was acquired by AMD in 2006.

### 3.5 Validating the results

We validate this approach by comparing the extracted entities to a manually-processed random sample of 500 dissertations, for which we meticulously identified funders by hand. Figure 5 compares the share of this sample with government, non-profit, and firm funding according to our manual and automated procedures. For this exercise, we include all validated funding organizations that were identified using our automated procedure. The results show a substantial degree of similarity, suggesting the automated LLM-based approach and matching process approximates manual extractions.

Figure 5: Comparison of Manual and LLM Entity Recognition




Notes: Figure compares manual and automatic entity recognition from dissertation texts. The figure shows the percentages of dissertations acknowledging funding support by organization type for a sample of 500 dissertations, by classification source. Note that a dissertation can acknowledge support from multiple organization types.

We undertake two additional validation exercises. In Appendix A.3, we collect data on all NSF Graduate Research Fellowships (GRFs) awarded since the beginning of the program, link awardees to PQDT, and evaluate how many of these graduates acknowledge NSF support in their dissertations. We find that among the matched NSF graduate fellows, 60.8% mention the NSF in their dissertation. Our LLM-based extraction was able to identify and link 98.6% of these mentions. These results provide further validation for our approach while also outlining its limitations—namely, that we rely on graduates to acknowledge funding support in their dissertation texts.

In Appendix A.4, we compare our sample to data from the NSF's Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). We find strong and highly significant correlations between the share of federally-funded PhD graduates in each field and year that we measure as being supported by six major U.S. federal agencies (DoD, DOE, HHS, NASA, NSF, and USDA) and the share of federally-funded graduate students at doctorate-granting institutions reporting primary support from these agencies in the GSS. Regression coefficients range from 0.75 to 1.42, with HHS and NSF support measures not statistically distinguishable from a one-to-one relationship. These results provide additional validating evidence for our approach to measuring research sponsors using the PQDT database.

The three validation exercises described above provide strong support for our measurement approach. First, compared to a manual extraction from 500 dissertations, we find that our automated approach accurately identifies a similar number of funding organizations. Second, our approach correctly identifies almost all GRF awardees that acknowledge the NSF in their dissertation as being NSF-supported. Third, our measured agency funding rates at the field and year level exhibit strong correlations with the rates reported in the GSS, despite differences in their sample populations. Although evidence from these exercises suggests that our approach may undercount funded dissertations, follow-up testing reveals that this is due to dissertations under-reporting their funding sources rather than limitations in our methodology.

### 3.6 Desriptive patterns

Figure 6(A) shows the share of dissertations supported by different types of organizations from 1950 to 2020. In 2020, approximately 45% of dissertations in our sample acknowledged outside support. U.S. federal agencies are the most common funders, supporting nearly 40% of dissertations. Roughly 13% of dissertations are supported by non-profit organizations, and 7.5% by private firms. These patterns have changed substantially over time: government support increased rapidly in the 1950s and 1960s, peaked around 1967-1968, retracted in the 1970s, and stabilized (at ≈35%) in 1980. It has since modestly recovered (2000 onwards). Non-profit and firm support has also fluctuated: the share of dissertations supported by firms, for example, peaked at 15% in the late 1950s, declined to near zero in the 1970s, and has currently settled in between.

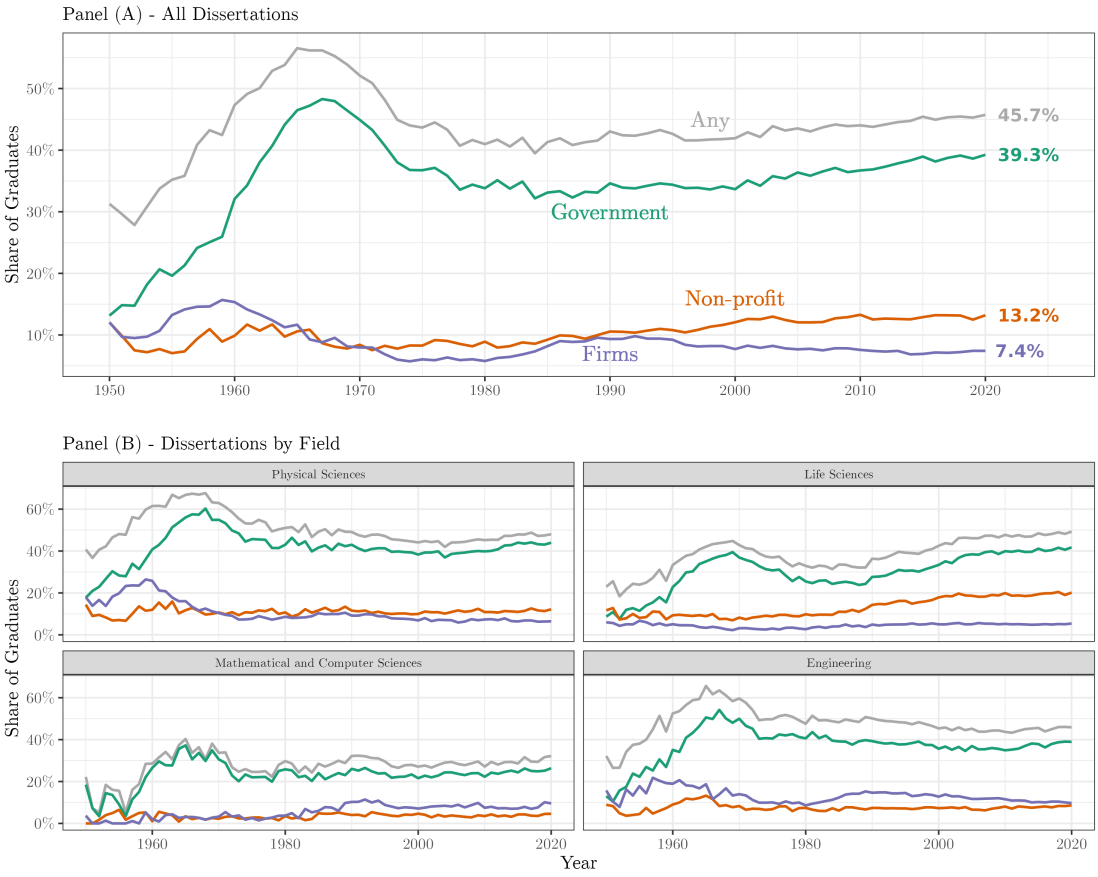
Figure 6(B) disaggregates these patterns by broad field. Disaggregation reveals that government support was traditionally especially high in the physical sciences and engineering, but has since been largely equalized across broad fields. Non-profits' support is somewhat higher in the life sciences, and firm support in engineering, though these patterns have also fluctuated: firms, for example, supported 25% of dissertations in the physical sciences in 1960—which follow-up analysis indicates was led by Du Pont in chemistry—but only 5% today; conversely, non-profits supported 10% of dissertations in the life sciences in 1980 and fund ≈20% today.[6]

Table 2 lists the top 20 funders by sector and their associated number of dissertations. The National Science Foundation (NSF), National Institutes of Health (NIH), Department of Energy (DOE), and various offices of the Department of Defense (DoD) rank high on this list, as do the National Aernautics and Space Administration (NASA) and Department of Agriculture (USDA). Top non-profit funders include academic societies and philanthropic foundations. The top firms tend to be large, research-intensive multinational corporations. IBM and Intel top this list, but the most represented industries are the chemical, oil & gas, and pharmaceutical industries—indicating that many major corporations in high-tech manufacturing industries invest in developing highly-trained researchers who might augment the stock of relevant knowledge or can contribute directly to their future R&D efforts as potential future employee researchers.

Figures 7 and 8 drill into federal funding agencies, grouping up federal funders into parent agencies and focusing on the four largest government supporters of PhD training: NSF, HHS, DoD, and DOE. Figure 7 characterizes the share of graduates in each of 17 major fields funded by these agencies. NSF provides relatively even support across subjects (except for health sciences, as a result of NSF's prioritization of basic science) and is an especially large sponsor of students in geological (i.e., earth, atmospheric, and ocean) sciences. Other agencies have a clear mission focus. HHS, for example, funds a large fraction of trainees in biological and health science and biomedical engineering, but few in other fields. DoD is the country's largest funder of aerospace engineering PhDs and a large supporter of engineering more broadly. DOE emphasizes physics, chemical engineering, and materials science engineering. Though not shown, similar patterns are present for other agencies—for example, USDA is a major sponsor of agricultural sciences; DOT, of civil engineering; and NASA, like DoD, of aerospace engineering.

[6] Appendix Figure B.1 presents corresponding counts of supported dissertations for 2005-2020 (as opposed to shares, which Figure 6 provides). One interesting pattern is the spike in funding for computer science and statistics from all sources starting in 2017—coinciding with surging enrollment in this field.

Figure 6: Funding of doctoral dissertations, by organization type (1950-2020)




Notes: Figure shows the share of PhD graduates supported by U.S. Federal Government agencies, firms, and non-profit organizations over time. Panel (A) presents overall results. Panel (B) presents results by broad field (physical sciences, life sciences, mathematical and computer sciences, and engineering).

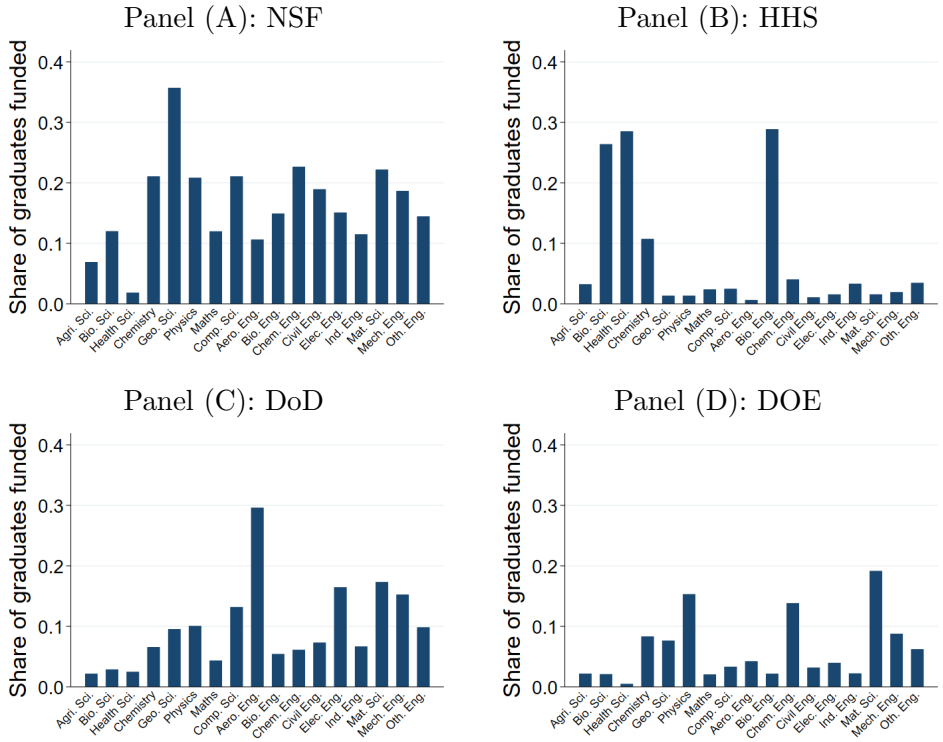
Table 2: Top 20 acknowledged funding organizations, by sector

| Federal Government | | Firms | | Non-profit | |
|---|---|---|---|---|---|
| Name | N | Name | N | Name | N |
| National Science Foundation | 143063 | IBM | 3978 | Sigma Xi | 5497 |
| National Institutes of Health | 73147 | Intel | 2389 | American Heart Association | 4255 |
| Department of Energy | 31996 | DuPont | 2189 | Howard Hughes Medical Institute | 4146 |
| National Aeronautics and Space Administration | 24356 | General Electric | 1547 | American Cancer Society | 3009 |
| Office of Naval Research | 15211 | Dow Chemical | 1529 | Geological Society of America | 2596 |
| Department of Agriculture | 15023 | Merck | 1502 | Ford Foundation | 2573 |
| Department of Health and Human Services | 14261 | Chevron | 1486 | Robert Wood Johnson Foundation | 2394 |
| Department of Defense | 14177 | ExxonMobil | 1478 | Welch Foundation | 2281 |
| Defense Advanced Research Projects Agency | 9428 | Google | 1308 | American Chemical Society | 1603 |
| Air Force Office of Scientific Research | 9155 | Microsoft | 1300 | Commonwealth Fund | 1490 |
| Environmental Protection Agency | 8606 | Pfizer | 1256 | Rockefeller Foundation | 1367 |
| National Cancer Institute | 7337 | Eli Lilly | 1234 | W. M. Keck Foundation | 1245 |
| Department of Education | 7234 | General Motors | 1076 | National Geographic Society | 1176 |
| National Oceanic and Atmospheric Administration | 7008 | Procter & Gamble | 1065 | David and Lucile Packard Foundation | 1164 |
| Army Research Office | 6822 | Ford Motor Company | 1002 | Andrew W. Mellon Foundation | 1035 |
| Public Health Service | 5939 | Monsanto | 953 | American Association of University Women | 951 |
| United States Air Force | 5145 | Boeing | 941 | Gordon and Betty Moore Foundation | 924 |
| National Institute of General Medical Sciences | 4638 | Texas Instruments | 888 | Electric Power Research Institute | 883 |
| Agency for International Development | 4285 | Shell | 887 | March of Dimes | 864 |
| Department of State | 3856 | 3M | 850 | W.K. Kellogg Foundation | 858 |

Notes: Table lists the top 20 government, non-profit, and commercial funders of dissertations in our sample as measured by our LLM-based procedure, along with their number of funded dissertations. Government funders listed in this table are presented as originally extracted to illustrate their range, but are consolidated into ultimate parent agencies in the analysis (e.g., defense agencies are grouped up to the U.S. Department of Defense (DoD); the National Cancer Institute is grouped into the National Institutes of Health, which is in turn grouped up to the U.S. Department of Health and Human Services; etc.).

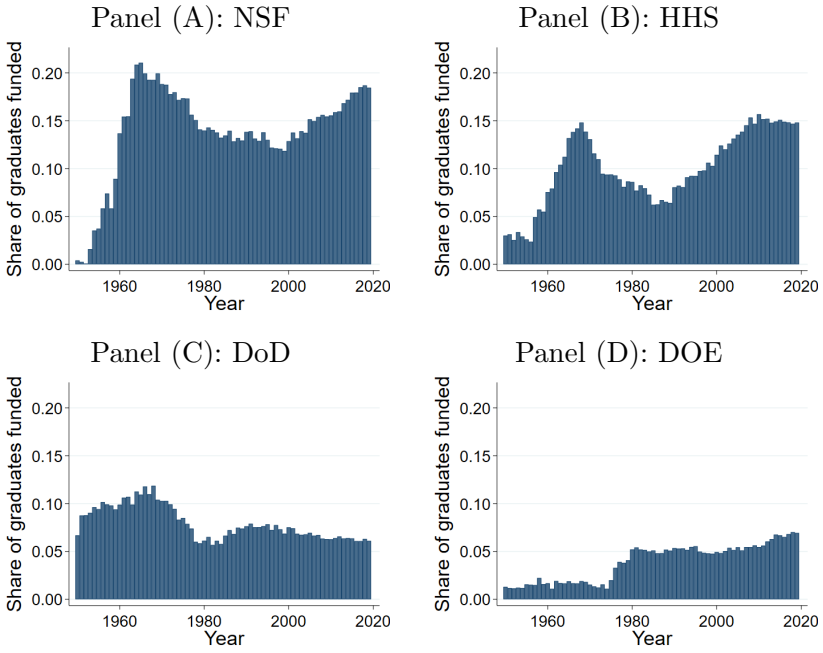
Figure 8 shows how this support has varied over time. The common pattern across most agencies is that their support for PhD training peaked in the late 1960s and in many cases has not recovered. The share of graduates which we identify with NSF support peaked in the mid-1960s at ≈20%, after which it steadily declined for thirty years; though by the late 2010s it had largely recovered, it remains below historical levels. HHS likewise peaked in the 1960s at around 15% of graduates, declined for twenty years, and then slowly recovered, returning to 1960s levels only around 2010. DoD and DOE show sharply different patterns: DoD investment in PhD training is now half of 1960s levels (down from ≈10% to ≈5% of graduates); conversely, whereas DOE supported 1-2% of students pre-1975, it has supported roughly 5% every year since.

Figure 7: Major fields' share of graduates supported by select federal agencies




Notes: Figure shows the share of PhD graduates in individual major fields supported by each agency listed (NSF, HHS, DoD, DOE), illustrating differences in subject matter priorities across agencies.

Figure 8: Share of all graduates supported by select federal agencies over time, 1950-2019




Notes: Figure shows the share of PhD graduates supported over time by each agency listed (NSF, HHS, DoD, DOE), illustrating long-run changes in these agencies' investments in PhD training.

This evidence indicates the mix of federal sponsors of PhD training has changed, especially the mix of "mission" sponsors. Together with the evidence that different agencies fund different kinds of science, the results suggest declining public investment in fields like physics and engineering, and increasing public investment in biomedical sciences.

# 4 Public Funding, PhD Production, and Scientific Careers

## 4.1 Effect of government funding on PhD production

When federal research funding programs expand, does PhD production grow? If so how much? To evaluate this question, we relate field-level annual PhD production to the number who acknowledge federal support. Formally, we estimate the following relationship:

$$PhdGraduates_{jt} = \beta \cdot SponsoredDissertations_{jt} + \alpha_j + \delta_t + \varepsilon_{jt} \tag{1}$$

where $j$ and $t$ index fields and years, and $\alpha_j$ and $\delta_t$ are associated fixed effects. We estimate this specification in levels and logs to evaluate counts and elasticities.

An important potential limitation is that this relationship may be confounded by unobserved factors. For example, PhD funding may target fields with high scientific potential, which attracts general enrollment—which would lead to overstating the effect of federal funding on PhD production. Conversely, federal agencies funding might seek to pull more students into fields that are important to their missions when they are undersubscribed, in which case the estimated relationship between federal funding and PhD production might be understated.

To overcome this challenge, we use a shift-share design to construct an instrument for sponsored dissertations (e.g., Goldsmith-Pinkham et al. 2020, Borusyak et al. 2022), exploiting agencies' annual number dissertations supported (the shift) and each field's share of agency dissertations in a base period (the share). This design exploits the variation in Figures 7 and 8—the combination of which creates variation in federal support in each field over time. For example, changes in total DoD and NIH research funding will (ostensibly) have different effects on aerospace engineering vs. biology. This is the variation that drives econometric identification.

Concretely, we (i) calculate field $j$'s share of sponsored dissertations from each funding agency $k$ in a base period $t_0$; (ii) interact it with the (actual and leave-one-out) number of funder $k$'s sponsored graduates in $t$, and (iii) take the sum across funders $k \in K$ to produce a field-specific predicted number of sponsored dissertations in each field-year $jt$:

$$\widehat{SponsoredDissertations}_{jt} = \sum_{k \in K} \omega_{jk} \cdot Dissertations_{kt} \tag{2}$$

where $\omega_{jk}$ is the share of graduates funded by agency $k$ in the base period and $Dissertations_{kt}$ is the number of dissertations funded by agency $k$ in year $t$, alternatively computed as a leave-one-out variant (excluding field $j$), $Dissertations_{kt}^{-j}$. For this analysis we examine 17 major fields (the ones shown in Figure 7) and restrict our sample to the 50 years of 1970 to 2019. We take three approaches to defining the base period for computing $\omega_{jk}$: in our preferred approach, we calculate shares for agency $k$ in year $t$ by averaging over the previous ten years, and in robustness checks we calculate shares from $t-10$ and from $t = 1970$. For this exercise we also condense our set of funding

agencies $K$ to the twelve largest funding agencies and an "other agencies" category, and to ensure we do not double-count dissertations with multiple federal sponsors we reallocate dissertations with multiple funding agencies to a 'multiple agencies' category.[7]

We thus estimate Equation (1) by ordinary least squares (OLS) and two-stage least squares (2SLS), in level and logged specifications, with unadjusted and leave-one-out instruments, and with three approaches to measuring the instrument (vis-à-vis how shares are computed). Our main results are presented in Table 3, where the instrument is computed using prior decade average shares (results for alternative constructions are statistically similar; see appendix). Columns (1) to (3) estimate outcomes in levels, and Columns (4) to (6) in logs (where we regress $ln(PhDGraduates)$ on $ln(\widehat{SponsoredDissertations})$). In each such sequence, we proceed from OLS, to 2SLS with a full-sample instrument, to 2SLS with a leave-one-out instrument.

We find that federal funding has a very large impact on PhD production. For every new PhD funded by the federal government, roughly two new PhD graduates are produced (Column 1). This result is statistically similar when estimated by OLS or 2SLS, where the instruments have very large predictive power. That this estimate is greater than one could reflect two possibilities. One is that federal funding crowds in other investments in PhD training; the other is that our data and methods undercount federally-sponsored dissertations, leading to over-estimated effects. For example, if there are latently two federally-funded dissertations for every one we observe, then a completely measured sample would have a regressor with twice the observed magnitude, and a $\hat{\beta}$ of half the magnitude (i.e., $\hat{\beta} \approx 1$). Columns (4) to (6) indicate that a doubling of federally-funded dissertations yields 85-90% more PhD graduates. These results are similarly stable across approaches; moreover, because natural logs are scale invariant, our estimation in logs is unaffected by any potential undermeasurement that may affect the magnitudes of $\hat{\beta}$ in previous columns.

Table 3: Effects of federal funding on PhD production

| | PhDs produced | | | Ln(PhDs produced) | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS |
| Govt-funded PhDs | 2.077*** | 2.028*** | 2.025*** | | | |
| | (0.029) | (0.031) | (0.033) | | | |
| Ln(Govt-funded PhDs) | | | | 0.829*** | 0.920*** | 0.921*** |
| | | | | (0.027) | (0.028) | (0.029) |
| N | 850 | 850 | 850 | 850 | 850 | 850 |
| $R^2$ | 0.99 | | | | | |
| F-stat | | 1906.08 | 730.45 | | 217.28 | 199.26 |
| Field FEs | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y |
| Leave-out | | | Y | | | Y |

Notes: Table estimates the relationship of annual PhD production in individual major fields to government-funded PhD graduates in those fields. Columns (1) and (4) estimate this relationship by OLS, and the remaining columns by two-stage least squares, using a shift-share instrument as described in the text. All specifications include fixed effects for field and year, and the estimation sample covers the post-1970 period. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

[7] The final list of federal funding agencies for this exercises is DOC, DoD, DOE, DOI, DOT, EPA, HHS, NASA, NSF, USDA, and the Departments of Education and State, to which we add a category for all other agencies and a category for dissertations funded by multiple agencies.

### 4.2 Long-run effects on the scientific workforce

Changes in public investment in PhD training have the potential for long-lasting effects: many newly-minted PhDs will remain in their PhD fields for the rest of their careers. Because careers are long, forces which shape scientific training (like funding patterns) could potentially impact the size and capabilities of the scientific workforce for many years after.

To evaluate whether investments in scientific training affect the long-run size and composition of the scientific workforce, we turn to the NSF's Survey of Doctorate Recipients (SDR)—a biennial survey of persons with a science or engineering doctoral degree from a U.S. institution. We specifically study the 2019 edition, which sampled roughly 120,000 doctoral degree-holders—around 10% of the estimated population of 1.15 million (NCSES 2021), the vast majority of whom are living and working in the United States. Based on this sample, the SDR public data provide (imputed) population counts of doctorate-holders by field and period, which are measured in 10-year intervals pre-1970 and 5-year intervals thereafter. We crosswalk degree fields to the SED major fields we have used throughout our analysis, and then merge the SDR data with our PQDT-based measures of dissertation funding. The result is a dataset of PhD cohorts by field and year, with PhD graduates since 1960, as seen in 2019. With the combined dataset, we can examine how the past 60 years of government support relates to the size of the modern workforce.

To do so, we adapt Equation 1 to estimate the relationship between the number of SDR scientists (in 2019) from a given field and cohort to the number of publicly-funded graduates in the field and time period they graduated. As before, we use a shift-share instrument to identify causal effects. We construct several variants of this instrument: our primary variant uses shares from the previous period, and in secondary approaches we apply shares from two periods prior and by averaging shares for the last two periods. We also develop leave-one-out variants.

Table 4 presents results in analogous format to Table 3. Put briefly, we find significant long-run effects of public investment in doctoral training on the size of the scientific workforce—a result which is likely a mechanical continuation of its effects on PhD production. The effects are directionally similar to but modestly attenuated compared to Table 3, potentially due to attrition from the workforce over time or incomplete sampling in the SDR. We provide several robustness checks in Appendix B.3, where we find consistent results using (i) different variants of the instrument, and (ii) a different sample derived from NSF's National Register of Scientific and Technical Personnel (NRSTP), an effort in the 1960s to produce a census of U.S. technical workers and which directly measures a much larger share of the PhD-holding population.

In Table 5 we examine the relationship between increases in PhD production and cohorts' long-term career outcomes, as reported by the SDR. We specifically evaluate the share of each cohort employed in the (i) academic tenure track, (ii) academic non-tenure track, and (ii) non-academic (e.g., industry or government) sectors. The first three columns of Table 5 do so in relation to the size of the overall cohort, and the latter three columns in relation to the number of government-funded graduates, in both cases instrumenting for the explanatory variable as before. We find that funding-fueled increases in cohort size on the margin increase the share of each cohort in non-tenure track academic and non-academic employment (the former statistically significant, the latter not), and significantly reduce the share of the cohort in academic tenure track employment. Although funding-fueled growth in PhD production increases the number of SDR-reported scientists in all

three tracks/sectors, the results indicate it disproportionately increases the number of contingent faculty relative to other categories—consistent with an outward shift of the doctoral labor supply curve without commensurate (offsetting) changes in demand.

Table 4: Effects of PhD funding on the future scientific workforce in 2019

| | Cohort size in SDR | | | Ln(Cohort size in SDR) | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS |
| Govt-funded PhDs | 1.374*** | 1.308*** | 1.221*** | | | |
| | (0.282) | (0.305) | (0.307) | | | |
| Ln(Govt-funded PhDs) | | | | 0.382*** | 0.580** | 0.583** |
| | | | | (0.097) | (0.228) | (0.249) |
| N | 154 | 154 | 154 | 140 | 140 | 140 |
| $R^2$ | 0.92 | | | | | |
| F-stat | | 274.66 | 117.14 | | 57.58 | 57.15 |
| Field FEs | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y |
| Leave-out | | | Y | | | Y |

Notes: Table estimates the relationship between the size of the scientific workforce in 2019 and the number of government-funded PhD graduates over the previous six decades. Unit of observation is a PhD cohort in 2019 (defined at the degree field x graduation period level, where periods are measured in 10-year intervals pre-1970 and 5-year intervals post-1970), and the outcome is the number of PhD-holding scientists in 2019 who graduated in these 5- and 10-year windows. Columns (1) and (4) estimate this relationship by OLS, and the remaining columns by two-stage least squares, using a shift-share instrument as described in the text. All specifications include fixed effects for field and year. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table 5: Effects of funding-driven growth in PhD production on tenure track employment

| | Share of cohort | | | Share of cohort | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | TT Job | Non-TT | Non-educ. | TT Job | Non-TT | Non-educ. |
| Ln(Graduating PhDs) | -0.127** | 0.066** | 0.061 | | | |
| | (0.061) | (0.033) | (0.052) | | | |
| Ln(Govt-funded PhDs) | | | | -0.086** | 0.045** | 0.041 |
| | | | | (0.040) | (0.022) | (0.034) |
| N | 98 | 98 | 98 | 98 | 98 | 98 |
| F-stat | 42.51 | 42.51 | 42.51 | 63.00 | 63.00 | 63.00 |
| Field FEs | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y |
| Y mean | 0.28 | 0.06 | 0.66 | 0.28 | 0.06 | 0.66 |

Notes: Table estimates the relationship between cohorts' career tracks in 2019 and the number of government-funded PhD graduates in their PhD cohort. Unit of observation is a PhD cohort in 2019 (defined at the degree field x graduation period level, where periods are measured in 5-year intervals post-1985), and the outcome is the share of each cohort with tenure track, non-tenure track, and non-educational employment. All columns estimate this relationship by two-stage least squares, using a shift-share instrument as described in the text. All specifications include fixed effects for field and year. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

### 4.3 Is the U.S. government subsidizing foreign science?

Conditional on producing more scientists, a distinct question is whether publicly-funded PhD scientists stay in the U.S.—and in turn, whether the human capital borne of U.S. government investment remains a resource for U.S. firms and universities (versus foreign ones). This question is particularly important in the current era of increasing geopolitical competition between the U.S. and China in science and high-tech industries, though also relevant broadly.

To answer this question, we are currently (and as of January 10, 2025, nearly finished) developing new dataset of PQDT graduates linked to their future scientific publications 5 to 15 years post-PhD, which can be geographically located through institutional affiliations. We have concurrently also been granted access to SED microdata (as of January 10, 2025), which report U.S. PhD graduates immediate post-graduation plans. Using these two sources, we plan to evaluate the propensity for U.S. PhD graduates to remain in the U.S. in both the short-run and long-run. Analysis is pending, and is expected to be completed by March 2025.

## 5 Discussion

There are several implications of these findings. First, who funds scientific training determines what scientists get trained and in what fields, and in turn what science is done. Second, investments in PhD training have long-run effects. These two insights are fundamental to understanding the science-based innovation system today as a product of past choices. Forward-looking funders may wish to be cognizant that investments in graduate training not only support current research, but will also shape science decades into the future. In addition, the fact that public funding expands the supply of PhDs, rather than crowding out other sources, suggests increased funding can be used to address shortages of scientific expertise in strategic areas. On the other hand, public subsidy can potentially also incentivize more graduate trainees than future job markets will demand at current wages, and result in marginal wage declines. Indeed, the evidence suggests that marginal graduates are more likely to end up underemployed relative to their training.

The evidence that financial support affects PhD production raises many more questions for future research. At the top of our list is representation in science. Science was historically an exclusionary institution on the basis of its high barriers to entry, especially vis-à-vis the high cost of training (Bush 1945). Several federal graduate fellowship programs have been designed to broaden participation in science—such as NSF graduate fellowships or National Defense fellowships (under the National Defense Education Act), both of which encouraged a broad geographic distribution. An outstanding question is whether these programs (or others) had the intended effect. Linking our data on graduate support to other resources like historical or modern census data, it is also possible to study characteristics of beneficiaries (e.g., parental income and education) and evaluate whether expansions in federal funds for PhD training enabled students from lower-income or otherwise underrepresented families to pursue scientific careers.

Other questions also remain. How does the level or source of funding change students' research or career trajectories? Are some fields or institutions higher-return targets for public funding than others? At what point do decreasing returns set in? How do direct financial versus in-kind support vary in their effects? And, ultimately, why is PhD training so expensive—and how might costs be contained? Answers to these questions will await the future.

# References


Antman, Francisca M, Xuechao Qian, Kirk Doran, and Bruce A Weinberg. 2023. *Innovation Nation: Evidence from Broadening Access to Ph.D. Training in the U.S.*. Working Paper.

Arora, Ashish, Sharon Belenzon, Larisa C Cioaca, Lia Sheer, and Hansen Zhang. 2023. "The Effect of Public Science on Corporate R&D,"Technical report, NBER Working Paper No. 31899.

Arrow, Kenneth. 1962. "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity: Economic and Social Factors*: Princeton University Press, pp. 609–626.

Bikard, Michaël, Fiona Murray, and Joshua S Gans. 2015. "Exploring trade-offs in the organization of scientific work: Collaboration and scientific reward," *Management Science*, Vol. 61, No. 7, pp. 1473–1495.

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.*: O'Reilly Media Inc.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb. 2020. "Are ideas getting harder to find?" *American Economic Review*, Vol. 110, No. 4, pp. 1104–44.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-experimental shift-share research designs," *The Review of Economic Studies*, Vol. 89, No. 1, pp. 181–213.

Buffington, Catherine, Benjamin Cerf, Christina Jones, and Bruce A Weinberg. 2016. "STEM training and early career outcomes of female and male graduate students: Evidence from UMETRICS data linked to the 2010 census," *American Economic Review*, Vol. 106, No. 5, pp. 333–338.

Bush, Vannevar. 1945. *Science, the Endless Frontier: A report to the President.*

Colby, Glenn. 2023. *Data Snapshot: Tenure and Contingency in US Higher Education.*. American Association of University Professors, available at https://www.aaup.org/article/data-snapshot-tenure-and-contingency-us-higher-education.

Cyranoski, David, Natasha Gilbert, Heidi Ledford, Anjali Nayar, and Mohammed Yahia. 2011. "Education: The PhD factory," *Nature*, Vol. 472, pp. 276–279.

Gat, Noam. 2023. "lm-format-enforcer," https://github.com/noamgat/lm-format-enforcer.

Gibbons, Michael T. 2022. *Universities Report Largest Growth in Federally Funded R&D Expenditures since FY 2011.* National Center for Science and Engineering Statistics (NCSES), Report no. NSF 23-303, available at https://ncses.nsf.gov/pubs/nsf23303.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift. 2020. "Bartik instruments: What, when, why, and how," *American Economic Review*, Vol. 110, No. 8, pp. 2586–2624.

Gross, Daniel P. and Bhaven N. Sampat. 2023. "The World War II crisis innovation model: What was it, and where does it apply?" *Research Policy*, Vol. 52, No. 9, p. 104845.

Jiang, Xuan, Joseph Staudt, and Bruce A Weinberg. 2023. *A tale of two fields? STEM career outcomes.* NBER Working Paper No. 31835.

Kim, Dahyun, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. "SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling."

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. "Efficient Memory Management for Large Language Model Serving with PagedAttention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lazear, Edward P. 2009. "Firm-specific human capital: A skill-weights approach," *Journal of Political Economy*, Vol. 117, No. 5, pp. 914–940.

Mansfield, Edwin. 1991. "Academic research and industrial innovation," *Research Policy*, Vol. 20, No. 1, pp. 1–12.

Mowery, David C and Bhaven N Sampat. 2006. "Universities in national innovation systems," in Fagerberg, Jan and David C Mowery eds. *The Oxford Handbook of Innovation*: Oxford University Press.

National Center for Science and Engineering Statistics (NCSES). 2021. *Survey of Doctorate Recipients, 2019*. NSF 21-320, available at https://ncses.nsf.gov/pubs/nsf21320.

National Science Foundation. 2022. *Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS)*. Available at https://ncses.nsf.gov/surveys/graduate-students-postdoctorates-s-e/2022.

Nelson, Richard R. 1959. "The simple economics of basic scientific research," *Journal of Political Economy*, Vol. 67, No. 3, pp. 297–306.

Nelson, Richard R. and Edmund S. Phelps. 1966. "Investment in Humans, Technological Diffusion, and Economic Growth," *American Economic Review*, Vol. 56, No. 1/2, pp. 69–75.

Nelson, Richard R and Nathan Rosenberg eds. . 1993. *National innovation systems: A comparative analysis*. New York: Oxford University Press.

Nguyen, Phuc, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. 2021. "SemTab 2021: Tabular Data Annotation with MTab Tool," in *SemTab@ISWC 2021*, Vol. 3103 of CEUR Workshop Proceedings, pp. 92–101: CEUR-WS.org, URL: http://ceur-ws.org/Vol-3103/paper8.pdf.

Pal, Arka, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. "Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive," *arXiv preprint arXiv:2402.13228*.

ProQuest. 2023. *ProQuest Dissertations & Theses Global*. Accessed September 2023.

Romer, Paul M. 1990. "Endogenous technological change," *Journal of Political Economy*, Vol. 98, No. 5, Part 2, pp. S71–S102.

Rosenberg, Nathan and Richard R Nelson. 1994. "American universities and technical advance in industry," *Research Policy*, Vol. 23, No. 3, pp. 323–348.

Smolka, Adam J, Perry V Halushka, and Elizabeth Garrett-Mayer. 2015. "The faculty costs to educate a biomedical sciences graduate student," *CBE—Life Sciences Education*, Vol. 14, No. 1, p. ar3.

Toole, Andrew A and Dirk Czarnitzki. 2010. "Commercializing science: Is there a university "brain drain" from academic entrepreneurship?" *Management Science*, Vol. 56, No. 9, pp. 1599–1614.

# Web Appendix

# A Data Appendix

## A.1 Sample Construction

Our goal in constructing the dataset is to obtain comprehensive coverage of physical, life sciences, and engineering doctoral dissertations from the U.S. Our starting point is the Proquest Dissertations & Theses Global (PQDT) database. At the time of extraction, the database included 5,708,925 Master's theses and doctoral dissertations spanning various fields and institutions worldwide.

We filter the PQDT database through a multi-step process to isolate U.S. doctoral dissertations in the natural sciences. First, we remove all dissertations from non-U.S. universities using the institution and degree metadata provided by PQDT. This initial filtering step reduces our dataset to 3,178,082 U.S. dissertations.

Next, we identify and exclude dissertations with degree descriptions containing the term "Master", such as "Master of Science" and "Master of Arts", to remove theses and non-doctoral dissertations. We then manually review all degree names present in the dataset and identify 111 unique degree names that explicitly correspond to doctoral degrees in the natural sciences. While almost all dissertations are marked as "Ph.D", some doctoral degrees are marked as "D.CS." (denoting "Doctor of Computer Science", "Chem. D." (denoting "Doctor of Chemistry"), and "Dr. Phil." (denoting "Doctor of Philosophy). some doctoral degrees, such as "S.T.D." (denoting "Doctor of Sacred Theology"), fall outside the scope of our dataset.

In the third step, we filter the doctoral dissertations based on their research field. For this, we leverage the subject terms provided by PQDT, which describe the content of each dissertation. Focusing on the first subject term, which represents the main subject field that best describes the dissertation's area of study, we identify 3,640 terms belonging to the natural sciences out of the 8,238 unique first subject terms found across all dissertations. We disregard dissertations related to 4,598 subject terms that represent the social sciences and humanities.

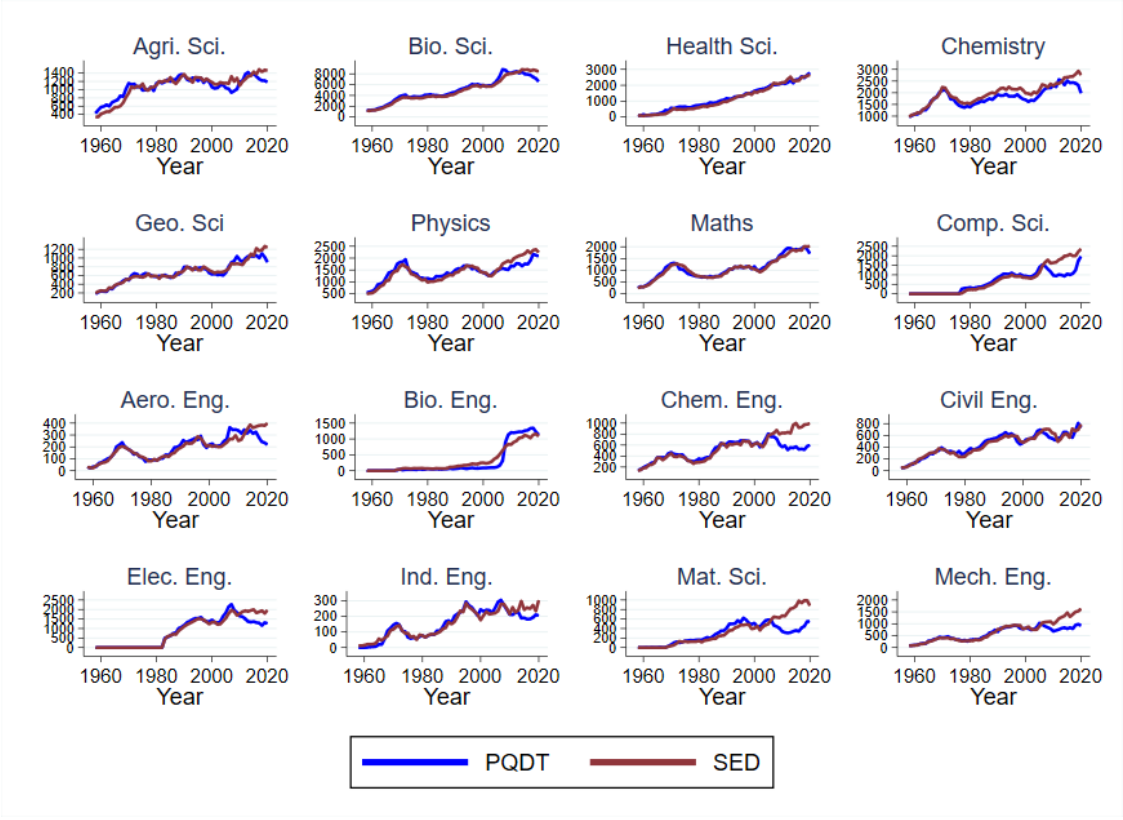
Lastly, we are interested in grouping the dissertations into top fields. Given the substantial number of unique first subject terms (3,640) in our data, we match them with the subject categories used in the Survey of Earned Doctorates (SED). Within the SED's Science and Engineering category, there are four major subjects: Life Sciences, Physical Sciences and Earth Sciences, Mathematics and Computer Sciences, and Engineering. These four major subjects are further divided into 18 minor subjects. We carefully map each of the 3,640 unique first subject terms to one of these 18 SED minor subjects to ensure consistency and comparability with established categories.

Our final sample consists of 1,246,495 U.S. doctoral dissertations in the natural sciences, spanning the years 1900 to 2022. Of these, we have the text available for 909,034 dissertations. Analyses presented in the paper focus on the years 1950-2020. In this time window, our sample includes 845,442 dissertation texts.

As a complement to Figure 1, Figure A.1 offers a detailed comparison between the dissertation counts from PQDT and the SED, categorized by minor subjects available in the SED. As outlined

in section 3.1, the SED annual counts serve as a benchmark for evaluating the coverage and representativeness of our sample. The analysis reveals that the PQDT counts (represented by blue lines) exhibit a close alignment with the SED counts (represented by red lines), mirroring the patterns observed in the total count comparison. However, it is important to acknowledge a slight divergence in coverage during the more recent years, which may be attributed to factors such as the scope of PQDT coverage, potential labeling errors, or temporal lags in data availability. Notwithstanding these minor discrepancies, the overall comparisons provide compelling evidence that our dissertation sample, sourced from PQDT, serves as a reliable proxy for the underlying population represented by the SED.

Figure A.1: ProQuest vs. SED dissertations in the hard sciences, by subject and year




Notes: Figure shows annual PQDT dissertation counts in the physical and life sciences and engineering (blue bars) and SED counts (red line) for comparison, from 1958 (the first year of the SED) to 2020, disaggregated across 16 SED major fields. We crosswalk PQDT dissertations to these major fields on their author-provided, first-listed subjects, using a manually-curated crosswalk developed from the concordance used by SED to map its PhD graduate survey responses to these fields.

## A.2 Identifying dissertation sponsors

Having established a dissertation sample and obtained their full text, the next step is to identify research sponsors in this text—a named entity recognition problem for which we turn to natural language processing tools. To do so, we have crafted a six-step pipeline:

Step 1. *Isolate sentences from within the text*

Step 2. *Identify named entities in these sentences*

Step 3. *Classify entities into sectors (i.e., government, private companies, non-profit), etc.*

Step 4. *Identify support type and extract grant identifiers, where available*

Step 5. *Consolidate named entities into parent organizations*

Step 6. *Match extracted organizations with external registries of funders*

The procedure we develop uses of large language models (LLMs) in several steps, particularly Upstage.ai's Solar 10.7B and Abacus.ai's Smaug 34B models (Kim et al. 2023, Pal et al. 2024).

### A.2.1 Isolating acknowledgment sentences in dissertation full text

To isolate sentences that acknowledge funding in the dissertation texts, we employ a rule-based string matching approach. First, we split the raw dissertation text into sentences using NLTK's Punkt Sentence Tokenizer (Bird et al. 2009). We modify the tokenizer to ignore common academic abbreviations such as "Dr.," "Prof.," and "univ." Next, we iterate through each sentence, searching for words that indicate a potential acknowledgement or support mention. After careful refinement, we arrived at the list of keywords as appears in Listing 1 below.

Listing 1: Keyword list for acknowledgement sentence extraction

```
keywords = ["contract", "fund", "grant", "financ", "fellow", "scholarship
    ", "internship", "award", "generous", "generos", "support", "thank", "
    grateful", "appreciate", "assist", "help", "mentor", "journey", "
    guidance", "effort", "encourag", "opportunity", "provide", "gratitude",
     "committee", "advisor", "sponsor", "appreciat", "paying", "investing",
     "invested", "indebt", "obtain", "money", "stipend", "facilitat", "
    blessed", "acknowledgements", "biography", "vita", "preface"]
```

### A.2.2 Identifying named entities in acknowledgments text

In steps 2 through 4, we utilize a Named Entity Recognition (NER) process employing a generative large language model (LLM). Specifically, we use Upstage.ai's Solar 10.7B Instruct v1.0 (Kim et al. 2023), a fine-tuned model based on the Mistral 7B model (which, in turn, is based on the Llama-2 architecture).

By carefully refining the prompt, we established a procedure that initially verifies if the sentence acknowledges support from an external entity. Subsequently, we prompt the model to identify the supporting entities mentioned in the text, categorize them by the type of organization and support, and document any grant and contract identifiers where applicable.

For batch inference, we utilize the VLLM Python package (Kwon et al. 2023) and the LM format enforcer (Gat 2023). These tools facilitate efficient inference and require the model to produce a well-structured JSON schema that we predefined. Our prompt to the model is structured as appears in Listing 2 below:

Listing 2: LLM prompt for extracting supporting entities

```
### System:
You are a helpful assistant. Always answer as helpfully as possible.

### User:
Below is a sentence taken from a doctoral dissertation.

Instructions:
1. Determine whether in the provided text, the author expresses gratitude
   or mentions support from some organization.
    - set ‘gratitude_support‘ to true if yes, and false otherwise.
2. Determine whether in the provided text is part of a CV or biography.
   Set ‘biography_vita‘ to true if yes and false otherwise.
    - Note that biographies typically list awards and recognitions.
3. Extract names of supporting organizations mentioned in the text. Notes:
    - Skip people’s names. I don’t care about advisors, mentors, committee
      members etc.
    - Ensure complete extraction of grant, contract, and award identifiers
      , if present.
    - Where necessary, infer the organization’s name and type based on the
      grant name.
    - Represent each financial support and funding source as a distinct
      entity.
    - Ignore doctoral committees, academic departments, colleagues, lab
      members, unaffiliated individuals, family and friends.
    - Ensure that each organization is listed separately.
    - Return an empty list if no supporting organizations are mentioned.

Text:
{sentence}

You MUST answer using the following json schema:
{extractionFormat.schema_json()}

### Assistant:
```

To validate the structure of the LLM output, we use two Pydantic models as described in Listing 3 below. The `orgFormat` model includes fields for the organization's name and type, as well as the type of support provided. The organization and support types are restricted to a set of predefined values. The `extractionFormat` model incorporates the `orgFormat` model as a list, allowing for multiple organizations to be associated with each analyzed sentence. Additionally, it includes boolean fields for indicating whether the sentence expresses gratitude or support, or is a part of a biography. These models serve to standardize the representation of extracted support entities and facilitate subsequent data processing steps.

Listing 3: Pydantic model for LLM output

```
class extractionFormat(BaseModel):
    gratitude_support: bool
    biography_vita: bool
    organizations: List[orgFormat]

class orgFormat(BaseModel):
    organization_name: str
    organization_type: Literal['government', 'federal agency', 'non-profit
        ', 'private company', 'foundation', 'academia', 'research lab', '
        consortium', 'center', 'association', 'other', 'unknown']
    support_type: Literal['funding', 'grant', 'financial support', '
        internship', 'scholarship', 'fellowship', 'training', 'materials',
        'facilities', 'data', 'in-kind support', 'technical support', '
        administrative support', 'employment', 'unspecified','other', '
        unknown']
    grant_contract_numbers: str
```

After running this procedure on our sample of dissertations, we extracted over 92 million sentences that potentially included acknowledgments of supporting entities. We then performed batch inference on these sentences, one at a time, using the language models as described above. The language model output indicated that 11 million sentences (about 12%) were indeed sentences that acknowledged support or were part of a biography section. Among all sentences, the language model extracted 9.3 million mentions of organizations, with 4.8 million of these mentions originating from sentences classified as support or biography. Aggregating these entities resulted in a list of 4 million unique organization names.

### A.2.3 Consolidate named entities and match to external registries

The next step in the data construction is to match the organization names in our sample to external registries of funding organizations. We rely on two main registries. The first is the Research Organization Registry (ROR), which provides persistent identifiers for research organizations. ROR aims to disambiguate institution names and connect research organizations to researchers and research outputs. It is an openly available identifier focused on identifying affiliations in scholarly metadata and is designed to be integrated into scholarly infrastructure. The second registry we use is Wikidata, a free and open knowledge base by Wikimedia that acts as central storage for structured

data. Wikidata provides a collaborative platform for storing and managing information about various entities, including research organizations. By leveraging the data available in Wikidata, we can further enhance the accuracy and completeness of our organization matching process.

**Government Entities**

We consolidate organizations that are part of the U.S. Federal Government and match them to ROR entities. First, we filter our sample for names that the LLM classified as "government" or "federal agency". We apply some standardization using regex (e.g., dropping the "U.S." for the beginning of strings) and cluster together similar names. Among 739,494 thousand clustered government organization names, we end up with 242,382 names are associated with at least one sentence that the LLM classified as related to support or part of the author's biography.

To further consolidate the list of clustered governmental names, we run them again through a large language model. For this task, we use Abacus.ai's Smaug 34B v0.1 model (Pal et al. 2024). This model is larger than the one we used for the first round. It requires more resources to run, but performs better in correctly identifying organizations and linking them to their parent organizations. Our prompt in this round appears below, in Listing 4. The required schema includes fields for the organization name, the parent organization, indicators for the Federal Government and independent agencies, and location.

Listing 4: LLM prompt for consolidating governmental organizations

```
### System:
You are a helpful assistant. Always answer as helpfully as possible.

### User:
Below is a name of a funding organization, most likely in the US.
Return a json schema that includes:
1. The official, clean, unabbreviated name of the organization.
2. The ultimate federal government department or agency (if relevant).
3. Indicators for whether the organization is a federal organization and/
   or an independent agency.
4. Country and state. If irrelevant return "-".

Don't make stuff up. If you are unsure, return "unknown".

You MUST answer using the following json schema:
{entityFormat.schema_json()}

Organization name: {organization}

### Assistant:
 """
```

Among the initial consolidated list of 242,382 entities, the model identified 90,655 as related to the U.S. federal government (either directly or as an independent agency). The model produced 3,550 unique parent organization names. We matched these names to a list of governmental organizations

from the ROR dataset. Weighted by the number of mentions, the match resolved 81% of the mentions of U.S. governmental organizations and matched them with a ROR identifier.

**Private Companies**

The list of extracted entities includes 448,745 unique names that are classified as private companies. Within them, 170,009 names are associated with at least one sentence that the LLM classified as related to support or part of the author's biography. We start the matching process by using the Smaug 34B LLM to consolidate these names. The prompt is presented in Listing 5 below. The results were further validated using an additional run through the LLM and manual comparisons.

Listing 5: LLM prompt for consolidating private companies

```
### System:
You are a helpful assistant. Always answer as helpfully as possible.

### User:
Below is a name of a grant or award.

Return a json schema that includes:
1. The ticker name for the firm. Return past ticker where relevant.
2. The name of the firm, associated with the ticker.
3. The relevant stock exchange.

Don't make stuff up. If you are unsure, return "unknown".

You MUST answer using the following json schema:
{entityFormat.schema_json()}

name: {organization}

### Assistant:
 """
```

The consolidation process resulted in 10,802 names. First, we used the ROR REST API to search for ROR identifiers that are associated with these names. Second, we used the SemTab API to find additional matches with the Wikidata identifiers (Nguyen et al. 2021). Third, we ran additional API searches using the original, unconsolidated list of names that were not previously matched. This step allowed for matches of firms that are not publicly traded and startups. Finally, we validated the list of resulting matches using similar techniques to those described above. The resulting match includes a ROR or Wikidata ID for 217,013 mentions of firms, corresponding to 123,366 dissertations. These matches are linked to 3,275 unique ROR identifiers and 7,638 unique Wikidata identifiers.

**Non-profit Organizations**

We matched mentions of non-profit organizations in our data with external identifiers. The

original list included 549,329 unique strings that were classified as foundations, philanthropic organizations, non-profits, associations or centers. Among them, 322,901 names were associated with at least one sentence that the LLM classified as related to support or part of the author's biography. For matching these mentions, we used a similar process to the one used for firms. Overall, we were able to match 333,603 mentions, corresponding to 157,884 dissertations. These matches are linked to 2,139 unique ROR identifiers and 8,126 Wikidata identifiers.

### A.3 Validating measures against NSF graduate fellowships

One opportunity for validating our measurement of dissertation sponsors is against NSF graduate research fellowship (GRF) awardees, which we retrieve from NSF's website from 1952 (the first year of the program) onwards.[1] Specifically, we wish to know how many of these awardees are in our PQDT sample, and how many we are able to measure as having received NSF support—and if not, why not: due to incomplete measurement or incomplete reporting?

To prepare for linking to PQDT, we first clean the PQDT and NSF data under common parameters, standardizing names and separating names into first names, surnames, and middle names/initials. We then crosswalk the subjects reported by each awardee to the SED major fields and institutions to which we have harmonized our PQDT data. Importantly, the utility of this information is limited by the fact that awardees' reported fields often represent intended (rather than actual) fields of study. Nevertheless, these fields have information, which we will selectively use.[2] The NSF data present other challenges, including that individuals are sometimes awarded multiple times (implying that they chose not to use the award in one year, and later applied again); to deal with these challenges we attempt to de-duplify the data by identifying awards to same-named individuals within ten years and restricting the sample to the latest award for every individual.

We then link NSF GRF awardees to PQDT in an iterative procedure. We first attempt to link on last name (LN), first name (FN), and middle initial (MI). Among the unlinked set, we then link on LN and FN; then, on LN, first initial (FI), and MI. At every step we keep only links where the PQDT graduation year was within 10 years of the NSF GRF award year, and we discard links where available, unused information would rule them out (e.g., when linking on LN and FN, if MI does not match we discard the link). When multiple links are made from the NSF GRF list to PQDT or vice versa, we use subjects to disambiguate, and we then attempt to manually disambiguate any remaining ambiguous cases. Our goal throughout this procedure is to produce a highly precise link for validation including at the expense of recall.

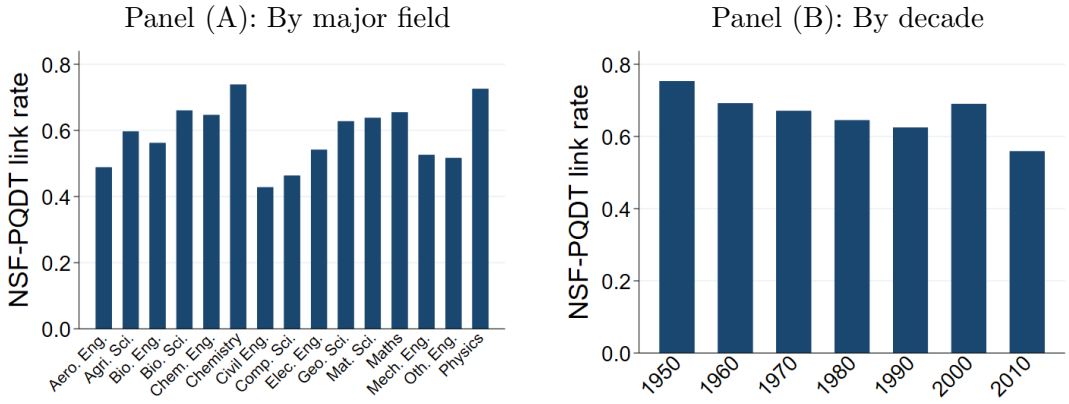
Through this process we successfully (unambiguously) link 65% of GRF awardees to our PQDT sample. Figure A.2 illustrates how this link rate varies across NSF GRF awardees' reported subjects and over time, where it varies between 40% and 75% across subjects and steadily declines over time,

[1] See https://www.research.gov/grfp/AwardeeList.do?method=loadAwardeeList.

[2] The NSF data also include reported institutions; however, because many GRF applicants are pre-PhD, awardees' current institutions are often their baccalaureate or master's program rather than their PhD program. We found the institution data to be too ambiguous to use in this exercise.

from ≈75% in the 1950s to ≈55% in the 2010s. There are several reasons why these link rates may be below 100%. The most salient among them are that the NSF sample can include awardees who are not accepted to PhD programs, decline to enroll, or never graduate. A second reason is residual ambiguity in name-based links, though because the NSF and PQDT samples are specific and narrow slices of the general population, sampled on a common feature (potential or actual PhD trainees in the hard sciences), and because full names and specializations are provided in both datasets, we believe this link is both precise and (mostly) complete.

Figure A.2: Link rates of NSF GRF awardees to our PQDT sample




Notes: Figure shows the share of NSF graduate research fellows that we are able to link the PQDT sample, by major field (Panel A) and by decade (Panel B). Link rates below one may reflect awardees who are not accepted to PhD programs, decline to enroll, or never graduate.

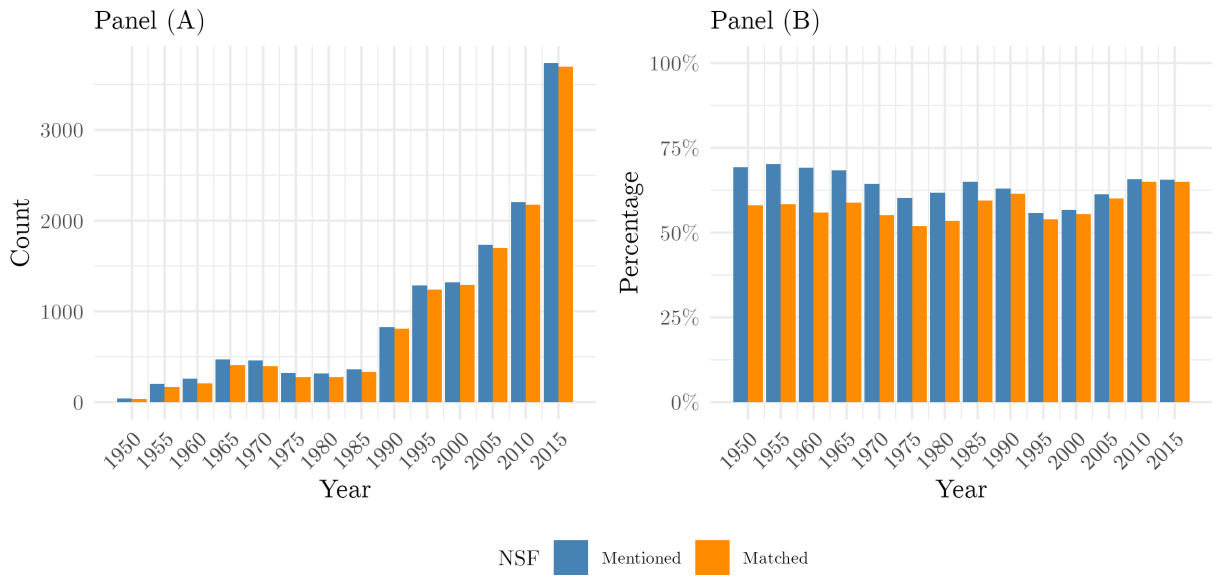
We next evaluate how our text-based measurement of dissertation sponsors performs within this linked sample. Within our sample of 909 thousand dissertations, we identify 25,039 dissertations whose authors are linked to the NSF GRF awardee list. Figure A.3 presents an analysis of NSF mentions and matching rates within this sample. Among the NSF awardees, 15,229 (60.8%) mention the NSF somewhere in the text.[3] Our matching process links 15,026 (98.6%) of these mentions to the NSF identifier in ROR. In addition, the process identifies another 746 dissertations as funded by the NSF, which the simple search missed. These results provide an indication for very high levels of extraction rates. Of course, our extraction is dependent on the dissertations authors to mention their funding sources.

## A.4 Validating measures against GSS

A second opportunity for validation is possible using the NSF's Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). The GSS is an annual census of U.S. institutions granting research-based master's degrees or doctorates in science and engineering (National Science Foundation 2022). Usefully for our purposes, the data report primary sources of support for graduate students at doctorate-granting institutions, providing annual counts of graduate students in each

[3]For this test, we perform a simple search of "national science foundation" or "nsf " in the sentences we extract from the dissertation texts.

Figure A.3: NSF Graduate Fellows Matched to NSF Funding




Notes: Figure shows the counts (Panel A) and shares (Panel B) of NSF Graduate Fellows mentioning NSF funding support (in blue) and matched to the NSF by our matching process (in orange). Sample restricted to NSF fellows which we can link to PQDT.

major field with specific sources of funding: DoD, DOE, NIH and other HHS funders, NASA, NSF, USDA, other government agencies, institutional support, other U.S. or foreign sources, and self-support. Our goal here is to compare the field-year share of students in GSS funded by each of these agencies to the share we measure in our PQDT-based dissertation sample.

Several limitations are worth noting. First is the obvious caveat that the GSS measures enrolled students, whereas our PQDT sample measures graduates. An implication is that the GSS sample in a given year is much larger than PQDT. To a first order approximation, however, we expect that doctoral students' sources of support in a given year will correlate strongly with graduates' sources of support. A second, potentially more important caveat is that although we restrict the GSS measures to doctorate-granting institutions, this sample will include master's and doctoral students at these institutions. Sources of financial support can differ significantly across degree levels, and federal funding is particularly focused on supporting doctoral training (whereas master's training is often self-funded). In light of this, we will focus on comparing agency shares of federally-funded students (rather than all students) in GSS and our PQDT sample. Finally, we restrict our attention to 2000 to 2019, when GSS reported supports for all six of our focal agencies for this exercise (DoD, DOE, HHS, NASA, NSF, USDA; DOE wasn't added until 1999).

Specifically, we regress (i) the share of federally-funded PhD graduates in a given field and year which we measure as being supported by each of these agencies, on (ii) the share of federally-funded GSS respondents at doctoral institutions primarily supported by these agencies. If the samples perfectly aligned, and our measurement were accurate, we would expect these to correlate roughly

one-for-one (i.e., $\beta = 1$). Given sampling differences, as well as the fact that GSS measures primary sources of funding and our methodology will measure all sources of support reported in dissertations, these shares may not vary one-for-one, but they should correlate.

Table A.1 provides the results. We see strong and highly significant correlations across the table. For every 10% of federally-funded students in GSS with DoD support, we see an 8.7% higher rate of DoD funding in our PQDT data, 12.8% for DOE, 9.8% for HHS, 14.2% for NASA, 9.9% for NSF, and 7.5% for USDA. For HHS and NSF (the two largest funders of doctoral training since 2000), we cannot reject that the GSS and PQDT-based measures correlate one-for-one. More broadly, we view the strong correlation between the GSS and PQDT-based measures in this table as validating evidence for our approach to measuring research sponsors.

Table A.1: Comparing PQDT-based measures to NSF GSS

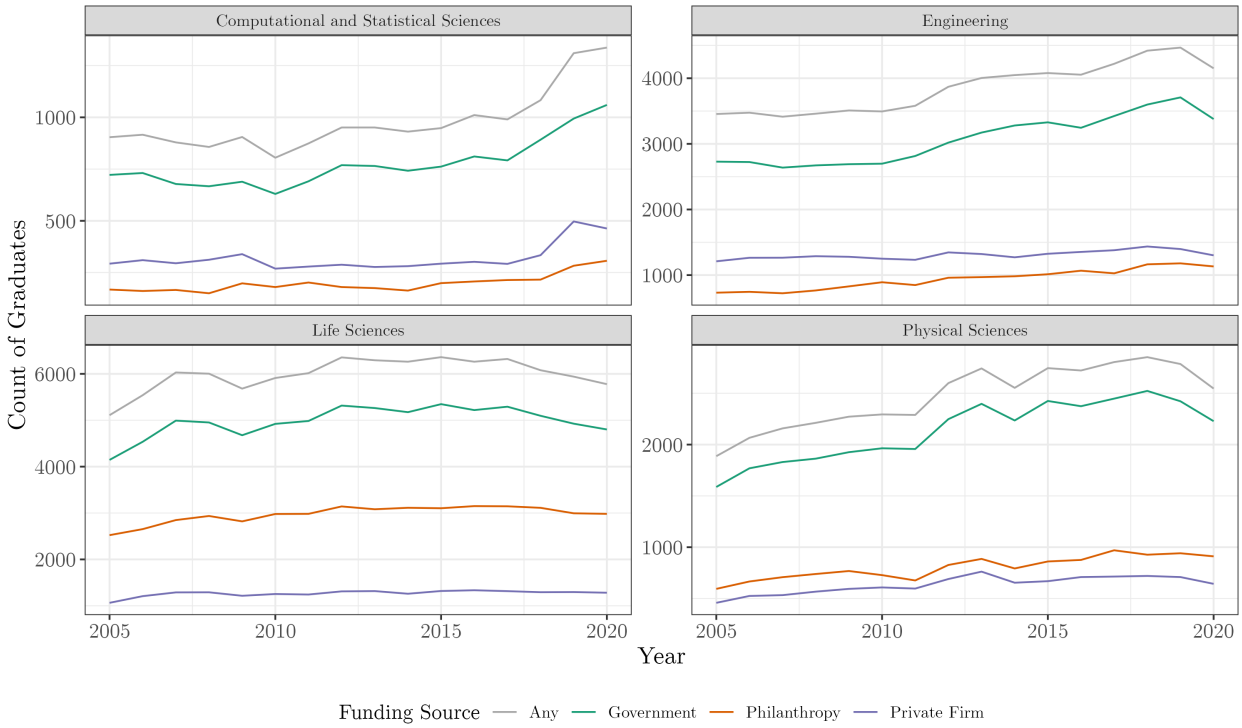
| | (1) DoD | (2) DOE | (3) HHS | (4) NASA | (5) NSF | (6) USDA |
|---|---|---|---|---|---|---|
| Share of GSS graduates | 0.866*** | 1.281*** | 0.981*** | 1.420*** | 0.987*** | 0.747*** |
| | (0.039) | (0.058) | (0.016) | (0.051) | (0.031) | (0.024) |
| N | 340 | 340 | 340 | 340 | 340 | 340 |
| $R^2$ | 0.72 | 0.72 | 0.96 | 0.86 | 0.71 | 0.93 |
| P-val: $\beta = 1$ | 0.00 | 0.00 | 0.26 | 0.00 | 0.66 | 0.00 |

Notes: Table correlates (i) the share of PhD graduates in each major field and year which we measure as having been supported by a given agency to (ii) the GSS-reported share of enrolled PhD students in that field and year reporting support from that agency. Sample includes 17 major fields measured between 2000 and 2019. Variables are not perfectly comparable because one measures PhD graduates and the other measures enrolled students, but if our dissertation-based measurement of PhD support is precise, we anticipate they will strongly correlate, and if complete, regression coefficients ($\hat{\beta}$) should be near 1. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

# B Supplementary Results

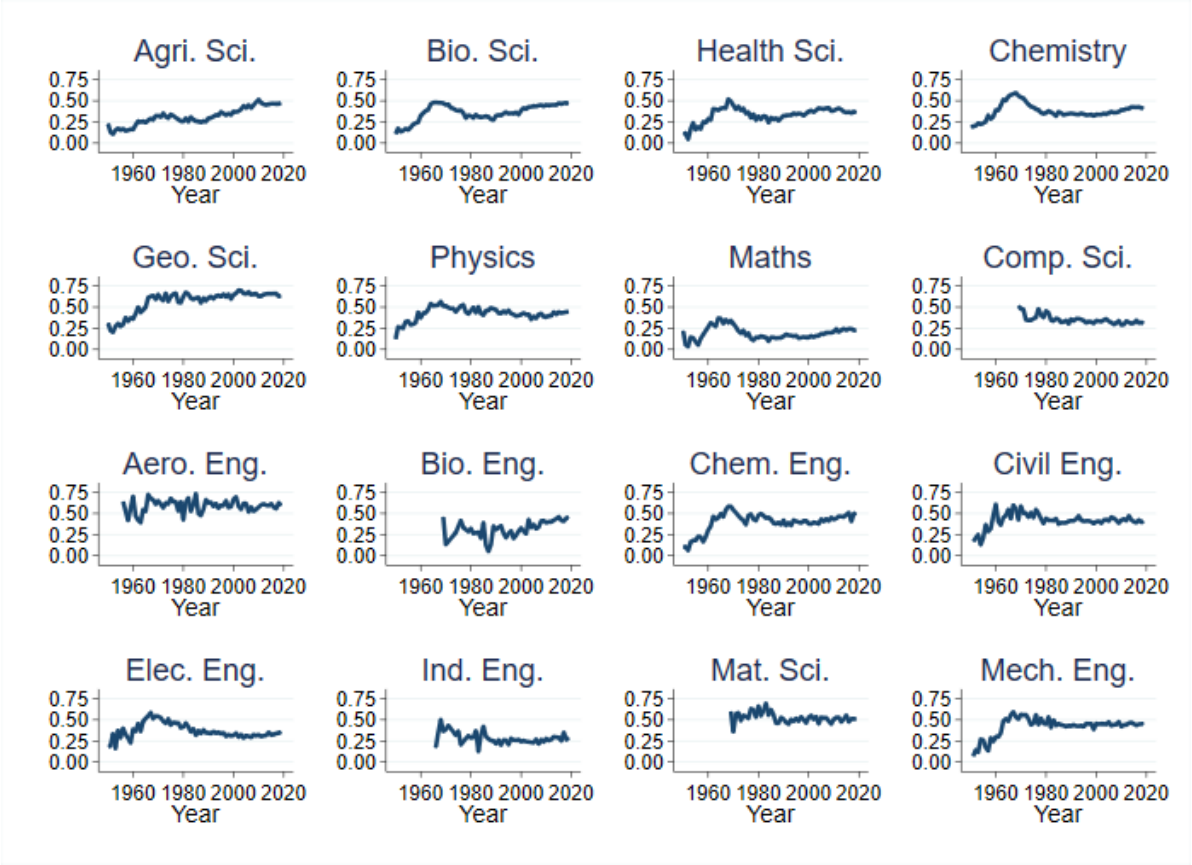
## B.1 Characteristics of the sample

Figure B.1: Counts of U.S. PhD graduates funded by U.S. federal agencies, private companies, and non-profit organizations, by broad field, 1950-2020




Notes: Figure illustrates the count of PhD graduates supported over time by government agencies, firms, and non-profit organizations, by broad research fields.

Figure B.2: U.S. government-funded share of dissertations, by field, 1950-2019




Notes: Figure shows government-funded share of dissertations by major field, from 1950 to 2019.

## B.2 Effects of government funding on PhD production

Tables B.1 and B.2 provide robustness checks on Table 3 using alternative constructions of the shift-share instrument, calculated from (i) prior decade average shares (as in Table 3), (ii) 10 year ago shares, and (iii) 1970 shares, all of which yield qualitatively (and often statistically) similar results—though the results based on 1970 shares tend to be noisier.

Table B.1: Effects of federal funding on PhD production; results from all instruments in levels

| | (1) OLS | (2) 2SLS-1 | (3) 2SLS-2 | (4) 2SLS-3 | (5) 2SLS-1 | (6) 2SLS-2 | (7) 2SLS-3 |
|---|---|---|---|---|---|---|---|
| Govt-funded PhDs | 2.077*** | 2.028*** | 2.001*** | 2.017*** | 2.025*** | 1.985*** | 2.385*** |
| | (0.029) | (0.031) | (0.034) | (0.051) | (0.033) | (0.039) | (0.286) |
| N | 850 | 850 | 850 | 850 | 850 | 850 | 850 |
| $R^2$ | 0.99 | | | | | | |
| F-stat | | 1906.08 | 612.07 | 65.58 | 730.45 | 291.78 | 4.49 |
| Field FEs | Y | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y | Y |
| Leave-out | | | | | Y | Y | Y |

Notes: Table estimates the elasticity of annual PhD production in individual major fields to government-funded PhD graduates in those fields. All columns provide estimations in levels. Column (1) estimates this relationship by OLS, Columns (2) to (7) do so by two-stage least squares using various shift-share instruments, constructed as follows: Column (2) using prior decade average shares, Column (3) using 10 year ago shares, Column (4) using 1970 shares, and Columns (5) to (7) repeating Columns (2) to (4) using leave-one-out shifters (see text for explanation). All specifications include fixed effects for field and year, and the estimation sample covers the post-1970 period. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table B.2: Effects of federal funding on PhD production; results from all instruments in logs

| | (1) OLS | (2) 2SLS-1 | (3) 2SLS-2 | (4) 2SLS-3 | (5) 2SLS-1 | (6) 2SLS-2 | (7) 2SLS-3 |
|---|---|---|---|---|---|---|---|
| Ln(Govt-funded PhDs) | 0.829*** | 0.920*** | 0.987*** | 0.747*** | 0.921*** | 1.001*** | 0.660*** |
| | (0.027) | (0.028) | (0.039) | (0.046) | (0.029) | (0.044) | (0.098) |
| N | 850 | 850 | 826 | 850 | 850 | 826 | 850 |
| $R^2$ | | | | | | | |
| F-stat | | 217.28 | 48.23 | 51.73 | 199.26 | 40.24 | 15.31 |
| Field FEs | Y | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y | Y |
| Leave-out | | | | | Y | Y | Y |

Notes: Table estimates the relationship of annual PhD production in individual major fields to government-funded PhD graduates in those fields. All columns provide estimations in logs. Column (1) estimates this relationship by OLS, Columns (2) to (7) do so by two-stage least squares using various shift-share instruments, constructed as follows: Column (2) using prior decade average shares, Column (3) using 10 year ago shares, Column (4) using 1970 shares, and Columns (5) to (7) repeating Columns (2) to (4) using leave-one-out shifters (see text for explanation). All specifications include fixed effects for field and year, and the estimation sample covers the post-1970 period. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

## B.3 Effects of government funding on scientific workforce

### B.3.1 Additional SDR-based results

Tables B.3 and B.4 provide robustness checks on Table 4 using alternative constructions of the shift-share instrument, calculated from (i) previous period (5/10-year) shares (as in Table 4), (ii) average shares from the previous two periods, and (iii) shares from two periods prior, all of which yield quantitatively and statistically similar results.

Table B.3: Effects of PhD funding on the future scientific workforce in 2019; SDR-based sample, results from all instruments in levels

| | (1) OLS | (2) 2SLS-1 | (3) 2SLS-2 | (4) 2SLS-3 | (5) 2SLS-1 | (6) 2SLS-2 | (7) 2SLS-3 |
|---|---|---|---|---|---|---|---|
| Govt-funded PhDs | 1.374*** | 1.308*** | 1.494*** | 1.199*** | 1.221*** | 1.479*** | 1.624*** |
| | (0.282) | (0.305) | (0.371) | (0.356) | (0.307) | (0.404) | (0.434) |
| N | 154 | 154 | 154 | 154 | 154 | 154 | 154 |
| $R^2$ | 0.92 | | | | | 0.56 | |
| F-stat | | 274.66 | 143.71 | 8.16 | 117.14 | 72.35 | 5.70 |
| Field FEs | Y | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y | Y |
| Leave-out | | | | | Y | Y | Y |

Notes: Table estimates the relationship between the size of the scientific workforce in 2019 and the number of government-funded PhD graduates over the previous six decades. Unit of observation is a PhD cohort in 2019 (defined at the degree field x graduation period level, where periods are measured in 10-year intervals pre-1970 and 5-year intervals post-1970), and the outcome is the number of PhD-holding scientists in 2019 who graduated in these 5- and 10-year windows. All columns provide estimations in levels. Column (1) estimates this relationship by OLS, Columns (2) to (7) do so by two-stage least squares using various shift-share instruments, constructed as follows: Column (2) using shares from the previous period, Column (3) using average shares from the previous two periods, Column (4) using shares from two periods prior, and Columns (5) to (7) repeating Columns (2) to (4) using leave-one-out shifters (see text for explanation). All specifications include fixed effects for field and year. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table B.4: Effects of PhD funding on the future scientific workforce in 2019; SDR-based sample, results from all instruments in logs

| | (1) OLS | (2) 2SLS-1 | (3) 2SLS-2 | (4) 2SLS-3 | (5) 2SLS-1 | (6) 2SLS-2 | (7) 2SLS-3 |
|---|---|---|---|---|---|---|---|
| Ln(Govt-funded PhDs) | 0.382*** | 0.580** | 0.613** | 0.360* | 0.583** | 0.625* | -0.574 |
| | (0.097) | (0.228) | (0.295) | (0.184) | (0.249) | (0.329) | (2.120) |
| N | 140 | 140 | 140 | 140 | 140 | 140 | 140 |
| $R^2$ | | | | 0.18 | | | |
| F-stat | | 57.58 | 41.14 | 46.10 | 57.15 | 34.19 | 0.53 |
| Field FEs | Y | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y | Y |
| Leave-out | | | | | Y | Y | Y |

Notes: Table estimates the relationship between the size of the scientific workforce in 2019 and the number of government-funded PhD graduates over the previous six decades. Unit of observation is a PhD cohort in 2019 (defined at the degree field x graduation period level, where periods are measured in 10-year intervals pre-1970 and 5-year intervals post-1970), and the outcome is the number of PhD-holding scientists in 2019 who graduated in these 5- and 10-year windows. All columns provide estimations in logs. Column (1) estimates this relationship by OLS, Columns (2) to (7) do so by two-stage least squares using various shift-share instruments, constructed as follows: Column (2) using shares from the previous period, Column (3) using average shares from the previous two periods, Column (4) using shares from two periods prior, and Columns (5) to (7) repeating Columns (2) to (4) using leave-one-out shifters (see text for explanation). All specifications include fixed effects for field and year. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

### B.3.2 Results from alternative data source: NRSTP

The SDR (Survey of Doctoral Recipients) has been administered as a survey of science, engineering, and health doctoral degree holders since 1973. Our analysis in this paper uses data from the 2019 edition of the SDR, but because it is produced from a roughly 10% sample of the estimated 1.15 million PhD recipients from U.S. institutions (through 2019; NCSES 2021), it uses survey weights to scale up to an approximate accounting of the population.

The predecessor to the SDR was a much more expansive NSF effort to produce a periodic census of the U.S. technical workforce, as required by the legislation which created the agency (the NSF Act of 1950, P.L. 81-507). This "National Register of Scientific and Technical Personnel" (NRSTP) was produced through a biennial survey administered to all members of U.S. academic societies in the 1950s and 1960s, and its high response rate among PhD-holders means that it approximates the contemporary population. The NRSTP measured many characteristics of the scientific workforce,

including these individuals' location, employer, employment function, salary, languages spoken, and more—and crucially for our purposes, education and field.[4]

We obtained the NRSTP raw response data from the U.S. National Archives for 1970 (the final year of the program) and carefully cleaned and harmonized the data. The NRSTP's coding scheme for scientific specialties enabled us to crosswalk degree subjects to the SED major fields we have used throughout our analysis. Using these data, we count up the number of PhD scientists in 1970 with degrees from U.S. universities by field and graduation year, focusing on 1950 to 1969 graduates.[5] The end result is a dataset of PhD cohorts by field and year.

We use these data for analogous analysis to that in Section 4.2 of the paper: merging these data with our PQDT-based measures of dissertation funding, we examine how government funding between 1950 and 1969 relates to the PhD workforce in 1970. As we did there, we adapt Equation 1 to estimate the relationship between the number of NRSTP scientists in 1970 and the number of publicly-funded PhD graduates produced in previous years, and as before, we use a shift-share instrument to explain the number of publicly-funded PhDs.

Table B.5 presents the results in analogous format to Table 4. Despite that the outcomes are measured from a fully independent data source (NRSTP, rather than SDR) and for a different time period (1970, vs. 2019), the effects are strikingly similar in sign, significance and magnitude to—and statistically within range of—the results in Table 4.

Table B.5: Effects of PhD funding on the future scientific workforce in 1970

| | Cohort size in NRSTP | | | Ln(Cohort size in NRSTP) | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | OLS | 2SLS | 2SLS | OLS | 2SLS | 2SLS |
| Govt-funded PhDs | 1.428*** | 1.381*** | 1.445*** | | | |
| | (0.143) | (0.138) | (0.136) | | | |
| Ln(Govt-funded PhDs) | | | | 0.298*** | 0.608*** | 0.736*** |
| | | | | (0.036) | (0.107) | (0.170) |
| N | 140 | 140 | 140 | 140 | 133 | 133 |
| $R^2$ | 0.96 | | | | | |
| F-stat | | 1007.33 | 450.37 | | 21.41 | 9.23 |
| Field FEs | Y | Y | Y | Y | Y | Y |
| Year FEs | Y | Y | Y | Y | Y | Y |
| Leave-out | | | Y | | | Y |

Notes: Table estimates the relationship between the size of the scientific workforce in 1970 and the number of government-funded PhD graduates over the previous two decades. Unit of observation is a PhD cohort in 1970 (defined at the degree field x year level), and the sample consists of PhD-holding scientists in the 1970 NRSTP survey who graduated between 1950 and 1969. Columns (1) and (4) estimate this relationship by OLS, and Columns (2)-(3) and (5)-(6) do so by two-stage least squares, using a shift-share instrument (i) constructed from previous-year shares in all columns and (ii) with leave-one-out shifters in Columns (3) and (6). All specifications include fixed effects for field and year. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

[4]The NRSTP was maintained for the purposes of locating technical talent that could be mobilized in a national emergency—motivated by the country's experience in World War II (e.g., Gross and Sampat 2023).

[5]We do so for non-engineering fields only, as engineering was not measured by the NRSTP in 1970, being instead the focus of a parallel National Engineering Register (NER) in the late 1960s—albeit less systematically and with less complete sampling. For these reasons, we exclude engineering fields in this analysis.

### B.3.3 Dissertations from Institutional Repositories

Table B.6: Coverage Data From Institutional Repositories

| University | # of GOID from PQDT | PDF avail. from PQDT | # missing PDF from PQDT | PDF in both PQDT and Univ Repo | Missing PQDT PDF found in Univ Repo | Missing coverage (%) | Coverage (%) |
|---|---|---|---|---|---|---|---|
| MIT | 27,546 | 0 | 27,546 | 0 | 16,052 | 58% | 58% |
| MSU | 15,907 | 9,005 | 6,902 | ? | 6,384 | 92% | 97% |
| UC Berkeley | 32,037 | 17,581 | 14,456 | 6,142 | 20 | 0.14% | 55% |
| UC Davis | 14,936 | 9,992 | 4,944 | 675 | 69 | 1.40% | 67% |
| UC Irvine | 7,596 | 7,230 | 366 | 2,236 | 93 | 25% | 96% |
| UCLA | 19893 | 13323 | 6570 | 4737 | 61 | 1% | 67% |
| UC Mered | 439 | 412 | 27 | 394 | 10 | 37% | 96% |
| UC Riverside | 5605 | 3749 | 1856 | 2350 | 75 | 4% | 68% |
| UC Sandiego | 239 | 211 | 28 | 61 | 1 | 4% | 89% |
| UCSB | 6529 | 5349 | 1180 | 1381 | 50 | 4% | 83% |
| UCSC | 3159 | 2518 | 641 | 1176 | 19 | 3% | 80% |
| UCSF | 4460 | 3294 | 1166 | 3062 | 882 | 76% | 94% |
| Cornell | 25061 | 12949 | 12112 | ? | 847 | 7% | 55% |
| Univ Minn | 23085 | 12356 | 10729 | ? | 197 | 2% | 54% |
| Univ Tex Austin | 15232 | 9147 | 6085 | ? | 840 | 14% | 66% |
| Univ Md | 18682 | 12499 | 6183 | ? | 288 | 5% | 68% |

## Appendix references

Gross, Daniel P. and Bhaven N. Sampat. 2023. "The World War II crisis innovation model: What was it, and where does it apply?" *Research Policy*, Vol. 52, No. 9, p. 104845.

National Center for Science and Engineering Statistics (NCSES). 2021. *Survey of Doctorate Recipients, 2019*. NSF 21-320, available at https://ncses.nsf.gov/pubs/nsf21320.