NBER WORKING PAPER SERIES

HOW DOES DATA ACCESS SHAPE SCIENCE? THE IMPACT OF FEDERAL STATISTICAL RESEARCH DATA CENTERS ON ECONOMICS RESEARCH

Abhishek Nagaraj
Matteo Tranchero

Working Paper 31372
http://www.nber.org/papers/w31372

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023, revised September 2024

This paper previously circulated as 'How Does Data Access Shape Science? Evidence from the Impact of U.S. Census's Research Data Centers on Economics Research.' We thank Ryan Hill for sharing code and training data to classify the style of economics articles. We also thank Brian Qi, JiYoo Jeong, Jai Singh, Randol Yao, and especially Cecil-Francis Brenninkmeijer for excellent research assistance. We acknowledge the financial support of the Alfred P. Sloan Foundation (Grant Number: G-2021-16965). We are grateful to seminar participants at the 2022 NBER Summer Institute, 2022 Research Data Center Annual Conference in Kansas City, 2022 Workshop on Big Data Analyses and New Developments in Research Data Centers at ZEW Mannheim, 2023 BITSS Annual Meeting, 2023 Columbia MAD Conference, 2023 DRUID Conference, 2023 Academy of Management Conference, as well as to seminar participants at UC Berkeley, Center for Economic Studies, and U Maryland Baltimore County. We are grateful to Andrea Cerrato, Wayne Gray, Lucia Foster, Jeff Furman, Bronwyn Hall, Julie Hotchkiss, and Bill Kerr for their feedback on this work. Any errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Does Data Access Shape Science? The Impact of Federal Statistical Research Data Centers on Economics Research

Abhishek Nagaraj and Matteo Tranchero

NBER Working Paper No. 31372
June 2023, revised September 2024
JEL No. C81,H00,L86,O33,O38

ABSTRACT

How does access to data shape the rate, quality, and policy relevance of academic research? To shed light on this question, we study the impact of access to confidential microdata through the progressive geographic expansion of the U.S. Census Bureau's Federal Statistical Research Data Centers (FSRDCs). FSRDCs boost the use of confidential microdata and help applied researchers publish more policy-relevant articles in top outlets. Besides direct data usage, this effect also reflects spillovers: exposed researchers are encouraged to explore topics related to FSRDC-enabled findings and employ similar empirical methods. Our findings underscore the importance of data access for scientific progress and evidence-based policy formulation.

Abhishek Nagaraj
Haas School of Business
University of California, Berkeley
2220 Piedmont Ave
Berkeley, CA 94720
and NBER
nagaraj@berkeley.edu

Matteo Tranchero
The Wharton School
3620 Locust Walk
2204 Sh-Dh
Philadelphia, Penn 19104
mtranc@wharton.upenn.edu

A data appendix is available at http://www.nber.org/data-appendix/w31372

1 Introduction

Modern science is largely empirical. In fields as diverse as astronomy, chemistry, and environmental sciences, researchers increasingly rely on large-scale, centralized datasets rather than on data curated for a single question (Hill and Stein, 2024a; Locarnini et al., 2018; York et al., 2000). Economics is no exception (Backhouse and Cherrier, 2017). The share of theoretical papers published in top economics journals decreased from 50.7% in 1963 to 19.1% in 2011 (Hamermesh, 2013), while empirical work has surged during this period (Angrist et al., 2020). A key factor behind this growth has been the availability of microdata on consumers and firms, often from administrative government records (Card, 2022; Heckman, 2001). Around 20% of recent articles in the five most prestigious economics journals use such data (Currie et al., 2020).

Data at the level of individual units are unique in their potential to inform evidence-based policies on a broad range of economic and social phenomena (Cole et al., 2020). Academic economists have, therefore, urged broad access to large-scale microdata, claiming significant scientific benefits (Card et al., 2010). However, the level of granularity afforded by microdata poses significant privacy risks, prompting government agencies to adopt tight confidentiality and security standards (Abowd and Lane, 2004; Foster et al., 2009). Understanding how and to what extent research access to confidential data entails benefits would complement current debates, which are primarily centered around ensuring privacy protection even at the cost of reduced accessibility (Abowd and Schmutte, 2019; Chetty and Friedman, 2019; Lane, 2021). Despite this urgent need, concrete evidence of the impacts of increasing data access remains thin.

It is not a given that broadening confidential data access will benefit economic research and policy-making. First, access must necessarily be coupled with tight security restrictions. From a practical perspective, these barriers might dramatically lower adoption and severely limit any potential benefits from data access. Second, access to the same pool of data might stimulate perverse competitive dynamics, such as prioritizing speed over rigor, ultimately lowering the quality of science (Hill and Stein, 2024a). Moreover, in the words of Bob Solow, the focus might shift from addressing substantive "puzzles that need to be explained" to exploring more marginal questions simply because an "enormous bunch of data" is available (Dizikes, 2019; Hoelzemann et al., 2024). Even if data access yields meaningful positive benefits, it is crucial to establish the magnitude of these benefits and understand the channels through which they operate. For instance, do benefits primarily accrue to direct users or also extend to non-adopters? Additionally, do the resulting findings influence policy decisions or remain confined to academic circles?

We shed light on these questions by studying how access to confidential microdata distributed by the U.S. Census Bureau shapes the quantity, quality, and policy impact of economic research. Our focus is

motivated by the fact that they are perhaps *the* pre-eminent source of restricted-access data in the United States. Examples include the Longitudinal Business Database (LBD) and the Longitudinal Employer and Household Dynamics (LEHD) database (Abowd et al., 2009; Jarmin and Miranda, 2002). Importantly, researchers wishing to analyze these data must be physically present at the Census Bureau's headquarters or in a secure data enclave, termed the "Federal Statistical Research Data Center" (FSRDC). A network of 31 FSRDCs was set up all over the country in a phased manner between 1994 and 2019. As multiple analyses will suggest, the timing and location of these openings were partly driven by factors governing geographic equity rather than by pre-existing trends in the use of microdata or research productivity. Therefore, the focus on FSRDC data provides a natural experiment to identify the effects of confidential data access on economics research.

We create a novel longitudinal dataset that measures the publication outputs of individual economists based on EconLit, and we pair this information with a host of other hand-curated sources to measure the diffusion and impact of FSRDC data. Our central regressions estimate the effects of FSRDC openings on empirical researchers with FSRDCs at their institutions relative to similar researchers without similar access. We include researcher and university-tier $\times$ year fixed effects in our preferred specification, effectively controlling for time-invariant researcher quality and time-varying institutional trends that might correlate with the inception of an FSRDC. We find that local access is critical to the diffusion of confidential data. Even though such data could be accessed by collaborating with those who already have access to them or traveling to another city, the opening of an FSRDC in the same institution increases usage by 116%–118% over the sample mean. Interestingly, we also find a large impact on the likelihood of citing past work based on confidential data. This suggests that opening an FSRDC also raises awareness about empirical results based on its data, potentially influencing the future research trajectory of non-adopting researchers.

Next, we explore how data access affects the productivity of empirical researchers. Contrary to views relating data access to more marginal research, we find that treated empiricists produce around 18% more publications in top-ranked outlets. This result becomes stronger when considering citation-weighted publications, suggesting that confidential data are particularly useful in boosting the scientific quality of empiricists' output. Moreover, FSRDCs lead empirical economists to publish research that receives more citations from policy documents. This result is not accompanied by a shift of emphasis toward more policy-relevant fields or topics. Instead, access to microdata leads to findings that are at the same time more scientifically impactful and relevant outside academic circles (Card et al., 2010; Chetty, 2012; Einav and Levin, 2014b).

Our findings are robust to a host of potential concerns. First, detailed event studies and formal tests confirm the validity of our identification strategy (Callaway and Sant'Anna, 2021; Rambachan and Roth, 2023).

Moreover, the qualitative evidence we gathered from interviews with FSRDC administrators suggested a few alternative research designs. We exploit them to identify within-institutional variation in exposure to data access, confirming our results when we compare treated empiricists with their colleagues in theoretical fields or applied domains that use other types of data. Second, university-level analyses show that FSRDCs are not systematically opened in institutions on a rising trend of research intensity or hiring spree. Third, we experiment with various definitions of being "treated" in terms of FSRDC access. We find consistent results when we exclude institutions hosting the data center on their premises and instead consider researchers in nearby institutions as treated. Our results get stronger the closer an economist is to an FSRDC, being the highest for researchers who enjoy access on their campus, even among institutions that jointly run a common FSRDC. Finally, several robustness tests rule out that our effects are due to the zero-sum sorting of more productive economists to institutions with better FSRDC access.

We investigate the mechanisms behind these findings. Our results are only partly driven by researchers who directly adopt confidential data, hinting at significant spillovers to those who do not use an FSRDC directly (Myers and Lanahan, 2022). Yet, spillovers apply only to economists who are more aware of FSRDC research, either because they cite FSRDC-based papers in their own work or have colleagues using FSRDC data. The effects disappear for departments without data adoption or authors not building on the findings enabled by FSRDC data. These patterns suggest that local data access shapes academic output by exposing economists to research based on confidential data. But how does exposure spill over into greater productivity? Our results suggest that increased awareness about FSRDC-based research might inspire a change in research direction that leads to more impactful work. First, empirical researchers are more likely to explore topics and fields commonly associated with using FSRDC data. Second, treated researchers increase their reliance on administrative datasets and quasi-experimental research designs. Taken together, it seems that the opening of an FSRDC improves the output of treated researchers by leading them to explore a newer set of topics using similar databases and robust methodologies.

Finally, we note that despite large increases in data diffusion prompted by FSRDCs, the absolute number of papers using their data remains limited. Accordingly, we explore how our findings can inform the design of the data infrastructure in the United States. Our analysis finds that geographical barriers disproportionally penalize researchers from less prestigious institutions or with fewer resources to access confidential data. While the optimal level of data access is shaped by social preferences around privacy (Abowd and Schmutte, 2019), there might be ways to increase the diffusion of confidential data holding the current regulatory environment constant. We discuss the possibility of increasing the number of data centers with a "back-of-the-envelope" cost-effectiveness calculation. While a complete cost-benefit analysis is beyond the scope of our paper, even our conservative estimates imply that a further expansion of the FSRDC network would be warranted.

Our work contributes to three different strands of research. New legislation in the U.S. (like the Evidence-Based Policy Act and the Federal Data Strategy) puts data access at the heart of policy-making. The assumption is that providing restricted-access confidential data will lead to the development of policy-relevant economics research (Lane, 2021; Chetty et al., 2018, 2024). Yet, while some have hailed current data access programs as successful, others have criticized them for being costly and cumbersome, with uncertain impact (Atrostic, 2007; Card et al., 2010; CES, 2017; Cole et al., 2020). Ironically, most of these debates have themselves been data-free. We contribute the first systematic investigation of the effects of confidential data access on academic research, providing key parameters for the important discussion around the privacy risks (Abowd and Schmutte, 2019; Foster et al., 2009). Further, by examining the policy consumption of research, our work highlights how confidential data access can benefit evidence-based policy-making (Hjort et al., 2021; Yin et al., 2022).

Second, we add to recent research that has investigated questions of relevance to the economics profession using bibliometric data. This includes past work documenting the empirical turn of economics (Backhouse and Cherrier, 2017; Currie et al., 2020; Einav and Levin, 2014a; Hamermesh, 2013). Our results show that increased access to high-quality data is an important factor driving the increase in the impact and credibility of empirical scholarship (Angrist et al., 2020; Brodeur et al., 2020). In addition, we contribute to the growing scholarship on labor markets for economists, including research on status dynamics, credit attribution, and editorial roles (Card et al., 2020, 2022; Feenberg et al., 2017; Heckman and Moktan, 2020; Sarsons et al., 2021; Carrell et al., 2024). Our results hint that confidential data pose asymmetric hurdles for researchers depending on their resource endowments. Efforts to further democratize data access could help level a field characterized by large inequities in status and resources (Wapman et al., 2022).

Finally, we contribute to the economics of the scientific process more broadly (Azoulay et al., 2019; Jones, 2009; Hill and Stein, 2024b,a; Wang and Barabási, 2021). A vibrant strand of research has studied the impact of access to research tools on the production of knowledge (Biasi and Moser, 2021; Furman and Stern, 2011; Furman and Teodoridis, 2020; Murray et al., 2016; Waldinger, 2016). While this work has primarily focused on how access to research material shapes academic output, it has rarely examined access to data, whose importance to research warrants careful examination (Hill et al., 2020). We add to this literature by investigating how data access shapes the rate, direction, and policy impact of scientific innovation (Nagaraj et al., 2020; Williams, 2013). Data can crucially determine which domains receive more attention from researchers (Hoelzemann et al., 2024; Myers, 2020; Truffa and Wong, 2022), and we are the first to document their effects on the advancement of science through spillovers to non-users. Finally, there is growing interest in examining how scientific progress can inform government and social policy (Hjort et al., 2021; Yin et al., 2021). Our work is among the first to link research inputs to policy-relevance of research outcomes.

The rest of the paper proceeds as follows. Section 2 provides some background on confidential data and the FSRDC program in the United States. Sections 3 and 4 describe our primary research design and data sources. Section 5 presents the key findings and robustness checks, while Section 6 explores the mechanisms that drive them. Section 7 presents evidence about the cost-effectiveness of FSRDCs and what factors hinder their usage. Section 8 concludes with a discussion of the implications and limitations of our findings.

2 Empirical Setting

2.1 Confidential Microdata in Economics Research

Microdata can be defined broadly as unit-level data that provide information about the characteristics of individual people or entities (Abraham et al., 2022; Dupriez and Boyko, 2010; Heckman, 2001). This type of data can be collected to answer specific research questions, like in the case of the Panel Study of Income Dynamics (PSID) curated by the University of Michigan. However, recent years have seen a growing diffusion of microdata not initially collected for research purposes (Goroff et al., 2018; Groves, 2011). Government agencies are big sources of such information, routinely storing administrative records during normal functioning. Typical examples include unemployment insurance claims, Medicare data, or tax records. Other agencies, such as the U.S. Census Bureau, collect individual-level information via statistical surveys and enumerations as part of their mandate to assemble timely data about the nation's demographic and economic trends (Foster et al., 2009).

There is little doubt that the increased availability of microdata has played an essential part in the empirical turn taken by economic scholarship in recent decades (Card, 2022; Backhouse and Cherrier, 2017). This is especially true in the case of government microdata, which usually encompass large samples that allow tracking individual units over time and with little attrition (Cole et al., 2020). Einav and Levin (2014b) provide anecdotal evidence on a few impactful studies based on government records studying topics ranging from broadband internet to Medicaid expansion (Akerman et al., 2015; Taubman et al., 2014). However, the same features that make microdata invaluable for research put the privacy of respondents at risk (Bowen, 2024). Distributing firm- or individual-level information such as wages, health records, or similarly sensitive information poses security and privacy concerns. Government agencies thus face a trade-off between providing research access to microdata and their duty to protect the confidentiality of the information entrusted to them (Abowd and Schmutte, 2019; Foster et al., 2009).

To address this conundrum, government agencies have experimented with several second-best solutions, from the release of anonymized public use samples to the development of synthetic data (Abowd and Lane, 2004; Kinney et al., 2011; Weinberg et al., 2007). Unfortunately, no approach comes close to the research

potential that the universe of individual-level information has (Cole et al., 2020; Goroff et al., 2018). Most statistical agencies have eventually resorted to providing direct access to confidential microdata through strong security barriers. These include providing access exclusively to vetted researchers for pre-approved projects and allowing only the release of results that have undergone careful review. Yet, where government agencies differ is the data access modality. Some countries, especially in Northern Europe, try to trade off control with ease of access by providing virtual private networks (VPNs) or dedicated remote devices that ensure biometric authentication of users. In others, such as the United States, access to confidential microdata is provided through physical presence at secure facilities on specialized devices where data use is closely monitored.[1] This data enclave model guarantees maximum security, but it imposes significant costs to universities and researchers.

2.2 The FSRDC Network

We focus on the FSRDC program created by the U.S. Census Bureau. This program traces its origin to the establishment of the Center for Economic Studies (CES) in 1982 to combine microdata collected during its routine activities and provide restricted access to interested researchers (Atrostic, 2007). The objective of the Census Bureau was to enable research that could improve its data programs while abiding by its legal mandate to protect data confidentiality.[2] However, interested scholars had to access the data in the Census headquarters near Washington, D.C., which was costly and inconvenient (McGuckin et al., 1993). To overcome this limitation, the CES spearheaded a major effort to set up additional secure facilities where confidential data could be accessed for research purposes. These data enclaves, known as the FSRDCs, are hosted by partner institutions often organized in local consortia to share operational costs (CES, 2017). Thanks in no small part to the financial support of the National Science Foundation (NSF), the FSRDC network expanded to 31 data centers as of 2019 (Davis and Holly, 2006; CES, 2019).[3]

Each FSRDC is a research facility that meets several physical and computer requirements to ensure the confidentiality of sensitive data. First, each branch has controlled access doors, security cameras, and a Census employee onsite. No data reside in the FSRDC since all the statistical analyses are carried out on

[1] In 2019, the Census Bureau began to provide remote virtual access to a select number of FSRDC researchers not working with records originating from the IRS. The pilot has been scaled up during the COVID-19 pandemic, but this development does not affect our analyses since it affects only projects that will be published outside of our sampling period. Appendix C provides more details on this recent development.

[2] The data collected by the U.S. Census Bureau is tightly regulated. Title 13 of the U.S. Code provides a legal framework to acquire, use, and protect confidential data, ensuring that they are only used by authorized personnel for statistical purposes. Moreover, the Bureau collects and integrates data from other government sources. These data are governed by similar confidentiality provisions in Title 26 of the U.S. Code (for IRS records) and in the Confidential Information Protection and Statistical Efficiency Act (for other statistical agencies). From a legal standpoint, the Census Bureau is authorized to give access to confidential microdata only insofar as it derives a clear benefit for its programs and as long as users are sworn to observe the limitations imposed by Title 13. A detailed discussion of the objectives of the Census Bureau in sharing their data for research is in Appendix C.1.

[3] The only FSRDCs to have closed are those opened at Carnegie Mellon University (CMU) in 1996 and at RTI International in 2010. For our empirical analysis, we consider CMU and RTI to have lost local access to the data after the FSRDC closed in 2004 and 2018, respectively. Additional FSRDCs have opened outside our sampling period (see Appendix Table C1). In 2018, the FSRDC network formally became part of the Center for Enterprise Dissemination without any change in its operations.

Census Bureau servers through a secure client physically located in the data center. Second, researchers can access an FSRDC only after submitting an application that outlines the research question and the data needed to answer it. Approval requires passing an evaluation of the merits, feasibility, disclosure risks, and benefits for the Census Bureau. Third, pre-approved researchers must receive a Special Sworn Status after thorough security checks. Special Sworn Status individuals take an oath of confidentiality and are subject to the same legal obligations and penalties as Census Bureau employees. Fourth, results produced in an FSRDC undergo a full disclosure review before they can be shared outside the research facility.

While the CES was originally established to provide access to data about the manufacturing sector (McGuckin et al., 1993), FSRDCs now permit the investigation of a large variety of economic phenomena via datasets that have become household names for economists such as the LBD or the LEHD (CES, 2017). Two interrelated developments have led to this expansion in scope. On one hand, CES staff and FSRDC users have contributed to improving or creating new databases as part of their research projects; recent examples include the re-design of the LBD (Chow et al., 2021) and the creation of the Management and Organizational Practices Survey (Bloom et al., 2019). On the other hand, the growth of the FSRDC network has led other federal agencies to make their confidential data available through the same infrastructure. The agencies partnering with the FSRDC network include the Agency for Healthcare Research and Quality, Bureau of Economic Analysis, Bureau of Justice Statistics, BLS, NCHS, and National Center for Science and Engineering Statistics.[4]

Several accounts suggest that the FSRDC network enabled path-breaking advances in economics (CES, 2017).[5] However, the strict security guardrails encoded in the Census Bureau's mandate have limited the diffusion of administrative data relative to other European countries (Cole et al., 2020). Researchers have even suggested that limited access to government microdata puts the U.S. at risk of losing leadership in cutting-edge empirical research (Card et al., 2010). Moreover, restrictive access to data is disproportionally penalizing for researchers with fewer resources or affiliated with lower-status institutions (Nagaraj et al., 2020). The gradual expansion of the FSRDC network was designed to tackle both these issues, but it is unclear how successful the program has been. Further, the costs of operating the program are significant. By some estimates, it costs over $7 million per year to maintain the FSRDC network as of 2021, not accounting for set-up expenses and the costs of renting space for the data enclaves themselves.[6] It is, therefore, timely and essential to empirically evaluate the effects of the FSRDC program on academic research and policymaking.

[4]We term the data collectively provided through this program as "FSRDC data" and use it to mean all confidential microdata distributed by the FSRDC network, regardless of how the dataset was created or the specific government agency that collected the original records. Different agencies started distributing different datasets at different points in time, but unfortunately, we cannot precisely track this expansion in our study. In our estimates, we rely on a tight set of time fixed effects to control for the introduction of new datasets. In Appendix F, we provide additional details on the databases most used in economics research.

[5]For instance, CES' establishment-level data is credited to have facilitated a generalized shift from "representative firm" thinking toward research that takes into account intra-industry heterogeneity (McGuckin, 1995; Davis et al., 1998).

[6]Estimates are taken from https://deepblue.lib.umich.edu/response_FSRDC_Directors_2021. A detailed cost-benefit assessment is provided in Section B and Appendix G.

3 Research Design

There are two key challenges in empirically assessing the impact of data access on economics research. First, researchers do not systematically cite data sources or could be influenced by data they do not use directly (Nagaraj, 2024). This makes it challenging to track data use and access. Second, any correlation between the use of specific data and publication quality is likely to be upward biased. Researchers might produce better research not because of their data access but because they have greater resources or are otherwise more creative. The empirical challenge is thus finding a research design that provides (a) a measure of access to confidential data independent from their use and (b) credible variation in the availability of the same data to otherwise comparable researchers.

In this paper, we employ the staggered geographical expansion of the FSRDC network as a source of variation in data availability for academic economists. Even though researchers could, in principle, access confidential FSRDC data through collaborators or by visiting the Census Bureau's headquarters, co-location to an FSRDC could make the researcher aware of the data or decrease the barriers to using them. In particular, our research design takes advantage of the requirement that FSRDCs' confidential microdata can only be analyzed in secure facilities. This allows us to proxy data access at the level of individual researchers as the presence of an FSRDC in their institution of affiliation. To avoid confounding effects from the generalized growth of empirical research, our main design compares empirical economists co-located to FSRDCs with empirical economists lacking local access.[7] Ruling out the presence of pre-trends in our estimates would further validate this research design.

A potential problem with our identification strategy is that the host institution's characteristics might drive the choice of FSRDC locations. However, if time-invariant university attributes such as research specialization, resource endowments, or status drive location choice, we can control for these with fixed effects estimation. More worrisome would be two other types of considerations: dynamic confounds, such as changes in research resources concurrent to the opening of a data center, and systematic sorting, such as the opening of data centers where their impact on scientific output could possibly be higher. While the former issue is a challenge to our identification of FSRDCs' effects, the latter would threaten the external validity of our estimates. To assess the plausibility of these concerns, we conducted nine interviews with users and administrators to learn the history and institutional details of the FSRDC network (see Appendix A.1).

Opening a new FSRDC requires universities to submit a competitive application to the NSF, which is then jointly evaluated by the NSF and the Census Bureau (Atrostic, 2007). This process has implications for

[7] One natural concern is that trends in the growth of empirical research might underlie both the increasing use and impact of confidential data. For instance, changes in editorial preferences in top journals might lead to productivity benefits for applied researchers even in the absence of local FSRDC openings. However, using empirical economists lacking FSRDC access as a control permits us to rule out the confounding effects of these trends.

both issues listed above. First, our interviews revealed a surprisingly large number of idiosyncratic factors behind the establishment of FSRDCs. For instance, interest in setting up an FSRDC arose after rotations in university leadership or when individual researchers started advocacy campaigns. Even after obtaining the NSF grant, FSRDC opening could be delayed by years, suggesting that the precise timing of openings was unrelated to concurrent confounds. Second, our interviewees indicated that the Census Bureau and especially the NSF were trying to balance researchers' demand with equitable geographical coverage across the United States. As a former FSRDC administrator explained to us, "Many institutions were interested in opening an RDC, but the NSF was interested in kind of parity across the U.S. so that researchers in one part of the country had the same access as researchers in another part of the country did" (interview T14). The presence of a nearby data center prevented even top-tier universities with large user bases from obtaining an FSRDC for many years, thus allaying concerns about sorting.[8]

Taken together, this qualitative evidence suggests that both the timing and the locations of FSRDCs were strongly influenced by idiosyncratic factors unrelated to underlying trends in research productivity or concurrent changes, lending support to our primary identification strategy. Nevertheless, we will directly test the validity of our results using alternate sources of variation (Section 5.3) and other robustness checks 5.4.

4 Data

To investigate our research questions, we need data on a few key dimensions: (a) identifying the set of relevant academic economists and their affiliations, (b) matching academics with the quantity, quality, and policy impact of their publication output, (c) identifying researchers who are empirically oriented, and (d) measuring the diffusion of FSRDCs and the adoption of confidential FSRDC data.

4.1 Building the Universe of Publishing Economists

The main data source we leverage is EconLit, a proprietary database of economic scholarship curated by the American Economic Association (AEA). Compared to other popular databases of scientific publications, EconLit has a broader coverage of economics journals and includes *Journal of Economic Literature* (JEL) codes that classify articles into economics fields. EconLit is increasingly used by researchers interested in studying economics research (Angrist et al., 2020; Card et al., 2022; Önder and Schweitzer, 2017).

Unfortunately, EconLit lacks unique author and affiliation identifiers, which prevents us from reliably linking researchers with their scientific output. To reconstruct authors' publication records, we need

[8]For example, Stanford University opened its FSRDC branch only in 2010 due to the presence of the relatively nearby Berkeley FSRDC. See Appendix A.2 for a case study.

to disambiguate publication metadata, a common but difficult and time-consuming task in bibliometric analyses. We disambiguate our data in several steps outlined below (and detailed in Appendix B). We start with the complete set of 839,513 scientific articles published in 1,856 journals between 1990 and 2019 in EconLit. While starting from such a large and heterogeneous body of articles makes the disambiguation task harder, including every paper is essential because we can use this information to detect mobility events from changes in academic affiliations in published work.

We then proceed in two steps. First, we standardize the name of the 178,798 affiliations appearing in EconLit to pin down researchers' location and, hence, treatment status over time, as well as to restrict the sample to U.S.-affiliated economists who are at risk of being co-located to an FSRDC. We retain in our sample all doctorate-granting institutions in the United States taken from the 2018 Carnegie Classification of Institutions of Higher Education, to which we add the most important institutions active in economic research (such as the IMF, Rand Corporation, World Bank, RTI, and all the regional FED offices). Using fuzzy matching and extensive manual checks, we standardize the 11,491 different spellings of the 439 U.S. research institutions appearing in our list of research-intensive institutions.

Second, we disambiguate researchers' names using a graph-theoretic disambiguation procedure (Önder and Schweitzer, 2017). This approach assumes that the combination of first, middle, and last names uniquely identifies each economist while at the same time being conservative in assigning ambiguous names that lack a clear middle name (Card et al., 2022). To avoid confounding effects from including researchers working in unrelated fields but occasionally publishing in economics journals, we match our data with 19 yearly lists of AEA members spanning 1993–2019 (Jelveh et al., 2024). We further exclude from our sample researchers who are or have been affiliated with the U.S. Census Bureau or any partner agency since these people might enjoy privileged connections and access to data.[9] This procedure results in a starting sample of 14,589 U.S.-based economists who have been AEA members, which we will later refine based on their methodological orientation.

4.2 Publication-Level Information

The EconLit database lists article metadata such as authors, journal, year of publication, JEL codes and, for the majority of articles, the abstract. We collected additional abstracts from websites like Google Scholar, EconStor, and JSTOR. Next, we augment EconLit by merging the yearly citation count for each article extracted from SSCI/Web of Science and information on funding sources acknowledged from Crossref. We base individual-level productivity metrics on all articles appearing in journals that are i) indexed in Web of Science, ii) published in English, iii) and listed in SCImago under the subject areas "Economics,

[9]Our results are almost identical if we include non-AEA members or researchers affiliated with the Census Bureau. Results are available from the Authors upon request.

Econometrics and Finance," or "Business, Management and Accounting." This results in a final set of 188,181 articles published in 158 journals in the period 1990–2019. We can match academic citation data for 97.2% of these articles.

In addition to scientific impact, we are also interested in directly assessing the impact of FSRDC access on the policy relevance of economics research. To do so, we leverage novel data from Overton (see Appendix E for a detailed discussion). In particular, Overton collects data on scientific articles cited by policy documents from a wide range of institutions, ranging from government reports to international organizations. Policy sources are mainly collected directly from organizations' websites and merged into the bibliographic records using metadata such as title, authors, and year of publication. We match these data to our sample from EconLit, resulting in a unique database that permits examining economic research consumption by policy sources (Yin et al., 2022).

We painstakingly compile a list of all articles that *directly* employ restricted-access data accessible only in an FSRDC. We carefully sift published records with several complementary strategies. Projects using confidential data are expected to be indicated as such in the paper's acknowledgments (see Appendix Figure A2). We perform keyword searches for common expressions denoting the use of FSRDC data using databases such as Web of Science, Scopus, JSTOR, and Google Scholar. In addition, we exploit the fact that the CES requires submitting a working paper for online publication upon project completion. We collect the metadata of all the working papers and manually match them with records of published work.[10] We also aim to capture how FSRDC data affects research *indirectly*, that is, by enabling findings that inspire subsequent research. We do so with two approaches. First, we record which articles cite the papers written using FSRDC data and thus explicitly build on their results.[11] Second, we tag all papers that include JEL codes that are the most representative of research using FSRDC data. In this way, we capture papers that are thematically close to research done in FSRDCs, regardless of whether they cite it.

Finally, we use the text of titles and abstracts of each paper to characterize its methodological approach. We use a machine learning classifier that outputs a score capturing the probability that an article is empirical, leveraging information about the journal, title, abstract, year of publication, and JEL codes.[12] Following Angrist et al. (2020), we classify a paper as empirical if it uses data to estimate an economically meaningful parameter even if it develops new methodological tools to do so (see Appendix D for more details). The

[10] Unfortunately, the publication of FSRDC output in the CES working paper series is a requirement hard to enforce, which is why we also relied on keyword searches as detailed in Appendix C.3. We cannot separately code papers stemming from internal Census projects that are not subjected to the FSRDC application procedure. However, this should not impact our analyses since we exclude researchers who have been formally affiliated with the U.S. Census Bureau from the analyses.

[11] We exclude papers directly using FSRDC data from the count since they are likely to mechanically cite other FSRDC papers with which they share the data.

[12] We are indebted to Ryan Hill for sharing the code and the training data originally used in Angrist et al. (2020). We use the same code to classify each paper in a field of economics research. This information is then used in Appendix Table H6 to identify scholars more likely to benefit from FSRDC access based on their research topic.

results of this classification effort are highly reliable, as validated in several manual checks. We use this publication-level classification to characterize the methodological orientation of each publishing economist in our sample. Then, we follow the approach of Currie et al. (2020) to code more detailed information about each paper's research design (Brodeur et al., 2020). We tag each paper that explicitly mentions the use of a particular method (e.g., difference-in-differences) or data type (e.g., administrative data) in the title or abstract. The complete list of keywords used is reported in the Appendix C.5.

4.3 Researcher-Level Information

Thanks to the host of article-level variables outlined above, we can compute several metrics that capture the research productivity of each researcher. In particular, we compute yearly publication counts weighted in three different ways: by the prestige of the outlet, by scholarly citations received, and by policy impact. The first metric sums the number of publications in the top field and top five economics journals in a year, which we collectively refer to as "top publications."[13] The second metric captures scientific impact by weighting publication counts by the citations received in a window up to five years after publication. Finally, we use the count of articles that receive mentions in policy documents as a third outcome.

One of our objectives is to measure changes in research trajectory after obtaining access to confidential microdata (Furman and Teodoridis, 2020). This is possible by leveraging the JEL codes, a hierarchical taxonomy of research topics in economics. We rely on 2-digit JEL codes to track the topical focus of research (Card et al., 2020), exploring the likelihood that a researcher writes a paper with JEL codes that she has never used in her previous work. While a potential shortcoming of this metric is that JEL codes are author-assigned, all our models will include researcher fixed effects, thus effectively controlling for different researchers' styles in assigning such codes (more details in Appendix C.5).

We rely on our article-level methodological classification (empirical or theoretical) to categorize each scholar according to their methodological orientation. For our primary analyses, an empiricist is defined as anyone whose majority of her first five publications are empirical in nature.[14] We carry out several checks. First, we adopt a case-control approach and validate the results of our classification for the editorial board

[13] These journals are especially important for career progression in economics. We rely on the list assembled by Heckman and Moktan (2020). The list includes top field journals (the *Journal of Development Economics*, the *Journal of Econometrics*, the *Journal of Financial Economics*, the *Journal of Economic Theory*, the *Journal of Health Economics*, the *Journal of Industrial Economics*, the *Journal of Labor Economics*, the *Journal of Monetary Economics*, the *Journal of Public Economics*, the *Journal of International Economics*, and the *Journal of Economic History*), high-profile generalist journals (the *Review of Economics and Statistics*, the *Journal of the European Economic Association*, the *Economic Journal*), and the so-called top five journals (the *American Economic Review*, the *Quarterly Journal of Economics*, the *Journal of Political Economy*, *Econometrica*, and the *Review of Economic Studies*). See also Appendix Figure B3.

[14] This measure has the advantage of being available for every publishing researcher in our sample. Results change little if we define empiricists based on all their lifetime publications or only on publications that predate co-location to an FSRDC. However, the latter measure would not be defined for researchers who start their careers at institutions hosting FSRDCs, which leads us to prefer the definition described in the main text. Appendix D provides additional details and shows the robustness of our classification of empirical economists.

members of some journals with a known methodological bend (e.g., the *Journal of Economic Theory* versus *AEJ: Economic Policy*). Second, we compile a list of all Ph.D. students who completed their doctorate in a U.S. university from the records published yearly by the JEL and compare our classification with their dissertation fields. Both tests confirm the face validity of our approach (Appendix D.2). In the results section, we discuss additional robustness checks where we repeat our analyses with progressively more stringent cut-offs to define empirical scholars, showing consistent results.

We use this classification to derive our main sample of 10,060 empirical economists, giving rise to an unbalanced sample of 155,720 researcher-year observations by imputing missing years between the first and the last year we see a researcher publishing. For years with missing publications, we have to impute institutional affiliation, which can lead to measurement error when an affiliation is observed to change in non-consecutive years with gaps in between. Our approach consists of attributing the old affiliation for the first one-third of the missing years and the new affiliation for the remaining two-thirds. Our data change little when we experiment with different imputation rules. See Appendix Figure B1 for a summary of how we built the author-year panel from bibliographic records.

4.4 University-Level Information

We can assess the details of FSRDC openings from the 15 CES research reports published between 2005 and 2020. Figure 1 synthesizes the expansion of the FSRDC network in time and space. Overall, the figure confirms the oral accounts testifying a conscious effort by the U.S. Census Bureau and the NSF to provide geographic balance in access. We see that the data centers spanned different regions of the United States, starting from some of the major centers of economic research (e.g., Boston, Berkeley, Los Angeles) but also leaving out until much later other illustrious universities until much later if relatively close-by alternatives were present (e.g., Stanford, Yale, Princeton). The complete list of 31 FSRDCs established in our sampling period and their opening date is reported in Appendix C.

Data on the research intensity of economics departments comes from Kalaitzidakis et al. (2003). Their ranking of academic institutions is based on the count of publications in the top journals weighted by each journal's prestige. This ranking fits well for our purposes because it considers publications in the five years from 1995 to 1999, thus predating the establishment of most FSRDCs. We use this data to classify each university into seven tiers based on the ranking of economics departments.[15] Using these data, we can see how less research-intensive institutions gained access at the same pace as their more prestigious peers (Appendix Figures H1). The remarkably balanced expansion of the FSRDC network that

[15] We chose the number of tiers to ensure a roughly equal number of observations for each, but the results do not change if we change this number. Note that the ranking of Kalaitzidakis et al. (2003) is limited to the first 200 research institutions worldwide, of which only 88 are in the United States. We assign all institutions missing from their list to a residual tier. See details about the ranking in Appendix B.2.

we observe (Appendix Figure H2) echoes what we learned from our interviews, and further excludes explicit prioritization of high-status universities.

4.5 Summary Statistics

Table 1 provides summary statistics of the dataset we assembled. Panel A provides summary statistics on the cross-section of the 10,060 empirical researchers who constitute our main sample, roughly corresponding to 63% of all U.S.-based publishing economists. Almost 4% of the applied economists we observe have directly used FSRDC data in published work, and 28.4% have cited at least one paper using FSRDC data. The average researcher publishes about 1.9 academic papers in top outlets during the time that they appear in our sample. However, the number of policy-relevant articles is much higher and close to 3.9, suggesting that many policy-relevant findings are not published in the most prestigious outlets. Additional details on our novel data on policy citations are presented in Appendix E.

As per our data, of the sample of 10,060 researchers, about 79% never had access to an FSRDC in the same institution where they were employed. Appendix Figure C1 breaks this sample down by those who always had access (532) and those who got access through an FSRDC opening in their city (928) or by moving to a city with an FSRDC (615). Panel B presents summary statistics on the unbalanced panel of researcher-years. The panel extends from 1990 to 2019 (inclusive), and the median year is 2007. As this table shows, the average researcher publishes about 0.13 papers every year in the set of top economics journals. For every researcher-year, 0.005 papers use FSRDC data, and 0.041 build on past research using them.

5 Results

Our basic specification is at the empirical researcher i, university j, and time t level and takes the form of the following equation to test the impact of local data access:

$$y_{i,j,t} = \alpha + \beta PostFSRDC_{j,t} + \mu_i + \delta_t \times \omega_j + \epsilon_{i,j,t}, \tag{1}$$

where the dependent variable $y_{i,j,t}$ is publications by researcher i affiliated to university j in year t. The main independent variable $PostFSRDC_{j,t}$ is a time-varying dummy that takes a value of one after a researcher has gained access to an FSRDC facility located in her university, and it is zero before. The specification includes researcher fixed effects (μ_i) that control for time-invariant differences across researchers in data use but also individual propensity to publish. Finally, university tier-by-year fixed effects ($\delta_t \times \omega_j$) control for time trends in productivity or resources that might be specific to universities of different research intensity.[16]

[16] We also show results with less stringent inclusion of separate university and time fixed effects. In Section 5.3, alternative identification strategies leveraging within-university variation enable the inclusion of tighter university-by-year fixed effects.

We present results with the treatment dummy coded as an absorbing state, which means that individuals do not lose the treated status even if they move to a non-treated institution. This approach accounts for potential delays in the publication process and makes it straightforward to interpret $\hat{\beta}$ as the result of a Wald-DID estimator.

5.1 Impact of FSRDC Openings on the Diffusion of Confidential Data

The first step to unpacking the effects of confidential data access on economics research is to examine how FSRDC openings affect their use. Note that even before being co-located with an FSRDC, it was always possible to collaborate with a Census Bureau researcher with access to the data or to commute to a center in another city to access the data; therefore, it is not obvious that being co-located with an FSRDC will lead to a substantial diffusion of confidential microdata.

Table 2 displays the results from this analysis. As Column 1 shows, FSRDCs boost the use of confidential microdata among empiricists: local access is associated with about 0.006 more papers using confidential data, an increase of roughly 118% compared to the baseline of 0.005 (Table 1). The coefficient is only marginally smaller when we include more stringent university tier-by-time fixed effects in Column 2. The results on data diffusion are similar if we employ alternative measures of FSRDC use that do not rely on published articles, such as the likelihood of having an FSRDC project approved or submitting a working paper to the CES (Appendix Table H1). Interestingly, we also find a large impact on the likelihood of citing past work based on FSRDC data (Columns 3 and 4). While the increase on the sample mean is smaller (around 26%–42%), this result is even more remarkable since lack of access should not impede learning about papers published in academic journals.[17] This suggests that opening an FSRDC is making local economists aware of past research using confidential data, potentially through interactions with others directly dealing with such data in their own work.

The validity of our regressions hinges on the assumption of a parallel trend between treated and control units in the absence of treatment. We empirically assess this by estimating the event study version of the results reported in Table 2. Specifically, we estimate $y_{i,j,t} = \alpha + \Sigma_z \, \beta_t \times 1(z) + \mu_i + \delta_t \times \omega_j + \epsilon_{i,j,t}$, where μ_i and $\delta_t \times \omega_j$ represent researcher and university tier $\times$ time fixed effects, as before. z represents the "lag," or the years elapsed since an empirical researcher first received access to confidential data via a local FSRDC. Figure 2 shows the effect of increased access to FSRDC data on their use (panel (i)) and citations to papers based on them (panel (ii)). Both charts confirm no pre-trends driving our effects, which might have been the case if FSRDCs were opened in locations where diffusion of confidential data was already rising. Further, the effects appear gradually and then grow in magnitude before stabilizing after three to five years, with a

[17] Note also that we exclude from the count of citing papers those that directly employ FSRDC data. Therefore, our estimates are conservative and not mechanically explained by the contemporary diffusion of confidential data.

pattern that fits publication lags and our general intuition of how data might diffuse in an academic setting.

5.2 The Effects of Data Access on Scientific Productivity

Having established that local FSRDC access increases the diffusion of confidential data, we turn to estimating the impacts of data access on researchers' productivity. Formally, we estimate a similar specification as before, except that now our outcome variable is not limited to economics research using FSRDC data. For each researcher, we count the number of papers published yearly in top field outlets and top five generalist journals, which carry outsized weight in determining career trajectories (Heckman and Moktan, 2020). To measure scientific impact, we use a second dependent variable based on the number of citations each paper receives up to five years following its publication. Finally, the count of publications that receive policy attention is used as a third outcome to capture the policy relevance of research.

Table 3 presents the results from this analysis. For each outcome variable, the first model reports results with university and year fixed effects, and the second model presents results including university tier-specific time trends. Across the board, the results suggest that FSRDC access boosts the productivity of empirical researchers. Empirical economists produce about 0.023 more publications in the top journals, which, compared to a baseline of 0.126, is an 18.3% increase in output. Columns 3 and 4 show that the increase in productivity does not come at the expense of reduced impact: citation-weighted publications increase by about 1.24–1.71 (25%–35% increase against a baseline of 4.9). This result suggests that access to confidential data leads to significantly higher-impact publications.

Next, we estimate the causal impact of FSRDCs on the policy impact of economics research. Columns 5 and 6 show that the work of applied researchers is more likely to be referenced by policy sources after becoming exposed to FSRDCs. The percentage increase over the sample mean is slightly bigger among U.S.-based policy sources, confirming the impact of the federal data infrastructure on evidence-based policy discussions in the United States (Appendix Table H2). Interestingly, our results do not seem to be driven by a shift in research topics toward more policy-relevant fields. If this were the case, we should see the use of language that emphasizes the policy implications of research or a direct shift to policy topics. However, we do not find either of these effects when directly looking at the language used in the abstracts or the JEL codes indicated by the authors (Appendix Table H3). Instead, our results could be consistent with increased policy relevance due to the higher scientific quality of applied research caused by access to better data.

We explore the reliability of our research design by empirically estimating pre-trends in the outcome variables. Figure 3 presents event study estimates similar to those presented in Figure 2 but for the measures of research productivity discussed above. In all three panels, publication output is flat before the treatment and remains low in the first few years but then gradually improves until it stabilizes on positive and significant

values. Appendix Table H4 and Figure H3 repeat the same analysis using the doubly robust DID estimator by Callaway and Sant'Anna (2021), confirming that our results are not biased by heterogenous responses to FSRDC openings at different times. Finally, we formally test the parallel trends assumption in Appendix Figure H4. Using the approach proposed by Rambachan and Roth (2023), we bound the slope change in the differential trend between treated and control economists that would be needed to make our estimates insignificant. The analysis reveals that productivity deviations from the trends preceding the FSRDC opening would need to be implausibly large to account fully for our results.[18]

5.3 Alternative Identification Strategies

Our main analysis identifies the effect of FSRDCs by comparing empirical researchers in institutions hosting them with empirical researchers lacking such local access. The divergence in confidential data usage and productivity following the co-location to a data enclave allows us to pin down the causal impact of increased data access. While the institutional setting and the absence of pre-trends bolster our confidence in this research design, our interviews with people familiar with FSRDCs' history suggested alternative sources of variation that we can use to validate our findings.

Academic economists tend to specialize along methodological lines and devote themselves primarily to empirical work or theoretical modeling (Backhouse and Cherrier, 2017). Rather than evaluate our research question among empirical researchers only, we can also confirm our baseline result using theoretical economists within the same university as a natural control group. Accordingly, we add theoretical economists to our sample and interact the independent variable with $Empiricist_i$, a dummy that only takes a value of one for the economists whom we classify as empiricists. This allows us to include university-by-year fixed effects ($\delta_t \times \omega_j$), providing even more fine-grained controls for university-level research dynamics and time trends. In particular, exploiting only within-university variation rules out potential university-wide confounders that might correlate with opening an FSRDC (e.g., a sudden influx of funding in the economics department).[19] The results of this analysis, presented in Appendix Table H5, confirm our main estimates.

Nevertheless, the concern with using theoretical economists as a control group is that it would not control for other empiricist-specific productivity shocks co-occurring with FSRDC openings. For instance, FSRDCs might be accompanied by increased support only for applied work at the hosting institutions, hence confounding our estimates. We address this concern by exploiting within-university variation in "exposure"

[18] For instance, when looking at the effects of FSRDC access on publications in top journals, we can reject a null effect unless we are willing to allow for the linear extrapolation across consecutive periods to deviate by more than 16% from the linear trend estimated in the pre-period.

[19] This specification differs from a triple-difference estimate because the term "Post" is not defined for never-treated units due to the staggered rollout of FSRDCs. In our case, we can sidestep the need for a matching estimator because empiricists are assessed against theorists receiving access simultaneously because they are affiliated with the same institution. Additional details are in the footnote of Appendix Table H5.

to FSRDCs based on researchers' topical interest. Our interviews highlighted that FSRDCs access was less relevant to academics specializing in empirical fields such as development or behavioral economics. We use this variation to identify our main result using less exposed empiricists within the same university as an alternative control group. While less powered, the results presented in Appendix Table H6 confirm that the benefits of FSRDC access are larger among applied researchers in environmental, international, labor, and public economics. Besides corroborating our findings, this alternative design reassures us that other potential changes correlated with local FSRDC openings and benefiting empirical researchers are unlikely to drive our estimates.

During our interviews, we also discovered that when a consortium submitted an FSRDC application, the decision on which member would host the data center was often the outcome of a compromise among the interested parties. This suggests that within-consortium variation in distance from the hosting institution could also be used to identify the effect of data access while holding constant unobserved local factors shared by consortium members. For instance, this would account for investments in empirical research made by consortium members around the same time or if they all show a rising demand for confidential microdata. Leveraging this intuition, Appendix Table H7 shows that our results are mostly robust to within-consortium estimates.[20] In sum, the combined evidence from these additional research designs confirms that access to confidential microdata increases the quantity, quality, and policy relevance of research for empirical economists.

5.4 Other Robustness Checks

We perform various additional tests to confirm the robustness of our main findings. First, we investigate concerns about endogeneity in the choice of FSRDC locations. Note that the absence of pre-trends in individual productivity and the qualitative evidence discussed in Appendix A alleviate the concern that our results might simply be due to systematic sorting of FSRDC staggered openings. To test this concern directly, we examine trends in productivity using a panel at the university-year level. Besides finding significant aggregate gains for the institutions opening FSRDCs (Appendix Table H8), Appendix Figure H6 is helpful to reassure that FSRDCs are not systematically opened in institutions on a rising trend of research intensity.

Second, we rule out that other empiricist-specific shocks might drive our results at the universities that open a data center. We do so by alternatively excluding from the sample the researchers or the institution directly involved in bringing an FSRDC to a given location. We code the recipients of NSF grants establishing each FSRDC, and Appendix Table H9 shows that the results are robust to excluding them. Likewise, we find similar results if we estimate the effects only for researchers at universities that are in the same city as an

[20] Additional estimates reported in Appendix Figure H5 illustrate how the effect of FSRDC access rapidly declines with distance, even among consortium members. We thank an anonymous referee for suggesting this alternative identification strategy.

FSRDC but do not host the data center on their premises (Appendix Table H10). This robustness test is particularly compelling because researchers in non-hosting institutions close to the FSRDC are unlikely to be directly involved in its establishment but can nonetheless benefit from increased data access.

Third, we experiment with alternative specifications of "exposure" to confidential data in geographical and intellectual space. Appendix Table H11 shows that an economist's likelihood of using confidential FSRDC data rapidly decays with distance. In general, our results get stronger the closer an economist is to a data center, being the highest for researchers who enjoy access directly on their campus. Similarly, the effects of access monotonically grow when we use increasingly stringent definitions of empirical researchers (Appendix Figure H7). The placebo regressions in Appendix Table H12 find precisely estimated zeros when we estimate FSRDCs impact on theoretical economists. This reassures us that our estimates are not an artifact of the threshold we use to classify someone as an empiricist.

One might also be worried that researcher mobility events are driving our results, especially since we impute the year of the move based on publication data. Appendix Tables H13 show that the results are robust to excluding researchers that gain access due to (potentially endogenous) mobility events. An additional concern is that FSRDCs might just induce productive economists to relocate close to them. If that were the case, the effects we document would result from a zero-sum reallocation with potentially small aggregate scientific gains. Appendix Table H14 directly addresses this point. The opening of an FSRDC is not associated with an increase in department size (Column 1) nor in the number of star researchers (Columns 2 and 3). Furthermore, institutions hosting FSRDCs do not change their composition by becoming more empirically-focused (Column 4) or increasing the quality of their hires relative to standing faculty (Column 5), reassuring that our estimates reflect actual increases in research productivity.

6 Mechanisms

This section examines potential mechanisms that drive the productivity effect stemming from confidential data access. In particular, we explore (a) to what extent results are driven by spillovers to non-adopters and (b) what explains the presence of research quality spillovers following FSRDC openings.

A. Spillovers from Data Access: So far we have shown that confidential data access leads to greater productivity for empirical economists, but note that we did not restrict the sample to those who adopt the data directly in their work. This leaves open the question of to what extent productivity effects are driven by researchers who directly employ FSRDC data. Two complementary possibilities are that co-localization to an FSRDC: (1) directly impacts research quality by allowing the use of confidential data to produce more impactful research; or (2) it also indirectly alters research quality through spillovers to non-adopting researchers.

We evaluate the role of spillovers in explaining our overall results by examining the extent to which direct users of FSRDC data are driving our results. Table 4 reports our main results when we exclude researchers whom we observe using data from FSRDCs from the analyses. Notably, coefficients are around 21%–25% smaller but are still large and significant. The implication is that even though FSRDC openings benefit researcher productivity by enabling direct data usage, there are substantial spillovers from data access even among those who do not directly adopt (Myers and Lanahan, 2022). This finding suggests an additional mechanism through which access to FSRDCs shapes economics: exposing researchers to research based on confidential microdata. Our result earlier that FSRDC openings lead to increased citations to FSRDC research points to the possibility that local data access is raising general awareness about research based on FSRDC data,[21] with potential downstream implications for research trajectories.

Figure 4 shows two tests supporting this awareness channel. First, we estimate the effect of FSRDC access interacted with an indicator variable for researchers who cite research based on FSRDC data. We find that the positive effects on productivity do not extend to all empirical researchers at treated institutions: instead, they are limited to researchers whose work is directly influenced by past work based on FSRDCs. Second, we were told in our interviews that researchers often learn about the potential of FSRDC data after seeing the work of their colleagues (Appendix A). We, therefore, separately examine the effects for those treated empirical researchers in departments with and without colleagues using confidential microdata. We find no positive spillovers for researchers with fewer colleagues directly using FSRDC data. Both results support the idea that spillovers from co-localization operate by making researchers aware of research based on restricted-access FSRDC data.

B. How Exposure to FSRDCs Increases Research Quality: While both direct data usage and spillovers drive the effect of FSRDCs, we explore two specific channels through which awareness of past research carried out with FSRDC data might improve research output. First, researchers might be inspired to formulate new research questions close to the topics investigated with FSRDC data. We test this idea by examining whether researchers exposed to FSRDCs are more likely to explore topics such as labor, trade, or firm productivity. Table 5 confirms that the likelihood of working on JEL codes commonly associated with FSRDC data goes up by 9.3% over the sample mean, while there is no statistically significant increase for the remainder of JEL codes.[22] This result is robust to alternative ways to select which JEL codes most represent FSRDC-based research and gradually attenuates as we make the selection less stringent (Appendix Table H15).

[21] Appendix Figure H8 shows that an economist's likelihood of using and citing confidential data rapidly decays with distance. Results get stronger the closer an economist is to a data center, being the highest for researchers who enjoy access on their campus.

[22] We also explore whether FSRDC access encourages scholars to explore new topics rather than doubling down on the same questions they were already working on. Column 1 of Table 5 seems to indicate a higher propensity to work on new JEL codes, but the results are marginally below the conventional thresholds for statistical significance.

Second, exposed researchers might learn from the research design of FSRDC studies, potentially leading them to adopt similar methods or data with similar characteristics. Table 5 shows that treated researchers increase mentions of quasi-experimental methods, such as DID or synthetic controls. The absence of a similar effect on laboratory experiments or randomized control trials implies that this result is not a byproduct of a broader "credibility revolution." Learning about FSRDC data might also inspire the researcher to search for new datasets with similar characteristics and lower bureaucratic hurdles. This theme also emerged in our interviews, with respondents noting how administrative data from foreign countries often offer similar advantages and are easier to access (Appendix A). Indeed, we find suggestive evidence that researchers are more likely to employ microdata from other administrative sources, while a similar increase is absent for traditional research surveys.

These two effects, a shift in research direction and methods, are consistent with past work showing that access to new research tools encourages researchers to change their research trajectory and explore novel approaches (Murray et al., 2016; Nagaraj et al., 2020; Furman and Teodoridis, 2020). Nonetheless, it could be that FSRDCs increase researchers' productivity by simply enabling access to other types of resources. For instance, having access to FSRDCs could make it easier to apply for external grants. We rule out this possibility by showing that treated researchers are not more likely to mention grant funding in the acknowledgment of their papers (Appendix Table H16). It could also be that FSRDCs shift collaboration patterns, either helping researchers find new collaborators or increasing the size of research teams. Both these eventualities would imply that the observed productivity increases are due to the rise of teamwork (Jones, 2009; Wang and Barabási, 2021). We thoroughly explore these conjectures in Appendix Table H17 but do not find support for them.

7 Increasing the Diffusion of Confidential Microdata

Our analysis paints a picture of how FSRDC openings drive improved scientific productivity and policy impact among applied researchers. However, the productivity gains we document beg the question of how to structure a data infrastructure that ensures broad research access to confidential data under confidentiality constraints. While the optimal level of data access is ultimately a function of social preferences around privacy (Abowd and Schmutte, 2019), there might be avenues to increase the diffusion of confidential data, even holding the current regulatory environment constant. One approach could be growing the number of researchers that use existing FSRDCs, but this requires understanding better what frictions might be slowing down data usage. A second approach could be increasing the number of FSRDC facilities, which, however, is a costly strategy that needs a cost-benefit assessment. While a complete treatment of these issues is beyond the scope of our paper, in what follows, we build on our reduced form estimates to provide some guidance on both these aspects.

A. What are the Barriers to Using Confidential Data?: The use of FSRDCs rapidly declines with geographical distance from the closest FSRDC, more than halving for a distance of 50 miles (Appendix Figure H8). Yet, we also noted that some scholars are able to write several papers using confidential data, even when affiliated with universities further away from an FSRDC. What could be driving this heterogeneity? Researchers' financial constraints are a natural candidate. We learned during our interviews that researchers customarily add research assistants to their FSRDC project and help them obtain Special Sworn Status. In this way, they can bypass the hurdle of physical presence in an FSRDC by relying on their assistant to access data in the enclave. This (legitimate) practice suggests that access constraints might be more binding for researchers lacking the resources to recruit assistants to aid with data access.

Figure 5 reports estimates of FSRDCs' effect on the use of confidential data at increasing distances from the hosting institution. Each panel separately shows the impact for researchers we classify as having access to more resources. While we do not have precise data on individual research budgets, we employ several proxies for researchers' resources: the affiliation to an FSRDC consortium[23] (Panel (i)), the prestige of researchers' affiliation (Panel (ii)), and the availability of external funding (Panel (iii)). All three panels paint a consistent image: geographical distance is especially binding for less resource-endowed researchers. Panel (iii) is especially stark in showing how usage outside a radius of five miles from the host institution is mainly confined to researchers who can secure external funding sources. Indeed, even scholars at prestigious academic institutions usually need external grants to cover the hiring costs of full-time research assistants.

Documenting the role of resource constraints in accessing confidential data has distributional implications, especially because access to data disproportionately benefits marginalized researchers (Nagaraj et al., 2020). The research impact of FSRDCs could be magnified by expanding how many researchers use confidential data, even without increasing the number of FSRDC facilities. These results help support recent efforts by the Census Bureau to provide virtual access to confidential Title 13 data. While this program seems valuable, the open question remains how to increase accessibility for the data that cannot be shared remotely.

B. Assessing the Cost-Effectiveness of FSRDCs: To understand if a further expansion of the FSRDC network would be justified, we conduct a "back-of-the-envelope" cost-effectiveness analysis. A complete cost-benefit assessment is beyond the scope of our paper: we lack data for disciplines outside economics, and we cannot measure the benefits occurring to the Census Bureau from FSRDC research. However, we can undertake a simple mapping of the productivity impact of FSRDCs into monetary benefits that can be assessed relative to operational costs or alternative uses of FSRDCs' budgets (details are discussed in Appendix G). Specifically, our analysis requires an estimate of (i) the cost of operating an FSRDC, (ii) the

[23] Researchers affiliated with consortium members usually have their data access fees waived, while non-affiliated users usually pay a fee of around $15,000-20,000 per year.

value of an economics publication, and (iii) the number of additional publications induced by FSRDCs.[24]

A rough assessment of FSRDCs' operational costs can be derived from official CES website documentation. We use the upper bound yearly estimates of $185,000 per data center, and we estimate that, on average, 77% of that amount pertains to economics research projects.[25] Evaluating the monetary value of a scientific publication is more challenging. In Appendix G.2, we discuss three alternative approaches. Our preferred estimate derives from the implicit value that universities attribute to an additional top publication as revealed by the salary of their faculty. This approach yields an estimated present value of $26,916 per additional publication, the most conservative among our potential estimates (Appendix Table G1).

We then compare the costs of FSRDCs' operations with the value of the additional publications they generate. We present our results as a function of two interrelated parameters: the effect of FSRDC access on researchers' productivity and the number of consortium partners that sustain the expenditures for the data enclave. Appendix Figure G2 shows how much a publication must be valued to justify opening an FSRDC. Using our estimates in Table 3, we find that the value of an FSRDC is evident for medium-sized consortia of four or five institutions despite conservative assumptions about the value of additional publications. When considering institutions with a larger number of applied scholars and treatment effects reflecting the dynamically increasing impact of FSRDCs documented in Figure 3, even individual universities could justify opening an FSRDC.

Strikingly, investments in FSRDCs seem more effective in fostering scientific production than alternative uses of those resources (Appendix Figure G3). Compared to estimates of the impact of research grants to economists, our estimates imply that resources invested in an FSRDC yield 1.7-6.5 times more publications.[26] Taken together, this analysis implies that FSRDCs' large value-added likely surpasses their operating costs despite their relatively limited direct use. Reassuringly, our calculations rely on conservative choices: the rationale for operating an FSRDC grows substantially for institutions that attach higher value to scientific publications, experience higher-than-average local use, or participate in larger consortia.

[24] This evaluation is necessarily speculative, so our preference is to make choices that provide a more conservative assessment of FSRDCs' benefits. Depending on the application, some parameters might be preferred to the ones we use. For this reason, we leave the full menu of choices in the Appendix G for interested readers.

[25] We include in the calculation the start-up funding of $300,000 usually offered by the NSF. However, these cost estimates do not include the in-kind goods the hosting institution provides (e.g., physical space). A detailed discussion of these choices is in Appendix G.1.

[26] The estimates of Arora and Gambardella (2005) suggest that allocating $100,000 of NSF funding to an economist yields 0.48 additional top publications, much less than the 3.14 additional ones that our estimates would imply for a mid-sized FSRDC consortium. Section G.4 in Appendix G details our procedure and alternative benchmarks.

8 Conclusion

We assemble a novel longitudinal dataset of U.S.-based academic economists and exploit the staggered diffusion of U.S. Census Bureau's FSRDCs to investigate how increased data access shapes economic science. We first find that researchers co-located to an FSRDC are much more likely to use or build upon work that uses confidential data. We then assess the consequences of data access on scientific output. In our setting, researchers are more likely to publish scientifically impactful and policy-relevant papers in prestigious journals after they gain access to FSRDC data. We explore the mechanisms behind this finding and document significant spillovers on applied economists who do not directly use these data. Researchers exposed to work using confidential data are more likely to build upon it, exploring novel questions that stem from it and adopting similar research designs. Back-of-the-envelope calculations suggest that FSRDCs are highly cost-effective investments to increase the scientific productivity of applied economists.

Our results have implications for ongoing policy discussions on using confidential microdata for academic research. To the best of our knowledge, we are the first to provide causal, empirical evidence for the debate around the growing role of these data in economics research. Our findings are consistent with the idea that increased access to confidential data is crucial to scientific progress, but also that the current regime of restricted access poses challenges for researchers with fewer resources. Harnessing the full potential to inform policy of confidential microdata collected by the U.S. Federal Government might require legislative changes to ensure broader research access. In the context of the FSRDC network, even holding current regulations and procedures constant, we document how a further expansion of the secure facilities where the data are accessible might be warranted. Further, our findings around spillovers show how evaluations of the impact of data access programs need to extend beyond those directly using the data. Data access can shape research by changing the topical and methodological focus of entire fields, leading to more impactful science across the board.

Our work has some limitations. First, our intention is not to provide a complete policy evaluation of the FSRDC network. Our analysis is only a modest start in this regard. To go deeper, we would need to expand our focus to other disciplines that benefit from FSRDC data, such as health policy and demography. Moreover, the Census Bureau's objective in creating the FSRDC network is to obtain direct benefits for its data programs (Foster et al., 2009). To assess the overall success of the network from the Census viewpoint, we would need to consider all such benefits, including the improvement of existing datasets (CES, 2017). Our work cannot capture all these aspects. Still, we establish that, at least in the case of economics, the researcher-level impact of these institutions is significant and directly leads to the production of higher-quality scientific output that receives more attention in policy documents. Even if it is likely to understate the total value generated by FSRDCs, our cost-benefit assessment documents high returns from

these data-access institutions.

Second, our analysis faces some data limitations. Our extensive robustness checks rule out the most obvious confounds, but we suffer from the common pitfalls of studies based on bibliographic data. In particular, our within-researcher estimates could not consider dynamic changes in connections to editors, which might directly affect scientific productivity, or in researchers' status, which might affect the impact of their work. These changes might be partially interlinked with FSRDC access. Further, we do not observe specific datasets that became available over time through FSRDCs and cannot examine their impact separately. Finally, our results are based on a particular type of confidential data that entails geographical constraints. While this is crucial to our research design, it also limits the generalizability of our results to contexts where researchers face similarly large barriers to accessing data.

Despite these caveats, our analysis helps inform the debate around confidential data access. Data are the lifeblood of modern economics, which means that the ability to understand and intervene in the economy is fundamentally "constrained by the extent and quality of the available data" (Griliches, 1994). While more research is needed to design a data access infrastructure that maximizes research use and ensures privacy protection, our results indicate that the payoffs of doing so are vast for both research and policy.

References

Abowd, J. M. and J. Lane (2004): "New approaches to confidentiality protection: Synthetic data, remote access and research data centers," in *International Workshop on Privacy in Statistical Databases*, Springer, 282–289.

Abowd, J. M. and I. M. Schmutte (2019): "An economic analysis of privacy protection and statistical accuracy as social choices," *American Economic Review*, 109, 171–202.

Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer, and S. Woodcock (2009): "The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators," in *Producer Dynamics: New Evidence from Micro Data*, University of Chicago Press, 149–230.

Abraham, K. G., R. S. Jarmin, B. Moyer, and M. D. Shapiro (2022): *Big Data for 21st Century Economic Statistics*, NBER Book Series Studies in Income/Wealth.

Akerman, A., I. Gaarder, and M. Mogstad (2015): "The skill complementarity of broadband internet," *The Quarterly Journal of Economics*, 130, 1781–1824.

Angrist, J., P. Azoulay, G. Ellison, R. Hill, and S. F. Lu (2020): "Inside job or deep impact? Extramural citations and the influence of economic scholarship," *Journal of Economic Literature*, 58, 3–52.

Atrostic, B. (2007): "The Center for Economic Studies 1982-2007: A brief history," CES working paper.

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2019): "Public R&D investments and private-sector patenting: Evidence from NIH funding rules," *The Review of Economic Studies*, 86, 117–152.

Backhouse, R. E. and B. Cherrier (2017): "The age of the applied economist: The transformation of economics since the 1970s," *History of Political Economy*, 49, 1–33.

Biasi, B. and P. Moser (2021): "Effects of copyrights on science: Evidence from the WWII Book Republication Program," *American Economic Journal: Microeconomics*, 13, 218–60.

Bloom, N., E. Brynjolfsson, L. Foster, R. Jarmin, M. Patnaik, I. Saporta-Eksten, and J. Van Reenen (2019): "What drives differences in management practices?" *American Economic Review*, 109, 1648–1683.

Bowen, C. M. (2024): "Government Data of the People, by the People, for the People: Navigating Citizen Privacy Concerns," *Journal of Economic Perspectives*, 38, 181–200.

Brodeur, A., N. Cook, and A. Heyes (2020): "Methods matter: P-hacking and publication bias in causal analysis in economics," *American Economic Review*, 110, 3634–3660.

Callaway, B. and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 225, 200–230.

Card, D. (2022): "Design-based research in empirical microeconomics," *American Economic Review*, 112, 1773–1781.

Card, D., R. Chetty, M. S. Feldstein, and E. Saez (2010): "Expanding access to administrative data for research in the United States," in *Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*, American Economic Association.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020): "Are referees and editors in economics gender neutral?" *The Quarterly Journal of Economics*, 135, 269–327.

——— (2022): "Gender differences in peer recognition by economists," *Econometrica*, 90, 1937–1971.

Carrell, S. E., D. Figlio, and L. Lusher (2024): "Clubs and networks in economics reviewing," *Journal of Political Economy*.

CES (2017): "Center for Economic Studies and Research Data Centers Research Report: 2016," Available on the U.S. Census Bureau's website.

——— (2019): "Center for Economic Studies and Research Data Centers Research Report: 2018," .

Chetty, R. (2012): "Time trends in the use of administrative data for empirical research," NBER Summer Institute presentation. Available at the author's website.

Chetty, R. and J. N. Friedman (2019): "A practical method to reduce privacy loss when disclosing statistics based on small samples," in *AEA Papers and Proceedings*, vol. 109, 414–20.

Chetty, R., J. N. Friedman, E. Saez, and D. Yagan (2018): "The SOI Databank: A case study in leveraging administrative data in support of evidence-based policymaking," *Statistical Journal of the IAOS*, 34, 99–103.

Chetty, R., J. N. Friedman, M. Stepner, and O. I. Team (2024): "The economic impacts of COVID-19: Evidence from a new public database built using private sector data," *The Quarterly Journal of Economics*, 139, 829–889.

Chow, M. C., T. C. Fort, C. Goetz, N. Goldschlag, J. Lawrence, E. R. Perlman, M. Stinson, and T. K. White (2021): "Redesigning the Longitudinal Business Database," NBER Working Paper w28839.

Cole, S., I. Dhaliwal, A. Sautmann, and L. Vilhuber (2020): *Handbook on Using Administrative Data for Research and Evidence-Based Policy*, JPAL and MIT Press.

Currie, J., H. Kleven, and E. Zwiers (2020): "Technology and big data are changing economics: Mining text to track methods," *AEA Papers and Proceedings*, 110, 42–48.

Davis, J. C. and B. P. Holly (2006): "Regional analysis using Census Bureau microdata at the Center for Economic Studies," *International Regional Science Review*, 29, 278–296.

Davis, S. J., J. C. Haltiwanger, and S. Schuh (1998): "Job Creation and Destruction," *MIT Press Books*.

Dizikes, P. (2019): "The productive career of Robert Solow," *MIT Technology Review*.

Dupriez, O. and E. Boyko (2010): "Dissemination of microdata files: principles procedures and practices," The World Bank, IHSN Working Paper No 005.

Einav, L. and J. Levin (2014a): "The data revolution and economic analysis," *Innovation Policy and the Economy*, 14, 1–24.

——— (2014b): "Economics in the age of big data," *Science*, 346, 1243089.

Feenberg, D., I. Ganguli, P. Gaule, and J. Gruber (2017): "It's good to be first: Order bias in reading and citing NBER working papers," *Review of Economics and Statistics*, 99, 32–39.

Foster, L., R. Jarmin, and L. Riggs (2009): "Resolving the tension between access and confidentiality: Past experience and future plans at the U.S. Census Bureau," *Statistical Journal of the IAOS*, 26, 113–122.

Furman, J. L. and S. Stern (2011): "Climbing atop the shoulders of giants: The impact of institutions on cumulative research," *American Economic Review*, 101, 1933–1963.

Furman, J. L. and F. Teodoridis (2020): "Automation, research technology, and researchers' trajectories: Evidence from computer science and electrical engineering," *Organization Science*, 31, 330–354.

Goroff, D., J. Polonetsky, and O. Tene (2018): "Privacy protective research: Facilitating ethically responsible access to administrative data," *The ANNALS of the American Academy of Political and Social Science*, 675, 46–66.

Griliches, Z. (1994): "Productivity, R&D, and the Data Constraint," *American Economic Review*, 84, 1–23.

Groves, R. M. (2011): "Three eras of survey research," *Public Opinion Quarterly*, 75, 861–871.

Hamermesh, D. S. (2013): "Six decades of top economics publishing: Who and how?" *Journal of Economic Literature*, 51, 162–172.

Heckman, J. J. (2001): "Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture," *Journal of Political Economy*, 109, 673–748.

Heckman, J. J. and S. Moktan (2020): "Publishing and promotion in economics: The tyranny of the top five," *Journal of Economic Literature*, 58, 419–470.

Hill, R. and C. Stein (2024a): "Race to the bottom: Competition and quality in science," Northwestern University and UC Berkeley.

——— (2024b): "Scooped! Estimating rewards for priority in science," *Journal of Political Economy*.

Hill, R., C. Stein, and H. Williams (2020): "Internalizing externalities: Designing effective data policies," *AEA Papers and Proceedings*, 110, 49–54.

Hjort, J., D. Moreira, G. Rao, and J. F. Santini (2021): "How research affects policy: Experimental evidence from 2,150 Brazilian municipalities," *American Economic Review*, 111, 1442–1480.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2024): "The streetlight effect in data-driven exploration," UC Berkeley, University of Vienna, and University of Pennsylvania.

Jarmin, R. S. and J. Miranda (2002): "The longitudinal business database," CES working paper.

Jelveh, Z., B. Kogut, and S. Naidu (2024): "Political language in economics," *The Economic Journal*, ueae026.

Jones, B. F. (2009): "The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?" *The Review of Economic Studies*, 76, 283–317.

Kalaitzidakis, P., T. P. Mamuneas, and T. Stengos (2003): "Rankings of academic journals and institutions in economics," *Journal of the European Economic Association*, 1, 1346–1366.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011): "Towards unrestricted public use business microdata: The synthetic longitudinal business database," *International Statistical Review*, 79, 362–384.

Lane, J. (2021): *Democratizing Our Data: A Manifesto*, MIT Press.

Locarnini, M., A. Mishonov, O. Baranova, T. Boyer, M. Zweng, H. Garcia, D. Seidov, K. Weathers, et al. (2018): "World ocean atlas 2018, volume 1: Temperature," NOAA Atlas NESDIS 81.

McGuckin, R. H. (1995): "Establishment microdata for economic research and policy analysis: Looking beyond the aggregates," *Journal of Business & Economic Statistics*, 13, 121–126.

McGuckin, R. H., R. H. McGukin, and A. P. Reznek (1993): "The statistics corner: Research with economic microdata: The Census Bureau's Center for Economic Studies," *Business Economics*, 52–58.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern (2016): "Of mice and academics: Examining the effect of openness on innovation," *American Economic Journal: Economic Policy*, 8, 212–52.

Myers, K. (2020): "The elasticity of science," *American Economic Journal: Applied Economics*, 12, 103–134.

Myers, K. R. and L. Lanahan (2022): "Estimating spillovers from publicly funded R&D: Evidence from the U.S. Department of Energy," *American Economic Review*, 112, 2393–2423.

Nagaraj, A. (2024): "A mapping lens for estimating data value," *Harvard Data Science Review*, https://hdsr.mitpress.mit.edu/pub/c84g7tz9.

Nagaraj, A., E. Shears, and M. de Vaan (2020): "Improving data access democratizes and diversifies science," *Proceedings of the National Academy of Sciences*, 117, 23490–23498.

Önder, A. S. and S. Schweitzer (2017): "Catching up or falling behind? Promising changes and persistent patterns across cohorts of economics PhDs in German-speaking countries from 1991 to 2008," *Scientometrics*, 110, 1297–1331.

Rambachan, A. and J. Roth (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90, 2555–2591.

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021): "Gender differences in recognition for group work," *Journal of Political Economy*, 129, 101–147.

Taubman, S. L., H. L. Allen, B. J. Wright, K. Baicker, and A. N. Finkelstein (2014): "Medicaid increases emergency-department use: Evidence from Oregon's Health Insurance Experiment," *Science*, 343, 263–268.

Truffa, F. and A. Wong (2022): "Undergraduate gender diversity and direction of scientific research," *American Economic Review*.

Waldinger, F. (2016): "Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge," *Review of Economics and Statistics*, 98, 811–831.

Wang, D. and A.-L. Barabási (2021): *The Science of Science*, Cambridge University Press.

Wapman, K. H., S. Zhang, A. Clauset, and D. B. Larremore (2022): "Quantifying hierarchy and dynamics in US faculty hiring and retention," *Nature*, 610, 120–127.

Weinberg, D. H., J. M. Abowd, P. M. Steel, L. Zayatz, and S. K. Rowland (2007): "Access methods for United States microdata," CES working paper.

Williams, H. L. (2013): "Intellectual property rights and innovation: Evidence from the human genome," *Journal of Political Economy*, 121, 1–27.

Yin, Y., Y. Dong, K. Wang, D. Wang, and B. F. Jones (2022): "Public use and public funding of science," *Nature Human Behaviour*, 6, 1344–1350.

Yin, Y., J. Gao, B. F. Jones, and D. Wang (2021): "Coevolution of policy and science during the pandemic," *Science*, 371, 128–130.

York, D. G., J. Adelman, J. E. Anderson Jr, S. F. Anderson, J. Annis, N. A. Bahcall, et al. (2000): "The Sloan Digital Sky Survey: Technical summary," *The Astronomical Journal*, 120, 1579.

9 Tables and Figures

Table 1: Summary Statistics

Panel A: Researcher-Level

	N	Mean	Std. Dev.	Median	Min	Max
Ever Had Access (0/1)	10060	0.206	0.40	0	0	1
Year of Access	2075	2009.988	5.94	2011	1994	2019
Ever Used FSRDC (0/1)	10060	0.038	0.19	0	0	1
Ever Cited FSRDC (0/1)	10060	0.284	0.45	0	0	1
Lifetime Top Publications	10060	1.944	3.62	1	0	57
Lifetime Cite-weighted Papers	10060	75.952	210.09	8	0	3776
Lifetime Policy-relevant Papers	10060	3.898	5.70	2	0	79

Panel B: Researcher-Year-Level

	N	Mean	Std. Dev.	Median	Min	Max
Post-FSRDC (0/1)	155720	0.108	0.31	0	0	1
Papers Using FSRDC Data	155720	0.005	0.08	0	0	3
Papers Citing FSRDC Papers	155720	0.041	0.22	0	0	6
Top Publications	155720	0.126	0.40	0	0	7
Cite-weighted Papers	155720	4.907	23.24	0	0	1285
Policy-relevant Papers	155720	0.252	0.58	0	0	11
New JEL Codes (0/1)	155720	0.303	0.46	0	0	1
Papers with FSRDC JEL (0/1)	155720	0.149	0.36	0	0	1
Papers without FSRDC JEL (0/1)	155720	0.201	0.40	0	0	1
Papers Mentioning Admin Data	155720	0.006	0.08	0	0	4
Papers Mentioning Survey Data	155720	0.018	0.14	0	0	6
Quasi-experimental Papers	155720	0.023	0.16	0	0	4
Experimental Papers	155720	0.007	0.09	0	0	5
Year	155720	2006.229	7.69	2007	1990	2019

Note: This table lists summary statistics at the researcher level for 10,060 publishing empirical economists (Panel A) and at the researcher-year level for an unbalanced panel of 155,720 observations (Panel B). Ever Had FSRDC Access: 0/1 = 1 for researchers who spent at least one year co-located to an active FSRDC. Year of Access: average year when a researcher becomes co-located to an active FSRDC. Ever Used FSRDC: 0/1 for researchers who published at least one paper using FSRDC data. Ever Cited FSRDC: 0/1 = 1 for researchers who published at least one paper that cited a publication based on FSRDC data. Lifetime Top Publications: sum of the papers in top economics journals. Lifetime Cite-Weighted Publications: sum of the papers weighted by the citations received up to the five years after publication. Lifetime Policy-relevant Publications: sum of papers cited at least once in a policy document. Empiricist: 0/1 = 1 for those researchers whose majority of first five publications are empirical in nature. Post-FSRDC: 0/1 = 1 after a researcher is first co-located to an active FSRDC. Papers Using FSRDC Data: count of papers using FSRDC data. Papers Citing FSRDC Papers: count of papers citing papers based on FSRDC data (the count excludes those directly using FSRDC data). Top Publications: count of papers in top economics journals. Cite-Weighted Papers: count of papers weighted by the citations received up to the five years after publication. Policy-Relevant Papers: count of papers cited at least once in a policy document. New JEL Codes: 0/1 = 1 for researchers who used JEL codes that they had not used before. Papers with FSRDC JEL: 0/1 = 1 for researchers who used JEL codes common among papers using FSRDC data. Papers with non-FSRDC JEL: 0/1 = 1 for researchers who use JEL codes common among papers not using FSRDC data. Papers Mentioning Admin Data: count of papers mentioning the use of administrative data in their title or abstract. Papers Mentioning Survey Data: count of papers mentioning the use of survey data in their title or abstract. Quasi-Experimental Papers: count of papers mentioning the use of quasi-experimental methods in their title or abstract. Experimental Papers: count of papers mentioning the use of experimental methods in their title or abstract. Year: average year of publication. See text for details.

Table 2: Effect of FSRDC Access on the Diffusion of Confidential Data

	Papers Using FSRDC Data		Papers Citing FSRDC Papers	
	(1)	(2)	(3)	(4)
Post-FSRDC	0.00590***	0.00582***	0.0175***	0.0107**
	(0.00128)	(0.00141)	(0.00411)	(0.00372)
Researcher FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	Yes	No
University FE	Yes	No	Yes	No
University Tier × Year FE	No	Yes	No	Yes
N	155615	155622	155615	155622

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data diffusion among empirical economists. Columns (1) and (2) report results from OLS models, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Columns (3) and (4) report results from OLS models, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution with an operating FSRDC. All models include individual fixed effects. Columns (1) and (3) include year fixed effects and university fixed effects, while columns (2) and (4) include year fixed effects interacted with university tier dummies. The number of observations changes slightly due to the inclusion of different fixed effects. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

Table 3: Effect of FSRDC Access on Research Output

	Top Publications		Cite-weighted Publications		Policy-relevant Publications	
	(1)	(2)	(3)	(4)	(5)	(6)
Post-FSRDC	0.0240***	0.0230**	1.706***	1.239**	0.0361**	0.0362**
	(0.00665)	(0.00839)	(0.43773)	(0.42415)	(0.01374)	(0.01139)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	No	Yes	No	Yes	No
University FE	Yes	No	Yes	No	Yes	No
University Tier × Year FE	No	Yes	No	Yes	No	Yes
N	155615	155622	155615	155622	155615	155622

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research output for empirical economists. Columns (1) and (2) of report results from OLS models, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals from Heckman and Moktan, 2020). Columns (3) and (4) report results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Columns (5) and (6) report results from OLS models, where the dependent variable is the count of articles with at least one citation from a policy document. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution with an operating FSRDC. All models include individual fixed effects. Columns (1), (3), (5) include year fixed effects and university fixed effects, while columns (2), (4), (6) include year fixed effects interacted with university tier dummies. The number of observations changes slightly due to the inclusion of different fixed effects. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

Table 4: Effect of FSRDC Access on Research Output for Non-FSRDC Users

	Top Publications		Cite-weighted Publications		Policy-relevant Publications	
	Baseline	Indirect	Baseline	Indirect	Baseline	Indirect
	(1)	(2)	(3)	(4)	(5)	(6)
Post-FSRDC	0.0230**	0.0181*	1.239**	0.930*	0.0362**	0.0280*
	(0.008)	(0.008)	(0.424)	(0.374)	(0.011)	(0.011)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	155622	149113	155622	149113	155622	149113

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research output for empirical economists after removing from the sample researchers using FSRDC data. Columns (1), (3), and (5) are our baseline models with the full sample of researchers, while Columns (2), (4), and (6) remove researchers using FSRDC data. Columns (1) and (2) report results from OLS models, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals from Heckman and Moktan, 2020). Columns (3) and (4) report results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Columns (5) and (6) report results from OLS models, where the dependent variable is the count of articles with at least one citation from a policy document. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

Table 5: Effect of FSRDC Access on the Direction and Design of Empirical Research

Panel A: Research Direction

	New JEL Codes (0/1)	FSRDC JEL Paper (0/1)	Non-FSRDC JEL Paper (0/1)
	(1)	(2)	(3)
Post-FSRDC	0.0141	0.0139*	0.0123
	(0.008)	(0.006)	(0.008)
Researcher FE	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes
N	155622	155622	155622

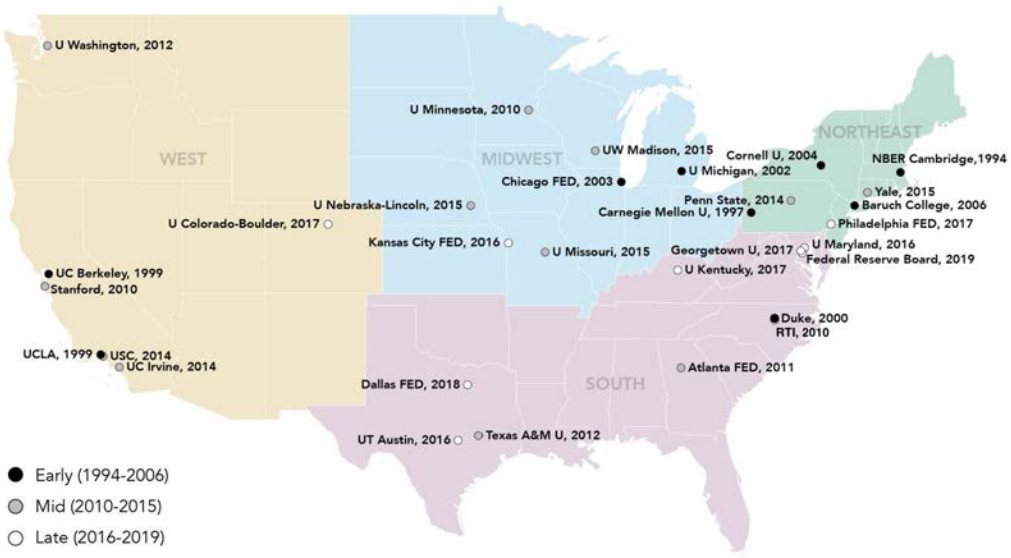
Panel B: Research Design

	Data		Research Methods	
	Admin	Survey	Quasi-Exp.	Experiment
	(1)	(2)	(3)	(4)
Post-FSRDC	0.00333*	0.00315	0.00632*	-0.00115
	(0.002)	(0.003)	(0.003)	(0.002)
Researcher FE	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes
N	155622	155622	155622	155622

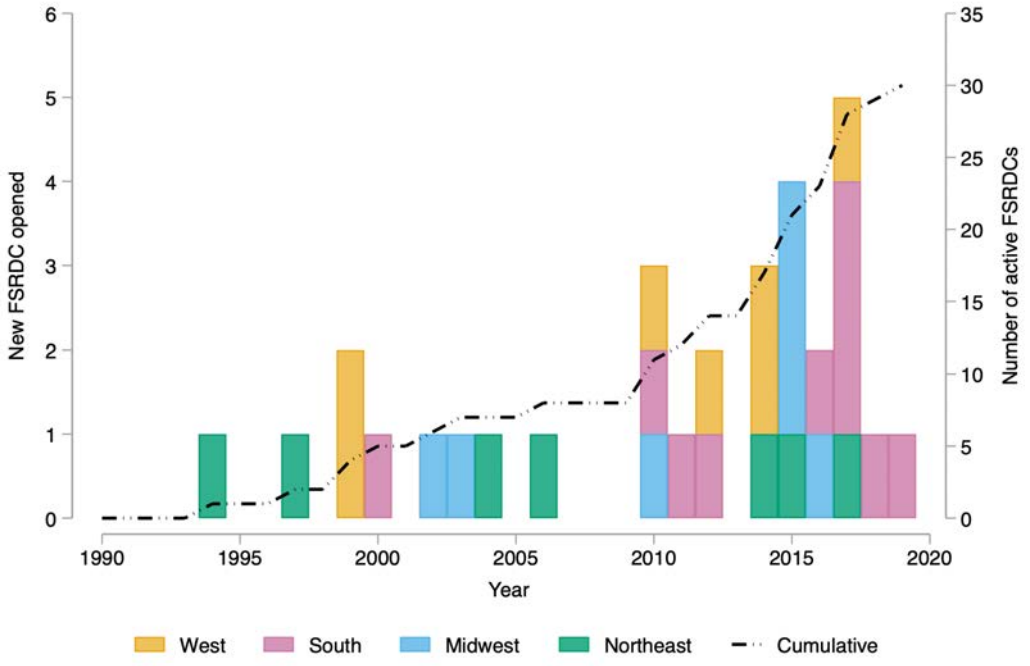
Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on the direction and design of empirical research. Column (1) of Panel A reports results from a linear probability model, where the dependent variable is an indicator equal to one if the researcher used at least one JEL code never used before. Column (2) of Panel A reports results from a linear probability model, where the dependent variable is an indicator equal to one if the researcher used at least one FSRDC JEL code during the year. Column (3) of Panel A reports results from a linear probability model, where the dependent variable is an indicator that equals one if the researcher used at least one non-FSRDC JEL code during the year. Columns (1) and (2) of Panel B report results from OLS models, where the dependent variable is the number of published articles that mention in their title or abstract keywords related to the use of administrative and survey data, respectively. Columns (3) and (4) of Panel B report results from OLS models, where the dependent variable is the number of published articles that mention in their title or abstract keywords related to the use of quasi-experimental or experimental methods, respectively. The regressions exclude papers directly using FSRDC data. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

Figure 1: Expansion of the FSRDC Program over Time and Space

Panel A: Geographic Expansion of FSRDCs



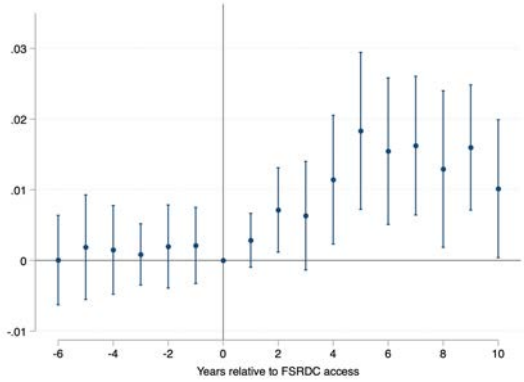
Panel B: Time Expansion of FSRDCs



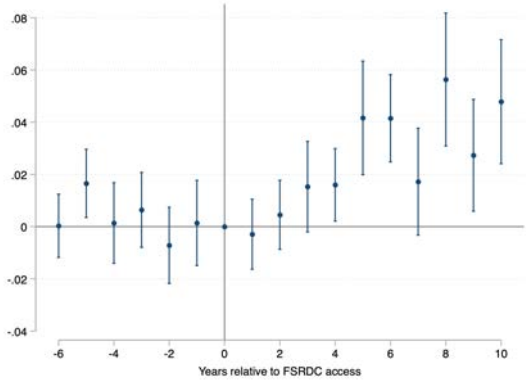
Note: This figure illustrates the timing and geographic scope of the expansion of the FSRDC network. See text for more details.

Figure 2: Time-Varying Estimates of the Impact of FSRDCs on the Diffusion of Confidential Data

(i) Papers Using FSRDC Data



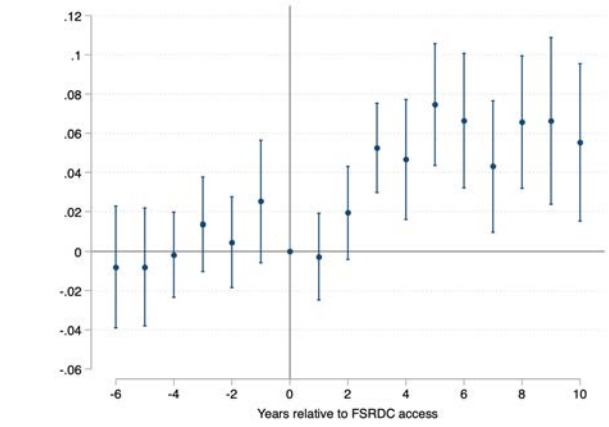
(ii) Papers Citing FSRDC Papers



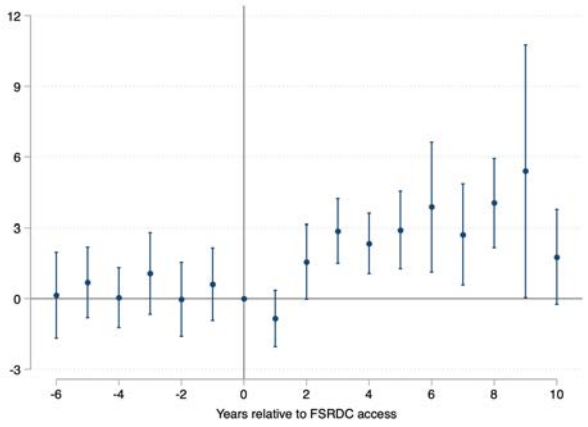
Note: This figure provides visual illustrations of the event study version of the main regression on confidential data adoption among empirical researchers. The main dependent variables are the number of papers written using FSRDC data (panel (i)) or the number of papers that cite papers using FSRDC data (panel (ii)). The chart plots values of β for different lags before and after obtaining access to an FSRDC with 95% confidence intervals. Regressions include researcher and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure 3: Time-Varying Estimates of the Impact of FSRDCs on Research Output

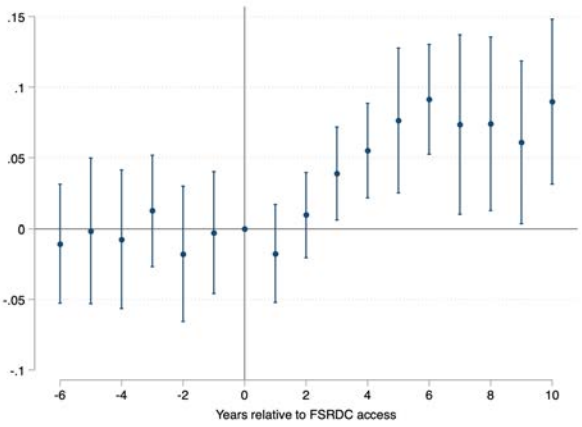
(i) Top Publications



(ii) Cite-Weighted Publications



(iii) Policy-Relevant Publications

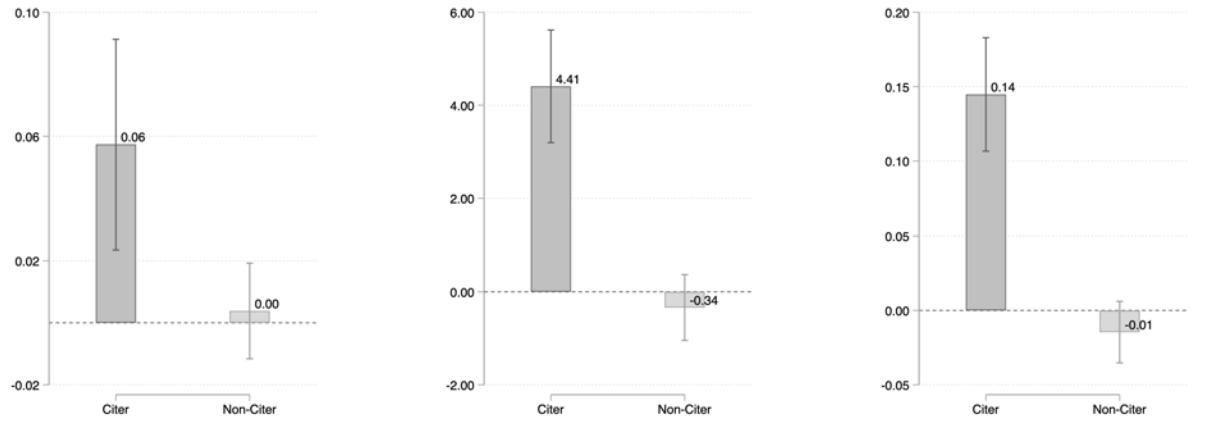


Note: This figure provides visual illustrations of the event study version of the main regressions evaluating the impacts of FSRDC access on measures of research output for empirical researchers. The main dependent variables are the count of papers published in the main economics journals (panel (i)), the count of papers weighted by the number of citations received up to five years following publication (panel (ii)), and the count of papers cited in policy documents (panel (iii)). The charts plot values of β for different lags before and after obtaining access to an FSRDC with 95% confidence intervals. Regressions include researcher and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure 4: Heterogeneous Effect of FSRDC Access

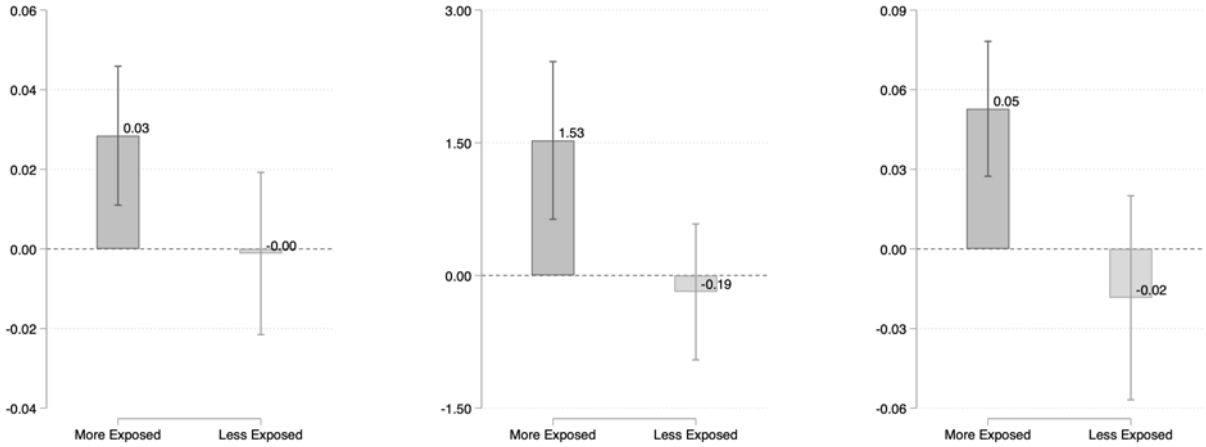
Panel A: by Researchers Citing FSRDC Research

(i) Top Publications *(ii) Cite-Weighted Publications* *(iii) Policy-Relevant Publications*



Panel B: by Researchers Exposed to FSRDC Research

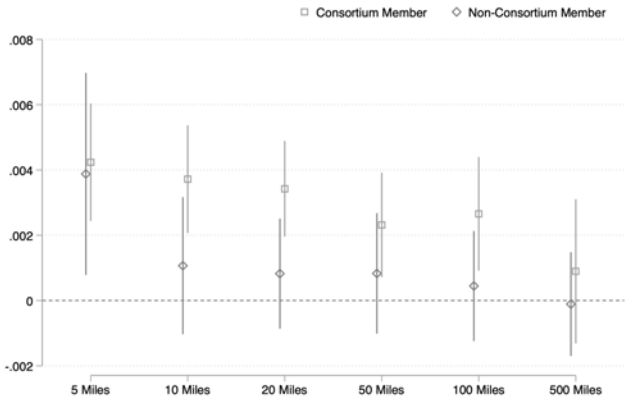
(i) Top Publications *(ii) Cite-Weighted Publications* *(iii) Policy-Relevant Publications*



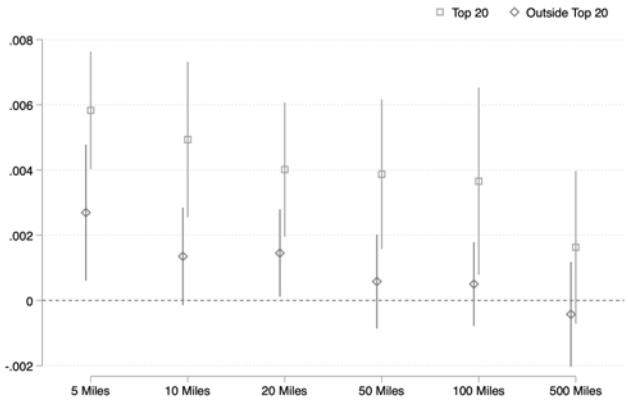
Note: This figure provides visual illustrations of the impacts of FSRDC access on measures of research output for different types of empirical researchers. Each chart is estimated from a single interaction-based regression and excludes economists who have used FSRDC data directly. Columns (i), (ii), and (iii) of Panel A show the effects of being affiliated with an institution with an FSRDC separately for researchers who cited a paper using FSRDC data versus those who have not. Columns (iv), (v), and (vi) of Panel B show the effects separately for researchers who had an above-median number of lifetime colleagues who directly used FSRDC data users versus those who did not. The main dependent variables are the number of top publications (columns (i) of Panels A and B), the citation-weighted number of publications (columns (ii) of Panels A and B), and the number of policy-relevant publications (columns (iii) of Panels A and B). Regressions include researcher and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure 5: Determinants of Confidential Data Use by Distance to the Closest FSRDC

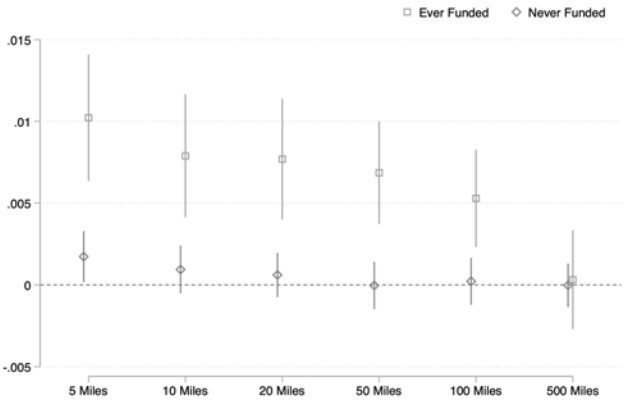
(i) Consortium Membership of Researcher



(ii) University Ranking of Researcher



(iii) External Funding Status of Researcher



Note: This figure explores which factors foster confidential data usage at different distances from the closest FSRDC. We chart estimates from OLS models evaluating the effect of distance on the individual count of papers using confidential data. We consider researchers treated if the geometric distance between their institution and the nearest FSRDC is within a certain threshold (ranging from 5 to 500 miles). The regressions are estimated separately per distance threshold. At each distance threshold, we add an interaction for empirical economists of differing consortium membership status (panel (i)), university rank (panel (ii)), and funding status (panel (iii)). Consortium Member: 0/1 = 1 if the researcher is affiliated with an FSRDC consortium partner institution at the time of treatment. Top 20: 0/1 = 1 if the researcher is affiliated with a top 20 institution at the time of treatment, per Kalaitzidakis et al. (2003). Ever Funded: 0/1 = 1 if the researcher acknowledges external funding sources. Regressions include researcher and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

How Does Data Access Shape Science? The Impact of Federal Statistical Research Data Centers on Economics Research

Online Appendix

Abhishek Nagaraj
UC Berkeley-Haas & NBER
nagaraj@berkeley.edu

Matteo Tranchero
The Wharton School
mtranc@wharton.upenn.edu

A Qualitative Evidence

A.1 Interviews

We conducted nine semi-structured interviews with FSRDC administrators and users to learn more about our empirical setting. The interviews have been crucial in providing us with institutional details on the FSRDC network, validating our research design, and suggesting additional sources of variation. Before our first interview, we consulted the Berkeley IRB to assess whether our questions would constitute human subjects research. The IRB assessed that our questions were factual and focused on the institution, thus not requiring formal IRB review. Below, we summarize the main topics that emerged during our conversations.

Expansion of the FSRDC network: We inquired about the history and institutional evolution of the FSRDC network. Three key factors shaped the progressive diffusion of research data centers: the presence of a large potential user community, the goal of geographical parity in access, and the advocacy of local research leaders. While demonstrating sufficient local demand eventually became necessary, it was not a sufficient condition to open an FSRDC. This is because the NSF, upon becoming involved in establishing new FSRDCs, explicitly tried to achieve an equitable geographical diffusion across the United States after taking into consideration local demand (interview T14). The NSF pursued this objective by carefully allocating the grants that help institutions set up a new data center. Obtaining the NSF grant was fairly competitive and "lots of places that have tried to get the NSF money haven't been able to" (interview D88). Given this allocation process, the likelihood of winning NSF's support often hinged on the distance from the closest FSRDC: "If you're close to an already existing location, that probably hurts your application" (interview S12). Moreover, submitting an FSRDC application hinged on the presence of individuals who would advocate opening a new data center (interview S94). The following quote well summarizes the importance of vocal advocacy:

> "And he persuaded the Dean that this would be a really good thing for the School. Well, it would be a really good thing for himself, which would be a really good thing for the School and

the University." (interview Y79)

In practice, a combination of many potential users and key individual advocates was fundamental for a successful application to the NSF:

> "[...] for an RDC to get established, you're going to need at least one and preferably a few sort of well-connected and very enthusiastic people who want to push for having an RDC. I don't think there's any case where you'd find that there were sort of 40 people who kind of felt it would be nice to have an RDC, and when number 41 said they'd kind of like to do it, then it all sort of magically happens. You've got to have, and in some cases they might have just been one influential senior person who could kind of really push for it and talk it up." (interview D88)

Moreover, our interviewees revealed idiosyncrasies behind each FSRDC opening. For instance, Boston's choice for the first FSRDC was in no small part due to the presence of a Census Regional Office willing to host it before it was eventually transferred to NBER premises (interview Y79). Another handful of FSRDCs was either opened because of the will of university administrators or because of specific collaborations between a faculty member and the Census Bureau (interview S94). In cases where a consortium opened the FSRDC, the choice of which consortium member would host the data center was often the result of a compromise (interview S12). Taken together, this qualitative evidence suggests that geographic and idiosyncratic factors strongly influenced both the timing and the locations of FSRDCs.

Scientific and policy impact of FSRDC data: All the people we talked to confirmed the revolutionary impact of U.S. Census's confidential data on economics research. Labor, productivity, health, trade, public, and environmental economics were regarded as the fields that benefited the most. While empirical researchers clearly benefited the most, there was a sense among our respondents that the linked datasets available in the FSRDCs "opened up fields of research that weren't possible before" (interview T14). One example is constituted by the work of John Haltiwanger and colleagues, whose impact transcended the traditional boundaries between labor and macroeconomics:

> "I think if you look at Haltiwanger's work, I think his work has transformed, I don't know whether you want to call it labor or macro or, you know, kind of the intersection of labor and macro. I think he's had a big impact." (interview Y79)

Besides research, our respondents felt that FSRDC data yielded findings with an outsized policy impact. Our respondents said that this was due to two characteristics of FSRDC data: representativeness and granularity. Without representative data, "you're only getting a sliver of the story", as Nick Bloom said on the occasion of

the Stanford FSRDC opening.[1] Without granularity, "special populations get lost in the noise of aggregated data sets," as Bill Maurer remarked during the opening of the UC Irvine FSRDC.[2] In practice, confidential microdata allow to study heterogeneous effects for sub-populations, which is what policymakers usually need (interview P39).

Productivity gains from data access: Among our respondents, there was a sense that confidential data were "incredibly good" for the careers of the researchers who had access to them (interview S94). Access to FSRDC data "opens up a lot of research questions" (interview S12) that translate into high-quality publications. Interestingly, our respondents also highlighted that the local availability of FSRDC data was likely to have heterogeneous benefits for different types of scholars. Most people agreed that the lengthy and uncertain approval process to work in an FSRDC posed risks for faculty with fewer resources. The whole process requires researchers to be "a little bit careful and a little bit lucky" (interview S12). However, working with these data is perceived to be an upfront investment that can offer large payoffs throughout a career, especially for researchers who started working in an FSRDC during their PhD (interviews S12, T14, S94). The following quote summarizes well the point:

> "I think the problem with being a junior professor doing that, it just takes a long time to get something done. So it'd be dangerous as a first year assistant professor to say, 'I'm going to start using an RDC and I'm going to get tenure' because that may not work out so well, but if you worked in it as a grad student and you've already got stuff going on that can be pretty, pretty effective." (interview D88)

A theme that emerged from our conversations was that even relatively small geographic barriers are a major hindrance to using an FSRDC. Several administrators indicated that, in their experience, a one-hour commute is enough to discourage a researcher from ever applying to the data (interviews P39, S94, T52). Not surprisingly, several users said they hire research assistants to physically access the data enclaves and conduct analyses (interview T14, F72). This highlights how the stringent security requirements of FSRDCs might lead to higher productivity benefits for researchers with the necessary complementary resources, such as funding from their department or external grants.

Awareness of data available in FSRDCs: Even among researchers in proximity to a data center, awareness about their availability and research potential is a crucial factor determining the diffusion of FSRDC data. For instance, administrators of the FSRDC often engage in outreach activities to inform prospective users that "there's all this great data" (interview S12). However, the main determinant of the diffusion of confidential data seems to be exposure to research that uses them because researchers might be inspired after hearing about their colleagues' research. One economist, who was not affiliated with the economics department, told

[1] Source: https://news.stanford.edu/news/2010/february1/census-data-center-020210.html
[2] Source: https://news.uci.edu/2015/new-uci-center-gives-researchers-link-to-us-census-data/

us that many colleagues from her department started using FSRDC data because of exposure to her work (interview S94). Another FSRDC director emphasized the importance of word of mouth in spreading the impact of local FSRDCs, after remarking that sheer co-location to a data center hardly suffices to increase data usage (interview D88).

Diffusion of other confidential databases: A few respondents noticed that submitting a project to the Census Bureau ultimately presented a risk-reward trade-off. In certain cases, the procedural uncertainty around the approval process might not be worth it: "the trade-off goes the other way where you say, well, I could be a little bit better with the RDC data, but do I really wanna wait nine months?" (interview S12). Moreover, an increasingly large number of comparable microdata from foreign countries are available to U.S.-based researchers. One researcher even suggested that the hurdles in using FSRDC data might have spurred the diffusion of foreign microdata (interview F72). In the words of one respondent:

> "I think at the moment there is this issue which people in the RDC community are sort of talking about, which is that the U.S. data is a lot harder to work with than German data or Dutch data or whatever. [...] And there's a sense in the U.S. perhaps that [...] the data is very good and those who are working with the data can do good stuff. But there's a lot of hurdles to get to it and it's not clear that the data is really even necessarily as good or better than some of the European sorts of data. [...] If you have to look at the U.S., then it is the best data available. But if you just want to test the theory and you don't really care whether it's for the U.S. or Germany or France, the U.S. data may not be quite as good." (interview D88)

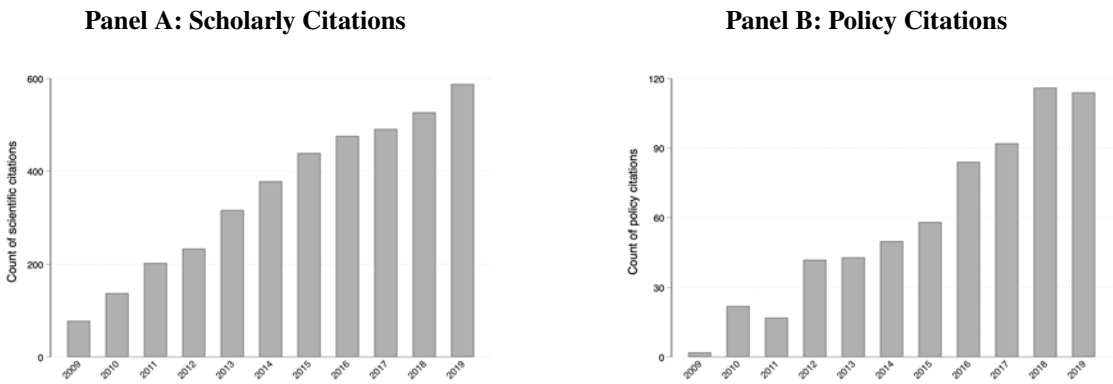
Interestingly, researchers have long feared that the United States might be losing the edge in applied research due to the limited access to administrative microdata (Card et al., 2010). Our interviews suggest that researchers might have pragmatically shifted to greater use of foreign administrative data with lower barriers to access (interviews F72, D88).

A.2 Case Study

In 2009, Chang-Tai Hsieh and Peter Klenow published the already-classic paper "Misallocation and manufacturing TFP in China and India" in the *Quarterly Journal of Economics* (hereafter, "Misallocation paper"). At the end of 2019, the Misallocation paper already had over 3,850 citations on Google Scholar, corresponding to over 352 citations every year. Moreover, the paper proved extremely influential in the policy debate (Figure A1). According to the data collected by Overton, over 640 policy documents have cited this paper up to 2019.

The paper's starting point is that firm heterogeneity and the allocation of resources across firms play a

Figure A1: Impact of the Misallocation Paper by Hsieh and Klenow (2009)



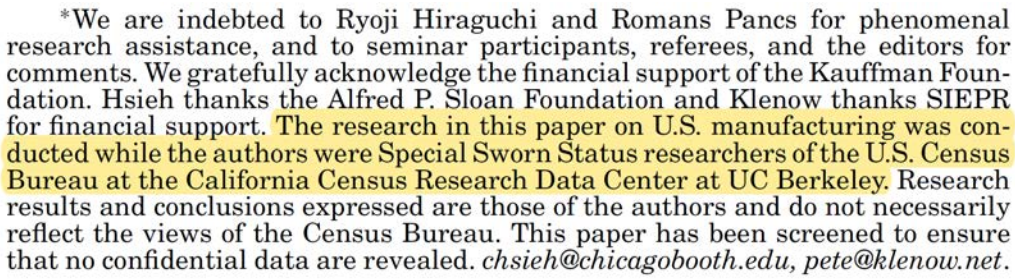
Note: This figure shows the impact of the Misallocation paper by Hsieh and Klenow (2009). Panel A reports the number of scholarly citations received by the Misallocation paper using data from Google Scholar. Panel B reports the number of policy citations received by the Misallocation paper using data from Overton. See text for more details.

crucial role in determining aggregate productivity. If the factors of production are not allocated to their most efficient use, that is, to the most productive firms, then aggregate productivity and welfare are reduced. Hsieh and Klenow (2009) showed that the extent of misallocation could be estimated from firm-level microdata. Their paper first develops a model of monopolistic competition with heterogeneous firms, which is then employed to measure the contribution of resource misallocation to aggregate manufacturing productivity in China and India versus the United States. Their results are striking: aggregate productivity would grow up to 50% in China and up to 60% in India should those countries re-allocate capital and labor similarly to what is observed in the U.S. Since its publication, this paper has become the backbone of a fertile strand of empirical and theoretical research (Hopenhayn, 2014).

The paper by Hsieh and Klenow (2009) was carried out using confidential microdata available only at the U.S. Census Bureau (Figure A2). We gathered additional information about this paper from U.S. Census Bureau's records. The misallocation project started in 2006 at the Berkeley FSRDC, where Hsieh was affiliated at the time. In the abstract of the project submitted to the CES, the authors delineated their objective of developing a new methodology to "help shed light on the underlying sources of productivity differences". To do so, they asked access to confidential U.S. company microdata from the Census of Manufactures (from 1963, 1967, 1972, 1977, 1982, 1987, 1992, 1997, and 2002) and the Annual Survey of Manufacturers (1973-2001). The granularity of the establishment-level data collected by the U.S. Census Bureau was crucial to study the misallocation of production factors across firms. Indeed, Hsieh told us it would not have been possible to do the same paper without FSRDC data (Hsieh, personal communication, 21st December 2020).

We asked both Hsieh and Klenow about the origins of their collaboration and how much the need to physically access FSRDC data shaped their work. Before starting the project, the two authors had been thinking about the potential idea but needed data to develop the methodology and carry out empirical validation. Then, one

Figure A2: Footnote of the Misallocation Paper by Hsieh and Klenow (2009)



*We are indebted to Ryoji Hiraguchi and Romans Pancs for phenomenal research assistance, and to seminar participants, referees, and the editors for comments. We gratefully acknowledge the financial support of the Kauffman Foundation. Hsieh thanks the Alfred P. Sloan Foundation and Klenow thanks SIEPR for financial support. The research in this paper on U.S. manufacturing was conducted while the authors were Special Sworn Status researchers of the U.S. Census Bureau at the California Census Research Data Center at UC Berkeley. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census Bureau. This paper has been screened to ensure that no confidential data are revealed. *chsieh@chicagobooth.edu, pete@klenow.net*.

Note: The figure reports the footnote of the paper by Hsieh and Klenow (2009), highlighting their usage of the Berkeley FSRDC.

day, Hsieh found out about the data available at the Berkeley FSRDC from one graduate student who was using the same data for her dissertations. This detail highlights how local FSRDCs can result in knowledge spillovers that substantially alter research trajectories. Interestingly, even if the Berkeley FSRDC had been in operation since 1999, Hsieh became aware of the potential of FSRDC data for his research only after seeing the same data used by someone else.

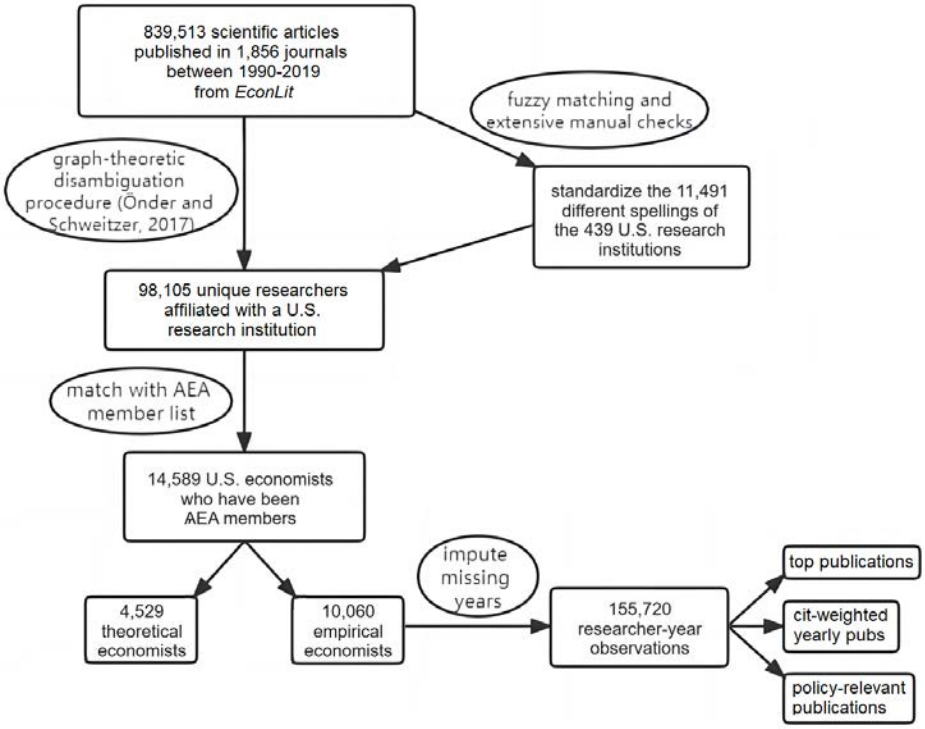
The datawork for the Misallocation paper was supervised mainly by Hsieh, who had easier access to the Berkeley FSRDC. Indeed, the physical distance between Berkeley and Stanford was a formidable barrier to the use of confidential FSRDC data: as noted by Klenow, "if I hadn't had a co-author at Berkeley, I don't think I would have started the Misallocation paper" (Klenow, personal communication, 8th December 2020). This is remarkable given that the campuses of UC Berkeley and Stanford are relatively close by, just about 40 miles apart, which prevented Stanford from obtaining its own FSRDC for many years. In 2010, Stanford was eventually allowed to establish a branch of the California FSRDC on its campus. The new FSRDC branch drastically reduced the geographical barriers faced by Stanford researchers. In the official press release of the opening, Nick Bloom was quoted saying that "having to go to Berkeley was an immense waste of time"[3].

B Additional Data Details

The primary data source we leverage is *EconLit*, a proprietary database of economic scholarship curated by the American Economic Association. Our version of EconLit includes 839,513 scientific articles published in 1,856 journals between 1990 and 2019 inclusive. Unfortunately, EconLit lacks unique author identifiers, preventing us from easily aggregating the publication data into an author-year-level panel. To reconstruct authors' publication records, we need to disambiguate the author names and affiliations, a complex and time-consuming task that we implement in several steps summarized in Figure B1.

[3]Source: https://news.stanford.edu/news/2010/february/census-data-center-020210.html

Figure B1: Steps Followed for Panel Construction



Note: This figure illustrates the steps we took to construct our panel of U.S.-based publishing economists. See text for more details.

B.1 Data Disambiguation

We disambiguate the names of researchers appearing in EconLit following three steps. First, we replace all non-English characters (e.g., "ó" is replaced by "o") and transform the most common name abbreviations into standardized names (e.g., "Ted" is replaced by "Edward"). Second, we apply the disambiguation procedure developed by Önder and Schweitzer (2017). The method employs a graph-theoretic approach that follows a hierarchical process. After identifying the set of name entries with identical surnames, the algorithm constructs a graph of the relationships of all the corresponding first names to each other. Names associated with a particular surname can be identical, different, or subsets of each other. For example, the first name "Michael J." is identical to "Michael J.", it is different from "Tom", and it is a subset of "Michael". The algorithm classifies "Michael J." and "Michael" as the same person if no other "Michael *x.*", with *x.* different from J., appears in the data. This approach is equivalent to assuming that the combination of first, middle, and last names uniquely identify each economist (Card et al., 2022), while at the same time being conservative in assigning ambiguous names that lack a clear middle name. Third, we performed extensive manual checks and corrected several misclassifications due to either misspellings in EconLit (e.g., "Tabellini, Gudio" instead of "Tabellini, Guido") or ambiguous names (e.g., "David Levine", that could refer to David I. Levine or David K. Levine). This three-step procedure results in a database of 434,938 unique researchers from the 552,570 names originally appearing in EconLit.

Next, we standardize the names of the 178,798 affiliations appearing in EconLit. This step is necessary to pin down researchers' treatment status through their affiliation and restrict the sample to U.S.-affiliated economists who are ever at risk of being co-located to an FSRDC. We begin with a list of all research universities in the United States taken from the 2018 Carnegie Classification of Institutions of Higher Education.[4] In particular, we consider all doctoral universities (corresponding to the codes 15, 16, and 17 in the Carnegie Classification), to which we add the main institutions active in economic research (such as the IMF, RAND Corporation, World Bank, RTI, and all the regional FED offices). The result is a list of 439 universities and research centers, which we merge with the EconLit record via fuzzy string matching. This is done by employing string partial ratio similarity to consider the information in the affiliation word ordering (e.g., to distinguish "University of Washington" from "Washington University"). We retain all matches achieving a partial ratio similarity score equal to or greater than 90 over 100, and we manually check them.[5] The result is a list of 11,491 different spellings of the 439 U.S. research institutions appearing in the Carnegie Classification.

B.2 Panel Construction

After the disambiguation of author names and affiliations, we can use bibliographic data to construct an author-year level database with an annual record for each economist who has published at least one paper in the journals included in EconLit (Moed et al., 2013). Out of the 434,938 unique researchers in our disambiguated data, we retain 98,105 scholars affiliated with a U.S. research institution for at least one year in our sample period. To further restrict our sample to academic economists, we match these names to nineteen yearly lists of members of the American Economic Association (AEA). This step allows us to avoid confounding effects arising from the inclusion of researchers working in unrelated disciplines but occasionally publishing in economics outlets. This leaves us with 14,589 publishing economists. Finally, since our main specification considers the effect of access to confidential FSRDC data on the research of empirically-minded economists, we use the machine learning classification procedure developed by Angrist et al. (2020) to classify each researcher in our sample based on their first five publications (see Appendix D for more details). We prefer this approach over a classification based on lifetime count since the latter is less robust to the (anecdotally unlikely) possibility that gaining access to an FSRDC shifts a researcher's methodological orientation. Furthermore, this allows us to still classify researchers who have only worked at an institution with an FSRDC, which would not be possible if the classification was based only on pre-FSRDC research. The final result is an unbalanced panel of 10,060 publishing empirical economists.

We use this set of authors to derive a panel of 155,720 researcher-year observations by imputing missing

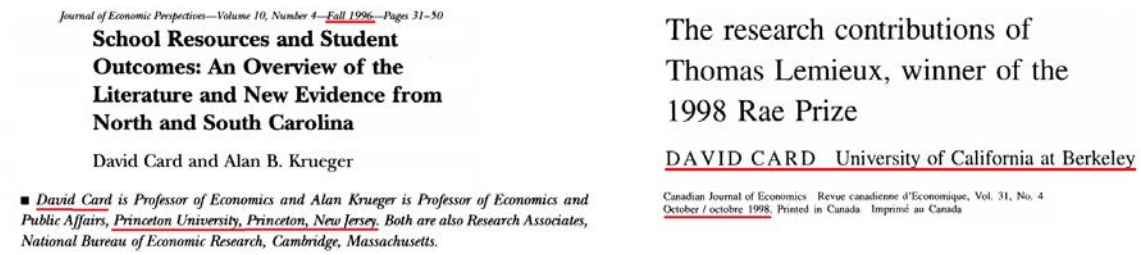
[4]The Carnegie Classification of Institutions of Higher Education is available online at https://carnegieclassifications.iu.edu/.

[5]The rate of false positives for matches achieving a score of 90/100 is around 86%, so we decided to drop affiliations with lower scores and consider them as non-U.S. universities.

years between the first and the last year we see a researcher publishing. In practice, we interpolate missing researcher-year pairs by assigning a zero in the count of scientific output. As noted by Moretti (2021), this interpolation choice leads to partially conflating the extensive margin (i.e., the probability of publishing at least one paper) with the intensive margin (i.e., the number of papers published in a given year, given a positive number of publications). This issue could meaningfully change the interpretation of estimates in contexts where observable outputs are relatively rare, such as in the case of inventors obtaining a patent (Moretti, 2021). However, this is less of a concern in our case because publishing a paper is a more common event than patenting, increasing our panel's granularity and reducing the need for imputations.

Our research design takes advantage of the requirement that confidential microdata can only be analyzed in FSRDC facilities. In practice, we can approximate the cost of data access by the location where the researcher works relative to the closest data center. We use the information on university affiliations in scientific articles to pinpoint scholars' location and mobility over time (see Figure B2 for an example). In cases when the affiliation changes in non-consecutive years with gaps in between, we attribute the old affiliation for the first third of missing years and the new affiliation for the remaining two-thirds of years. Fortunately, EconLit has a broad coverage of economics journals, allowing us to record scientific mobility with high precision, even for less prolific authors.

Figure B2: Example of Researcher Mobility from Bibliographic Records


Journal of Economic Perspectives—Volume 10, Number 4—Fall 1996—Pages 31–50

School Resources and Student Outcomes: An Overview of the Literature and New Evidence from North and South Carolina

David Card and Alan B. Krueger

■ *David Card is Professor of Economics and Alan Krueger is Professor of Economics and Public Affairs, Princeton University, Princeton, New Jersey. Both are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts.*

The research contributions of Thomas Lemieux, winner of the 1998 Rae Prize

DAVID CARD University of California at Berkeley

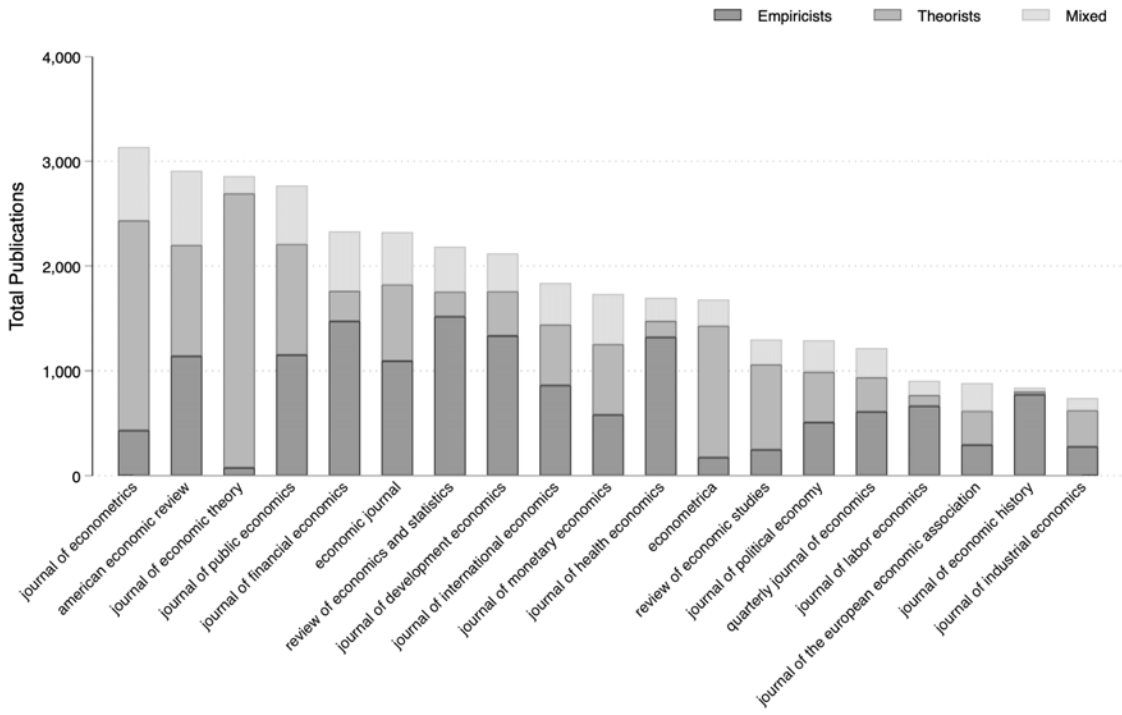
Canadian Journal of Economics Revue canadienne d'Economique, Vol. 31, No. 4
October / octobre 1998. Printed in Canada Imprimé au Canada

Note: This figure shows how bibliographic records can be leveraged to construct a panel of publishing economists. In particular, the affiliation details permit the detection of mobility events over time and co-location to an FSRDC. In this case, we would be imputing the new affiliation for 1997.

We construct productivity metrics at the researcher-year level by counting the number of yearly published articles. We consider as "top publications" all papers appearing in top five and top field journals as defined by Heckman and Moktan (2020). Figure B3 shows the number of articles recorded in EconLit for each main outlet. Next, we augment EconLit by merging each article with its yearly citation count extracted from SSCI/Web of Science. We are able to match 97.2% of our data with the corresponding citation records to construct citation-weighted metrics of academic productivity. We use this information to adjust yearly paper counts by the number of citations each article receives up to five years following publication. We merge the articles in our data with the list of mentions in policy documents assembled by Overton. Detailed description of this data source and summary statistics are presented in Appendix E.

Information on the ranking of economics departments is taken from Kalaitzidakis et al. (2003). We use this information to code seven tiers of universities when estimating models that include a tier-specific time trend. We divide the North American institutions reported by Kalaitzidakis et al. (2003) into five groups with a roughly similar number of observations in our sample. The top tier includes the first five U.S. universities (Harvard, Chicago, MIT, Northwestern, Penn), the second tier the following five (Yale, Princeton, Stanford, UC Berkeley, NYU), the third tier the following ten (from Columbia to Boston U), the fourth tier the following eighteen (from Brown to U Virginia), while the fifth tier includes the remaining ones (from Wash U to Stony Brook). The two residual tiers are constituted, respectively, by researchers affiliated with foreign universities or U.S.-based research institutions not covered by the ranking.

Figure B3: Count of Articles in the Main Economics Journals (1990-2019)



Note: This figure shows the count of papers in the most prestigious economics journals as recorded by EconLit. We decompose the total yearly publication count into publications authored exclusively by empiricists, publications authored exclusively by theorists, and publications with a mixed authorship team. The list of top five and top field journals is from Heckman and Moktan (2020). See text for more details.

B.3 Data Checks

A possible concern of our data construction is that it might lead to measurement errors in our dependent variables if the bibliographic records employed are incomplete. To ensure that EconLit provides good coverage of published research, we randomly drew 15 researchers from our panel and manually searched their publication records. For 14 of them, we could find a PDF version of their CV with a list of their scholarly work. We recorded all the publications listed in their CV as the "ground truth" against which

we assessed the coverage of EconLit. These researchers collectively published eight articles in a top five journal and 21 articles in a top field journal during our sample period. Except for one article appearing in the *Journal of Economic History*, all other papers were correctly recorded in our data, suggesting that EconLit offers reliable coverage.

It is also important to note that we rely on published records to infer researchers' careers and mobility patterns. This could lead to measurement error in our independent variable for researchers who gain employment in a university with a local FSRDC but do not immediately publish articles with the new affiliation. We expect this issue to induce, at most, a downwardly biased coefficient, since it would translate into some researchers-year observations being wrongly coded as "non-treated" while having local access. Our estimates should be more conservative due to this type of measurement error.

However, the downward bias could be significantly larger for junior economists who might need several years to see their first papers published after completing their studies. To assess the plausibility of this concern, we merged our panel with data on researchers' careers from Sarsons et al. (2021). The result is a subset of 349 economists for which we have the precise date of PhD conferral and tenure decision alongside their publications. We find that the mean time from PhD conferral to first publication is 1.42 years (SD=2.88), with the large majority of researchers publishing their first paper within three years of graduation. The mean time from PhD conferral to tenure is 7.13 years (SD=2.65), similar to the estimate of Heckman and Moktan (2020). Overall, this implies that while junior scholars will enter our panel with some delay, the resulting attenuation bias in our estimates should be reasonably small.

C Measuring Diffusion and Impact of FSRDCs

This Appendix provides details on the inception and diffusion of FSRDCs, as well as the measurement of their research impact.

C.1 The Objectives of the FSRDCs Program

The U.S. Census Bureau constantly collects individual and business information via statistical surveys and enumerations as part of its mandate to assemble timely data about the nation's demographic and economic trends. A crucial challenge in achieving its mission involves earning the trust of households and businesses to elicit complete and truthful answers (Foster et al., 2009; Goroff et al., 2018). For this reason, the data collected by the U.S. Census Bureau is tightly regulated by the U.S. Code. The legal framework regulating the acquisition, use, and protection of confidential data is outlined in Title 13.[6] Among its stipulations is

[6]Data from other Federal Agencies are often integrated and merged with Census records. For instance, the LEHD database was built by linking Census data with IRS records. Such data are governed by the confidentiality provisions outlined in Title 26 of the

a requirement that Census Bureau employees take an oath to protect data confidentiality and ensure their protection.

The U.S. Census Bureau effectively faces a dual mandate to publish accurate statistics based on its microdata while preserving their confidentiality (Abowd and Schmutte, 2019). From a legal standpoint, Title 13 authorizes the Census Bureau to give access to its confidential information only as it derives a clear benefit for its data programs (McGuckin et al., 1993). The objective of the FSRDC network is thus to maximize the benefit to Census' programs subject to its confidentiality constraints, and not necessarily maximizing confidential data diffusion (Foster et al., 2009).[7] Title 13 also requires that anyone accessing Census data be sworn to observe their confidentiality for life or face severe penalties (including fines of up to $250,000 and up to five years of federal prison), precisely as if they were Census employees. This is why external researchers are required to undergo background checks and face an application procedure that necessarily involves a certain degree of bureaucratic hurdles (and several months).

Therefore, it must be noted that the intention of this study is not to provide a complete policy evaluation of the FSRDC network. First, we would need to expand our focus to other disciplines that benefit from FSRDC data, including demography and especially health policy, which are outside of the scope of our dataset. Second, we would need to consider all benefits occurring to the Census' data programs beyond those encapsulated in peer-reviewed publications. Significant examples include improvements to existing datasets, refining survey questionnaires, or even designing de novo surveys (Bloom et al., 2019; Chow et al., 2021). Both these considerations suggest that our analysis can be seen as a likely lower bound of the actual value generated by the FSRDC program to the Census Bureau and society.

Nonetheless, our analysis is important in providing the first data-driven assessment of FSRDCs' success in at least one of their objectives: fostering the diffusion and use of their confidential microdata for research purposes. While the optimal level of data access is ultimately a matter of social preferences (Abowd and Schmutte, 2019), gauging the risks of privacy losses is usually easier than assessing the benefits from data access (Lane, 2021). We contribute to this important debate by quantitatively documenting the significant benefits to research and policy-making that confidential data access entails, beyond the mostly anecdotal evidence existing to date (Bowen, 2024; Card et al., 2010; Einav and Levin, 2014b; Lane, 2007).

U.S. Code. For most other statistical data, the legal framework is given by the Confidential Information Protection and Statistical Efficiency Act.

[7]We thank an anonymous referee for underscoring this point and prompting us to add this important qualification.

C.2 Access to FSRDCs Over Time

Table C1 shows the list of the 31 Federal Statistical Research Data Centers opened until 2019.[8] The first was opened in the Census Bureau's Boston Regional Office (Atrostic, 2007). After that, Carnegie Mellon University and the CES pioneered a new institutional model where the data center would have been located and operated by a research institution, with the Census Bureau keeping an oversight role. This institutional arrangement became the standard model followed by all subsequent FSRDCs. Interestingly enough, the FSRDC at CMU is also the only one that later closed because of low usage.[9] This episode is consistent with the evidence from our interviews: FSRDCs usually open thanks to the leadership of a small number of researchers, but sometimes this might not translate into broader usage patterns. The closure of the CMU FSRDC prompted a change in the model in which the network is operated: most FSRDCs are now sponsored by consortia of local institutions partnering up to split the costs of running the data center. Such institutions include universities, regional branches of the Federal Reserve System, and other research centers active in the social sciences, such as the RAND Corporation or the Russell Sage Foundation.

Our primary analysis explores the dynamics of scientific productivity after gaining access to an FSRDC. We define all empirical researchers working at a university with an active FSRDC as "treated". This is consistent with evidence that confidential data diffuse only in the vicinity of the data enclave (Appendix Figure H8). Nevertheless, we also document that our results are robust when we exclude institutions hosting the data center on their premises and instead consider researchers in non-hosting institutions that are very close-by as treated (Appendix Table H10). Appendix Table H11 further shows the robustness of our main results with alternative definitions of the treatment, such as being affiliated with an institution in the same city as an FSRDC or with an FSRDC consortium member.

Importantly, according to our definition, researchers can gain access to FSRDC data in two ways: when a new data center opens at their current affiliation or when they change employer and move to a new institution with an active FSRDC. Figure C1 provides a breakdown of the researchers in our sample by the modality in which they received access (see also Appendix Figure H2). One might worry that the individuals who advocated opening a local FSRDC might be interested in using it, potentially inflating our estimates. Appendix Table H9 shows that our results are robust to excluding the researchers who appeared in the NSF grants that led to establishing the data centers. An alternative concern is that researchers who gain access by moving to an institution with an active data center might do so precisely because they want to work on confidential

[8] In 2019, the Census Bureau began to provide remote virtual access to a select number of FSRDC researchers working only with Title 13 data. The pilot has been scaled up during the COVID-19 pandemic, resulting in 83 approved projects by mid-2021. However, the possibility of remote access is not extended to data originating from the IRS – which includes several of the most popular databases such as LBD and LEHD. More details on these exciting developments are available at: https://www.census.gov/about/adrm/fsrdc/about/secure-remote-access.html.

[9] The other exception is the RTI International FSRDC, which was closed as part of a reorganization of the Triangle Research Data Center with Duke University and the University of North Carolina at Chapel Hill.

Table C1: Federal Statistical Research Data Centers Opened during our Sample Period

Year opened	Institution of location	City	State
1994	NBER Cambridge	Boston	Massachussets
1997*	Carnegie Mellon U	Pittsburgh	Pennsylvania
1999	UC Berkeley	Berkeley	California
1999	UCLA	Los Angeles	California
2000	Duke	Durham	North Carolina
2002	U Michigan	Ann Arbor	Michigan
2003	Chicago FED	Chicago	Illinois
2004	Cornell	Ithaca	New York
2006	Baruch College-CUNY	New York	New York
2010*	RTI International	Research Triangle Park	North Carolina
2010	U Minnesota	Minneapolis	Minnesota
2010	Stanford	Stanford	California
2011	Atlanta FED	Atlanta	Georgia
2012	Texas A&M U	College Station	Texas
2012	U Washington	Seattle	Washington
2014	UC Irvine	Irvine	California
2014	USC	Los Angeles	California
2014	Penn State	University Park	Pennsylvania
2015	U Missouri	Columbia	Missouri
2015	U Nebraska-Lincoln	Lincoln	Nebraska
2015	UW Madison	Madison	Wisconsin
2015	Yale	New Haven	Connecticut
2016	U Maryland	College Park	Maryland
2016	FED Kansas City	Kansas City	Missouri
2017	UT Austin	Austin	Texas
2017	U Colorado Boulder	Boulder	Colorado
2017	U Kentucky	Lexington	Kentucky
2017	Philadelphia FED	Philadelphia	Pennsylvania
2017	Georgetown U	Washington	DC
2018	Dallas FED	Dallas	Texas
2019	Federal Reserve Board	Washington	DC

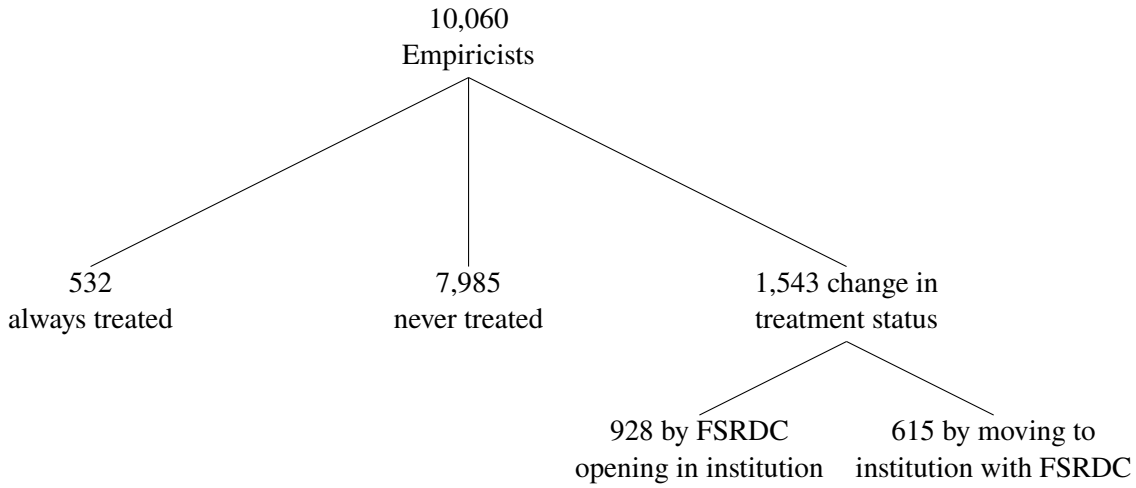
Note: The FSRDC at Carnegie Mellon University closed in 2004, while the FSRDC at RTI International closed in 2018. Additional FSRDCs opened at the University of Utah (2020), University of Illinois Urbana-Champaign (2020), University of Florida (2022), and UNC-Chapel Hill (2022) are outside of our sampling period. To the best of our knowledge, the year of opening reported refers to the date when the FSRDC became operative, as it could be inferred from CES Yearly Reports and university press releases.

data. This endogenous geographical sorting could bias our estimates upward. To rule out this concern, we repeated the primary analyses, excluding all scholars who had become co-located to a data center because they had changed affiliations. All the main results are robust to this test (Appendix Table H13). In addition, qualitative evidence from our interviews suggests that this concern is unlikely to hold in practice. While a local FSRDC could, in principle, convince a scholar to accept a job offer, this is usually true only for previous users whose research agenda crucially hinges on having an FSRDC nearby (interview S94).

C.3 Articles using FSRDC Data

We assembled a novel dataset of all articles that *directly* employ restricted-access microdata accessible only in U.S. Census' secure facilities. When we started our project, there was no official bibliographic record of research using FSRDC data, so we constructed one by carefully searching the bibliome using several complementary strategies.

Figure C1: U.S.-Based Economists by Modality of FSRDC Access



Note: This figure summarizes the treatment status of all empirical economists included in our Panel. See text for more details.

As a starting point, we exploited the fact that papers using U.S. Census confidential data are expected to indicate it clearly in the acknowledgment of the published version of the paper. We started by collecting all the most commonly used sentences appearing in the acknowledgments of a sample of FSRDC papers (such as "Census Research Data Center", "do not reflect the views of the Census Bureau", and "Special Sworn Status researchers"). Then we searched for them in the main databases of published research, Web of Science and Scopus, which recently started collecting the acknowledgment sections of journal articles. However, we found that many papers do not report the standard disclaimers required by the Census Bureau. We tried to overcome this limitation with additional searches in databases that allow full-text searches, such as JSTOR,[10] Google Scholar and the NBER working paper repository.

We further expanded our search by exploiting the fact that projects approved by the FSRDC are expected to submit a final working paper to the CES for online publication.[11] We collected the metadata of 1,081 working papers and matched them to EconLit through a combination of fuzzy title matching and extensive manual checks. Overall, just about half of these papers were ever published, and only 455 papers could be linked to the corresponding EconLit record.[12] The high share of CES working papers that never get published suggests that output-based assessments of FSRDCs will understate the actual use of the network by academics (see Section C.4 for some alternative approaches).

Finally, we requested access to all approved FSRDC projects via the Freedom of Information Act (FOIA) request to the Census Bureau (FOIA ID No. DOC-CEN-2020-001640). Notably, the results of our request

[10]The main limitations of JSTOR are that its coverage is unreliable for more recent years and that it does not encompass journals published by Elsevier.

[11]The papers are available online at the following link: https://ideas.repec.org/s/cen/wpaper.html.

[12]Published papers that do not appear in EconLit have either appeared in journals not covered (e.g., the *Strategic Management Journal*) or as a book chapter (mostly in NBER-edited books).

have graciously been published online and are being updated regularly for everyone interested in tracking the use of FSRDC data.[13] We manually searched the publication records of each researcher whose projects were approved to be carried out in an FSRDC. In total, we were able to find a total of 861 papers published in peer-reviewed journals that could be matched with EconLit. Once we restrict our sample to U.S. researchers affiliated with the AEA, the final sample of papers employing FSRDC data consists of 587 articles written by 509 economics researchers.[14] A more detailed exploration of the characteristics of these papers is provided in Nagaraj et al. (2024).

C.4 Other Measures of FSRDC Impact

One potential shortcoming of our measures of FSRDC impact is that they are based on research in peer-reviewed economics journals. However, we might be understating the actual diffusion of FSRDC data if many projects are not published. This would be a concern if papers using FSRDC have a lower propensity to be published, perhaps because they constitute riskier research with a wider variance in scientific quality. To sidestep this type of concern, we carry out robustness tests where the measure of data adoption is given by whether a scholar has a project approved to be carried out in an FSRDC. We obtained this information thanks to our FOIA request, which resulted in a list with details about all projects approved by the U.S. Census Bureau. We supplement these data with similar records from the other agencies participating in the FSRDC program, all obtained with additional FOIA requests.[15] These data give us an upstream measure of data usage that is not dependent on the project's outcomes.

Moreover, projects using Census confidential data must be submitted as working papers to the CES once completed. We record the count of CES working papers up to 2019 inclusive as an additional outcome variable to measure confidential data usage encompassing articles not published in peer-reviewed journals. Interestingly, only 48.4% of CES working papers were published as of June 2020. This figure is close to recent estimates that around 50-74% of NBER working papers are eventually published (Baumann and Wohlrabe, 2020; Lusher et al., 2021). Most CES working papers appeared in a peer-reviewed journal included in EconLit, with just 13% of them appearing as book chapters or in journals of other disciplines. Again, this number is very close to the finding that 12% of published NBER working papers do not appear in economics journals (Lusher et al., 2021).

[13] https://www.census.gov/about/adrm/fsrdc/about/ongoing-projects.html

[14] During our data construction, we also found several articles from other disciplines, such as sociology, demography, and especially health policy (Foster et al., 2009). We excluded them because they are outside the scope of this analysis.

[15] We obtained data from the AHRQ (FOIA ID No. 2021-00311-FOIA-PHS), the BEA (FOIA ID No. DOC-BEA-2021-000222), the BLS (FOIA ID No. 2021-F-00826) and the NCHS (FOIA ID No. 21-00102-FOIA).

C.5 Spillovers from Local Access to FSRDCs

A. Articles citing FSRDC papers: Besides using confidential data available only in an FSRDC, we aim to capture how access to confidential microdata might also affect research that does not directly use FSRDC data. Local access to FSRDCs might be shaping the topics or questions researchers decide to work on, for instance, by increasing awareness of past research done with such data or being exposed to the findings of colleagues working in the FSRDC. We measure this type of influence by recording which articles build on the findings of the papers written using FSRDC data. This is done by taking the list of papers using FSRDC data (identified with the procedure of Appendix Section C.3) and downloading their references from the Social Sciences Citation Index (SSCI) curated by Web of Science. We exclude papers directly using confidential microdata from the count since they are likely to mechanically cite other FSRDC papers (for instance, in their data section). The result is a set of papers that build directly on the results made possible by the FSRDC network, even if they do not use its confidential microdata.

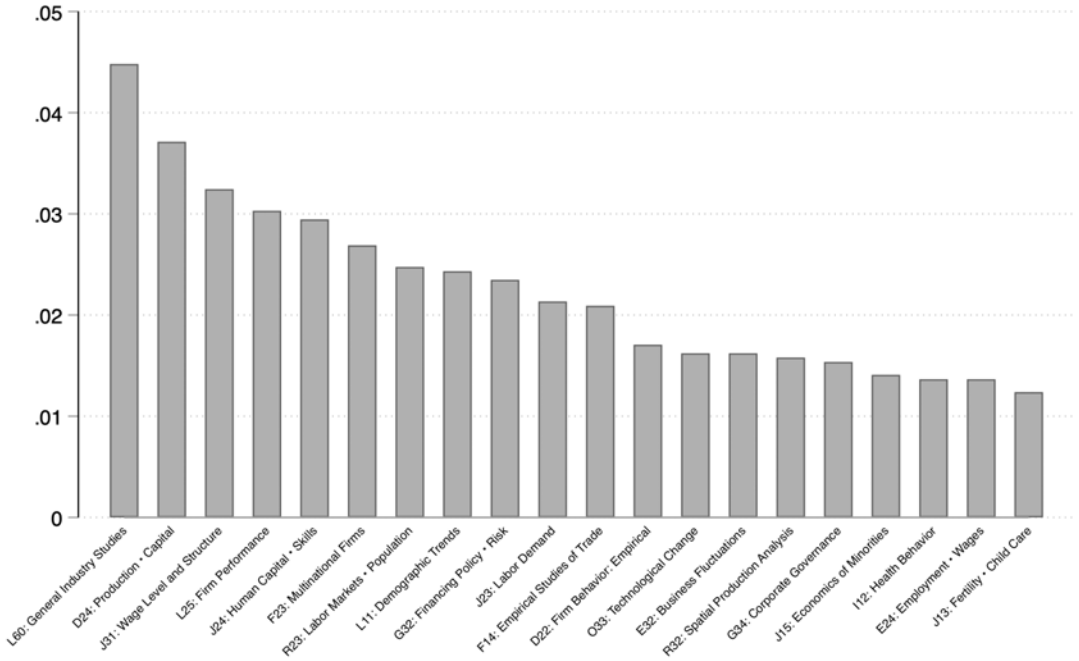
B. JEL codes: Another approach to measuring spillovers from FSRDC access is to consider its impact on publications that pertain to topics commonly associated with using FSRDC data. Researchers working in areas such as labor, productivity studies, trade, and environmental economics should benefit more from research exposure based on FSRDC data. We capture similarity in research topics using the paper-level JEL codes recorded in EconLit. This is conceptually similar to the approach of Azoulay et al. (2019), who leverage the keyword assigned to all life-science publications by the National Library of Medicine to define proximate research in topical space.

We use the list of FSRDC papers to find the JEL codes most likely to be associated with FSRDC data. Figure C2 displays in Panel (i) the 20 JEL codes with the greatest frequency in papers written in an FSRDC. The figure confirms anecdotal evidence that underscores the considerable potential of these data for firm-level analyses. Other common topics include studying labor markets, demographic trends, and international trade. Similarly, Panel (ii) of Figure C2 shows the 20 JEL codes that appear with the greatest frequency in research *not* conducted in FSRDCs. These are more likely to encompass topics like financial markets, development, monetary economics, and theoretical investigations of market efficiency and uncertainty. Appendix Table H15 shows the robustness of our main results to alternative definitions of which JEL codes are the most representative of research enabled by access to FSRDCs.

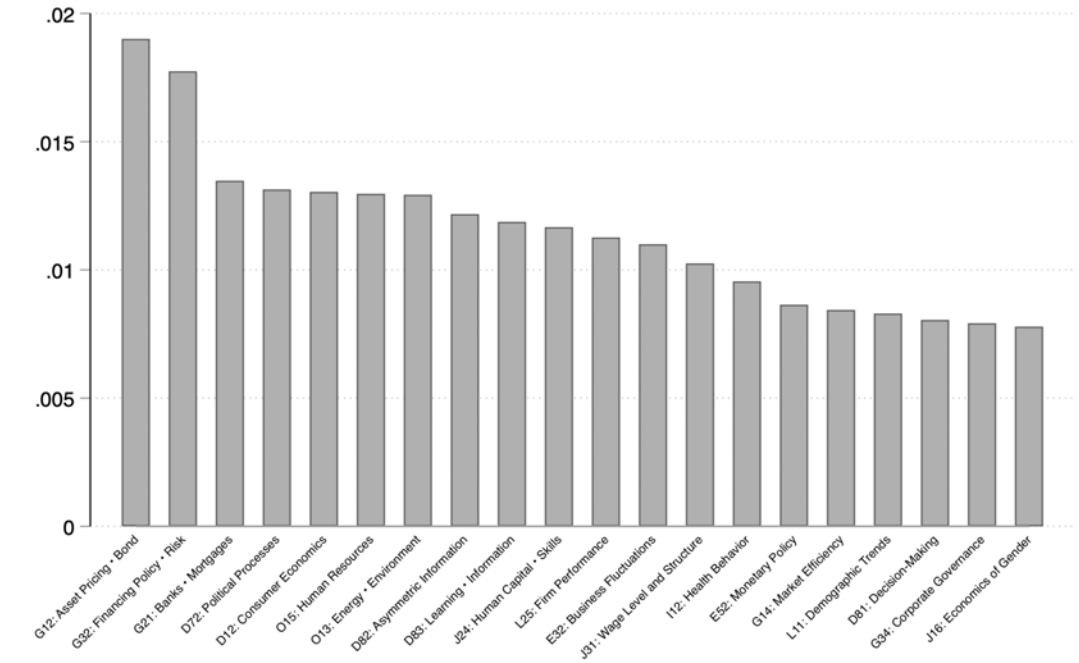
C. Text-based measures of research design: Finally, we compile data to assess potential spillovers on the adoption of empirical methods and administrative data from other sources. Following Currie et al. (2020), we use a series of regular expression searches to find keywords in the titles and abstracts that identify the use of a certain method (e.g., difference-in-differences) or type of data (e.g., survey data). The list of

Figure C2: List of JEL Codes Associated to Papers using FSRDC Data

(i) JELs most representative of FSRDC research



(ii) JELs most representative of non-FSRDC research



Note: This figure shows the JEL codes most distinctive of research carried out with and without confidential microdata available only in FSRDCs. Each bar represents the share of papers that include that JEL code for each of the two groups of papers. Panel (i) reports the 20 JEL codes with the greatest frequency in papers written in an FSRDC. Panel (ii) reports the 20 JEL codes with the greatest frequency in papers not written in an FSRDC. JEL codes are sorted from largest to smallest difference in frequency. See text for more details.

Table C2: Keywords Used to Tag Articles' Research Design and Data

Category	Keywords (inspired by Currie et al., 2020)
Administrative data	"administrative data" "administrative record" "administrative evidence" "administrative earning" "administrative sample" "administrative report" "administrative panel" "administrative microdata" "administrative regist" "registry data" "register data" "archival" "archives" "tax record" "payroll record" "social security record" "wealth record" "voting record" "earnings record" "personnel record"
Survey data	"survey data" "survey panel data" "survey result" "survey response" "longitudinal survey" "preference survey" "contingent valuation survey" "labor force survey" "values survey" "consumption survey" "population survey" "happiness survey" "national survey" "phone survey" "survey administered" "household survey" "survey evidence"
Quasi-experimental methods	"difference in diff" "differenceindiff" "differences in diff" "differencesindiff" "d-in-d" "diff in diff" "diffindiff" "event stud" "eventstud" "staggered adoption" "regression discontinuit" "regressiondiscontinuit" "regression kink" "regressionkink" "rd resign" "rdresign" "rd estimat" "rdestimat" "rd model" "rdmodel" "rd regression" "rdregression" "rd coefficient" "rdcoefficient" "rk design" "rkdesign" "rdd" "rkd" "instrumental variable" "instrumentalvariable" "two stage least squares" "twostage least squares" "2sls" "tsls" "valid instrument" "exogenous instrument" "iv estimat" "ivestimat" "iv regression" "ivregression" "iv strateg" "ivstrateg" "we instrument" "i instrument" "exclusion restriction" "paper instruments" "weak first stage" "simulated instrument"
Experimental methods	"randomized controlled trial" "randomized field experiment" "randomized controlled experiment" "randomised controlled trial" "randomised control trial" "randomised field experiment" "randomised controlled experiment" "social experiment" "rct" "laboratory experiment" "lab experiment" "public good game" "public goods game" "ztree" "orsee" "showup fee" "laboratory participant" "lab participant" "pre analysis plan" "preanalysis plan" "pre registered" "preregistered" "preregistration" "dictator game" "ultimatum game" "trust game"

Note: This table presents the keywords used to tag each paper as belonging to one of the following non-mutually exclusive categories: administrative data, survey data, quasi-experimental methods, and experimental methods. The list of keywords is built on Currie et al. (2020).

n-grams we use is presented in Table C2 and builds on the list in Appendix A of Currie et al. (2020). These keywords allow us to tag each paper as belonging to one of the following non-mutually exclusive categories: administrative data, survey data, quasi-experimental methods, and experimental methods.

Table C3 explores what keywords are more likely to be associated with papers using FSRDC data. Confirming our priors, these papers are five times more likely to mention the use of administrative records, while the same is not true for words denoting the use of traditional surveys compiled for research purposes. Further, Table C3 matches our intuition that research carried out in an FSRDC does not employ experimental research design (such as RCTs or behavioral laboratory experiments). Including year fixed effects helps alleviate the concern that these findings are just the reflection of a wider "credibility revolution" in the years of FSRDC activity.

Table C3: Characteristics of Papers Using FSRDC Data

	Administrative Data	Survey Data	Quasi-Exp. Methods	Experimental Methods
	(1)	(2)	(3)	(4)
FSRDC Paper	0.0364***	0.00589	0.0139	-0.0131***
	(0.00811)	(0.00351)	(0.00850)	(0.00041)
Year FE	Yes	Yes	Yes	Yes
N	188181	188181	188181	188181
Mean of DV	0.00644	0.00208	0.0310	0.0112

Note: This table presents estimates from OLS models evaluating the likelihood that papers using FSRDC data mention the keywords listed in Table C2. *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Robust standard errors are reported in parentheses.

D Classification of Research by Methodological Orientation

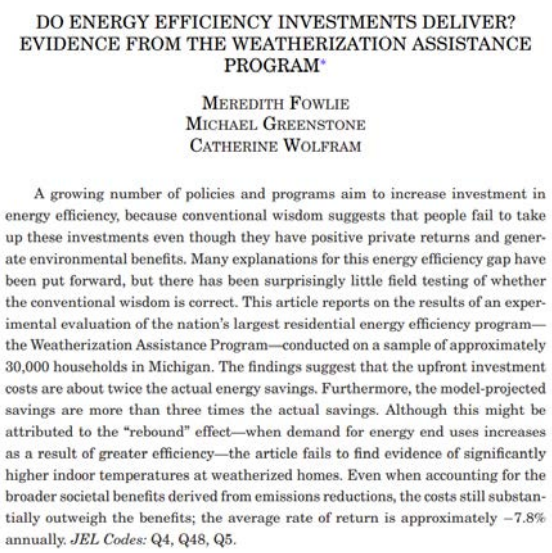
In this Appendix, we explain the machine learning procedure used to classify the methodological orientation of each EconLit record. We then discuss how we infer the specialization of each researcher from her published work and explore the robustness of our results to this classification.

D.1 Classification Algorithm

We employ the machine learning classification procedure developed by Angrist et al. (2020) to classify each economics paper as either empirical or theoretical. The algorithm aims to separate research that produces data-based estimates of economically meaningful parameters from purely theoretical or methodological papers. We also classify papers that develop new methods or models but apply them to produce substantively meaningful estimates as empirical. The classifier developed by Angrist et al. (2020) is a logistic ridge regression that fits a dummy variable indicating empirical papers using article titles, journals, JEL codes, publication years, and abstracts as inputs. The algorithm is trained on a set of 5,469 hand-classified papers and has an 87% accuracy for entries with the abstract, so we scraped the internet to fetch the abstracts of all articles that do not have it reported in EconLit.

Figure D1: Example of Article Classification using Machine Learning

Panel A: Empirical Article



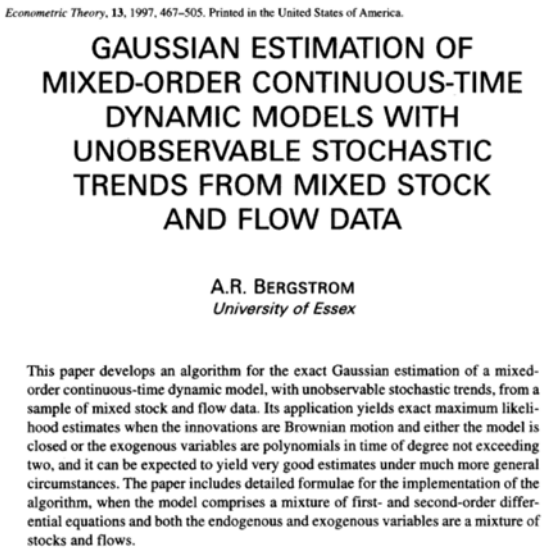
DO ENERGY EFFICIENCY INVESTMENTS DELIVER? EVIDENCE FROM THE WEATHERIZATION ASSISTANCE PROGRAM*

MEREDITH FOWLIE
MICHAEL GREENSTONE
CATHERINE WOLFRAM

A growing number of policies and programs aim to increase investment in energy efficiency, because conventional wisdom suggests that people fail to take up these investments even though they have positive private returns and generate environmental benefits. Many explanations for this energy efficiency gap have been put forward, but there has been surprisingly little field testing of whether the conventional wisdom is correct. This article reports on the results of an experimental evaluation of the nation's largest residential energy efficiency program—the Weatherization Assistance Program—conducted on a sample of approximately 30,000 households in Michigan. The findings suggest that the upfront investment costs are about twice the actual energy savings. Furthermore, the model-projected savings are more than three times the actual savings. Although this might be attributed to the "rebound" effect—when demand for energy end uses increases as a result of greater efficiency—the article fails to find evidence of significantly higher indoor temperatures at weatherized homes. Even when accounting for the broader societal benefits derived from emissions reductions, the costs still substantially outweigh the benefits; the average rate of return is approximately −7.8% annually. *JEL Codes:* Q4, Q48, Q5.

P(Empirical) = 96.9%

Panel B: Theoretical Article



Econometric Theory, 13, 1997, 467–505. Printed in the United States of America.

GAUSSIAN ESTIMATION OF MIXED-ORDER CONTINUOUS-TIME DYNAMIC MODELS WITH UNOBSERVABLE STOCHASTIC TRENDS FROM MIXED STOCK AND FLOW DATA

A.R. BERGSTROM
University of Essex

This paper develops an algorithm for the exact Gaussian estimation of a mixed-order continuous-time dynamic model, with unobservable stochastic trends, from a sample of mixed stock and flow data. Its application yields exact maximum likelihood estimates when the innovations are Brownian motion and either the model is closed or the exogenous variables are polynomials in time of degree not exceeding two, and it can be expected to yield very good estimates under much more general circumstances. The paper includes detailed formulae for the implementation of the algorithm, when the model comprises a mixture of first- and second-order differential equations and both the endogenous and exogenous variables are a mixture of stocks and flows.

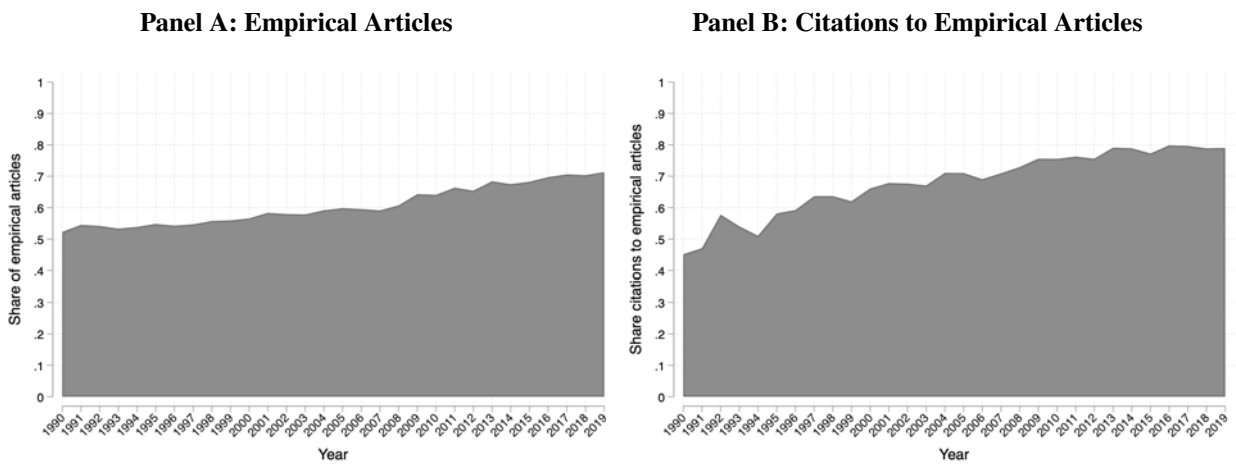
P(Empirical) = 16.4%

Note: This figure shows the results from the machine learning classification algorithm of Angrist et al. (2020) for two papers in our sample. See text for more details.

This procedure results in a paper-level score (between 0 and 1) that captures the probability that an article is empirical. Figure D1 shows an example of two articles and the results from the machine learning classification algorithm. The algorithm also performs remarkably in cases where theoretical papers mention

keywords usually associated with empirical papers, such as "data" or "estimates" in their abstract or title. We classify as empirical all papers with a predicted score larger than 0.5.[16] Our data show an increasing share of empirical publications over time, from 52.3% in 1990 to 71.3% in 2019 (Figure D2). Relative to the findings of Angrist et al. (2020), we document a slightly larger share of empirical research. Possible reasons for this discrepancy are the better coverage of abstracts in our data (that allows for a more precise classification) and the fact that we consider a broader set of journals in our analyses.

Figure D2: Publications and Citations to Economics Papers by Methodological Orientation



Note: This figure shows the share of empirical articles (Panel A) and citations received by empirical articles (Panel B) over time. See text for more details.

D.2 Robustness Checks

We define as empiricist all researchers with more than half of their first five publications classified as empirical. This measure has the advantage of being available for every researcher in our sample of publishing economists. In total, we classify as empiricists 64% of U.S.-based economists in our sample. Figure D3 shows the distribution of the share of empirical articles for the researchers in our sample.

We collected additional data to show the face validity of our classification. First, we adopted a case-control approach and checked our classification results for the Editorial Boards of some high-profile economics journals. In particular, we collected data for five journals that publish a broad spectrum of work, from mostly empirical (e.g., *AEJ: Economic Policy*) to mostly theoretical (e.g., *Journal of Economic Theory*). Table D.2 presents the total number of Editors reported on the website of each of these journals as of November 2022. We matched these Editors with our sample of U.S.-affiliated economists and checked how many of them our procedure classified as applied researchers. Confirming our priors, our procedure classified 96% of *Journal of Economic Theory*'s Editors as theoretical researchers and most components of the Editorial Boards of *AEJ: Applied Economics* and *AEJ: Economic Policy* as empirical researchers. Other

[16] Unlike Angrist et al. (2020), we do not separately classify econometrics articles into an ad-hoc category.

Table D1: Methodological Classification for the Editorial Board of Selected Economics Journals

Journal	Number of Editors	Ever U.S.-affiliated	Share Empirical
Journal of Economic Theory	62	47	4.26%
AEJ: Microeconomics	21	19	26.32%
AEJ: Macroeconomics	19	18	61.11%
AEJ: Economic Policy	37	35	94.29%
AEJ: Applied Economics	36	33	96.97%

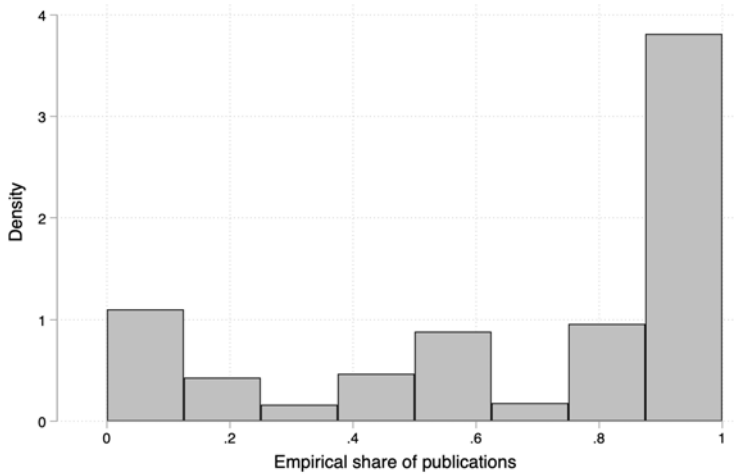
Note: The Table reports the composition of the Editorial Boards of five economics journals (as of November 2022) and the share of them that our machine learning procedure classified as empirical researchers.

journals encompass a greater variety of expertise, with *AEJ: Microeconomics* tilted towards theory-minded Editors and *AEJ: Macroeconomics* showing a preponderance of empirical Editors.

Second, we digitized the lists of North American Ph.D. graduates published yearly in the last issue of the *Journal of Economic Literature*. Figure D4 shows the topics of the doctoral dissertation of the 5,209 researchers we could match to our data, divided by whether we classified the researcher as an empiricist or a theorist. The figure shows that theorists are most likely to have written their doctoral thesis in Microeconomics, Macroeconomics, or Mathematical Methods. At the same time, the most popular fields for empiricists are Labor, Health, and Development Economics.

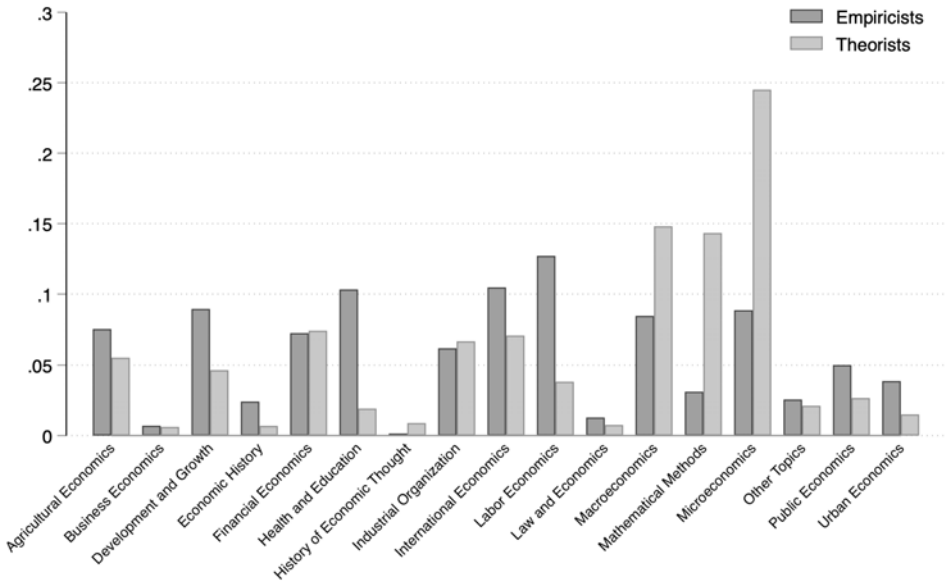
We also performed robustness checks to show the stability of our results to different thresholds in defining empirical researchers. In particular, we ran our main regressions with increasingly stringent definitions of empirical researchers. Figure H7 plots the coefficients of these additional regressions. Our results are driven by the most empirically-minded researchers in our sample, with the magnitude of the coefficients generally

Figure D3: Empirical Articles as a Share of First Five Published Works for Researchers in our Sample



Note: The figure reports the distribution of the share of empirical articles from the first five publications for researchers in our sample. Our main analyses classified scholars as empiricists if their share > 0.5. See text for more details.

Figure D4: Distribution of PhD Dissertation Fields for Researchers Classified as Empiricists or as Theorists



Note: The figure reports the distribution of dissertation topics for each researcher in our sample that we could match to the *Journal of Economic Literature* (JEL) yearly lists of graduates. We matched 4,903 economists, of which 3,438 were classified as empiricists and 1,1465 as theorists. See text for more details.

growing when using more stringent thresholds to define empiricists.

E Measuring the Policy Impact of Academic Research

E.1 Overton Data

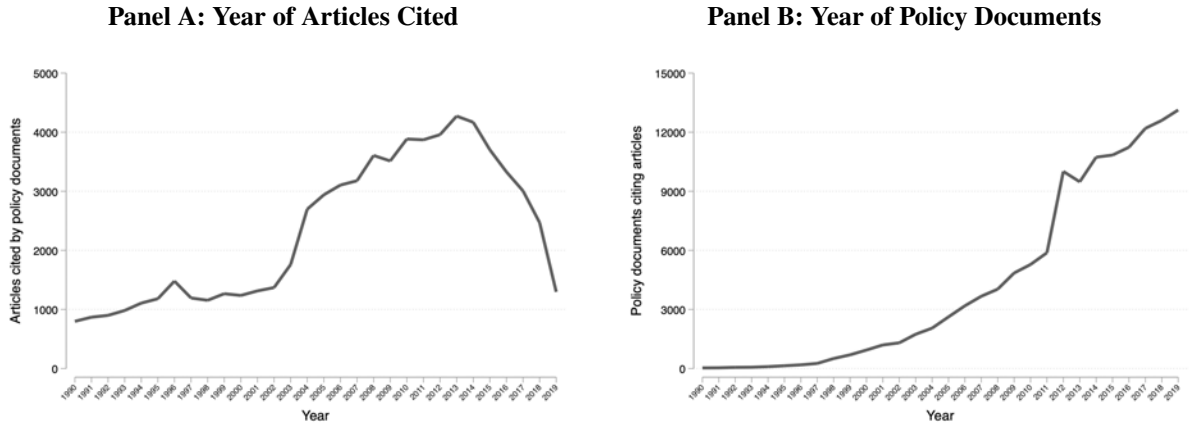
Citations from policy documents (or "policy citations" for brevity) are helpful in recording the influence of academic research in sectors outside the ivory tower. Yet, the use of this type of bibliometric data in academic studies is still rare, with a few notable exceptions (Yin et al., 2022). For a long time, the major obstacle has been the paucity of databases that reliably record citations from unstructured policy documents and merge them with unique identifiers for scientific articles. More recently, the increasing availability of open data on the scientific bibliome and the possibility of large-scale scraping of policy sources from the internet has enabled the first systematic collections of policy citations.

In this paper, we measure the policy impact of economics articles using data from Overton, a company that maintains a large database of policy documents linked to the academic research they cite. In particular, we leverage the fact that Overton indexes over 10 million policy documents from 1,700 sources in over 180 countries.[17] The company parses data within policy documents going as back as the 1920s, thus capturing impact that might take several years to materialize. The type of documents tracked range from white papers

[17]Source: https://help.overton.io/article/what-is-overtons-coverage-and-how-does-it-compare-to-other-systems/

of international development organizations to parliamentary transcripts and think tanks' reports. To the best of our knowledge, very few papers have used similar kinds of policy document citations to date (Haunschild and Bornmann, 2017).

Figure E1: Year of Publication of Articles Receiving Policy Citations and of the Citing Policy Documents in the Overton Data



Note: This figure shows the publication year of the economics articles that received at least one policy citation (Panel A) and the publication year of the policy sources that cite them (Panel B). See text for more details.

We merge our publication-level dataset to the Overton data using the articles' DOI. In total, we find 750,842 citations from 129,042 policy documents to 70,895 economics articles (38% of our sample).[18] Figure E1 shows the distribution of the publication year of the articles cited (left panel) and the documents citing them (right panel). Starting with the latter, one notices that the coverage of Overton is skewed towards recent policy sources. This is likely to explain why the left panel shows that scientific papers published after the mid-2000s have a much higher likelihood of being cited by policy documents. However, insofar as the recording of policy references is not systematically biased within the year of article publication, the inclusion of year fixed effects should account for the time-varying propensity of receiving a policy cite.

The largest single country of origin of policy documents is the United States, which alone makes up almost 20% of all policy citations (Figure E2). However, an even larger share of policy documents comes from intergovernmental organizations (IGOs) well-known for their policy advocacy, which altogether account for almost 40% of all the policy citations received by the economics articles in our sample (Figure E2). Figure E3 shows that articles appearing in more prestigious journals (e.g., the *Quarterly Journal of Economics* or the *AEJ: Applied Economics*) or in more policy-oriented outlets (e.g., the *World Bank Research Observer* or the *IMF Economic Review*) have a higher share of articles that receive policy cites.

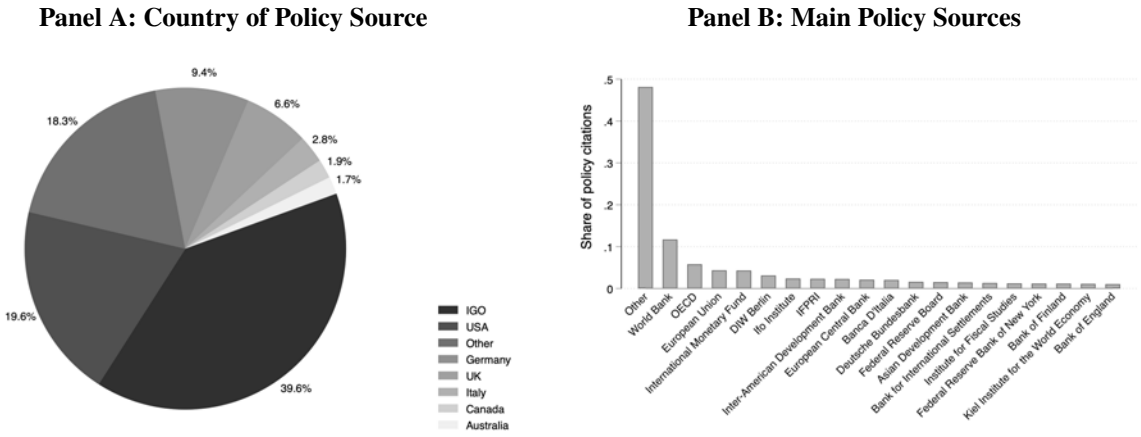
We empirically explore the policy impact of applied work compared to theoretical economics articles. Table E1 shows that empirical articles receive a much larger number of cites from policy documents, an effect

[18] We exclude from the Overton data citations from NBER, CEPR, and IZA sources since they are the working paper versions of academic articles and not policy documents.

proportionally larger when considering only U.S. sources. This effect is robust to the inclusion of journal and year fixed effects, thus effectively isolating the higher policy relevance of empirical research. We find that papers mentioning the use of administrative data in their abstract receive, on average, twice as many policy citations (Columns 3 and 4 of Table E1). Interestingly, we find that the effect increases substantially in the case of papers that employ confidential data available only in FSRDCs: the average number of policy citations from U.S. sources grows by more than five times. The fact that the increase of citations from outside the United States is proportionately much smaller fits well the intuition that evidence deriving from FSRDC data should be particularly salient to inform policy-making in the U.S. (Einav and Levin, 2014b).

Figure E4 shows the considerable heterogeneity in the consumption of economic research by field: papers in development, labor, and urban tend to be very influential in the policy debate, unlike fields such as history or microeconomic theory. Finally, the Pearson correlation coefficient between the count of citations from policy sources and academic papers is 0.53, a level of association consistent with past work (Yin et al., 2022). In unreported analyses, we confirm that all the results of this Appendix are robust to explicitly controlling

Figure E2: Origin of Policy Citations to Economic Scholarship in the Overton Data



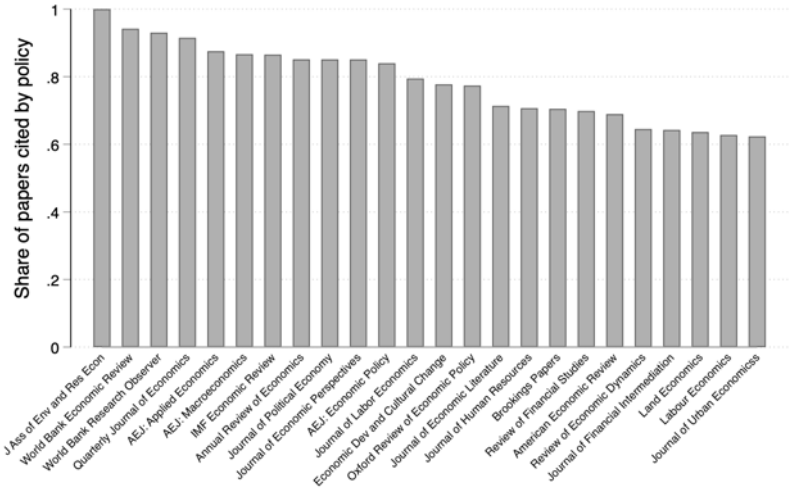
Note: This figure shows the countries of origin of the policy documents citing economics research (Panel A) and their main sources (Panel B). See text for more details.

Table E1: Statistical Association Between Research Design and Policy Impact of Economics Papers

	US Cites	Non-US Cites	US Cites	Non-US Cites	US Cites	Non-US Cites
	(1)	(2)	(3)	(4)	(5)	(6)
Empirical (0/1)	0.577***	2.182***				
	(0.04772)	(0.21677)				
Administrative (0/1)			0.697*	0.423		
			(0.28719)	(0.47368)		
FSRDC Use (0/1)					4.137***	5.473**
					(0.47620)	(1.98045)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Journal	Yes	Yes	Yes	Yes	Yes	Yes
N	188181	188181	188181	188181	188181	188181
Mean of DV	0.784	3.215	0.784	3.215	0.784	3.215

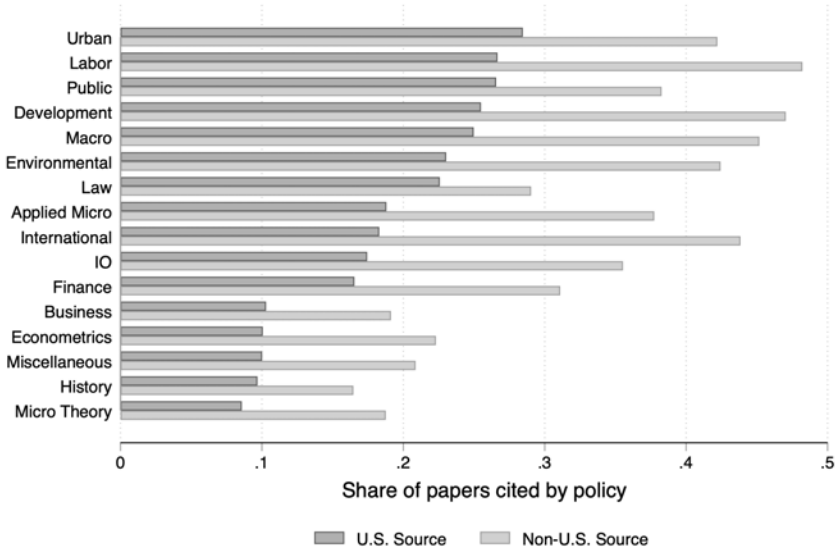
Note: This table presents estimates from OLS models evaluating the average increase in policy citations for papers that are empirical (columns 1 and 2), mention administrative data in their abstract (columns 3 and 4), and directly use FSRDC data (columns 5 and 6). Papers in columns 3 and 4 are tagged using the keywords listed in Table C2. *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Standard errors are clustered by year of publication.

Figure E3: Share of Articles Receiving at Least One Citation from Policy Documents by Journal



Note: The figure reports the share of published articles that received at least one policy citation for each journal. Only the journals with the highest shares are shown in the graph.

Figure E4: Share of Articles Receiving at Least One Citation from Policy Documents by Field



Note: The figure reports the share of published articles that received at least one policy citation by field. Articles are classified into fields using the method of Angrist et al. (2020). As a comparison, 38% of economics articles in our sample receive mentions from policy sources.

for academic citations.

E.2 Example: Greenstone et al. (2010)

To exemplify the type of policy citations captured by the Overton data, consider the paper "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings" published in 2010

by Michael Greenstone, Richard Hornbeck, and Enrico Moretti (henceforth, "Million Dollar Plant paper"). In this influential paper (1,216 citations on Google Scholar as of 2022), the authors quantify the extent of localized spillovers on productivity. The research design exploits the opening of large manufacturing plants to compare changes in total factor productivity of incumbent plants in "winning" and "losing" counties. To carry out their analyses, the authors accessed several confidential datasets hosted in the FSRDC: the Standard Statistical Establishment List (SSEL), the Annual Survey of Manufactures (ASM), and the Census of Manufactures (CM). Using these data sources, the authors find that incumbents' productivity is 12 percent higher in winning counties five years after the plant openings. Such findings have important implications for policies aimed at local economic development, as remarked by the authors themselves in the introduction of the paper.

Indeed, the paper by Greenstone et al. (2010) received 101 citations from policy documents up to 2019. The documents citing the paper come from a variety of countries: United States (36), IGOs (34), Germany (7), UK (7), Ethiopia (4), and nine others. The first citation came in 2010, the same year as the publication of the paper, in a Discussion Paper published by the Hamilton Project at the Brookings Institute[19]. The paper highlights Greenstone et al. (2010) as providing promising evidence of industrial cluster effects, and assesses whether this may provide the basis for a national subsidy policy. Two years later, the Million Dollar Plant was cited in a policy report prepared by London Economics (a policy consultancy) for the Department for Business, Innovation and Skills of the U.K. Government[20]. The document contains a literature review of the evidence regarding the productivity spillovers of investment in intangible assets. In particular, the report cites the Million Dollar Plant paper as one of the few studies quantifying the magnitude of agglomeration spillovers without depending on the usage of patent citations. Similarly, the "5 Year Productivity Review" compiled by the Australian Productivity Commission in 2017 lists this paper as evidence for agglomeration economies.[21]

Some policy documents appear to cite the paper to extrapolate the likely effect of place-based policies aimed at firms, and less for its academic contribution in causally showing the existence of agglomeration spillovers. The 2019 report "How to Solve the Investment Promotion Puzzle: A Mapping of Investment Promotion Agencies in Latin America and the Caribbean and OECD Countries" by the Inter-American Development Bank cites the Million Dollar Plant paper as showing the efficacy of place-based policy aimed at attracting manufacturing investments.[22] Other documents using the findings of Greenstone et al. (2010) as evidence of place-based policy effectiveness include books from Bruegel,[23] policy briefs from the Brooking Institute,[24]

[19] Source: https://www.brookings.edu/wp-content/uploads/2016/06/10_job_creation_bartik.pdf
[20] Source: https://gov.uk/government/uploads/12-793-investment-intangible-assets-on-productivity-spillovers.pdf
[21] Source: https://apo.org.au/sites/2017-10/apo-nid115951_1.pdf
[22] Source: https://publications.iadb.org/how-solve-investment-promotion-puzzle
[23] Source: https://www.bruegel.org/sites/2016/01/Blueprint-XXIV.pdf
[24] Source: https://www.brookings.edu/2018/ES_THP_CompetitionFacts.pdf

and policy reports from the OECD.[25]

F Evidence on Articles using FSRDC Datasets

This Appendix briefly describes the nature and expansion of datasets available to researchers in FSRDCs. Next, we show descriptive evidence of the characteristics of the economic articles that directly use FSRDC data. Using manually coded information on the specific confidential datasets used in each paper, we provide suggestive evidence on the lifecycle of research data sources.

F.1 The Development of New Datasets at the FSRDC

Cognizant of the immense research potential of its microdata, in 1982, the U.S. Census Bureau established the CES to make these resources accessible to researchers in economics (CES, 2017). The initial focus of the CES was the creation of data resources on the manufacturing sector (McGuckin, 1995). The first matched dataset, developed in 1984, was called the Longitudinal Establishment Database (LED) and consisted of pooling the manufacturing data from 1972 to 1981. The significant contribution of that database was the creation of Permanent Plant Numbers (PPNs) that enabled merging survey waves from different years (Atrostic, 2007). Following this first step, CES staff members continued expanding the LED with data from the Economic Censuses and the Annual Survey of Manufactures, eventually creating what is known as the Longitudinal Research Database (LRD) (McGuckin et al., 1993). The LRD became a favorite of academic researchers, allowing them to conduct pathbreaking empirical research on business dynamics and business demographics (Davis et al., 1998).

In the late 1990s, CES began developing a new database, later called LBD (Jarmin and Miranda, 2002). According to Chow et al. (2021), creating the LBD was prompted by the desire to test if the results based on the manufacturing data of the LRD also applied to other sectors. The LBD was created by merging the Standard Statistical Establishment List with the Economic Census, thus creating a source that contains basic information on the universe of all U.S. business establishments. This impressive data effort was joined by the concurrent development of the LEHD data (Abowd et al., 2004). The LEHD merges worker and employer records from Census Bureau surveys with state unemployment insurance claims to create matched employer-employee panel data. Together, the LBD and LEHD constitute unique research tools to investigate the dynamics of the U.S. economy.

Over time, the range of confidential datasets FSRDC offers has expanded. Several other statistical agencies have started making their microdata available through the FSRDC network, most notably the National Center for Health Statistics (NCHS), the Bureau of Labor Statistics (BLS), and the Agency for Healthcare Research

[25]Source: https://www.oecd/OECD-Trade-Policy-Paper-242

and Quality (AHRQ). The advantages of doing so are clear since they can leverage the existing system of data centers without having to muster the resources for creating a similar infrastructure (CES, 2017). Moreover, one farsighted feature of the FSRDC is allowing approved users to merge existing microdata and collaborate in developing new databases. For example, the recent development of the Management and Organizational Practices Survey (MOPS) has been spearheaded by academic researchers, who have collaborated with the U.S. Census Bureau to design a novel survey instrument (Bloom et al., 2019). As a result, survey and administrative microdata available in FSRDC are increasingly powerful in investigating a broad swath of economic, social, health, and policy questions.

F.2 Characteristics of FSRDC Papers

We begin by assessing the characteristics of papers that employ confidential Census microdata. Table F1 shows the strength of the statistical association between the use of FSRDC data and several indicators of research quality after controlling for the year of publication. Across the board, one notices that research using confidential microdata is much more likely to be published in highly prestigious outlets, with a large share of them featured in one of the so-called top five economics journals. Regarding impact, the average FSRDC paper receives ten more citations and is twice as likely to be referenced in policy documents.

Table F1: Statistical Association Between Proxies of Research Quality and Use of FSRDC Data

	Top Field (0/1)	Five-Year Cites	Policy Relevant (0/1)
	(1)	(2)	(3)
FSRDC Use (0/1)	0.209***	10.57***	0.365***
	(0.02069)	(1.34000)	(0.02199)
Year FE	Yes	Yes	Yes
N	188181	188181	188181
Mean of DV	0.196	12.31	0.377

Note: This table presents estimates from OLS models at the article level evaluating the association between proxies of research quality and the use of FSRDC data. *, **,*** denote significance at 5%, 1% and 0.1% level respectively. Standard errors are clustered by year of publication.

Figure F1 illustrates the steady growth of FSRDC articles published yearly. The growing diffusion of FSRDC data in economics closely mimics the expansion of the FSRDC network exploited in the primary analyses of this paper. Yet, a potential downside of lowering access costs is that it could increase the quantity of research produced at the expense of its quality. When confidential data are hard to access, researchers might be using them only for ideas of high quality, while easier access could increase scientific production mainly in the lower tail of the quality distribution. However, the evidence we present is inconsistent with this line of reasoning. Figure F1 shows that the share of those articles appearing in the top field or top five journals is relatively constant over time. This suggests that the expansion in data access is not leading to articles of lower quality, at least as proxied by the prestige of the journals where they get accepted (Nagaraj et al., 2024).

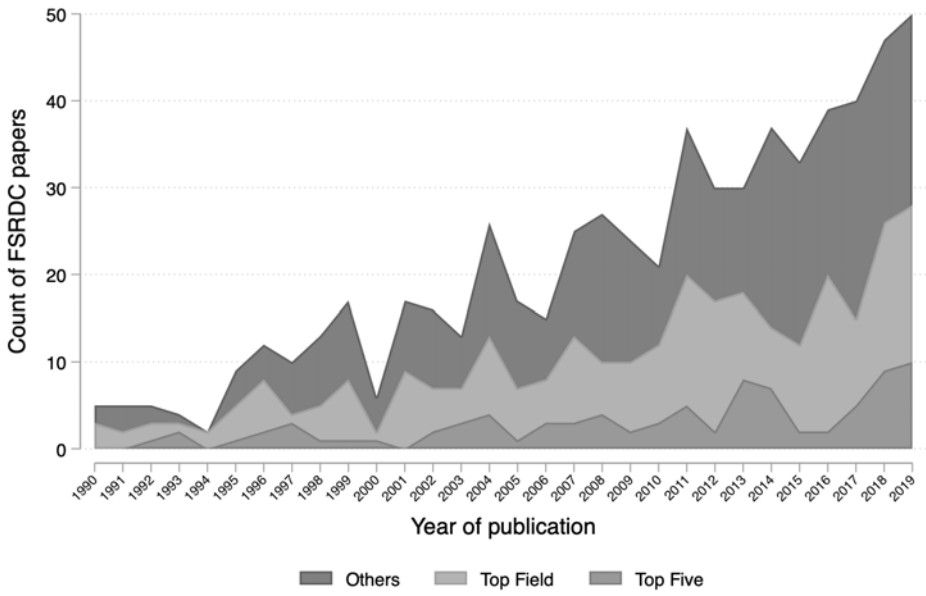
F.3 Descriptive Evidence at the Dataset-level

Our primary analyses exploit the staggered roll-out of FSRDC across the United States to estimate their causal impact on the rate and direction of academic economics research. However, one might wonder how different the effect of an FSRDC that opened in the 1990s could be relative to those that opened in the 2010s. On the one hand, researchers getting access earlier can easily exploit the data to their full potential, unlike those who get access later. This logic would imply diminishing returns to additional FSRDCs because many research lines that could be explored with confidential microdata will be exhausted over time. On the other hand, researchers who gain access later might be able to build off a larger pool of metadata and tacit knowledge about FSRDC datasets. Moreover, one of the objectives of the FSRDC program is to enable the creation of new research data, expanding the menu of available datasets over time. Therefore, it is not clear that the value of FSRDC access should be decreasing over time.

We painstakingly coded all the specific datasets employed in each FSRDC article to shed light on these alternative explanations.[26] The result of this manual effort is a unique opportunity to explore the "lifecycle" of individual confidential datasets and provide novel information on which FSRDC datasets are the most commonly used in economic research. Figure F2 shows the most popular dataset used in our sample of FSRDC articles. One notices immediately that the LBD and LRD account for almost 30% of all economics

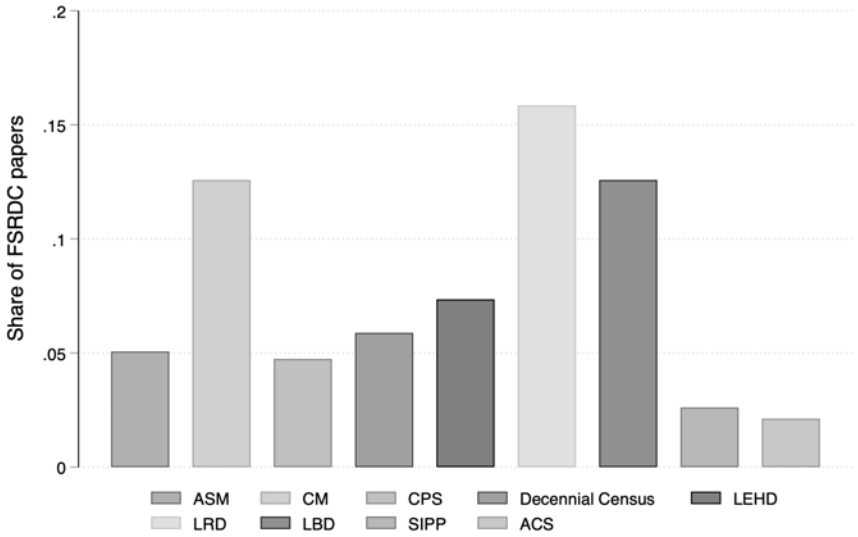
[26] We thank Buyi Geng and Jiamei (Jasmine) Xu for their precious help in manually coding the confidential dataset used by each FSRDC paper.

Figure F1: Count of Yearly FSRDC Articles over Time



Note: The figure reports the yearly count of economics articles using FSRDC data. The area of the plot is colored in different shades of grey to indicate articles published in a top field journal or one of the top five journals.

Figure F2: Confidential FSRDC Datasets Most Frequently Used in Economics Research



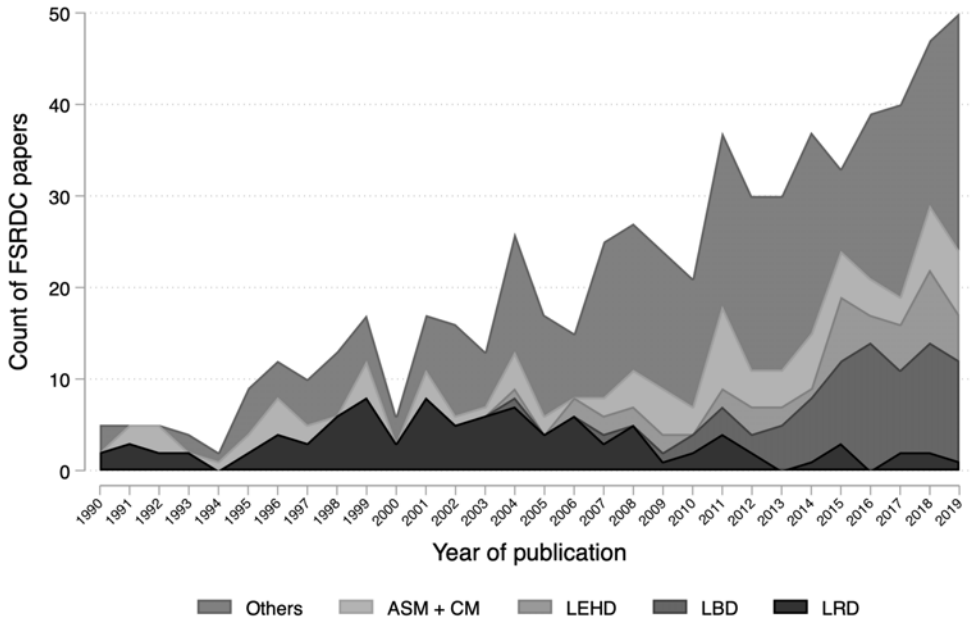
Note: The figure reports the share of articles using FSRDC data that use each of the most common datasets. The datasets shown in the figure include the Annual Survey of Manufactures (ASM), the Census of Manufactures (CM), the Current Population Survey (CPS), the Decennial Census, the Longitudinal Employer-Household Dynamics dataset (LEHD), the Longitudinal Research Database (LRD), Longitudinal Business Database (LBD), the Survey of Income and Program Participation (SIPP), and the American Community Survey (ACS). For this figure, we code the article as using only the LRD or LBD when the CM or ASM data are used in conjunction with the LRD or the LBD, respectively.

papers written in FSRDCs, suggesting that the CES' efforts in creating these resources for research have been highly successful. Data on manufacturing plants are among the most popular in economic research. Datasets that provide details on demographic trends, such as the Decennial Census or the Current Population Survey, are used in a smaller number of papers but are likely to be among the most relevant for other disciplines, such as sociology and demography.

Figure F3 breaks down the time series of FSRDC papers by the type of dataset used. A few trends emerge clearly. First, the LRD constituted the main data source for the first two decades of our sample period, consistent with the accounts of Atrostic (2007). Second, in recent years, the LRD has been replaced by the LBD, which is now the most popular confidential dataset provided by the Census Bureau. Third, the LEHD has grown in popularity recently, but its usage is still much lower than the LBD. Potential reasons for this are its partial coverage of U.S. states and the availability of similarly matched employer-employee data from Scandinavian countries. Other sources, such as the Census of Manufactures and the Annual Survey of Manufactures, are primarily used with the LRD and LBD.

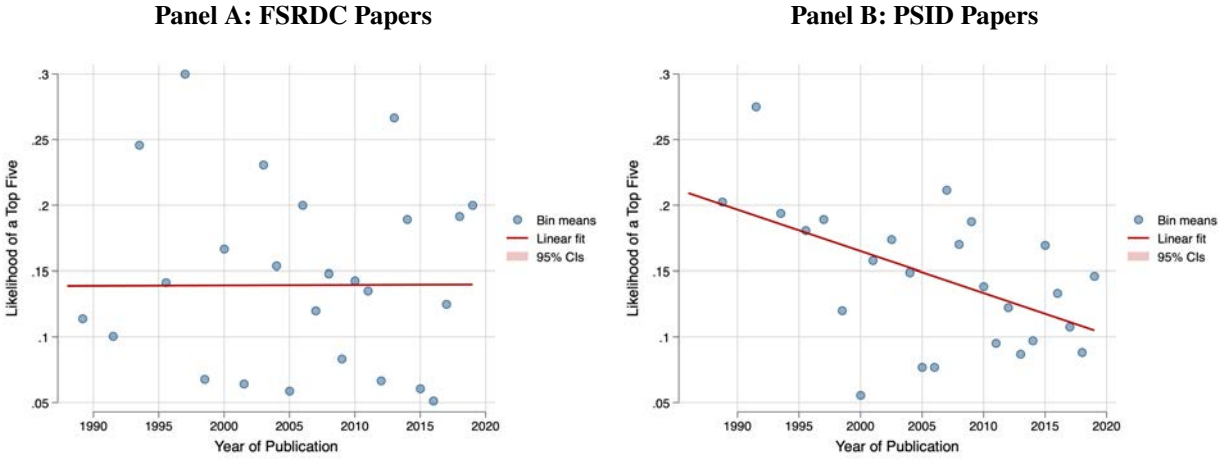
Finally, we directly assess whether there are decreasing returns to using FSRDC data over time. Panel A of Figure F4 plots the likelihood that a paper will be published in a top five journal over time, conditional on using data accessible only in an FSRDC. The figure shows some year-to-year variation but no specific time

Figure F3: Usage of Specific Confidential Datasets over Time



Note: The figure reports the yearly count of economics articles using FSRDC data, divided by the specific confidential dataset used. The figure shows the most common datasets on the manufacturing sector, namely the Longitudinal Research Database (LRD), Longitudinal Business Database (LBD), Longitudinal Employer-Household Dynamics dataset (LEHD), the Annual Survey of Manufactures (ASM) and the Census of Manufactures (CM). For this figure, we code the article as using only the LRD or LBD when the CM or ASM data are used in conjunction with the LRD or the LBD, respectively.

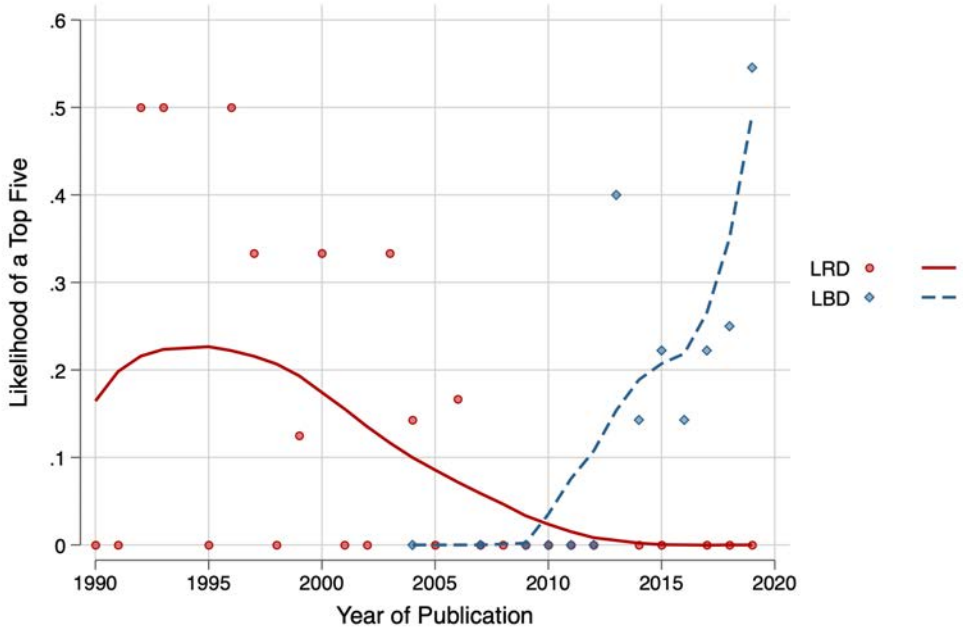
Figure F4: Likelihood of Being Published on a Top Five Journal, FSRDC Papers vs. PSID Papers



Note: This figure shows a binned scatterplot of the likelihood that an article appears in a top five economics journal over time, separately for papers using confidential FSRDC data (Panel A) and the Panel Study of Income Dynamics (PSID) data (Panel B). See text for more details.

trends. As a suggestive comparison, Panel B shows the same for articles that employ data from the Panel Study of Income Dynamics (PSID). In this case, we can see a clear downward trend, suggesting that the same dataset is less likely conducive to top publication over time. How can this difference be explained? One possibility is that adding new confidential datasets in the FSRDCs counteracts any given data source's natural

Figure F5: Likelihood of Being Published on a Top Five Journal Conditional on Using FSRDC Data (LRD vs. LBD)



Note: This figure shows a non-parametric binned scatterplot of the likelihood that an article using FSRDC data appears in a top five economics journal over time, separately for papers using the Longitudinal Research Database (LRD) and the Longitudinal Business Database (LBD). See text for more details.

decline in productivity.[27] Figure F5 shows some evidence consistent with this supposition. Considering only the LRD, we see a similar pattern to the PSID. However, the creation of the LBD in 2002 opened up new possibilities for scholars, offering an alternative to the progressive exhaustion of LRD's potential. In sum, the progressive addition of new confidential datasets offers a potential explanation for the continuing research value of access to a local FSRDC.

[27]The purpose of this comparison is merely suggestive, as we cannot fully account for many differences between PSID and FSRDC datasets.

G Cost-Effectiveness of FSRDCs

This paper estimates the effect of the FSRDC network on economic research. In this Appendix, we leverage those estimates to evaluate both the costs and benefits of FSRDC-based provision of data to researchers. We take the viewpoint of a research institution that is assessing whether to open an FSRDC, comparing the likely benefits with the financial costs of running an FSRDC.[28] Albeit necessarily involving some arbitrary assumptions, our objective is to reasonably assess the magnitude of these costs relative to the research gains.

G.1 The costs of opening an FSRDC

We evaluate the decision to open a new FSRDC over a horizon of twenty years. To do so, we leverage detailed internal estimates from the Census Bureau on the costs for an institution to open an FSRDC location.[29] We use the upper bound estimates, $185,000 annually, which corresponds closely to universities' reported payments.[30] According to Figure G1, the primary cost consists of the salary and overhead of FSRDC administrators, which are Census Bureau employees who oversee the functioning of the RDC. The hosting institution also provides the physical space for the data center, with an estimated $150,000 in infrastructure costs to comply with Census and IRS security specifications.[31] Finally, the hosting institution must cover a portion of the ongoing costs of disclosure avoidance reviews, a process that screens research results for sensitive information.

Figure G1: Census Bureau Estimates of Opening an FSRDC Location

	Year 1	Year 2	Year 3	Year 4	Year 5	Annual Average Cost
Disclosure services	$ 6,000	$ 12,000	$ 18,000	$ 24,000	$ 24,000	$ 16,800
Infrastructure costs	$ 50,000	$ 50,000	$ 50,000	$ -	$ -	$ 30,000
Administrator Salary & Overhead	$ 129,891	$ 134,220	$ 138,550	$ 142,881	$ 142,881	$ 137,684
Total yearly cost	$ 185,891	$ 196,220	$ 206,550	$ 166,881	$ 166,881	$ 184,484

Note: These costs are obtained from Attachment B in https://www.census.gov/documents/Guidelines_for_RDC_Development_and_Operations_FY2020.pdf. We take the top-range estimate for all administrator-related expenses.

[28] We are refraining here from quantifying the full benefits occurring to the Census Bureau (see discussion in Appendix C.1). However, we note that the legal requirements of Title 13 mandating a benefit for the Census Bureau necessarily imply that each approved project necessarily has a net positive value for its data programs. We are also unable to assess social benefits, such as directly quantifying the improvement in policy-making, or social costs, such as the privacy risks of making confidential data more broadly available. For more details about the debate on privacy protection issues, see Abowd and Schmutte (2019); Potok (2024).

[29] See Attachment B in https://www.census.gov/documents/Guidelines_for_RDC_Development_and_Operations_FY2020.pdf.

[30] The Executive Director of the Boston FSRDC reported a figure of approximately $187,500 yearly costs. See: https://deepblue.lib.umich.edu/ACDEB_response_FSRDCDirectorsFinal2021.pdf.

[31] This amount, spread over three years, does not include the in-kind contribution of the physical space by the host institution. However, the Bureau will supply much of the technology and equipment for the facility, including the router to connect to the FSRDC Virtual Private Network.

Notably, these costs are largely fixed and independent of researchers' actual use of FSRDCs. This means that the expenses incurred in operating FSRDCs may not be recovered if researchers do not use their data, but also that costs do not scale proportionally with increased use.[32] This explains why universities often apply to open an FSRDC location as a consortium, with one consortium member serving as the hosting institution: in addition to sharing the burden of the operational costs, this guarantees a larger potential user base. As it will be clearer later, we will consider the size of the consortium as a critical input into the analysis when evaluating the cost-effectiveness of FSRDCs.

We make several assumptions when modeling costs. First, we assume that the cost reported for Year 5 will be repeated every following year. Second, we assume the hosting consortium avails itself of the maximum grant support from the NSF to establish the FSRDC, amounting to a total of $300,000 to help cover start-up costs.[33] Finally, we make one adjustment to the internal cost estimates. Economists are not the only beneficiaries of FSRDC data, as sociologists, demographers, and public health scholars also use FSRDCs. For a fair comparison, we estimate the share of total costs of FSRDCs attributable to economists, assuming that it is proportional to usage. We manually classify the Census Bureau's full list of active and completed FSRDC projects since the inception of the program by discipline.[34] Out of 771 projects, we classify 597 as relating to economics research and hence attribute approximately 77% of the costs of opening an FSRDC to economists.

G.2 The monetary benefits of opening an FSRDC

We describe three alternative methods to obtain a plausible dollar value for an additional scientific publication. While the resulting estimates can be used to contextualize our findings, our cost-benefit analysis will allow the reader to factor in their priors on the dollar value of an economics publication.

Method 1: Revealed Value to the University. Our first method leverages the fact that an additional publication will likely increase the probability that an assistant professor is awarded tenure during their first employment spell. The basic idea is to use the increase in salary coming from the promotion to associate professor to estimate the revealed monetary value of an additional top publication for the university. For this method, we rely on the estimate from Sarsons (2015) that an additional publication is associated with a 5.7% increase in the likelihood that a professor's first tenure spell is successful. We then leverage the transparency of the University of California system, which publicly releases salary information for all its employees and allows us to pinpoint the salary premium from getting tenure in a higher-status campus relative to the likely

[32] This consideration is net of eventual congestion costs that will become binding when a facility is used at full capacity. However, this issue never came up in any of our interviews, thus suggesting that FSRDCs are not currently used at their full capacity.

[33] See for instance: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1955355&HistoricalAwards=false.

[34] Available at https://www.census.gov/about/adrm/fsrdc/about/ongoing-projects.html. We exclude from the calculation the Census Headquarters branch, which is less representative of usage in the partner institutions hosting FSRDCs.

counterfactual of having to move to a lower-ranked institution.

Specifically, each academic salary at the University of California consists of a scale component (UC-wide) and an off-scale component (employee-specific). The salary scale is based on the rank of the professor (i.e., assistant professor, associate professor, and full professor) and the number of years in the position.[35] In contrast, the off-scale component varies systematically across UC campuses: for instance, internal documents suggest the off-scale amounts paid to UC Berkeley faculty leads to 18% higher salaries on average than UC Irvine faculty.[36] We construct two hypothetical earnings timelines, each with a length of 35 years (the typical career of an economist), inspired from the data in Heckman and Moktan (2020).

In the first timeline, an assistant professor is promoted to associate professor after a 7-year spell and then to full professor after another 7-year spell. In this base case, we derive the professor's annual salaries from the salary scale. In the second timeline, the assistant professor is denied tenure after the first 7-year spell and moves to a lower-tiered institution to progress immediately to a tenured position. After a successful second 7-year spell, they are promoted to full professor. To adjust for the downward mobility in this timeline, we apply the off-scale wage differential between UC Berkeley and UC Irvine as a conservative earnings penalty. We calculate the present value of the annual salary difference between the two timelines to be $472,205 using a 5% discount rate. Note that this estimate does not include the option value of more consulting opportunities or book deals. In conjunction with the estimate from Sarsons (2015), this implies the present value of a publication to be $26,916. If we allow for a 2-year lag before the researcher is promoted to associate professor in the second timeline, the value of a paper increases to $32,043.

Method 2: Private Value for the Researcher. Our second method is more direct and just evaluates the researchers' salary raise associated with the marginal top publication. However, this approach relies entirely on the validity of the estimated relationship between publication output and annual salary. For this exercise, we take this estimate from Sauer (1988), which finds that one additional publication in a top 20 journal increases a professor's annual salary by at least $851 in 1982 dollars, equivalent to $2,255 in 2019 dollars. This corresponds closely to our measure of top publications, which considers any publication in top journals. To convert this measure into the lifetime value of a publication, we once more assume a 35-year career length and a 5% discount rate. Since the average economist in our sample is 13 years into their career, this implies the present value of a top publication is $31,167.

We note that the estimate from Sauer (1988) is contextual with more recent estimates from other related fields. For instance, Swidler and Goldreyer (1998) find that a publication in a top finance journal increases a finance professor's salary by $1,952 in 1993 dollars, which is equivalent to $3,453 in 2019 dollars. Mittal

[35] Economics faculty follow the salary scale for ladder-rank faculty in Business, Economics, and Engineering. See: https://www.ucop.edu/academic-personnel-programs/oct-2023-acad-salary-scales.pdf.

[36] See: https://academicsenate.ucdavis.edu/sites/archive/08_ucd_uc_salaries.pdf.

Table G1: Summary of methods to estimate the value of a publication.

Method	Preferred Estimate
1. Present value of the increased probability of tenure	\$26,916
2. Present value of the salary increases	\$31,167
3. Benchmark from another social science field (EO)	\$29,828

et al. (2008) estimate that a Tier 1 marketing journal publication increases a marketing professor's salary by \$2,176 in 2001 dollars, which is equivalent to \$3,143 in 2019 dollars. Finally, Li et al. (2023) estimate that a top US management journal publication increases a strategy professor's salary by anywhere from \$2,200-\$3,800. Taken together, this shows that our preferred estimate is quite conservative.

Method 3: Using Estimates of Social Value from Other Disciplines. The third and final method is adopting an existing estimate of social value from other disciplines. Unfortunately, there are currently few estimates because estimating social value requires strong assumptions. One recent exception is Morretta et al. (2022), who estimate the marginal social value of publications that use imagery from the Earth Observation (EO) satellites constellation of the Italian Space Agency. The study estimates the marginal cost of production of an EO publication, calculated by dividing annual researcher salaries (adjusted for the share of time dedicated to producing research) by annual productivity and adding in the publishing costs borne by publishers. The study then factors in the social value of a citation, which is calculated by augmenting the marginal cost of a publication by the share of time dedicated to reading research. Using this method, the study estimates that 1,235 publications from 1998-2018 produced \$36,837,104 in societal value, or \$29,828 per publication.

G.3 Cost-benefit analysis of opening an FSRDC location

While our three methods produce comparable estimates of the dollar value of publication (Table G1), we are hesitant to select any one of them. Instead, we pursue an alternative strategy: calculating the required dollar value of a publication for an investment in an FSRDC to break even. In other words, how much would a paper be worth to justify opening a new FSRDC location? This has the advantage that individuals or institutions can evaluate the merit of opening an FSRDC, depending on the value they attribute to a top publication. In practice, we solve for $Price$ in the equation:

$$\sum_{t=1}^{20} \frac{Price \times Number\ Publications_t}{1.05^{t-1}} = \sum_{t=1}^{20} \frac{0.77[(Total\ Costs_t) - NSF\ Grant_t]}{1.05^{t-1}} \tag{2}$$

Because universities typically apply for an FSRDC as a consortium, we evaluate the decision to open an FSRDC at the consortium level, assuming that costs are largely fixed. We calculate the number of additional

publications due to the FSRDC as follows:

$$Number\ Publications_t = 1.035 + (24 \cdot Consortium\ Effect) \cdot Number\ Partners \tag{3}$$

The first term consists of the additional top publications that accrue to the hosting institution, which we calculate by multiplying the median number of AEA-affiliated economists at the hosting institutions in our sample (45) by the treatment effect size for hosting institutions (0.0230). The second term consists of the additional top publications that accrue to all non-hosting consortium partners, which we calculate analogously by multiplying the median number of economists at the consortium partners in our sample (24) by the treatment effect size for consortium partners and by the number of consortium partners. Hence, we effectively calculate the required break-even dollar value per publication, conditional on the consortium size and effect size. We discount all costs and benefits using a 5% discount rate.

The results of this analysis are presented in Figure G2. To interpret this table, consider the cell (0.015, 4). This cell can be read as implying that the dollar value of a top publication would have to be at least $47,094 for a consortium to justify opening an FSRDC location, assuming the consortium consists of one hosting institution and four partners and the treatment effect for the partners is 65% as large as the effect for hosting institutions. This is beyond the upper limit of our estimated dollar publication value, although it could fall within an acceptable range for some institutions. If the same consortium could increase FSRDC adoption at the non-hosting partners, the dollar value of a paper would fall near the range of our estimated dollar value of a publication. In practice, the extent of the effect for any consortium member will depend on its geographical distance from the hosting institution (Appendix Figure H5). Alternatively, if the same consortium could include four additional members, then the dollar value of a paper would only have to be $29,772. This is within the range of our estimated dollar value of a publication.

Figure G2: Required Dollar Value of a Publication for Break-even of FSRDC Funding

Consortium effect \ Number of consortium partners	1	2	3	4	5	6	7	8	9
0.005	$100,916	$91,418	$83,554	$76,936	$71,289	$66,415	$62,164	$58,425	$55,110
0.01	$91,418	$76,936	$66,415	$58,425	$52,151	$47,094	$42,931	$39,444	$36,481
0.015	$83,554	$66,415	$55,110	$47,094	$41,114	$36,481	$32,787	$29,772	$27,265
0.02	$76,936	$58,425	$47,094	$39,444	$33,932	$29,772	$26,521	$23,909	$21,766
0.025	$71,289	$52,151	$41,114	$33,932	$28,887	$25,147	$22,265	$19,976	$18,113

Note: These costs are obtained from Equation 2 at increasing values of the treatment effect size for consortium partners (from 0.005 to 0.025) and the number of consortium partners (from 1 to 9). Each cell reports how much a publication would need to be valued to justify opening an FSRDC with that effect size and number of consortium members. See text for details.

While our analysis rests on various assumptions, there are several takeaways from this exercise. First, while increasing access to confidential microdata can result in higher research productivity, the decision to open an FSRDC is not unconditional for university administrators. If the consortium cannot guarantee sufficient

engagement with the FSRDC, then the required dollar value might be prohibitively large to justify the FSRDC. On the upside, the cost-benefit analysis in Table G2 would look much more favorable if universities managed to increase their user base, which we assumed to be fixed. When considering institutions with a larger number of applied scholars and treatment effects reflecting the dynamically increasing impact of FSRDCs documented in Figure 3, even individual universities could justify opening an FSRDC.[37] Second, this exercise highlights multiple potential strategies to maximize the value of an FSRDC. Either the consortium can prioritize efficiency, increasing the number of papers that employ FSRDC data from its current roster of faculty (perhaps by removing barriers to adoption, see Section A), or the consortium can prioritize scale, increasing the potential user base while holding engagement constant.

G.4 Cost-effectiveness of opening an FSRDC location

In a second exercise, we evaluate whether the same amount of dollars invested in an FSRDC could instead have a greater research impact if allocated as an unrestricted research grant. To facilitate this comparison, we turn to the literature for studies estimating the elasticity of research output to funding. Table G2 presents the complete list of benchmarks, all normalized to $100,000 of funding.

Table G2: Estimates of Papers per Dollar of Funding from Other Academic Studies

Academic Study	Sattari et al. (2022)	Wahls (2018)	Payne and Siow (2003)	Whalley and Hicks (2014)	Rosenbloom et al. (2015)	Arora and Gambardella (2005)
Field	Biomedicine	Biomedicine	All	All	Chemistry	Economics
Papers per 100k dollars of funding	0.8	0.53-0.87	1	0.83	0.6-1.9	0.48

The first study is Sattari et al. (2022), which looks at how much biomedical research was stimulated by science funding from the National Institutes of Health (NIH) at 72 university campuses in the United States. Using an event-study design, the study finds that an additional $100,000 in funding produces 0.8 additional papers. A related study from Wahls (2018) finds that each $1 million of NIH funding to prestigious institutions produced 5.3 papers, while $1 million to less prestigious institutions produced 8.7 papers. A third study by Payne and Siow (2003) examines how federal research funding affects the total quantity of research produced at a university. Using alumni representation on the congressional appropriations committees to instrument for the share of federal R&D funding, the study finds that an additional $1 million in funding results in 10 more articles. Similarly, the fourth study from Whalley and Hicks (2014) estimates how spending by universities spending affects total research output. Since universities are constrained in their spending by the market value of their endowment, the study uses shocks in the stock market to instrument spending. The authors find that an additional $1.1 million in expenditures produces about nine additional papers. In the context of chemical research, Rosenbloom et al. (2015) find that every additional million of Federal R&D

[37] Note also that our count of potential users does not include doctoral students, whose usual lack of publications during the PhD program prevents from being included in our sample. FSRDCs have an important educational value and enable several PhD dissertations every year. Considering this aspect would further increase the appeal of opening an FSRDC.

funding produces 6-7 more papers, while their IV estimates (which uses federal R&D for mathematics and physics as an instrument) imply a greater increase of 19 articles.

Finally, in the setting closest to our own, Arora and Gambardella (2005) look at NSF grants' impact on US-based economists' research output.[38] However, we must first perform some imputations to facilitate a direct comparison. Firstly, their measure of research output is not publications but rather quality-adjusted publication units, with one publication in the highest-ranking journal corresponding to 100 publication units and any publication not among the top 50 journals corresponding to 1 publication unit. We consider 14.63 units equivalent to one top publication.[39] Secondly, the study only reports separate elasticities by academic tenure, so we weight them by the shares of NSF grants received by researchers of different seniority in their sample to reconstruct the aggregate elasticity. With this method, we calculate that adding $100,000 to an NSF grant would result in 7 additional quality-adjusted "publication units" per year, equivalent to half a top publication.

Figure G3: Annual Top Publications per $100,000 of Funding in FSRDCs

Consortium effect \ Number of consortium partners	1	2	3	4	5	6	7	8	9
0.005	0.90	0.99	1.09	1.18	1.27	1.37	1.46	1.55	1.65
0.01	0.99	1.18	1.37	1.55	1.74	1.93	2.11	2.30	2.49
0.015	1.09	1.37	1.65	1.93	2.21	2.49	2.77	3.05	3.33
0.02	1.18	1.55	1.93	2.30	2.67	3.05	3.42	3.79	4.17
0.025	1.27	1.74	2.21	2.67	3.14	3.61	4.07	4.54	5.01

Note: These costs are obtained from Equation 4 at increasing values of the treatment effect size for consortium partners (from 0.005 to 0.025) and the number of consortium partners (from 1 to 9). Each cell reports the additional number of publications per $100,000 of funding in an FSRDC with that effect size and number of consortium members. See text for details.

Then, we calculate the number of additional publications per $100,000 of funding invested in opening and operating an FSRDC. The formula for the number of top publications is identical to before, which means we keep $Consortium\ Effect$ and $Number\ Partners$ free and estimate one elasticity per cell. This time, however, we divide the number of publications by the steady state cost of operating an FSRDC (i.e., the cost from the fifth year onwards). This is the maximum yearly cost in our schedule since the NSF grant support will have already subsided. We then normalize the figure to $100,000 of funding. Hence, the cost-effectiveness of an FSRDC is calculated as follows:

$$Papers\ per\ \$100,000 = [1.035 + (24 \cdot Consortium\ Effect) \cdot Number\ Partners] \cdot \frac{100,000}{166,881 * 0.77} \tag{4}$$

The results are presented in Figure G3. These results suggest that opening an FSRDC location is a highly cost-effective method of producing more science compared to other existing funding initiatives. Even in

[38] This is a particularly salient comparison, given NSF's role in seeding the establishment of FSRDCs.

[39] To recall, we defined as a top publication the highest-ranked general purpose and top field journals. We consider a top publication as amounting to 14.63 because this is the minimum number of units corresponding to any of the top 20 journals in Arora and Gambardella (2005).

smaller consortia with more limited adoption of FSRDC data (i.e., the top left corner), the number of additional publications per $100,000 of funding generally exceeds 1, above most estimates from existing studies (even in fields generally more productive than economics). This holds when directly compared with NSF funding for economists: the most "apples-to-apples" comparison from Arora and Gambardella (2005) suggests an increase in only half a top publication for the same amount of funding. Furthermore, as for the earlier analysis, the cost-effectiveness of an FSRDC increases much more as the consortium grows in size and as adoption grows among its membership base.

Appendix References

Abowd, J. J., J. Haltiwanger, and J. Lane (2004): "Integrated longitudinal employer-employee data for the United States," *American Economic Review*, 94, 224–229.

Abowd, J. M. and I. M. Schmutte (2019): "An economic analysis of privacy protection and statistical accuracy as social choices," *American Economic Review*, 109, 171–202.

Angrist, J., P. Azoulay, G. Ellison, R. Hill, and S. F. Lu (2020): "Inside job or deep impact? Extramural citations and the influence of economic scholarship," *Journal of Economic Literature*, 58, 3–52.

Arora, A. and A. Gambardella (2005): "The impact of NSF support for basic research in economics," *Annales d'Economie et de Statistique*, 91–117.

Atrostic, B. (2007): "The Center for Economic Studies 1982-2007: A brief history," CES working paper.

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2019): "Public R&D investments and private-sector patenting: Evidence from NIH funding rules," *The Review of Economic Studies*, 86, 117–152.

Baumann, A. and K. Wohlrabe (2020): "Where have all the working papers gone? Evidence from four major economics working paper series," *Scientometrics*, 124, 2433–2441.

Bloom, N., E. Brynjolfsson, L. Foster, R. Jarmin, M. Patnaik, I. Saporta-Eksten, and J. Van Reenen (2019): "What drives differences in management practices?" *American Economic Review*, 109, 1648–1683.

Bowen, C. M. (2024): "Government data of the people, by the people, for the people: Navigating citizen privacy concerns," *Journal of Economic Perspectives*, 38, 181–200.

Callaway, B. and P. H. Sant'Anna (2021): "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 225, 200–230.

Card, D., R. Chetty, M. S. Feldstein, and E. Saez (2010): "Expanding access to administrative data for research in the United States," in *Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas*, American Economic Association.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2022): "Gender differences in peer recognition by economists," *Econometrica*, 90, 1937–1971.

CES (2017): "Center for Economic Studies and Research Data Centers Research Report: 2016," Available on the U.S. Census Bureau's website.

Chow, M. C., T. C. Fort, C. Goetz, N. Goldschlag, J. Lawrence, E. R. Perlman, M. Stinson, and T. K. White (2021): "Redesigning the Longitudinal Business Database," NBER Working Paper w28839.

Currie, J., H. Kleven, and E. Zwiers (2020): "Technology and big data are changing economics: Mining text to track methods," *AEA Papers and Proceedings*, 110, 42–48.

Davis, S. J., J. C. Haltiwanger, and S. Schuh (1998): "Job Creation and Destruction," *MIT Press Books*.

Einav, L. and J. Levin (2014): "Economics in the age of big data," *Science*, 346, 1243089.

Foster, L., R. Jarmin, and L. Riggs (2009): "Resolving the tension between access and confidentiality: Past experience and future plans at the US Census Bureau," *Statistical Journal of the IAOS*, 26, 113–122.

Goroff, D., J. Polonetsky, and O. Tene (2018): "Privacy protective research: Facilitating ethically responsible access to administrative data," *The ANNALS of the American Academy of Political and Social Science*, 675, 46–66.

Greenstone, M., R. Hornbeck, and E. Moretti (2010): "Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings," *Journal of Political Economy*, 118, 536–598.

Haunschild, R. and L. Bornmann (2017): "How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data," *Scientometrics*, 110, 1209–1216.

Heckman, J. J. and S. Moktan (2020): "Publishing and promotion in economics: The tyranny of the top five," *Journal of Economic Literature*, 58, 419–470.

Hopenhayn, H. A. (2014): "Firms, misallocation, and aggregate productivity: A review," *Annual Review of Economics*, 6, 735–770.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 124, 1403–1448.

Jarmin, R. S. and J. Miranda (2002): "The longitudinal business database," CES working paper.

Kalaitzidakis, P., T. P. Mamuneas, and T. Stengos (2003): "Rankings of academic journals and institutions in economics," *Journal of the European Economic Association*, 1, 1346–1366.

Lane, J. (2007): "Optimizing the use of microdata: An overview of the issues," *Journal of Official Statistics*, 23, 299–317.

——— (2021): *Democratizing Our Data: A Manifesto*, MIT Press.

Li, C., J. Ahn, J. Bu, and K. E. Meyer (2023): "The value of publishing in JIBS," *Journal of International Business Studies*, 54, 1688–1699.

Lusher, L. R., W. Yang, and S. E. Carrell (2021): "Congestion on the information superhighway: Does economics have a working papers problem?" NBER Working Paper w29153.

McGuckin, R. H. (1995): "Establishment microdata for economic research and policy analysis: Looking beyond the aggregates," *Journal of Business & Economic Statistics*, 13, 121–126.

McGuckin, R. H., R. H. McGukin, and A. P. Reznek (1993): "The statistics corner: Research with economic microdata: The Census Bureau's Center for Economic Studies," *Business Economics*, 52–58.

Mittal, V., L. Feick, and F. Murshed (2008): "Publish and prosper: The financial impact of publishing by marketing faculty," *Marketing Science*, 27, 430–442.

Moed, H. F., M. Aisati, and A. Plume (2013): "Studying scientific migration in Scopus," *Scientometrics*, 94, 929–942.

Moretti, E. (2021): "The effect of high-tech clusters on the productivity of top inventors," *American Economic Review*, 111, 3328–3375.

Morretta, V., D. Vurchio, and S. Carrazza (2022): "The socio-economic value of scientific publications: The case of Earth Observation satellites," *Technological Forecasting and Social Change*, 180, 121730.

Nagaraj, A., F. Stipanicic, and M. Tranchero (2024): "U.S. Census Data in Economic Research: Adoption, Diffusion, and Impact," UC Berkeley, University of Oslo, and University of Pennsylvania.

Önder, A. S. and S. Schweitzer (2017): "Catching up or falling behind? Promising changes and persistent patterns across cohorts of economics PhDs in German-speaking countries from 1991 to 2008," *Scientometrics*, 110, 1297–1331.

Payne, A. and A. Siow (2003): "Does federal research funding increase university research output?" *Advances in Economic Analysis Policy*, 3, 1018.

Potok, N. (2024): "Data usage information and connecting with data users: US mandates and guidance for government agency evidence building." *Harvard Data Science Review*, Special Issue 4.

Rambachan, A. and J. Roth (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90, 2555–2591.

Rosenbloom, J. L., D. K. Ginther, T. Juhl, , and J. A. Heppert (2015): "The effects of research development funding on scientific productivity: Academic chemistry, 1990-2009," *PloS one*, 10, e0138176.

Sarsons, H. (2015): "Gender Differences in Recognition for Group Work," *Working Paper*.

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021): "Gender differences in recognition for group work," *Journal of Political Economy*, 129, 101–147.

Sattari, R., J. Bae, E. Berkes, and B. A. Weinberg (2022): "The ripple effects of funding on researchers and output," *Science Advances*, 8.

Sauer, R. D. (1988): "Estimates of the returns to quality and coauthorship in economic academia," *Journal of Political Economy*, 96.

Swidler, S. and E. Goldreyer (1998): "The value of a finance journal publication," *The Journal of Finance*, 53, 351–363.

Wahls, W. P. (2018): "High cost of bias: Diminishing marginal returns on NIH grant funding to institutions," *BioRxiv*, 367847.

Whalley, A. and J. Hicks (2014): "Spending wisely? How resources affect knowledge production in universities." *Economic Inquiry*, 52, 35–55.

Yin, Y., Y. Dong, K. Wang, D. Wang, and B. F. Jones (2022): "Public use and public funding of science," *Nature Human Behaviour*, 6, 1344–1350.

H Appendix Tables and Figures

Table H1: Effect of FSRDC Access on Projects Approved and Working Papers

	FSRDC Project Approval (0/1)		CES Working Paper (0/1)	
	(1)	(2)	(3)	(4)
Post-FSRDC	0.00644***	0.00633***	0.00357**	0.00237*
	(0.00135)	(0.00133)	(0.00129)	(0.00113)
Researcher FE	Yes	Yes	Yes	Yes
Year FE	Yes	No	Yes	No
University FE	Yes	No	Yes	No
University Tier × Year FE	No	Yes	No	Yes
N	155615	155622	155615	155622

Note: This table presents estimates from linear probability models evaluating the impact of FSRDC access on additional measures of confidential data adoption. Columns (1) and (2) report the results from linear probability models, where the dependent variable is an indicator of whether the researcher has any new project approved to be carried out in an FSRDC. Columns (3) and (4) report the results from linear probability models, where the dependent variable is the number of CES working papers published. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution with an operating FSRDC. All models include individual fixed effects. Columns (1) and (3) include year fixed effects and university dummies, while columns (2) and (4) include year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively. See details in Appendix C.4.

Table H2: Effect of FSRDC Access on Policy Impact By Country of the Policy Source

	Policy-Relevant Publications (1)	U.S. Policy-Relevant Publications (2)	Non-U.S. Policy-Relevant Publications (3)
Post-FSRDC	0.0362**	0.0242**	0.0304**
	(0.01139)	(0.00834)	(0.01089)
Researcher FE	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes
N	155622	155622	155622
Mean of DV	0.252	0.164	0.224

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on the relevance of economic research to US and Non-US policymaking. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles with at least one cite in a policy document. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles with at least one cite in a US-based policy document. Column (3) reports the result from an OLS model, where the dependent variable is the count of articles with at least one cite in a non-US-based policy document. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the researcher level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively

Table H3: Effect of FSRDC Access on the Policy Orientation of Empirical Researchers

	Mentions of Policy (1)	Policy JELs (all) (2)	Policy JELs (subset) (3)
Post-FSRDC	-0.00125	0.000421	0.00141
	(0.00650)	(0.00176)	(0.00149)
Researcher FE	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes
N	155622	155622	155622

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of policy orientation of empirical researchers. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that mention the words "policy" or "policies" in the title or abstract. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that report JEL codes including the word "policy". Column (3) reports the result from an OLS model, where the dependent variable is the count of articles that report JEL codes most associated with government, labor, and public policies. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the researcher level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively

Table H4: Effect of FSRDC Access on Research Output Using the Callaway-Sant'Anna Doubly-Robust DID Estimator.

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
ATT	0.00474*	0.0236***	0.0195	1.743*	0.0454*
	(0.002)	(0.006)	(0.016)	(0.854)	(0.021)
Observations	150,567	150,567	150,567	150,567	150,567

Note: This table presents ATT estimates from models evaluating the impact of FSRDC access on measures of confidential data adoption and research output using the doubly-robust difference-in-difference estimator developed by Callaway and Sant'Anna (2021). Column (1) reports the result from a model where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from a model where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from a model where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan, 2020). Column (4) reports the result from a model where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Column (5) reports the result from a model where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC (ATT) equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects, year fixed effects, and university fixed effects. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H5: Comparing Empiricists with Theorists (Within-University Variation)

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC × Empiricist	0.00852***	0.0251***	0.0445***	2.294***	0.0549***
	(0.00178)	(0.00479)	(0.00914)	(0.53370)	(0.01320)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University × Year FE	Yes	Yes	Yes	Yes	Yes
N	232145	232145	232145	232145	232145

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data adoption and research output using theorists at the same university as the control group. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received during the five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC: 0/1 = 1 after a researcher has been affiliated to a research institution with an operating FSRDC. Empiricist: 0/1 = 1 for researchers whose first five publications are mostly empirical in nature, as assessed by our machine-learning based paper classifier. All models include individual fixed effects and year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H6: Comparing Empiricists in Differentially Impacted Fields (Within-University Variation)

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC × Impacted Field	0.0107***	0.0309***	0.0356*	2.773***	0.0165
	(0.00204)	(0.00675)	(0.01511)	(0.81346)	(0.01748)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University × Year FE	Yes	Yes	Yes	Yes	Yes
N	152988	152988	152988	152988	152988

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data adoption and research output for empiricists in fields impacted by FSRDC data, using empiricists in less impacted fields but at the same university as the control group. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received during the five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC: 0/1 = 1 after a researcher has been affiliated to a research institution with an operating FSRDC. Impacted Field: 0/1 = 1 if the majority of the researcher's work is classified under environmental economics, international economics, labor economics, applied microeconomics, or public economics, as assessed by our machine-learning based paper classifier. All models include individual fixed effects and year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H7: Comparing Empiricists to Other Members of the Same Consortium (Within-Consortium Variation)

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC	0.00605**	0.0167*	0.0252**	1.306*	0.0222
	(0.00203)	(0.00711)	(0.00908)	(0.61967)	(0.01838)
Researcher FE	Yes	Yes	Yes	Yes	Yes
Consortium × Year FE	Yes	Yes	Yes	Yes	Yes
N	63096	63096	63096	63096	63096

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data adoption and research output, using empiricists in non-hosting institutions but in the same consortium as the control group. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received during the five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC: 0/1 = 1 after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with consortium dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H8: Effect of FSRDC Access for Universities

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC	0.208**	0.883**	1.718**	69.90*	3.055**
	(0.07831)	(0.28269)	(0.65360)	(35.43269)	(1.11295)
University FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	8323	8323	8323	8323	8323

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data adoption and research output at the university level. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan, 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Column (5) reports the result from an OLS model where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC equals one in all years after the university has an operating FSRDC. All models include university fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H9: Effect of FSRDC Access Excluding NSF Grant Applicants

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC	0.00589***	0.0104**	0.0219**	1.205**	0.0359**
	(0.00139)	(0.00386)	(0.00811)	(0.41970)	(0.01108)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	155039	155039	155039	155039	155039

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data adoption and research output, excluding researchers who were the applicants of the NSF grant leading to an FSRDC. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H10: Effect of FSRDC Access Excluding Authors in Institutions Hosting FSRDCs

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Within 10 Miles of FSRDC	0.00300**	0.00951**	0.0135*	0.973*	0.0223**
	(0.00091)	(0.00307)	(0.00660)	(0.41020)	(0.00832)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	122695	122695	122695	122695	122695

Note: This table presents estimates from OLS models evaluating the impact of FSRDC proximity on measures of confidential data adoption and research output, excluding researchers directly affiliated with an institution with an operating FSRDC center. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received during the five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Within 10 miles: 0/1 = 1 after a researcher has been affiliated with a research institution located within 10 miles of an operating FSRDC but not hosting it. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H11: Alternative Definitions to Measure FSRDC Access

	Top Pubs	Cite-weighted	Policy-relevant	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)	(6)
Post-FSRDC (City)	0.0146*	0.885*	0.0232**			
	(0.00637)	(0.39640)	(0.00794)			
Post-FSRDC (Consortium)				0.00653	0.961*	0.00742
				(0.00621)	(0.39377)	(0.00806)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	155622	155622	155622	155622	155622	155622

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research output using alternative definitions of access. Columns (1) and (4) report the results from OLS models, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan, 2020). Columns (2) and (5) report the results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Columns (3) and (6) report the results from OLS models, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC (City): 0/1 = 1 after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Post-FSRDC (Consortium): 0/1 = 1 after a researcher has been affiliated to a research institution belonging to a consortium operating an FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H12: Effect of FSRDC Access on Theorists

	Using FSRDC (1)	Citing FSRDC (2)	Top Pubs (3)	Cite-weighted (4)	Policy-relevant (5)
Post-FSRDC	-0.000489	0.00252	-0.00131	0.563	0.0104
	(0.00111)	(0.00400)	(0.01423)	(0.63503)	(0.00853)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	77538	77538	77538	77538	77538

Note: This table presents estimates from an OLS model evaluating the impact of FSRDC access on measures of confidential data adoption and research output when looking at a sample of theorists only. Column (1) reports the result from an OLS model, where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received during the five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H13: Effect of FSRDC Access Excluding Authors Treated After Mobility Events

	Using FSRDC	Citing FSRDC	Top Pubs	Cite-weighted	Policy-relevant
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC	0.00593***	0.0122*	0.0277**	1.183**	0.0297*
	(0.00153)	(0.00491)	(0.00922)	(0.42848)	(0.01375)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	146491	146491	146491	146491	146491

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of confidential data adoption and research output excluding researchers who were treated after moving to an institution with an operating FSRDC (i.e., considering only researchers treated by the FSRDC opening). Column (1) reports the result from an OLS model where the dependent variable is the count of articles that directly use confidential data accessible only in an FSRDC. Column (2) reports the result from an OLS model, where the dependent variable is the count of articles that cite a paper using confidential data accessible only in an FSRDC (excluding papers that directly use confidential FSRDC data). Column (3) reports the result from an OLS model, where the dependent variable is the count of articles published in the most prestigious economics journals (the list of top five and top field journals is from Heckman and Moktan 2020). Column (4) reports the result from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received during the five years following their publication. Column (5) reports the result from an OLS model, where the dependent variable is the count of articles cited in at least one policy document. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H14: Effect of FSRDC Access on University Hiring and Researchers' Composition

	Net Inflow	Net Inflow (Stars)	Net Inflow (Superstars)	Share Empiricists	Difference in Pubs
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC	-0.0926	-0.0415	-0.0442	-0.0196	-0.146
	(0.18602)	(0.04727)	(0.02537)	(0.02009)	(0.19047)
University FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	7917	7917	7917	8323	8323

Note: This table presents estimates from an OLS model at the university level that tests whether universities experience changes in faculty composition after establishing an FSRDC. Column (1) reports the result from an OLS model at the university level, where the dependent variable is the number of researchers that join the faculty, minus the number of researchers that leave the faculty. Column (2) reports the result from an OLS model, where the dependent variable is equal to the number of researchers in the top 10% of cumulative citations up to that year ("stars") that join the faculty, minus the number of stars that leave the faculty. Column (3) reports the result from an OLS model, where the dependent variable is equal to the number of researchers in the top 5% of cumulative citations up to that year ("superstars") that join the faculty, minus the number of superstars that leave the faculty. Column (4) reports the result from an OLS model at the university level, where the dependent variable is the share of empiricists out of all economists affiliated with the university. Column (5) reports the result from an OLS model, where the dependent variable is the rolling difference in average lifetime top publications between new arrivals and incumbent faculty (to test whether new hires are more accomplished than standing faculty). Post-FSRDC equals one in all years after a university has an operating FSRDC. All models include university fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H15: Effect of FSRDC Access on Alternative Measures of Research Direction

	Top 20 JELs (0/1)	Top 40 JELs (0/1)	Top 60 JELs (0/1)	Top 100 JELs (0/1)	Top 200 JELs (0/1)
	(1)	(2)	(3)	(4)	(5)
Post-FSRDC	0.0139*	0.0147*	0.0138	0.0143	0.0138
	(0.00590)	(0.00736)	(0.00749)	(0.00850)	(0.00868)
Researcher FE	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes	Yes
N	155622	155622	155622	155622	155622
Mean of DV	0.149	0.201	0.228	0.271	0.304

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research direction. We experiment with increasingly less stringent definitions of what constitutes a JEL code associated with FSRDC papers. Column (1) reports the result from an OLS model, where the dependent variable is an indicator equal to one if the researcher used at least one FSRDC JEL code during the year, which for this model we consider to be any of the top 20 JEL codes that appear with the greatest frequency in FSRDC papers. Column (2) uses the same model, but we consider any of the top 40 JEL codes that appear with the greatest frequency in FSRDC papers. The remaining columns are specified analogously. Post-FSRDC: 0/1 = 1 after a researcher has been affiliated with a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively

Table H16: Effect of FSRDC Access on External Grant Funding

	Funding Sources (1)	Funding Sources (0/1) (2)	New Funding Sources (3)	New Funding Sources (0/1) (4)
Post-FSRDC	0.00384	0.00356	0.00478	0.00373
	(0.00723)	(0.00330)	(0.00532)	(0.00276)
Researcher FE	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes
N	155622	155622	155622	155622

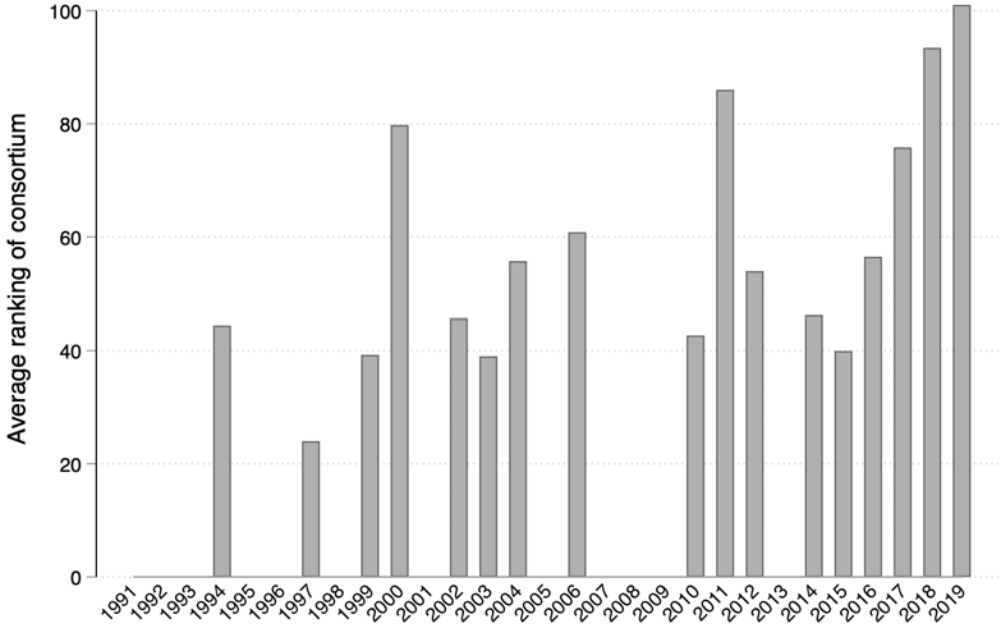
Note: This table presents estimates from OLS models evaluating whether researchers report additional funding sources after gaining access to an FSRDC. Column (1) reports the result from an OLS model, where the dependent variable is the count of funding sources acknowledged in all the publications of a given year. Column (2) reports the result from a linear probability model, where the dependent variable is equal to one if the researcher acknowledges any funding source in a given year. Column (3) reports the result from an OLS model, where the dependent variable is the count of new funding sources acknowledged in all the publications of a given year. Column (4) reports the result from a linear probability model, where the dependent variable is equal to one if the researcher acknowledges any new funding source in a given year. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. All models include individual fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively.

Table H17: Effect of FSRDC Access on Patterns of Collaboration

	Mean Number of Co-Authors	Mean Number of Unique Co-Authors	Mean Number of New Co-Authors	Mean Number of Team-Papers
	(1)	(2)	(3)	(4)
Post-FSRDC	0.105	0.101	0.0421	0.0310
	(0.07287)	(0.06535)	(0.05163)	(0.02069)
University FE	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	Yes	Yes	Yes
N	8323	8323	8323	8323

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of collaboration among researchers at the university level. Column (1) reports the result from an OLS model, where the dependent variable is the mean number of co-authors for faculty members who publish at least one paper that year. Column (2) reports the result from an OLS model, where the dependent variable is the mean number of unique co-authors for faculty members who publish at least one paper that year. Column (3) reports the result from an OLS model, where the dependent variable is the mean number of new co-authors for faculty members who publish at least one paper that year. Column (4) reports the result from an OLS model, where the dependent variable is the mean number of team papers with four or more authors for faculty members who publish at least one paper that year. Post-FSRDC equals one in all years after the university has an operating FSRDC. All models include university fixed effects and year fixed effects interacted with university tier dummies. Standard errors are in parentheses, clustered at the university level. *, **,*** denote significance at 5%, 1% and 0.1% level respectively

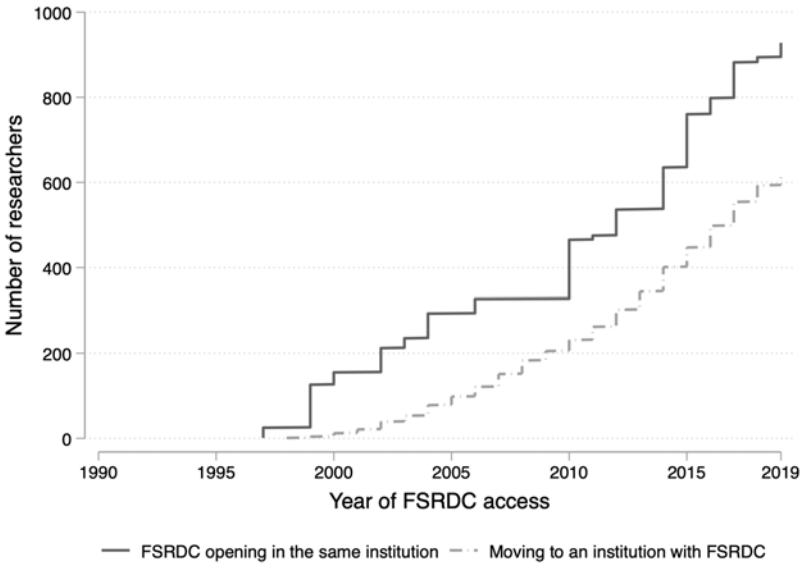
Figure H1: Average Rank of Consortium Members By Time of FSRDC Opening



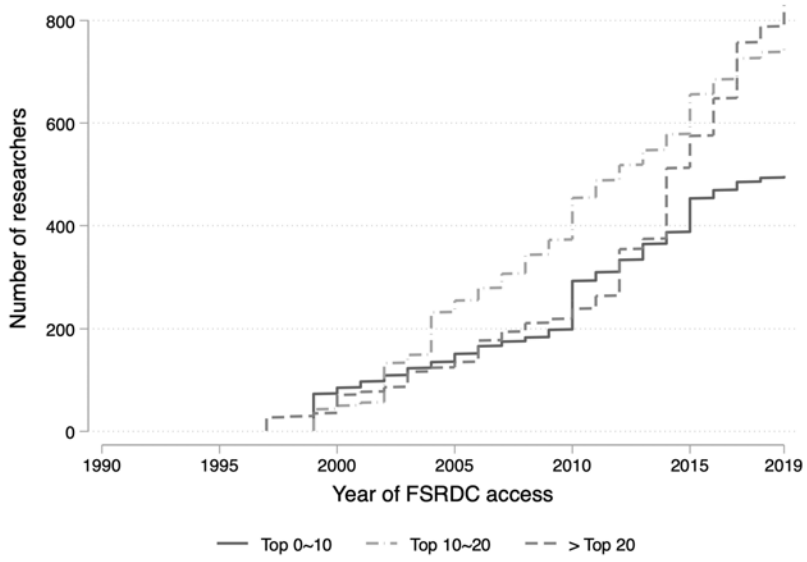
Note: This figure reports the average rank of all members in each FSRDC funding consortium, which is then averaged across all consortia that opened an FSRDC in a given year. When calculating the average rank of each consortium, we include all universities that were a member at any point during the sample period. Information on the ranking of economics departments is taken from Kalaitzidakis et al. (2003). Since Kalaitzidakis et al. (2003) only report information for the top 100 North American institutions, we assign the same residual rank of 101 to all other institutions in our sample. Hence, lower numbers correspond to departments ranked higher. See text for more details.

Figure H2: Cumulative Number of Researchers Gaining FSRDC Access over Time

(i) By Modality of Researchers' Access



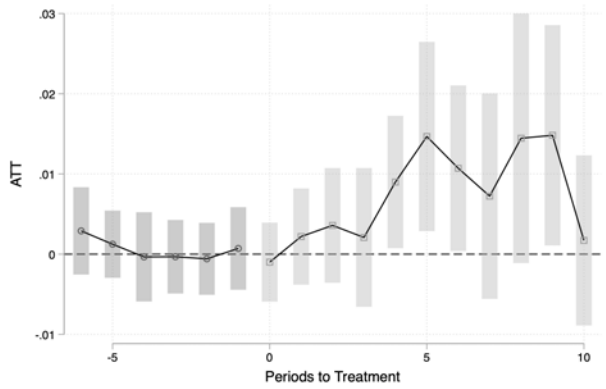
(ii) By Ranking of Researchers' Institution



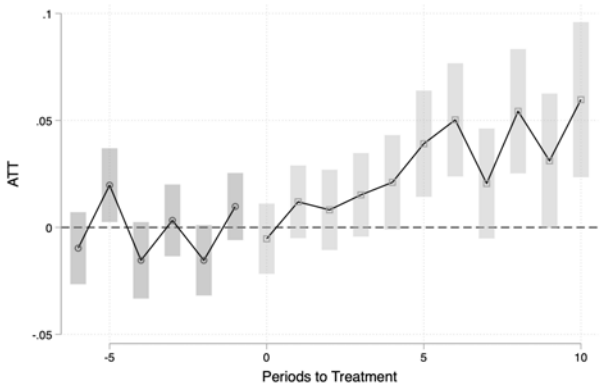
Note: This figure plots the cumulative number of researchers who gain access to an FSRDC through a hosting institution. The cumulative frequencies are split by the modality in which researchers become co-located to an FSRDC (panel (i)) and the ranking of their institution of affiliation (panel (ii)). See text for more details.

Figure H3: Time-Varying Estimates of the Impact of FSRDCs on Research Output Using the Callaway-Sant'Anna Doubly-Robust DID Estimator.

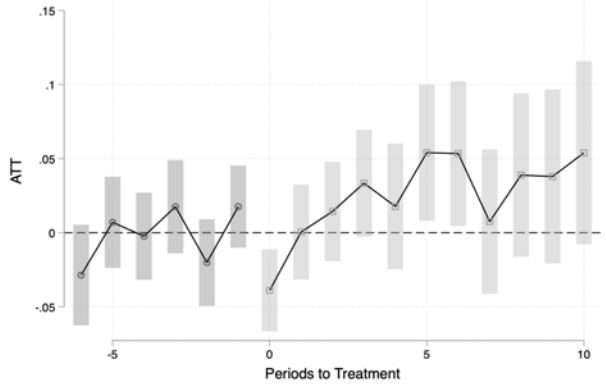
(i) Papers Using FSRDC Data



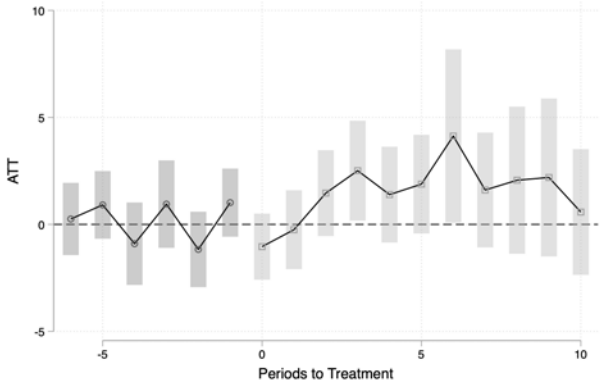
(ii) Papers Citing FSRDC Papers



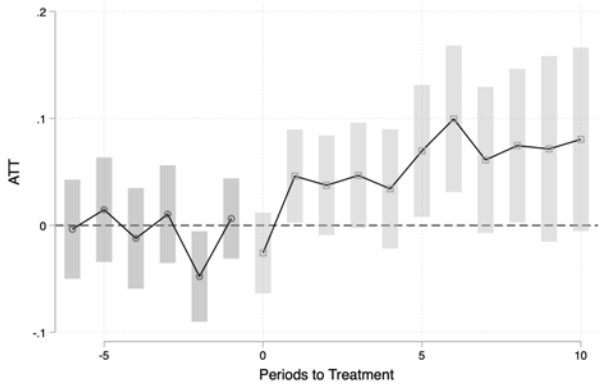
(iii) Top Publications



(iv) Cite-Weighted Publications



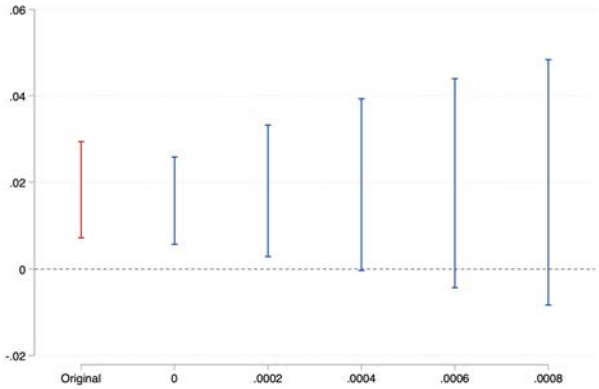
(v) Policy-Relevant Publications



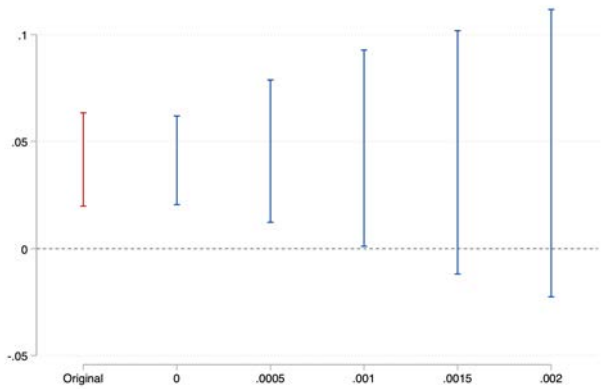
Note: This figure provides visual illustrations of the event study version of the main regressions evaluating the impacts of FSRDC access on measures of research output, estimated using the doubly-robust difference-in-difference estimator developed by Callaway and Sant'Anna (2021). The main dependent variables are the number of papers written using FSRDC data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), the citation-weighted number of publications (Panel (iv)), and the number of policy-relevant publications (Panel (v)). The charts plot values of β for different lags before and after obtaining access to an FSRDC with 95% confidence intervals. Regressions include individual, university, and year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure H4: Sensibility to Violations of Parallel Trends using the HonestDiD Approach from Rambachan and Roth (2023)

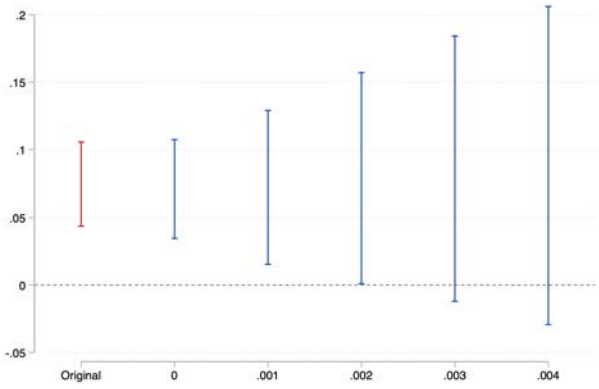
(i) Papers Using FSRDC Data



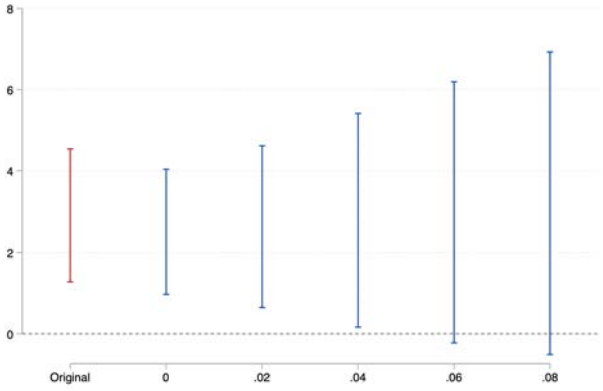
(ii) Papers Citing FSRDC Papers



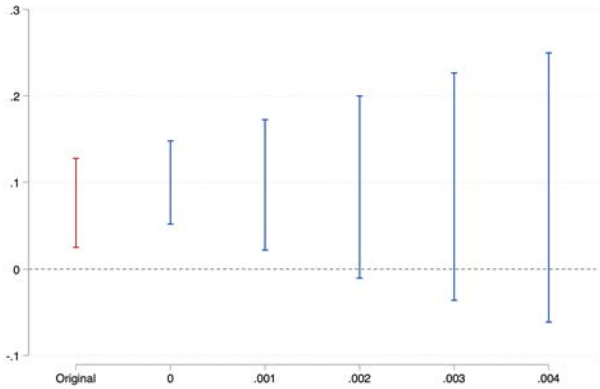
(iii) Top Publications



(iv) Cite-Weighted Publications



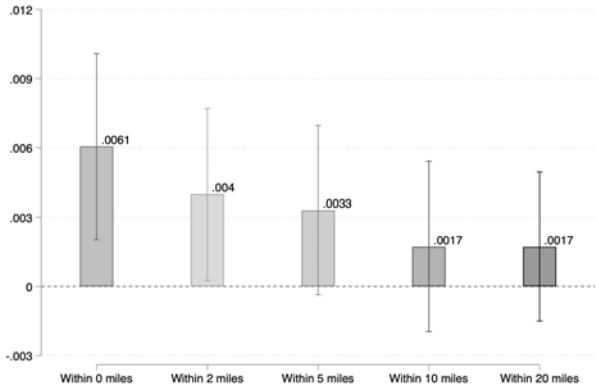
(v) Policy-Relevant Publications



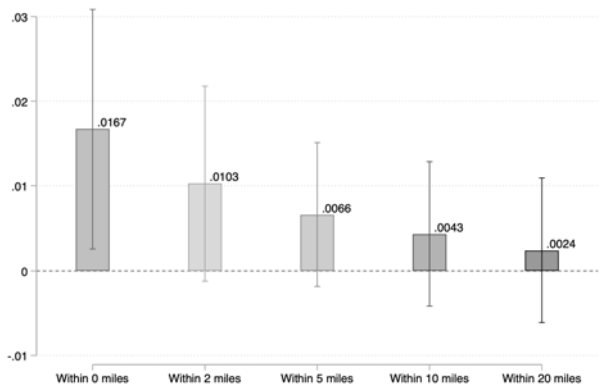
Note: In this figure, we implement the sensitivity analysis from Rambachan and Roth (2023) to test whether our results are robust to violations of parallel trends. In particular, we apply the "smoothness" restriction, which allows for a linear extrapolation of pre-trends to be off by some constant M. For each panel, we report coefficients and 95% confidence intervals for our original specification (in red), for M=0 (which corresponds to a perfectly linear extrapolation of pre-trends), and for progressively larger values of M (which allow for greater deviations from linearity). The main dependent variables are the number of papers written using FSRDC data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), the citation-weighted number of publications (Panel (iv)), and the number of policy-relevant publications (Panel (v)). All models include individual and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure H5: Effect of FSRDC Access for Consortium Members by Distance (Within Consortium Variation)

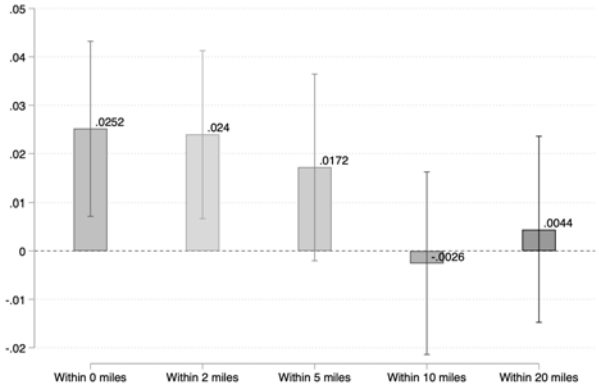
(i) Papers Using FSRDC Data



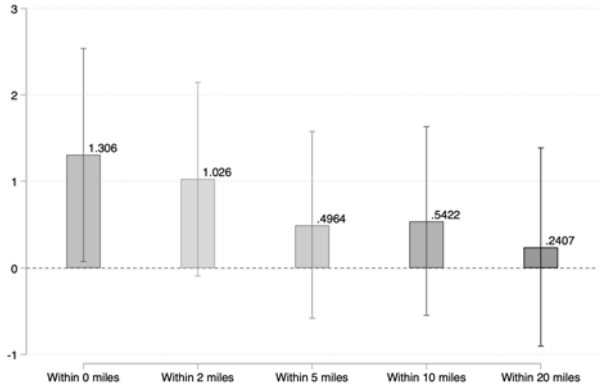
(ii) Papers Citing FSRDC Papers



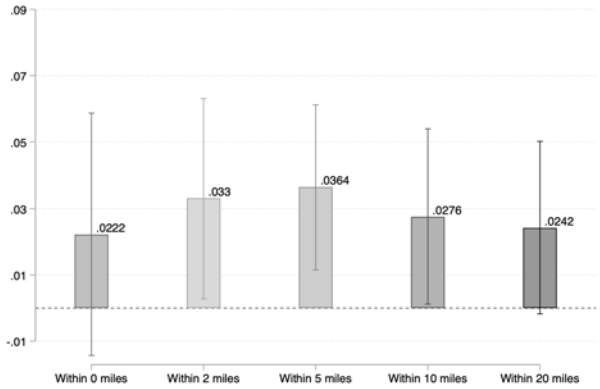
(iii) Top Publications



(iv) Cite-Weighted Publications



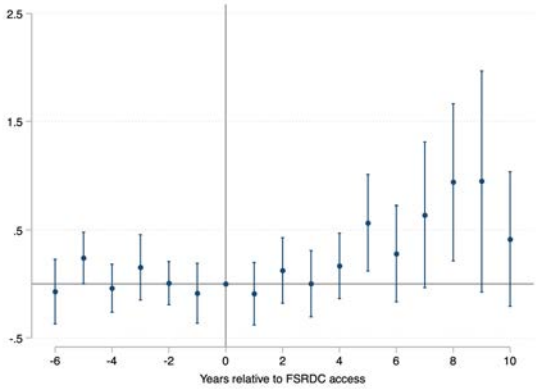
(v) Policy-Relevant Publications



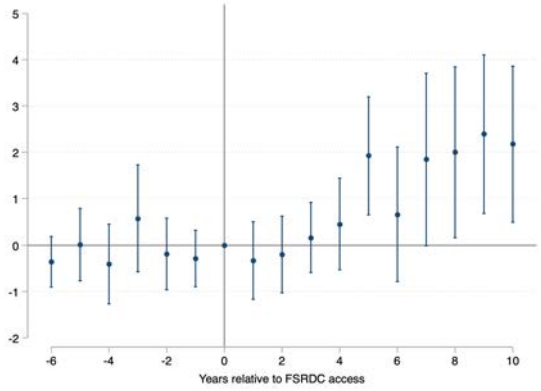
Note: This figure provides visual illustrations of the effect of data access at different distances, using empiricists in non-hosting institutions but in the same consortium as the control group. Distance is the geometric distance between the institution of the researcher and the closest operating FSRDC. The main dependent variables are the number of papers written using FSRDC data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), the citation-weighted number of publications (Panel (iv)), and the number of policy-relevant publications (Panel (v)). The chart plots values of β from different regressions where researchers are considered as treated if they are affiliated to institutions within a progressively larger radius from an FSRDC. All models include university and consortium $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure H6: University-Level Time-Varying Estimates of the Impact of FSRDCs

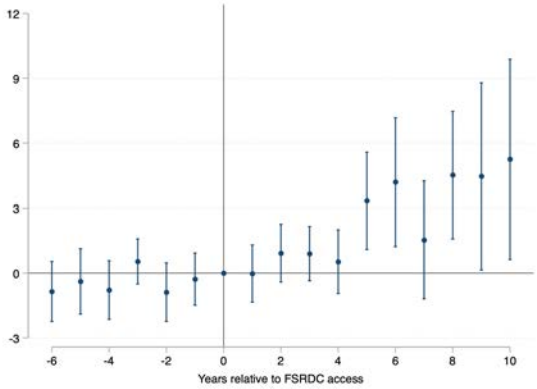
(i) Papers Using FSRDC Data



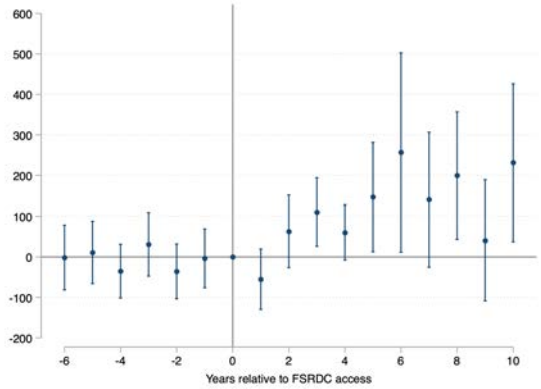
(ii) Papers Citing FSRDC Papers



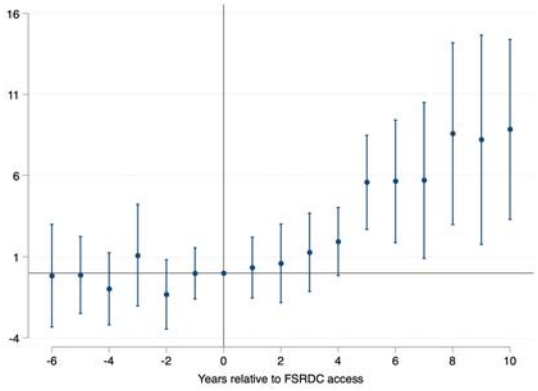
(iii) Top Publications



(iv) Cite-Weighted Publications



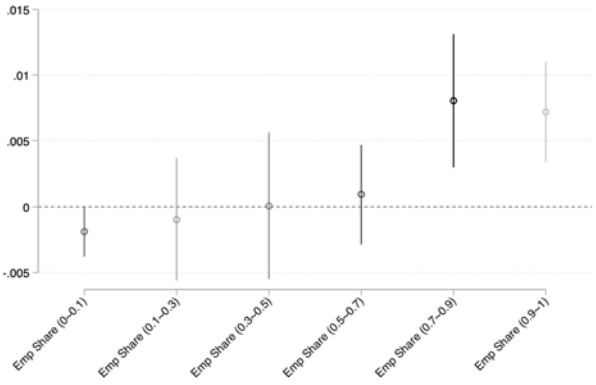
(v) Policy-Relevant Publications



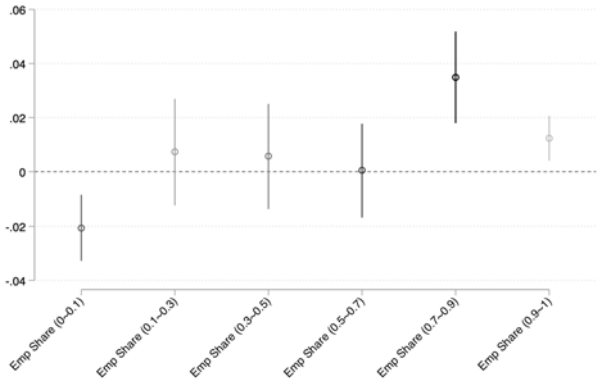
Note: This figure provides visual illustrations of the event study version of the regression evaluating the impact of FSRDC access on measures of research output at the university level. The main dependent variables are the number of papers written using FSRDC data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), the citation-weighted number of publications (Panel (iv)), and the number of policy-relevant publications (Panel (v)). All models include university and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure H7: Effect of FSRDC Access by Share of Empirical Work

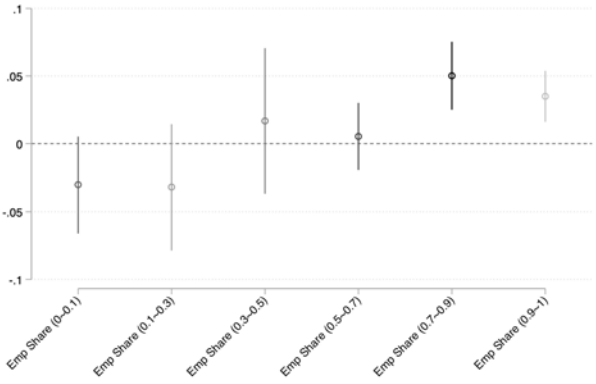
(i) Papers Using FSRDC Data



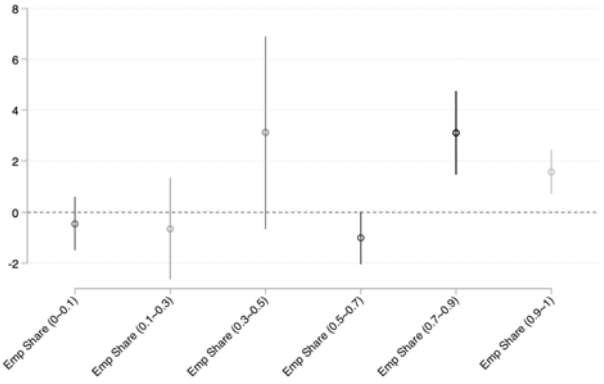
(ii) Papers Citing FSRDC Papers



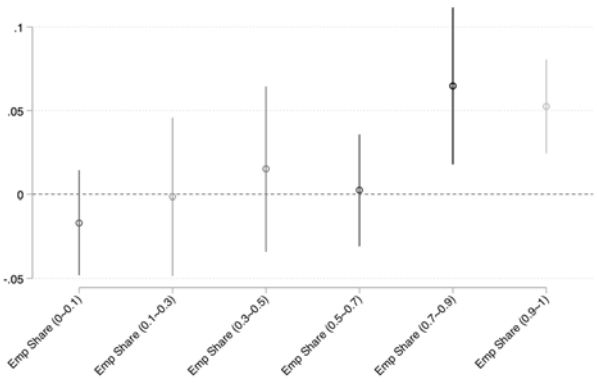
(iii) Top Publications



(iv) Cite-Weighted Publications



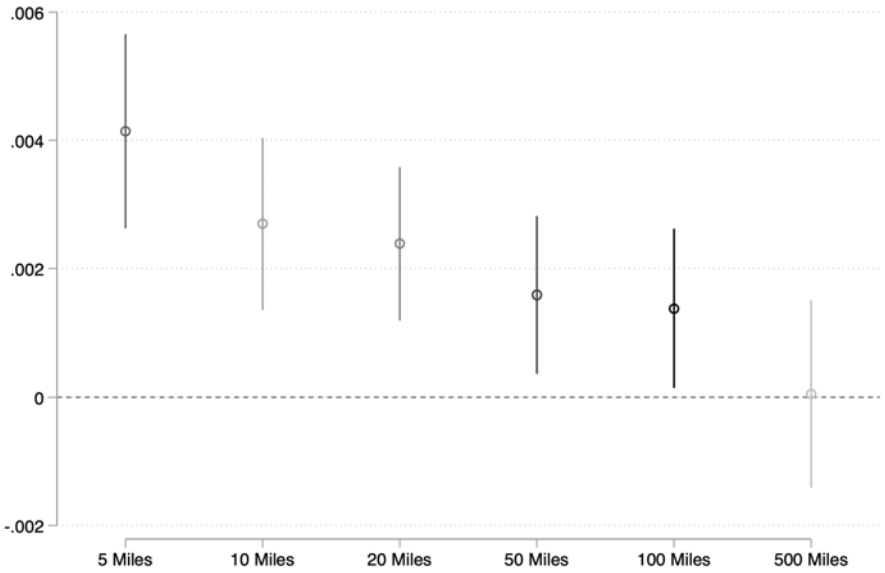
(v) Policy-Relevant Publications



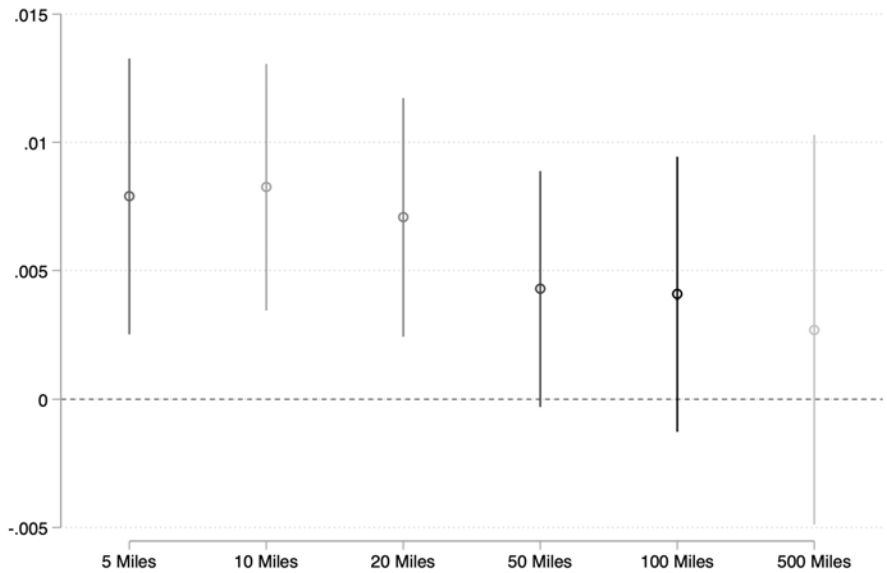
Note: This figure provides visual illustrations of the effect of data access for researchers with different methodological orientations. Empirical share is the proportion of a researcher's first five publications that is empirical in nature. The main dependent variables are the number of papers written using FSRDC data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), the citation-weighted number of publications (Panel (iv)), and the number of policy-relevant publications (Panel (v)). Each chart is estimated from a single interaction-based regression, where each researcher is classified into one of six mutually exclusive categories. All models include individual and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure H8: Effect of FSRDC Access on Confidential Data Diffusion by Distance

(i) Papers Using FSRDC Data



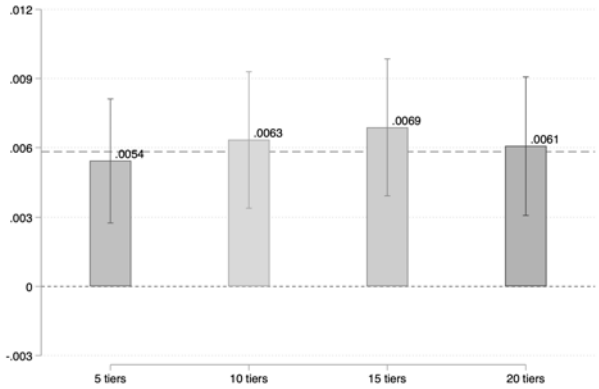
(ii) Papers Citing FSRDC Papers



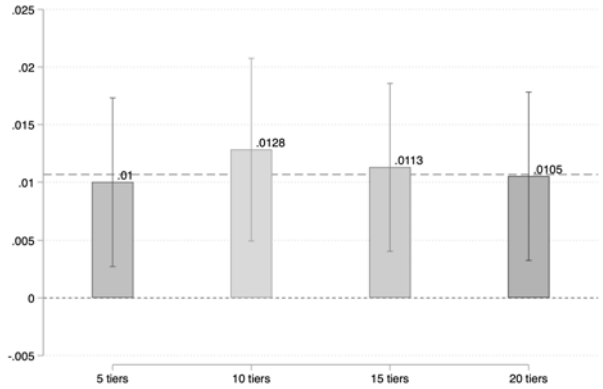
Note: This figure provides visual illustrations of the effect of data access on confidential data diffusion at different distances. Distance is the geometric distance between the institution of the researcher and the closest operating FSRDC. The main dependent variables are the number of papers written using FSRDC data (panel (i)) and the number of papers that cite FSRDC papers (panel (ii)). The chart plots values of β from different regressions where researchers are considered as treated if they are affiliated to institutions within a progressively larger radius from an FSRDC. Regressions include individual and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.

Figure H9: Sensitivity to Alternative Definitions of University Ranking Tiers Fixed Effects

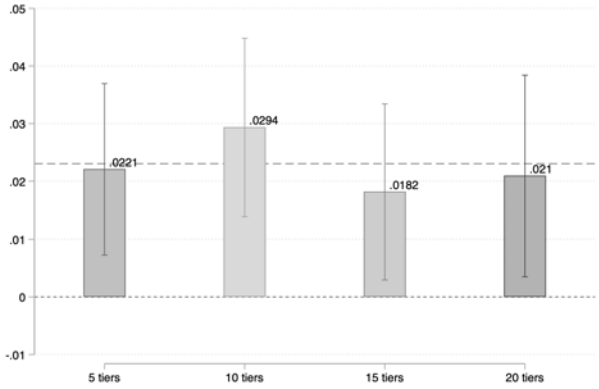
(i) Papers Using FSRDC Data



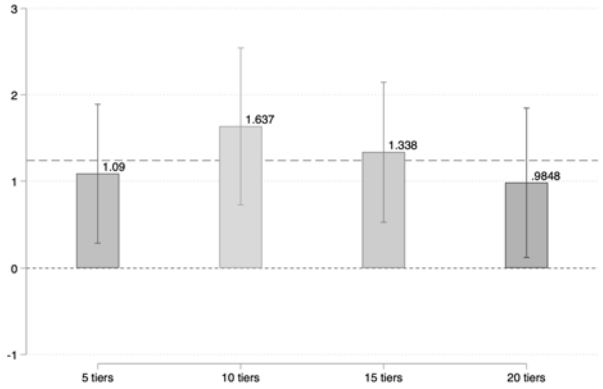
(ii) Papers Citing FSRDC Papers



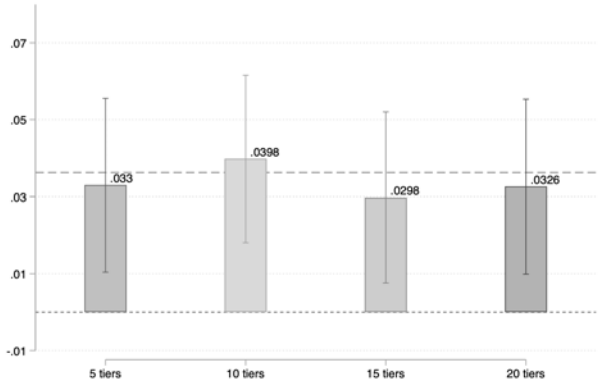
(iii) Top Publications



(iv) Cite-Weighted Publications



(v) Policy-Relevant Publications



Note: This figure provides visual illustrations of the effect of data access when we alter the definition of the university tiers used in our fixed effects structure. Unlike our main specification, where we use tiers of equal size (see B.2 for more details), we show robustness to alternative arbitrary categorizations. We construct the tiers by dividing the 100 ranked universities into equal groups, beginning with five ranking tiers and incrementing by 5 until there are 20 tiers, while keeping researchers affiliated with foreign universities or U.S.-based research institutions not covered by the ranking in two residual tiers. The chart plots values of β from different regressions where we redefine the university tier $\times$ year FEs. We also include the main estimate as a dashed horizontal line. The main dependent variables are the number of papers written using FSRDC data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), the citation-weighted number of publications (Panel (iv)), and the number of policy-relevant publications (Panel (v)). All models include individual and university tier $\times$ year fixed effects. Standard errors are clustered at the university level. See text for more details.