

HOW DOES DATA ACCESS SHAPE SCIENCE?
EVIDENCE FROM THE IMPACT OF U.S. CENSUS'S
RESEARCH DATA CENTERS ON ECONOMICS RESEARCH *

Abhishek Nagaraj Matteo Tranchero
UC Berkeley-Haas UC Berkeley-Haas
nagaraj@berkeley.edu tranchero@berkeley.edu

February 10, 2023

Abstract

This paper studies the impact of access to confidential administrative data on the rate and direction of economics research. We estimate how the progressive geographic expansion of the U.S. Census Bureau's Federal Statistical Research Data Centers (FSRDCs) shapes the research productivity of affected economists. We find that increased access to administrative data increases its diffusion in local universities and helps empirical researchers publish 24% more articles in top-ranked outlets and increase citation-weighted publication metrics by 40%. These effects are driven both by the direct adoption of administrative data as well as spillovers to non-adopters through increased awareness of Census data-based research. Spillovers operate by leading researchers to explore a new set of topics using more robust designs and accrue only to those citing research using Census data. Our work highlights the role of data access in shaping scientific advances in the social sciences and informs the design of data infrastructures to support evidence-based policies.

*We thank Ryan Hill for sharing code and training data to classify the style of economics articles. We also thank JiYoo Jeong, Jai Singh, Brian Qi, and especially Randol Yao 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 NBER Summer Institute, Haas Macro-MORS Research Lunch, 2022 Research Data Center Annual Conference in Kansas City, and Workshop on Big Data Analyses and New Developments in Research Data Centers at ZEW Mannheim as well as to Wayne Gray, Lucia Foster, Jeff Furman, Bronwyn Hall, Julie Hotchkiss, and Bill Kerr for their feedback on this work. Any opinions and conclusions expressed herein are those of the authors only, and any errors are our own.

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, 2021; Berman et al., 2000; 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). An important driver of this increase has been the use of confidential administrative from data government sources; approximately 20% of articles recently published in the five most prestigious economics journals employ data derived from administrative sources (Currie et al., 2020).

Administrative data are unique in enabling evidence-based policies on a broad range of economic and social phenomena (Cole et al., 2020). For this reason, academic economists have urged for the further diffusion of large-scale microdata from administrative sources (Card et al., 2010; Chetty, 2012). This call comes despite significant privacy risks from the dissemination of these data and the accompanying need for tight security standards (Foster et al., 2009; Abowd and Lane, 2004). However, merely increasing the availability of administrative data is not guaranteed to result in more policy-relevant research. Researchers might find the confidentiality and security requirements too cumbersome and may refrain from adopting such data in their work. Or, even if they do adopt, the urge to use administrative data might become an end in itself, leading to the pursuit of projects of lower impact—a trend where articles are written not because “there is a problem that needs to be solved, or a puzzle that needs to be explained” but because of coming upon an “enormous bunch of data,” in the words of Bob Solow (Dizikes, 2019). Therefore, exactly how access to data can shape the development of policy-oriented research, especially when accompanied by tight access restrictions, remains an open question.

We study how access to confidential administrative data from the U.S. Census shapes the quantity, quality, and direction of economic research. We define these “Census data” to include datasets created by the U.S. Census such as the Longitudinal Business Database (LBD) and the Longitudinal Employer and Household Dynamics (LEHD) database (Abowd et al., 2009; Jarmin and Miranda, 2002) as well as microdata made available by other agencies, such as the Bureau of Labor Statistics (BLS) and the National Center for Health Statistics (NCHS) to the U.S. Census. Our focus on Census data is motivated by the fact that they are perhaps the pre-eminent source of administrative data in the United States. Further, 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 30 FSRDCs were set up all over the country in a phased manner between 1994 and 2019. As we will show through multiple

means, the timing and location of these openings was driven by factors governing geographic equity rather than by pre-existing trends in the use of administrative data or research output. The focus on Census data therefore provides a natural experiment to identify the effects of administrative 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 Census data. Our central regressions estimate the effects of FSRDC access on empirical researchers as compared to theoretical researchers affiliated with the same institution. We include both researcher and university \times year fixed effects in our preferred specification, effectively controlling for time-invariant researcher quality and time-varying university trends that might correlate with the opening of an FSRDC. We first find that local access is critical for the diffusion of Census data. Even though such data could be accessed by collaborating with those who already have access to them or traveling to another city with an FSRDC, the opening of an FSRDC in the same city increases local access by 111%–131% over the sample mean. Interestingly, we also find a large impact on the likelihood of citing past work based on Census data. This suggests that an FSRDC opening raises awareness about empirical results based on Census data, potentially shaping future research trajectories.

Next, we explore how data access affects the productivity of empirical researchers. We find that treated empiricists produce around 24% more publications in top-ranked outlets. This result becomes stronger when considering citation-weighted publications, suggesting that administrative data are particularly useful to boost the quality of empiricists’ output. Likewise, the results grow larger when considering right-tail outcomes, such as the likelihood of publishing in a top five journal or authoring a highly cited article. Additional analyses show that the increase in research quality is mostly confined to senior economists affiliated with high-status universities.

Our results are robust to a host of potential concerns. First, time-varying specifications of our difference-in-difference (DID) regressions, as well as accounting for heterogeneous treatment effects over time, confirm the validity of our research design (Callaway and Sant’Anna, 2021). Additional university-level analyses show that FSRDCs are not systematically opened in institutions on a rising trend of research intensity. The results are also robust to excluding researchers who become treated after gaining employment in a city with a local FSRDC, ruling out that this type of sorting drives our results. Second, we experiment with various alternative specifications of “exposure” to administrative data. Our results get stronger the closer an economist is to a data center, being the highest for researchers who enjoy access on their campus. Similarly, the effects of access grow monotonically when we use increasingly stringent definitions of empirical researchers. Finally, we control for empiricist-specific time trends to account for the concern that our results simply reflect general

trends in economic research and find that our results are largely robust (Backhouse and Cherrier, 2017).

We investigate the mechanisms behind these findings. Our results are only partly driven by researchers who directly adopt administrative data, hinting at significant spillovers from data access to those who do not use an FSRDC directly. However, these results are restricted only to those economists who have colleagues who use Census data or those who cite papers based on Census data, but they disappear for departments with no adoption or for authors who do not cite FSRDC-enabled work. These patterns suggest that local data access shapes academic output by exposing economists to research based on administrative data. The increased awareness about such research might inspire a change in research direction or the adoption of research designs that lead to more impactful work.

Our results are suggestive of such a mechanism at play. First, we find that empirical researchers are more likely to explore new topics rather than doubling down on topics they were already working on. In particular, there is an increase in the likelihood of working on topics commonly associated with using Census data. Second, using keyword searches in the abstracts (Currie et al., 2020), we find that treated researchers increase mentions of administrative datasets and quasi-experimental methods, such as DID or natural experiments. We do not find similar increases in the mentions of survey data or laboratory experiments. Taken together, it seems that the opening of an FSRDC boosts the output of treated researchers by leading them to explore a newer set of topics using robust methodologies.

Our work contributes to three different strands of research. New legislation in the U.S. (like the Evidence-Based Policy Act, the OPEN Government Data Act, and the Federal Data Strategy¹) is putting data policy at the heart of economic policy-making. The assumption is that providing restricted-access confidential administrative data will lead to the development of policy-relevant economics research (Lane, 2021). Yet, while some have hailed current data access programs as important, others have criticized them for being costly and cumbersome, with uncertain impact (Atroistic, 2007; Card et al., 2010; CES, 2017; Cole et al., 2020). Ironically, most of these debates have themselves been data free. We contribute by providing, to our knowledge, the first systematic investigation of the effects of administrative data access on academic research. Our paper documents the use and downstream impacts of access to FSRDC facilities. Insofar as academic insights are a key driver of government policy (Hjort et al., 2021; Yin et al., 2021), our work highlights how data access, even under restrictive conditions, can have a positive impact on research output and potentially on economic policy itself.

Second, we add to recent research that has investigated questions of relevance to the economics profession using bibliometric data. This includes past work that has documented the empirical turn taken by economics (Backhouse and Cherrier, 2017; Brodeur et al., 2020; Currie et al., 2020; Einav and Levin, 2014a; Hamer-

¹<https://strategy.data.gov>

mesh, 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 also contribute to the growing research on labor markets for economists, including research on status dynamics, credit attribution, and editorial roles (Card et al., 2020, 2022; Feenbergs et al., 2017; Heckman and Moktan, 2020; Sarsons et al., 2021). Our results hint that democratizing access to data alone might not be enough to level a field characterized by large inequities in status and resources. In fact, data access seems to reinforce the advantage of established researchers instead of leveling the field.

Finally, we contribute to the economics of the scientific process more broadly (Azoulay et al., 2019; Jones, 2009; Hill and Stein, 2020, 2021; 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 Teodidis, 2020; Murray et al., 2016; Waldinger, 2016). While this work has primarily focused on how access to research material shapes academic output in the basic sciences, it has rarely examined access to data, whose importance to research warrants more careful examination (Hill et al., 2020). We add to a burgeoning literature in this space by investigating how data access shapes the rate and direction of scientific innovation (Hoelzemann et al., 2022; Nagaraj et al., 2020; Nagaraj, 2022; Williams, 2013).

The rest of the paper proceeds as follows. Section 2 provides some background on administrative data and the FSRDC program in the United States. Sections 3 and 4 describe our key data sources and research design. Section 5 presents the key findings, while Section 6 explores the mechanisms that drive them. Section 7 concludes with a discussion of the implications and limitations of our findings.

2 Empirical Setting

2.1 Administrative Data and Economics Research

Administrative data can be defined broadly as any record not originally collected for research purposes (Cole et al., 2020; Goroff et al., 2018; Groves, 2011). Government agencies are big sources of such data, routinely storing information during their normal functioning. Typical examples include unemployment insurance claims, Medicare data, or tax records. Other agencies, such as the U.S. Census Bureau, collect information via statistical surveys and Census enumerations as part of their mandate to assemble timely data about the nation’s demographic and economic trends (Foster et al., 2009). Both of these types of data encompass large samples with a degree of granularity that allows tracking individual units over time (Einav and Levin, 2014b). Unlike traditional surveys, plagued by low response rates and small samples, administrative data have little attrition and are often available for the entire population of interest at limited marginal cost

(Abraham et al., 2022; Jarmin and O’Hara, 2016; Meyer et al., 2015).

These features bestow on government administrative data the unique potential to enable policy-relevant research findings even though they were not originally collected for this purpose (Abraham et al., 2017; Card et al., 2010). Indeed, the increased availability of administrative data has played an important part in the empirical turn taken by economic scholarship in recent decades (Backhouse and Cherrier, 2017; Cole et al., 2020; Einav and Levin, 2014b; Groves, 2011). The share of research in top journals using administrative data is almost 70% when looking at studies on high-income countries and is 20% for all countries (Chetty, 2012; Currie et al., 2020). Einav and Levin (2014b) provide anecdotal evidence on a few impactful studies based on confidential government records studying diverse topics like broadband internet, teacher quality, and medicaid expansion (Taubman et al., 2014; Akerman et al., 2015; Chetty et al., 2014). Taken together, there is little doubt that the diffusion of administrative microdata has contributed to the increase in the incidence and impact of applied economic research (Angrist and Pischke, 2010; Angrist et al., 2020; Card, 2022).

However, the same features that make administrative records invaluable for research could put the privacy of respondents at risk. It is not possible to publicly and openly distribute firm- or individual-level wages, identification records, addresses, contact information, and so forth due to both security and privacy concerns. Statistical agencies thus face a trade-off between providing research access to microdata and their duty to protect the confidentiality of the data entrusted to them (Abowd and Schmutte, 2019; Foster et al., 2009). Over the years, 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 respondent-level information has.

To address this conundrum, some statistical agencies have erected strong security barriers around the data they distribute. This includes providing access exclusively to a set of vetted researchers for pre-approved projects and then allowing only the release of research results that have undergone careful review (Desai et al., 2016). Most notably, access is provided only through a physical presence at secure facilities on specialized devices where data use is closely monitored. This model guarantees maximum security because researchers analyze anonymized data in the facility, with no records ever leaving the data enclave. However, this approach also imposes significant costs both to the facility’s set-up and to researchers in terms of flexibility and ease of access.

Statistical agencies in different countries have often adopted alternative arrangements. For instance, European agencies are experimenting with providing access through (secure) portable computers, even outside of the home country, trying to trade off security and control with ease of access. On the other hand, other

agencies such as the U.S. Census Bureau have been reluctant to move away from a more restrictive approach based on physical access.² Given the central importance that data policy is playing in economic policy, the goal of our study is to provide some empirical evidence on the empirical effects of access under the current restrictive regime on follow-on outcomes.

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 was to enable research that could improve the data programs of the Census Bureau while preserving confidentiality.³ However, interested scholars had to relocate 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. This program, known as the FSRDC network, led to the creation of 30 data centers between 1994 and 2019 (Davis and Holly, 2006; CES, 2017).⁴

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 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 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.

FSRDCs permit access to some datasets that have become household names for economists. Among the

²In 2019, the Census Bureau began to provide remote virtual access for a select number of FSRDC researchers. The pilot has been scaled up during the COVID-19 pandemic, resulting in 83 projects using virtual access by mid-2021. This development does not affect our analyses since it affects only projects that will be published outside of our sampling period.

³The data collected by the U.S. Census Bureau is tightly regulated by Title 13 of the U.S. Code. Title 13 provides a legal framework for the Census Bureau 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 additional 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). For the purposes of our paper, we collectively refer to all confidential data that require onsite access in a secure facility managed by the U.S. Census Bureau as confidential “Census data.”

⁴The only FSRDC to have closed is the one opened at Carnegie Mellon University (CMU) in 1996 (Davis and Holly, 2006). For the purposes of our empirical analysis, we consider CMU as losing local access to the data after the FSRDC closed in 2004. Additional FSRDCs have opened outside of our sampling period (see Appendix Table C1). In 2018, the FSRDC network formally became part of the Center for Enterprise Dissemination but without any change in its operations.

most popular are the LBD, which links the Standard Statistical Establishment List to the Economic Census, and the LEHD program, which merges worker and employer records from Census Bureau surveys with state unemployment insurance claims (Abowd et al., 2004; Jarmin and Miranda, 2002). In several cases, FSRDC users have contributed to creating new databases as part of their research project; 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). Moreover, the success of the FSRDC network has led other federal agencies to make their own 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. Given this diversity, while the CES was originally established to give access to economic data for the manufacturing sector, FSRDCs now permit the investigation of a large variety of economic phenomena (McGuckin et al., 1993; CES, 2017).⁵

Anecdotal evidence suggests that the CES and the FSRDC network enabled path-breaking advances in economics (CES, 2017). For instance, the availability of establishment-level data is credited to have contributed to a generalized shift from “representative firm” thinking toward research that takes into account intra-industry heterogeneity (McGuckin, 1995; Coase, 1995; Davis et al., 1998; Bernard and Jensen, 1999). However, despite having enabled impactful research, the emphasis on security has limited the diffusion of government administrative data relative to other European countries (Cole et al., 2020). Researchers have even suggested that limited access to government administrative data puts the U.S. at risk of losing leadership in cutting-edge empirical research (Card et al., 2010). Moreover, restrictive access to data is disproportionately penalizing for early career researchers and scholars 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 due to the significant access restrictions that remain in place. 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 the additional costs of creating FSRDCs in the first place, and the costs of renting space for the data enclaves themselves.⁶ There is significant interest in expanding the FSRDC program, or creating similar alternatives under recent legislation, but challenges about the utility of the current system could hamper progress. It is therefore timely and important to empirically evaluate the effects of the FSRDC program on academic research with a view to shaping the design of new policies and institutions in this area.

⁵We term the data collectively provided through this program as “Census data” and use it to mean data distributed by the FSRDC program rather than data originating exclusively from the U.S. Census. Different agencies started distributing different datasets at different points in time, but we are unable to 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.

⁶Estimates are taken from https://deepblue.lib.umich.edu/response_FSRDC_Directors_2021.

3 Research Design

There are two key challenges in empirically assessing the impact of data access on economics research. First, it is usually very difficult to measure data use and access since we do not know who has access to a certain dataset and since researchers do not systematically cite data sources. Second, any correlation between the use of specific data and publication quality is likely to be problematic. Researchers who have access to certain datasets might produce better research not because of data access but because they have greater resources or are more creative (Nagaraj et al., 2020). The empirical challenge is thus finding a research design that provides (a) a measure of access to administrative 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 Census data through collaborators or by visiting the Census Bureau’s headquarters, co-location to an FSRDC should make the researcher aware of the data and decrease the barriers to using them. In particular, our research design takes advantage of the requirement that U.S. Census confidential data can only be analyzed in secure facilities. On the one hand, this allows us to approximate data access by measuring the distance between the location where the researcher works and the data center closest to them. On the other hand, the need to be physically co-located with a secure facility is equally binding for researchers with differing resources and abilities. This is a marked difference from other databases, for which access is usually constrained by financial resources or connections.

Consider Figure 1, which clarifies our research design. Here we use our data to plot the number of papers written in an FSRDC by researchers at the University of Michigan, Ann Arbor (panel i) and Stanford University (panel ii). The vertical line shows the year in which the FSRDC at these two institutions opens. As is clear from the figure, the number of FSRDC papers increases substantially earlier in Michigan as compared to Stanford even though both are R1 research universities and have active economics departments. Note also that by the time the Stanford FSRDC opens in 2010, confidential Census data were already commonly used in research, and yet almost no Stanford researcher was using them. This example suggests that the opening of a local FSRDC offers a credible natural experiment to understand how researchers are shaped by access to confidential administrative data.

A potential problem with our identification strategy is that the hosting institution’s characteristics might drive the choice of FSRDC locations. However, if time-invariant university attributes such as endowments or status drive location choice, we can control for these with university fixed effects. More worrisome would be dynamic considerations, such as a change in revenues or research intensity in institutions that open a

data center, which would confound the effect of FSRDC openings. For example, in the example discussed before, if an FSRDC at the University of Michigan opens due to an influx of money that also allows research facilities to be upgraded, then our estimates of the effects of the FSRDC opening would be biased upwards.

To address this type of concern, we exploit an additional source of variation for our inference. Academic economists tend to specialize along methodological lines and either devote themselves mostly to empirical work or theoretical modeling (Backhouse and Cherrier, 2017). In particular, we can exploit individual-level variation in “exposure” to FSRDCs by methodological orientation, i.e., distinguishing between theoretical and applied economists within the same university. For example, in Figure 1, as the color coding indicates, almost all of the papers in this set come from empirical researchers (rather than pure theorists). This suggests a natural control group of theoretical economists affiliated with the same institution against which empirical researchers can be assessed while simultaneously controlling for time-varying university trends that might correlate with FSRDC openings. An empirical assessment of the absence of pre-trends in our estimates would further validate this research design.

Equally problematic for our analysis would be the sorting of FSRDCs in the institutions where their impact on research could possibly be higher. To assess the plausibility of this concern, we conducted several interviews to learn the history and institutional details of the FSRDC network (see Appendix A.1). Opening a new FSRDC requires universities to submit a formal application to the National Science Foundation (NSF) through their competitive grant application process, which is then jointly evaluated by the NSF and the Census Bureau (Atrostic, 2007). 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 one of our interviewees, a former FSRDC administrator, explained, “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 from obtaining local data access for many years, a fact that appears to allay concerns about the sorting of FSRDC locations.⁷

Finally, our interviews revealed a surprisingly large number of idiosyncratic factors behind the establishment of most FSRDCs (Appendix A.1). For instance, some FSRDCs were either opened because of the will of one high-ranking university administrator or because of specific collaborations between some faculty members and the Census Bureau. Other times, the presence of an individual researcher advocating to open a new data center was enough to receive the NSF grant, even in the absence of a broad potential user community. In cases where a consortium of universities opened the FSRDC, the choice of which consortium member would actually host the data center was often the result of a compromise. Taken together, this qualitative evidence

⁷One recurrent example in our interviews is the case of Stanford University, which opened its FSRDC branch only in 2010 due to the presence of the relatively close-by Berkeley FSRDC. See Appendix A.2 for a case study.

suggests that both the timing and the locations of FSRDCs were strongly influenced by idiosyncratic factors unrelated to underlying trends in research productivity, further lending support to our identification strategy.

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 and quality of their publication output, (c) measuring the each researcher’s methodological orientation, and (d) measuring the diffusion of FSRDCs and the adoption of confidential Census 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 wider 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 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 full set of articles and journals in EconLit regardless of their prestige or centrality in the field: 839,513 scientific articles published in 1,856 journals between 1990 and 2019. 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 three 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.⁸ Using fuzzy matching and extensive manual checks, we standardize the 11,466 different spellings of the 438 U.S. research institutions appearing in our list of research-intensive institutions. Second, we disambiguate the resulting names using a graph-theoretic disambiguation procedure (Önder and Schweitzer, 2017). This approach

⁸We retain in our sample all doctorate-granting institutions in the United States taken from the 2018 Carnegie Classification of Institutions of Higher Education (<https://carnegieclassifications.iu.edu/>), to which we add the most important institutions active in economic research (such as the IMF, Rand Corporation, World Bank, and all the regional FED offices). We exclude from our sample researchers who are or have been affiliated to the U.S. Census Bureau or any partner agency since there people might enjoy privileged connections and access to data.

assumes that the combination of first, middle, and last names uniquely identifies each economist (Card et al., 2022) while at the same time being conservative in assigning ambiguous names that lack a clear middle name. To avoid confounding effects arising 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., 2018).

This procedure results in 15,750 U.S.-based economists who have been AEA members. We use this list to derive an unbalanced panel of 246,711 researcher-year observations by imputing missing years between the first and the last year in which we see a researcher publishing. For years with missing publications, we also 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 in 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 data.

4.2 Publication-Level Information

In addition to listing article metadata such as authors, journal, year of publication, and JEL codes, EconLit also includes the abstract for a large share of articles. We collected additional abstracts from websites like Google Scholar, EconStor, and JSTOR. Next, we augment EconLit by merging the yearly citations count for each article extracted from SSCI/Web of Science. 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, 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 citation data for 97.2% of these articles.

We then use journal, title, abstract, year of publication, and JEL codes to classify papers as either empirical or theoretical in style.⁹ We use a machine learning classifier that outputs a score capturing the probability that an article is empirical. 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. All the details are reported in Appendix D. The 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.

Next, we assemble data to measure the diffusion of administrative data available through the FSRDC network. We painstakingly assemble a list of all articles that *directly* employ restricted-access microdata accessible

⁹We are indebted to Ryan Hill for sharing the code and the training data originally used in Angrist et al. (2020).

only in an FSRDC. Since no official bibliographic record is available, we carefully sift published records with several complementary strategies (detailed in Appendix C.2). Projects using confidential U.S. Census data are expected to be indicated as such clearly in the paper’s acknowledgments (see Appendix Figure A1). We perform keyword searches for the most common expressions denoting the use of Census data using databases such as Web of Science, Scopus, JSTOR, and Google Scholar. We then exploit the fact that the CES requires submitting a working paper for online publication upon completion of the project. 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 Census data affects research *indirectly*, that is, by enabling findings that shape or inspire subsequent research. We do so with three approaches. First, we record which articles cite the papers written using Census data. This set of papers explicitly builds on the results based on confidential data.¹¹ Second, we tag all papers that include JEL codes that are the most representative of research using Census data. In this way, we capture papers that are thematically close to research done in FSRDCs. Finally, 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 certain method (e.g., DID) or type of data (e.g., survey data) in the title or abstract. The complete list of keywords used is reported in the Appendix C.4.

4.3 Researcher-Level Information

Thanks to the host of article-level variables outlined above, we can compute several metrics that capture the yearly research output of each researcher. To reflect productivity, we sum the yearly number of publications in top field and top five economics journals, which we collectively refer to as “top publications.”¹² We capture research impact by weighting publication counts by the citations received in a window of time up to five years after publication. To account for top tail outcomes, we also code the number of journal articles published in the top five economics journals and the count of papers in the top 95th percentile of the most cited articles published in any given year.

We rely on our article-level methodological classification (empirical or theoretical) to categorize each scholar according to their methodological orientation. For our main analyses, an empiricist is defined as anyone

¹⁰Unfortunately, we cannot separately code papers stemming from internal Census projects, but this should not impact our analyses since we exclude from the analyses researchers who have been formally affiliated with the U.S. Census Bureau.

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

¹²We 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.

with more than half of their publication output classified as empirical. This measure has the advantage of being available for every publishing researcher in our sample. We carry out several validations and checks. First, we adopt a case-control approach and check the results of our classification for the editorial board members of some journals with a clear methodological bend (e.g., the *Journal of Economic Theory* versus *AJ: 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.

Finally, we aim to measure changes in research trajectory for the economists in our sample. Following Furman and Teodoridis (2020), we do so with two complementary approaches: leveraging hierarchical taxonomies of research topics and using data-driven methods based on papers' abstracts. First, we rely on JEL codes to track the topical focus of research, exploring the likelihood that a researcher writes a paper with JEL codes that she has never used in her previous work. Second, we use an unsupervised machine learning algorithm to sidestep shortcomings of author-assigned JEL codes (details in Appendix C.4). We consider all the abstracts of articles published in a given year as the textual footprint of the topics spanned by the researcher during that year. In particular, we use the well-known bag-of-words algorithm called Latent Dirichlet Allocation (LDA) to generate 20 clusters of words that are found to appear together in the input text with a high probability (i.e., topics). We code researchers as working on a new topic if the model classifies at least 10% of their work as pertaining to a topic not featured in their past work.

4.4 University-Level Information

We can assess the details of FSRDC openings from the 14 CES research reports published online between 2005 and 2019. Figure 2 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 when a relatively close-by alternative was present (e.g., Stanford, Yale, Princeton). The complete list of 30 FSRDCs established in our sampling period and their opening date is reported in Appendix C.

The information 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 is based on publications in the

five-year period from 1995 to 1999, thus predating the establishment of most FSRDCs. Using these data, we can see how several less research-intensive institutions gained access earlier than their more prestigious peers. Appendix Figures E1 and E2 show that the average position in the ranking of treated institutions and treated researchers is relatively constant over time. The remarkably balanced expansion of the FSRDC network confirms oral accounts that exclude explicit prioritization of high-status universities.

4.5 Summary Statistics

Table 1 provides summary statistics of the dataset that we have assembled. Panel A provides summary statistics on the cross-section of 15,750 researchers who appear in the data. We classify 73% of them as empiricists. Almost 3% of the publishing economists who we observe have directly used Census data in published work, and 23.4% have cited at least one paper using Census data. The average researcher publishes about 2.3 academic papers in top outlets during the time that they appear in our sample. Slightly less than a quarter of the economists in the sample are observed publishing in at least one of the top five economics journals over the time period.

As per our data, of the sample of 15,750 researchers, about 60% never had access to an FSRDC in the city in which they were employed. Appendix Figure C1 breaks this sample down by those who always had access (2,104) and those who got access by an FSRDC opening in their city (2,735) or by moving to a city with an FSRDC (1,586). 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, on average, the average researcher publishes about 0.15 papers every year in the set of top economics journals. For every researcher-year, 0.003 papers use Census data and 0.033 build on past research using Census data.

5 Results

Our basic specification is at the individual researcher i , city j , and time t level and takes the form of a DID equation to test the impact of local data access:

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

where the dependent variable $y_{i,j,t}$ is publications at the researcher-year level. The main independent variable $PostFSRDC_{j,t}$ is a dummy that takes a value of one after a researcher has gained access to an FSRDC facility in time t and city j , and it is zero before. Motivated by the intuition that data access should mostly benefit applied researchers, we focus on estimating the local impact of FSRDC openings on empiricists relative to theorists. We do so by interacting the independent variable with $Empiricist_i$, a dummy that

only takes a value of one for the economists who we classify as empiricists. This allows us to control for university-by-year fixed effects ($\delta_t \times \omega_j$), which provides fine-grained control for university-level research dynamics, in addition to more general 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). μ_i are individual researcher fixed effects that control for time-invariant differences across researchers in data use but also individual propensity to publish.

We present results where the treatment dummy is coded as an absorbing state, which means that individuals do not lose the treated status even if they move to a non-treated city. This approach accounts for potential delays in the publication process and makes it straightforward to interpret β as the result of a Wald-DID estimator. However, including university-by-time fixed effects means that the main effect of $PostFSRDC_{j,t}$ will be identified only from researchers who move from an institution with local access to an institution that never gains local access.¹³ To provide a more meaningful estimate of the main effect that exploits variation between institutions, we relax our fixed effects structure and estimate additional models with university tier \times year fixed effects. We do so by classifying each university into seven tiers based on the ranking of economics departments from Kalaitzidakis et al. (2003).¹⁴

5.1 Impact of FSRDC Openings on Data Use

A first step to unpacking the effects of data access on economics research is to examine how FSRDC openings affect confidential data use. Note that even before being co-located to 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 to an FSRDC will lead to a substantial diffusion of confidential administrative data.

Table 2 displays the results from this analysis. As is clear from Column 1, local access to an FSRDC boosts the use of administrative Census data among empiricists but not among theorists (see also Appendix Figure E3). Local access is associated with about 0.004 more papers using Census data, an increase of roughly 131% compared to the baseline of 0.003. The coefficient is only slightly smaller when we include more stringent university-by-time fixed effects in Column 2. The results 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 E1). Interestingly, we also find a large impact on the likelihood of citing past work (from any university) based on Census data (Columns 3 and 4). While the increase on the sample mean is much smaller (around 53%–57%), this result is even more

¹³Said otherwise, there can be variation in treatment status within university-year cells only if there are researchers who have been exposed to FSRDC in other locations and later moved to a place without access.

¹⁴We chose the number of tiers to ensure a roughly equal number of observations for each of them, but the results do not change if we change this number. See details in Appendix B.2.

remarkable since geography should not be an impediment to learning about papers published in academic journals.¹⁵ This suggests that the opening of an FSRDC is enculturing local researchers into research using administrative data, potentially through interactions with others who are directly dealing with such data in their own work.

The validity of DID 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 $y_{i,j,t} = \alpha + \sum_z \beta_t \times 1(z) + \mu_i + \delta_t \times \omega_j + \epsilon_{i,j,t}$, where μ_i and $\delta_t \times \omega_j$ represent researcher and university \times time fixed effects, as before. z represents the “lag,” or the number of years that have elapsed since an empirical researcher first received access to Census data via a local FSRDC. Figure 3 shows the effect of increased access to confidential Census data on their use (panel (i)) and citations to papers based on them (panel (ii)). Both charts confirm that there are no pre-trends driving our effects, which might have been the case if FSRDCs are opened in locations where interest in using administrative data is rising. Further, the effects appear gradually and then grow in magnitude before stabilizing after five to six years, with a 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 Census data in the focal city, we turn to estimate the impacts of data access on researchers’ productivity. Formally, we estimate a similar specification as before except that now our outcome variables measure the quantity and impact of economics research. We focus on the most prominent journals in economic disciplines because recent work has documented that they carry outsized weight in determining career trajectories (Heckman and Moktan, 2020). For each researcher, we count the number of papers published yearly in top field outlets and top five generalists journals. To measure quality, we construct a measure of scientific impact based on the number of citations each paper receives up to five years following its publication. This citation-weighted publications count constitutes our second dependent variable.

Table 3 presents the results from this analysis for both outcome variables. The first model reports results with university tier \times year fixed effects, and the second model presents results including university-specific time trends. Across the board, the results suggest that local FSRDC access boosts the productivity of empirical researchers. Empirical economists produce about 0.035 more publications in the top journals, which, compared to a baseline of 0.145, is a 24% increase in publication output. These estimates are large and meaningful, but they are against a small denominator and so do not represent implausible shifts in

¹⁵Note also that we exclude from the count of citing papers those that directly employ FSRDCs. Therefore, our estimates are conservative and not mechanically explained by the contemporaneous diffusion of Census data.

absolute volume. 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 1.71–1.74 (40%–41% increase against a baseline of 4.3). This result suggests that access to administrative data leads to significantly higher-impact publications.

Further, we explore the effects of top tail outcomes, namely the likelihood of publishing in a top five journal or authoring highly cited articles. As shown in Table 4, the results are larger in magnitude, which supports anecdotal evidence that administrative data lead to breakthrough findings (Einav and Levin, 2014b). Next, we explore the reliability of our research design by empirically estimating pre-trends in the outcome variables. Figure 4 presents event study estimates similar to those presented in Figure 3 but for the measures of research productivity discussed above. In all four 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 E2 and Appendix Figure E4 repeat the same analysis using the doubly robust DID estimator by Callaway and Sant’Anna (2021), confirming that our results are not biased by different timings of FSRDC openings. The combined evidence from the regression and event study estimates confirm that access to administrative data can greatly boost the quantity and quality of research for empirical economics researchers.

Additional analyses explore the heterogeneity of these results using split-sample regressions. First, we estimate separate regressions for senior and junior researchers (i.e., within seven years from their first publication; see Appendix B.3 for details). Second, we estimate separate regressions for researchers affiliated with universities of different tiers (Kalaitzidakis et al., 2003). We show the main results in graphical form in Appendix Figure E5. The increase in research quality is mostly confined to senior economists affiliated with high-status universities. It is interesting to compare this result with recent evidence showing that democratizing access to data is especially beneficial to marginalized researchers (Nagaraj et al., 2020). Despite drastically reducing barriers to use, a local FSRDC might still prove cumbersome to use due to the onerous confidentiality requirements and bureaucratic uncertainties. Our conversations with FSRDC users pointed out that applying to use confidential data is especially risky for researchers on a tenure clock. Moreover, it is customary for researchers to hire research assistants who physically access the data center to carry out the datawork, which helps explain why researchers in resourceful universities are heavy users (see Appendix A for additional details on the interviews).

5.3 Robustness Checks

We perform a variety of 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 directly test this concern, we examine trends in productivity using a panel at the university-year level (Appendix Table E3). Appendix Figure E6 shows 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 be driving our results at the universities that open a data center. We do so excluding from the sample either the researchers or the institution that were directly involved in bringing an FSRDC to a given location. We code the recipients of NSF grants establishing each FSRDC, and Appendix Table E4 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 FSRDC but do not host the data center on their premises (Appendix Table E6). 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 Table E5 shows that the results are robust to excluding researchers treated due to mobility events.

Third, we experiment with a variety of alternative specifications of “exposure” to administrative data in geographical and intellectual space. In general, our results get stronger the closer an economist is located to a data center, being the highest for researchers who enjoy access directly on their campus (Appendix Tables E7 and E8). Appendix Figure E7 shows that an economist’s likelihood of using confidential Census data rapidly decays with distance despite this pattern being less pronounced on the other outcome variables we explored. Similarly, the effects of access monotonically grow when we use increasingly stringent definitions of empirical researchers (Figure E8). This confirms our research design and reassures that our results are not an artifact of the threshold we use to classify someone as an empiricist.

Finally, one alternative explanation for our results is that secular trends in empirical research might underlie the increasing use and impact of administrative data, even in the absence of local FSRDC openings. To rule out this concern, we re-estimate our specifications adding an additional set of time trends for empirical researchers (i.e., empiricists-by-time fixed effects, coded in bins of five years each). Appendix Figure E9 presents the time-varying estimates of the effect of data access, both with and without controlling for empiricist-specific time trends. The event study plots look similar to the main ones, albeit noisier due to lower statistical power. Including empiricist-specific time trends attenuates our effects on top journal publications (Appendix Table E9), but the results remain robust and significant when weighting publications by their impact.

6 Exploration of the Mechanisms

It is clear that an FSRDC opening increases the diffusion of Census data and the productivity of empirical researchers in affected areas. We now examine potential mechanisms driving the productivity effect. 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 and (2) indirectly alters research quality through spillovers to those not using the data. Note that the first option would benefit only researchers directly using confidential data, while the second could benefit a broader swath of exposed researchers.

We evaluate these two channels by first checking the extent to which FSRDC users are driving our results. Appendix Table E10 reports all of our main results excluding from the analyses of researchers who we observe using FSRDCs. Coefficients are around 24%–39% smaller but are still large and significant. The implication is that even though FSRDC openings clearly benefit researcher productivity by enabling direct usage of the data, there are also hints of significant spillovers from data access even among those who do not directly adopt.

This finding suggests an alternative channel through which access to FSRDCs shapes economics: exposing researchers to research based on administrative data. 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 Census data, with potential downstream implications for researcher productivity. Figure 5 shows two additional tests that provide support for this awareness channel. First, we implement split-sample regressions separating effects by those who cite research based on Census data as compared to those who do not. We find that the positive effects on productivity do not extend to all empirical researchers in affected regions—instead they are limited to researchers who cite past work based on FSRDCs. This test suggests that FSRDC openings raise awareness about insights derived from FSRDC research that can have positive spillovers for those building on this work.

Second, we were told in our interviews that researchers often learn about the potential of administrative data after seeing the work of their colleagues (Appendix A). We therefore separately examine effects for those empirical researchers with and without colleagues using confidential Census data in their own work. We find no positive spillovers for researchers affiliated with departments where nobody directly uses Census data. Both results show that spillovers are concentrated among researchers who were exposed to FSRDC research. In other words, the evidence supports the conclusion that spillovers from co-localization operate by making researchers aware of research based on confidential administrative Census data.

We next propose two specific channels through which awareness of past research carried out with Census data might improve research output. First, researchers might be inspired to formulate new research questions

that build on research they were not familiar with. Second, they might learn from the research design of these studies, potentially adopting similar methods or data with similar characteristics. We provide suggestive evidence on each of these channels below.

We first test whether empirical researchers exposed to FSRDCs are more likely to explore new topics rather than doubling down on the same questions they were already working on. In particular, the expectation is that spillovers could operate by leading researchers to pick topics commonly studied using Census data such as labor, trade, or firm productivity. Column 1 of Table 5 shows that empirical researchers are more likely to publish papers on new topics, as proxied by the usage of JEL codes that they never used in their previous work. Column 2 shows that the result is robust to using unsupervised LDA topics to measure changes in research trajectory. Notably, exploration in topical space is directed toward topics commonly associated with using Census data. Columns 3 and 4 indicate that the likelihood of working on what we dub “FSRDC JELs” goes up by 16.7%, while for the remainder JEL codes, the increase is only 8.2%.¹⁶ These elasticities are derived by comparing our coefficient estimates with the average likelihood of working on FSRDC and non-FSRDC JELs reported in Table 1.

Next, we assess the possibility that exposure to research using Census administrative data induces researchers to adopt similar research designs and data (albeit not necessarily those available in an FSRDC). Table 6 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 wider “credibility revolution.”¹⁷ Learning about Census data might also inspire the researcher to search for new datasets with similar characteristics but lower bureaucratic hurdles. This theme also emerged in several of our interviews, with respondents underlying that administrative data from foreign countries are often easier to use (Appendix A). Indeed, we find that researchers are more likely to employ microdata from administrative sources, while a similar increase is absent for traditional research surveys. Reassuringly, we do not find an effect when considering generic mentions of “big data” or even just “data” (Appendix Table E12).

Combined, our analysis can paint a picture of *how* FSRDC openings drive improved productivity among applied researchers. First, there is a cadre of researchers who can now access these data at lower cost and use them in their own work to great effect. However, given the significant costs of direct access, this group is only partially responsible for driving the overall effects. Spillovers to non-adopters are also important: FSRDC openings tends to raise awareness of insights derived from administrative data, as shown by the increase in citations to past FSRDC work. Researchers exposed to FSRDC research benefit by shifting their

¹⁶This result is robust to alternative ways to select which JEL codes are the most representative of FSRDC-based research; see Appendix Table E11.

¹⁷As an additional falsification test, we also code articles mentioning the use of a “natural experiment” in their abstracts, and we find a significant increase that we do not see for mentions of a “laboratory experiment” (Column 3 of Appendix Table E12).

research focus to become more explorative, working on topics typically studied with administrative data, and by adopting other types of administrative data and quasi-experimental research designs in their own work.

7 Conclusion

We assemble a novel longitudinal dataset of U.S.-based academic economists and exploit the staggered diffusion of U.S. Census Bureau 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 directly use or build upon work that uses confidential Census data. We then explore the consequences of data access on scientific output. In our setting, researchers are more likely to publish highly impactful papers in prestigious journals after they gain access to Census 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 like Census data are more likely to build upon it, exploring novel questions that stem from it and adopting similar research designs.

Our results have implications for ongoing policy discussions on the use of confidential administrative data 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 economic research. Our findings are consistent with the idea that increased access to confidential data is crucial to scientific progress, to the point that it might call into question the current tight regime of restricted access designed to protect privacy. 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 document 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 on research, 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. To do so, we would need to expand our focus to other disciplines that benefit from Census data, such as health policy and sociology. Moreover, the Census Bureau’s objective in creating the FSRDC network is to obtain benefits for its data programs (Foster et al., 2009). To assess the overall success of the network, we would need to take into account all such benefits, including the improvement of its datasets and the creation of new statistical surveys (CES, 2017). Our work cannot capture all these aspects, but we establish that, at least in the case of economics, the researcher-level impact of these institutions is large and directly leads to the production of higher-quality scientific output.

Second, we suffer from the common pitfalls of studies based on bibliographic data. In particular, assessing

research quality is often confounded by the researcher's status and social networks. Our within-researcher estimates would not take into account dynamic changes in collaboration networks or connections to editors that might directly affect scientific productivity. Finally, our results are based on a particular type of data that entail geographical barriers to use. While this is crucial to our research design, it also limits the generalizability of our results to contexts where researchers face large impediments to accessing data.

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 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., R. HASKINS, S. GLIED, R. GROVES, R. HAHN, H. HOYNES, J. LIEBMAN, B. MEYER, P. OHM, N. POTOK, ET AL. (2017): “The promise of evidence-based policymaking,” *Report of the Commission on Evidence-Based Policymaking*.
- 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.
- ANGRIST, J. D. AND J.-S. PISCHKE (2010): “The credibility revolution in empirical economics: How better research design is taking the con out of econometrics,” *Journal of Economic Perspectives*, 24, 3–30.
- 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.
- 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.
- BERMAN, H. M., J. WESTBROOK, Z. FENG, G. GILLILAND, T. N. BHAT, H. WEISSIG, I. N. SHINDYALOV, AND P. E. BOURNE (2000): “The protein data bank,” *Nucleic Acids Research*, 28, 235–242.
- BERNARD, A. B. AND J. B. JENSEN (1999): “Exceptional exporter performance: Cause, effect, or both?” *Journal of International Economics*, 47, 1–25.
- 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.
- 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," Nobel Memorial Lecture.
- 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. DELLA VIGNA, 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*.
- CES (2017): "Center for Economic Studies and Research Data Centers Research Report: 2016," Available on the U.S. Census Bureau's website.
- 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., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, 104, 2633–79.
- 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.
- COASE, R. H. (1995): *Essays on Economics and Economists*, University of Chicago Press.
- 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*.
- DESAI, T., F. RITCHIE, AND R. WELPTON (2016): "Five Safes: Designing data access for research," Economics Working Paper Series 1601, University of the West of England.
- DIZIKES, P. (2019): "The productive career of Robert Solow," *MIT Technology Review*.
- 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 US 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.

- GENTZKOW, M., B. KELLY, AND M. TADDY (2019): “Text as data,” *Journal of Economic Literature*, 57, 535–74.
- 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.
- 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. 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 (2020): “Scooped! Estimating Rewards for Priority in Science,” Northwestern University and UC Berkeley.
- (2021): “Race to the bottom: Competition and quality in science,” Northwestern University and UC Berkeley.
- 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. MANZO, A. NAGARAJ, AND M. TRANCHERO (2022): “The streetlight effect in data-driven exploration,” UC Berkeley and University of Vienna.
- 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,” Available at SSRN 2128793.
- JARMIN, R. S. AND A. B. O’HARA (2016): “Big data and the transformation of public policy analysis,” *Journal of Policy Analysis and Management*, 35, 715–721.
- JELVEH, Z., B. KOGUT, AND S. NAIDU (2018): “Political Language in Economics,” Columbia Business School.
- 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, C. PAVER, I. SMOLYAR, ET AL. (2018): “World ocean atlas 2018, volume 1: Temperature,” NOAA Atlas NESDIS 81.

- 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.
- MEYER, B. D., W. K. MOK, AND J. X. SULLIVAN (2015): “Household surveys in crisis,” *Journal of Economic Perspectives*, 29, 199–226.
- 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.
- 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.
- NAGARAJ, A. (2022): “The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry,” *Management Science*, 68, 564–582.
- 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.
- SARSONS, H., K. GÖRXHANI, 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.
- 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.
- 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., 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, J. BAKKEN, R. BARKHouser, S. BASTIAN, E. BERMAN, ET AL. (2000): “The Sloan Digital Sky Survey: Technical summary,” *The Astronomical Journal*, 120, 1579.

8 Tables and Figures

Table 1: Summary Statistics

Panel A: Researcher Level						
	N	Mean	Std. Dev.	Median	Min	Max
Ever Had FSRDC Access (0/1)	15750	0.408	0.49	0	0	1
Year of Access	6425	2007.517	6.83	2008	1994	2019
Ever Used FSRDC (0/1)	15750	0.028	0.17	0	0	1
Ever Cited FSRDC (0/1)	15750	0.234	0.42	0	0	1
Lifetime Top Publications	15750	2.278	4.41	1	0	93
Lifetime Cite-weighted Papers	15750	66.896	187.79	6	0	5536
Ever Top 5 Papers (0/1)	15750	0.236	0.42	0	0	1
Ever Top 5% Cited Papers (0/1)	15750	0.242	0.43	0	0	1
Rank of Institutions (avg)	15750	14.307	18.93	6	0	86
Empiricist (0/1)	15750	0.731	0.44	1	0	1

Panel B: Researcher-Year Level						
	N	Mean	Std. Dev.	Median	Min	Max
Post-FSRDC (0/1)	246711	0.258	0.44	0	0	1
Papers Using FSRDC	246711	0.003	0.06	0	0	3
Papers Citing FSRDC	246711	0.033	0.20	0	0	6
Top Publications	246711	0.145	0.43	0	0	9
Cite-weighted Papers	246711	4.271	19.79	0	0	1285
Top 5 Papers	246711	0.047	0.24	0	0	6
Top 5% Cited Papers	246711	0.044	0.23	0	0	5
New JEL Codes (0/1)	246711	0.389	0.49	0	0	1
New LDA Topics (0/1)	246711	0.297	0.46	0	0	1
Papers with FSRDC JEL (0/1)	246711	0.105	0.31	0	0	1
Papers without FSRDC JEL (0/1)	246711	0.337	0.47	0	0	1
Papers Mentioning Admin Data	246711	0.003	0.06	0	0	2
Papers Mentioning Survey Data	246711	0.005	0.08	0	0	3
Quasi-experimental Papers	246711	0.019	0.14	0	0	4
Experimental Papers	246711	0.036	0.23	0	0	10
Year	246711	2005.953	7.71	2007	1990	2019

Note: This table lists summary statistics at the researcher level for 15,750 publishing economists (Panel A) and at the researcher-year level for an unbalanced panel of 246,711 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 Census data. Ever Cited FSRDC: 0/1 = 1 for researchers who published at least one paper that cited a publication based on Census 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. Ever Top 5 Papers: 0/1 = 1 for researchers who published at least one paper in a top five journal. Ever Top 5% Cited Papers: 0/1 = 1 for researchers who published at least one paper in the top 95th percentile of the citation distribution. Rank of Institution: average rank of the institution of affiliation. Empiricist: 0/1 = 1 for those researchers whose majority of lifetime publications are empirical in nature. Post-FSRDC: 0/1 = 1 after a researcher is first co-located to an active FSRDC. Papers Using FSRDC: count of papers using Census data. Papers Citing FSRDC: count of papers citing Census 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. Top 5 Papers: count of papers in a top five journal. Top 5% Cited Papers: count of papers in the top 95th percentile of the citation distribution by year of publication. New JEL Codes: 0/1 = 1 for researchers who used JEL codes that they had not used before. New LDA Topics: 0/1 = 1 for researchers who publish at least 10% of their scholarship in an LDA topic that they had not published in before. Papers with FSRDC JEL: 0/1 = 1 for researchers who used JEL codes common among papers using Census data. Papers without FSRDC JEL: 0/1 = 1 for researchers who did not use JEL codes common among papers using Census 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 Administrative Data

	Papers Using FSRDC		Papers Citing FSRDC	
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.000329 (0.00069)	0.00216 (0.00114)	-0.00884** (0.00333)	-0.00235 (0.00463)
Post-FSRDC \times Empiricist	0.00392*** (0.00107)	0.00332** (0.00113)	0.0189*** (0.00425)	0.0174*** (0.00444)
Researcher FE	Yes	Yes	Yes	Yes
University Tier \times Year FE	Yes	No	Yes	No
University \times Year FE	No	Yes	No	Yes
N	246532	245556	246532	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data diffusion. Columns (1) and (2) report results from OLS models, where the dependent variable is the count of articles that directly use confidential microdata 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 Census data). Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning-based paper classifier. All models include individual fixed effects. Columns (1) and (3) further include year fixed effects interacted with university tier dummies, and columns (2) and (4) further include year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the researcher 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	
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.0139 (0.010)	-0.00712 (0.012)	-0.676* (0.336)	-0.170 (0.435)
Post-FSRDC × Empiricist	0.0353*** (0.011)	0.0352** (0.011)	1.714*** (0.425)	1.737*** (0.445)
Researcher FE	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	No	Yes	No
University × Year FE	No	Yes	No	Yes
N	246532	245556	246532	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research output. 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 is 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. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning-based paper classifier. All models include individual fixed effects. Columns (1) and (3) further include year fixed effects interacted with university tier dummies, and columns (2) and (4) further include year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the researcher level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

Table 4: Effect of FSRDC Access on the Top Tail of Research Impact

	Top Five Pubs		Top 5% Cite	
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.0131* (0.006)	-0.00994 (0.007)	-0.0193*** (0.005)	-0.0147* (0.006)
Post-FSRDC × Empiricist	0.0218** (0.007)	0.0231*** (0.007)	0.0285*** (0.006)	0.0287*** (0.006)
Researcher FE	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	No	Yes	No
University × Year FE	No	Yes	No	Yes
N	246532	245556	246532	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on additional measures of research output. Columns (1) and (2) report results from OLS models, where the dependent variable is the number of top five publications (*AER*, *JPE*, *QJE*, *ECA*, *RES*). Columns (3) and (4) report results from OLS models, where the dependent variable is the number of publications whose number of citations is in the top 5% of the citations distribution for the year in which they were published. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning-based paper classifier. All models include individual fixed effects. Columns (1) and (3) further include year fixed effects interacted with university tier dummies, and columns (2) and (4) further include year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at the 5%, 1%, and 0.1% level, respectively.

Table 5: Effect of FSRDC Access on the Direction of Research

	New JEL Code (0/1) (1)	New LDA Topic (0/1) (2)	FSRDC JELs (0/1) (3)	Non-FSRDC JELs (0/1) (4)
Post-FSRDC	-0.00686 (0.011)	-0.00444 (0.011)	-0.00332 (0.007)	-0.0125 (0.010)
Post-FSRDC \times Empiricist	0.0475*** (0.010)	0.0339*** (0.010)	0.0175** (0.006)	0.0278** (0.009)
Researcher FE	Yes	Yes	Yes	Yes
University \times Year FE	Yes	Yes	Yes	Yes
N	229886	229886	245556	245556

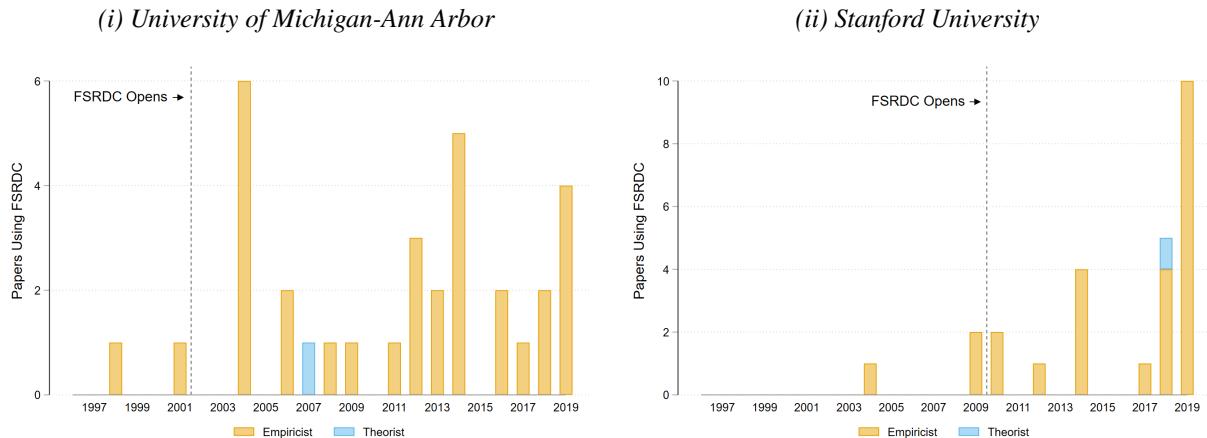
Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research direction. Column (1) reports results from a linear probability model, where the dependent variable is an indicator equal to one if the researcher used at least one JEL code never used before during the year. Column (2) reports results from a linear probability model, where the dependent variable is an indicator equal to one if at least 10% of the scholarship of a researcher in a given year is classified into an LDA topic that she never researched before. Column (3) 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 (4) 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. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

Table 6: Effect of FSRDC Access on the Design of Empirical Research

	Data		Research Methods	
	Admin (1)	Survey (2)	Quasi-Exp. (3)	Experiment (4)
Post-FSRDC	-0.00243* (0.001)	0.00220 (0.001)	0.000395 (0.003)	0.00372 (0.005)
Post-FSRDC × Empiricist	0.00328*** (0.001)	0.000364 (0.001)	0.00786** (0.003)	0.000294 (0.005)
Researcher FE	Yes	Yes	Yes	Yes
University × Year FE	Yes	Yes	Yes	Yes
N	245556	245556	245556	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on text-based proxies of research design. Columns (1) and (2) 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) 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. Post-FSRDC equals one in all years after a researcher has been affiliated with a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at the 5%, 1%, and 0.1% level, respectively.

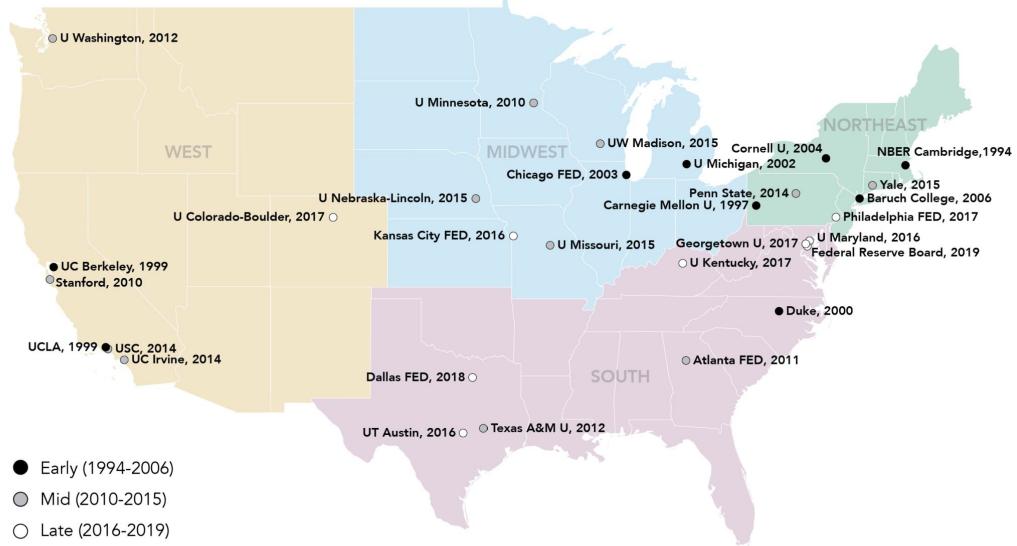
Figure 1: Yearly FSRDC Papers Written by Researchers Affiliated with the University of Michigan and Stanford University.



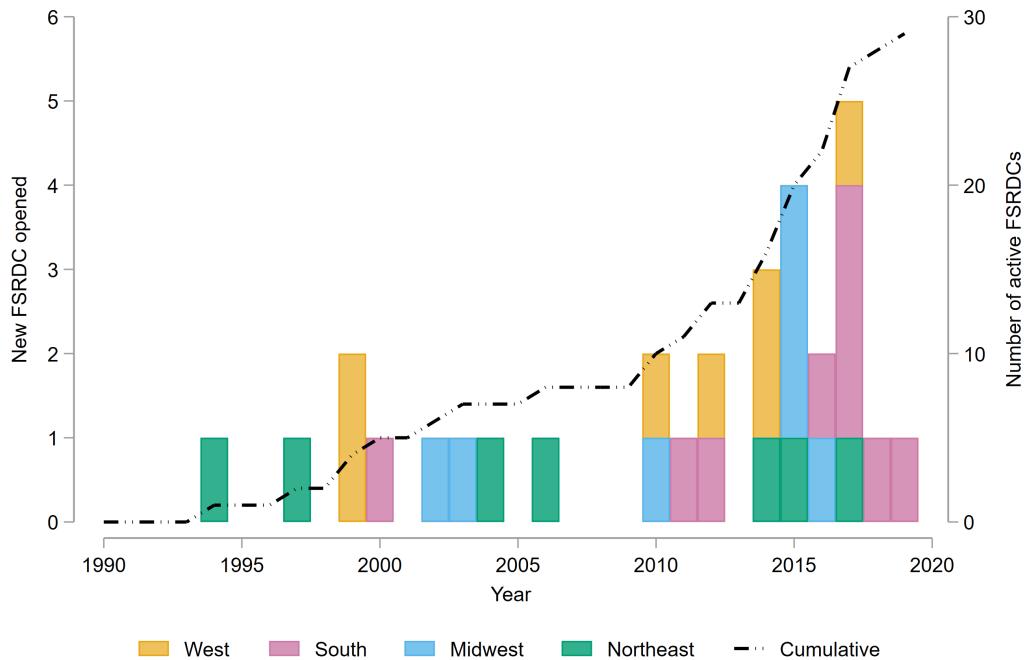
Note: This figure shows the number of papers using Census data published by researchers at the University of Michigan-Ann Arbor and Stanford University, respectively. The blue portion of each bar represents papers co-authored by at least one theoretical researcher. See text for more details.

Figure 2: Expansion of the FSRDC Program over Time and Space

Panel A: Geographic Expansion of FSRDCs

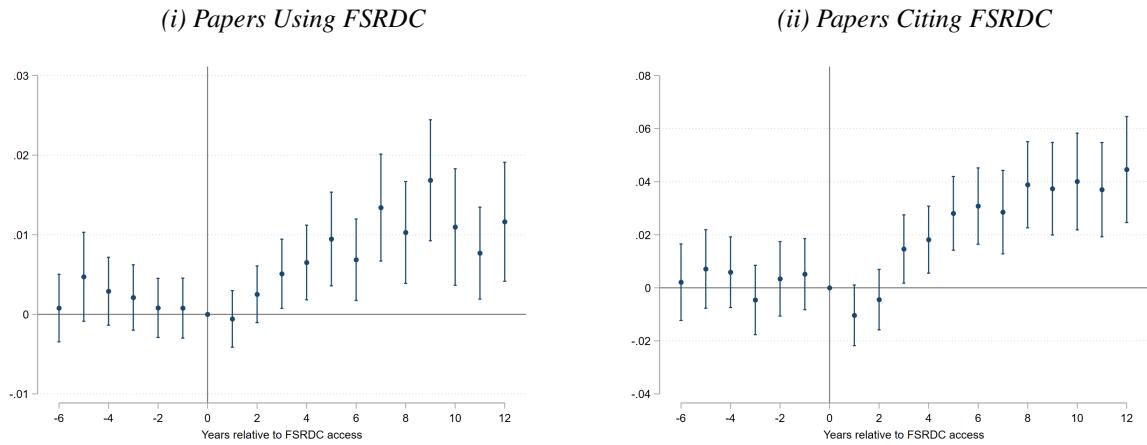


Panel B: Time Expansion of FSRDCs



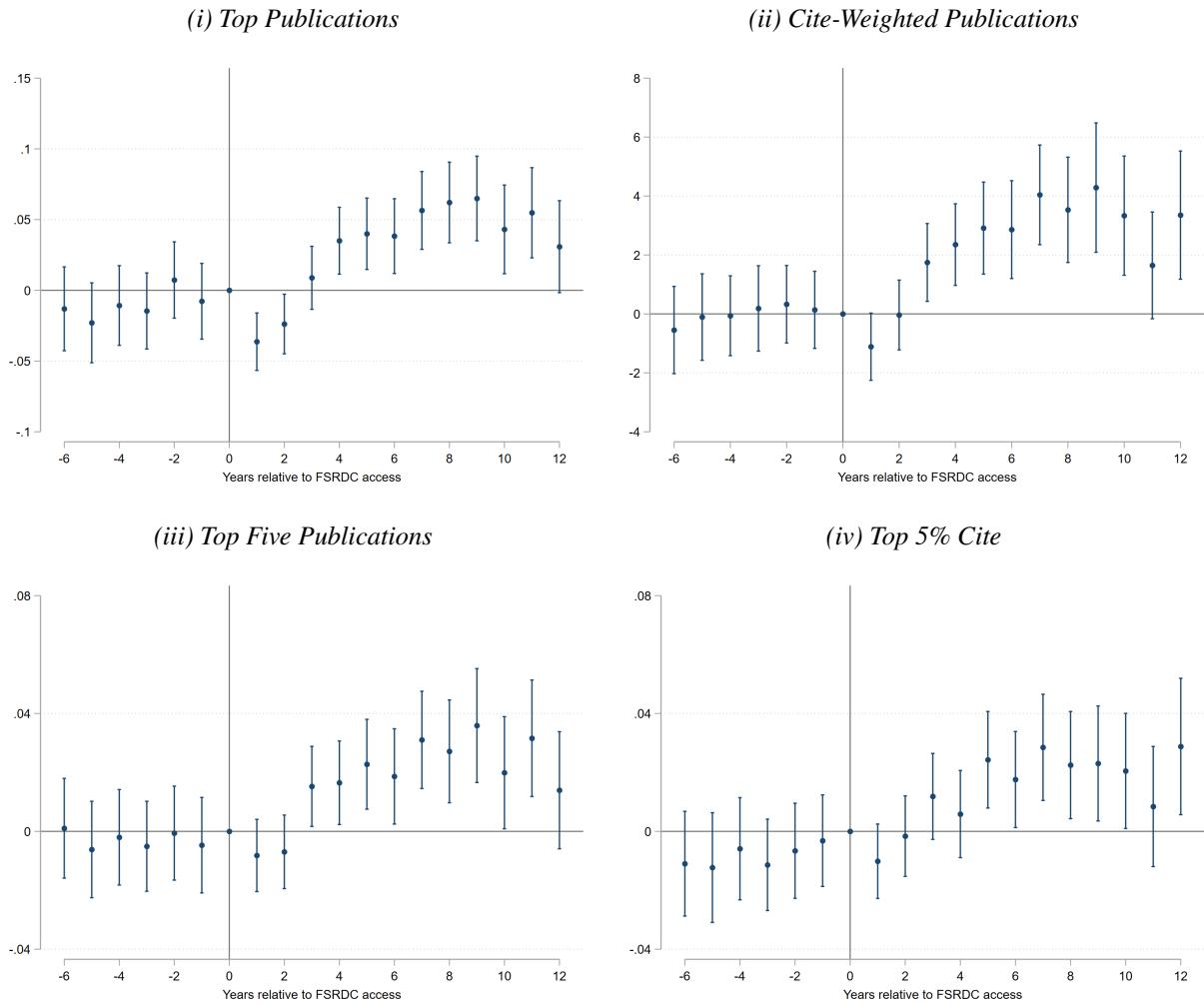
Note: This figure provides an illustration of the timing and geographic scope of the expansion of the FSRDC network. See text for more details.

Figure 3: Time-Varying Estimates of the Impact of FSRDCs on the Diffusion of Administrative Data



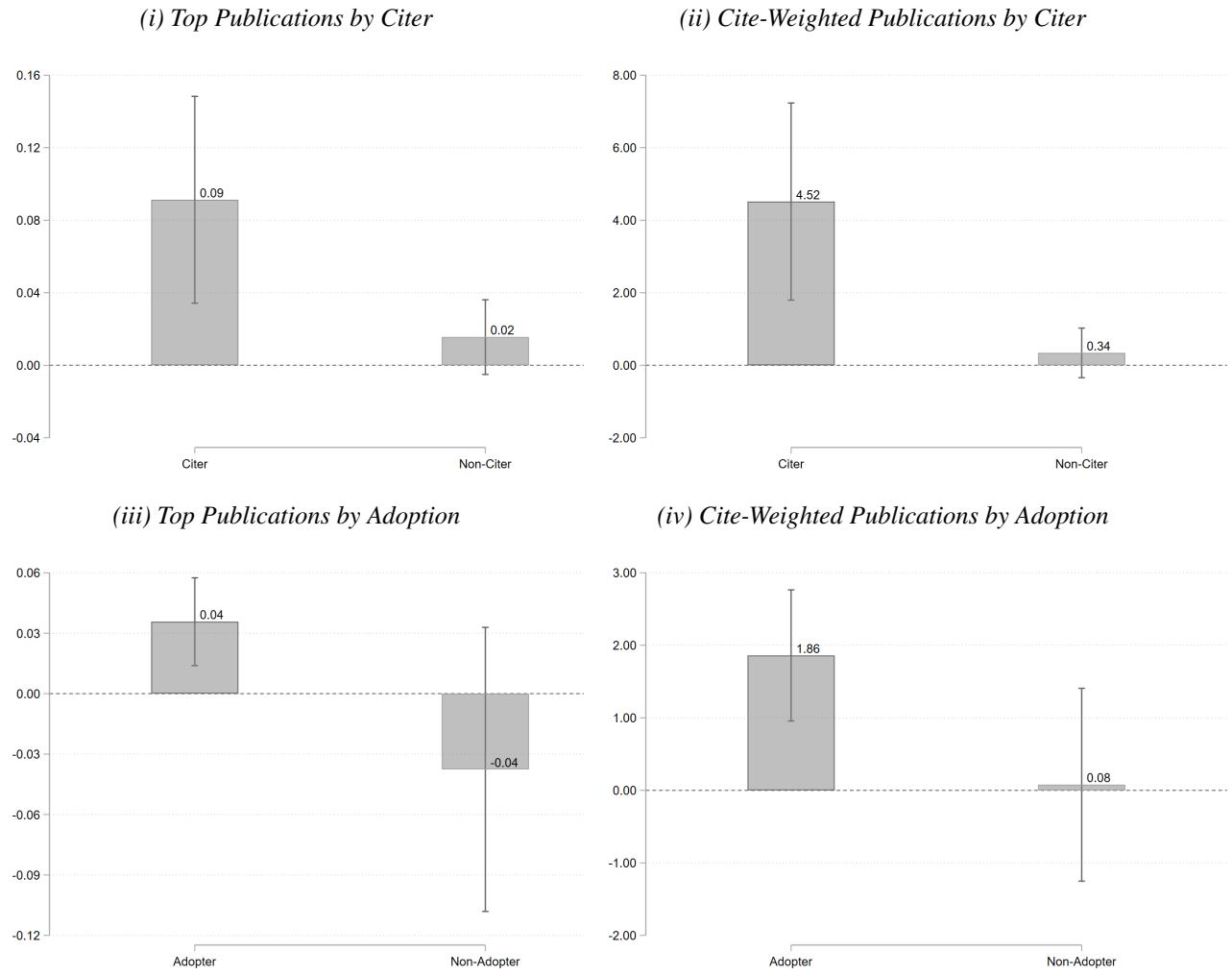
Note: This figure provides visual illustrations of the event study version of the main regression on administrative data adoption. The main dependent variables are the number of papers written using Census data (panel (i)) or the number of papers that cite papers using Census data (panel (ii)). The chart plots values of β for different lags before and after obtaining access to an FSRDC. Regressions include researcher and university \times year fixed effects. Standard errors are clustered at the researcher level.

Figure 4: Time-Varying Estimates of the Impact of FSRDCs on Research Output



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. 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)), the count of papers published in top five journals (panel (iii)), and the count of papers that are in the top 5% of the citations distribution for the year in which they were published (panel (iv)). The charts plot values of β for different lags before and after obtaining access to an FSRDC. Regressions include researcher and university \times year fixed effects. Standard errors are clustered at the researcher level.

Figure 5: Heterogeneous Effect of FSRDC Access by Intensity of Adoption



Note: This figure provides visual illustrations of the impacts of FSRDC access on measures of research output for different type of researchers using split-sample regressions. Panels (i) and (ii) show the effects separately for researchers who ever cited a paper using Census data versus those who have not. Panels (iii) and (iv) show the effects separately for researchers who work in a department with users of Census data versus those who do not. The main dependent variables are the number of top publications (panels (i) and (iii)) and the citation-weighted number of publications (panels (ii) and (iv)). Regressions include researcher and university \times year fixed effects. Standard errors are clustered at the researcher level.

Online Appendix

A Qualitative Evidence

A.1 Interviews

We carried out nine semi-structured interviews with FSRDC administrators and users to learn more about our empirical setting. The interviews have been crucial to provide us with institutional details on the FSRDC network and to validate our research design. 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.

Impact of U.S. Census Bureau’s confidential data: All the people we talked to confirmed the revolutionary impact that U.S. Census’s confidential data had on economics research. Labor, productivity, health, trade, public, and environmental economics were regarded as the fields that benefited the most. However, even more importantly, there is a sense 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)

But besides research, our respondents felt that Census data yielded findings with an outsized policy impact. Our respondents said that this was due to two characteristics of Census 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.¹⁸ 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.¹⁹ In practice, confidential microdata allow to study heterogeneous effects for sub-populations, which is what policymakers usually need (interview P39).

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

¹⁸Source: <https://news.stanford.edu/news/2010/february1/census-data-center-020210.html>

¹⁹Source: <https://news.uci.edu/2015/new-uci-center-gives-researchers-link-to-us-census-data/>

of a potential user community, the goal of geographical parity in access, and the advocacy of local research leaders. Upon becoming involved in establishing new FSRDCs, the National Science Foundation (NSF) explicitly tried to achieve an equitable geographical diffusion across the United States (interview T14). The NSF was able to pursue 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, the sheer presence of a potential user community was not enough in the absence of individuals that would advocate to open 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, which would be really good for the University.” (interview Y79)

In practice, a combination of potential users and individual advocates was key 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 a surprising amount of idiosyncrasies behind each FSRDC opening. For instance, the choice of Boston 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 some faculty members and the Census Bureau (interview S94). In cases where the FSRDC was opened by a consortium of universities, 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 both the timing and the locations of FSRDCs were strongly influenced by geographic and idiosyncratic factors, thus lending support to our identification strategy.

Career consequences: 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 Census

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 Census 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 junior faculty. The whole process requires researchers with a tenure clock 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 that 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)

Awareness of data available in FSRDCs: 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). However, even among researchers in proximity to a data center, a crucial factor determining the diffusion of Census data is awareness about their availability and research potential. For instance, administrators of the FSRDC often engage in outreach activities to inform prospective users that “there’s all this great data” (interview S12). But 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 happened not to be affiliated with the economics department, told us that many colleagues from her department started using Census data because of exposure to her work (interview S94). Another FSRDC director emphasized the importance of word of mouth to spread the impact of local FSRDCs (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, there is an increasingly large number of comparable microdata from foreign countries available to U.S.-based researchers. One researcher even suggested that the hurdles in using Census 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” on the *Quarterly Journal of Economics*. At the end of 2021, the paper was already counting over 5,600 citations on Google Scholar, corresponding to over 460 cites every year. The starting point of the paper is that firm heterogeneity and the allocation of resources across firms play a key 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, than 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 A1). 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 that it would not have been possible doing the same paper without Census data (Hsieh, personal communication, 21st December 2020).

Figure A1: Footnote from the Hsieh and Klenow (2009) Misallocation paper

*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.

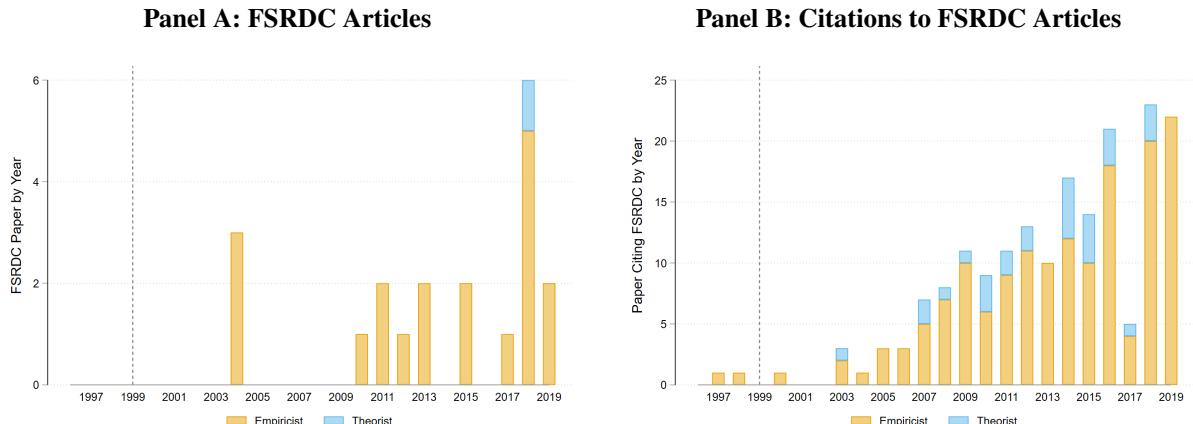
We asked to both Hsieh and Klenow about the origins of their collaboration, and how much the need to physically access Census data shaped their work. The datawork for the Misallocation paper was mostly supervised by Hsieh, who had an easier access to the Berkeley FSRDC. Indeed, the physical distance between Berkeley and Stanford was a formidable barrier to the use of confidential Census 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 was the reason that prevented Stanford from obtaining its own FSRDC for many years.

Before starting the project, the two authors had been thinking about the idea and potential methodology, but needed data for empirical validation. Then, one 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 1999²⁰, Hsieh became aware of the potential of Census data for his own research only after seeing the same data used by someone else. Figure A2 confirms this dynamics, showing a sizable increase in citations to past FSRDC work concomitantly to the increase of direct use of Census data.

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

²⁰Our interviews highlighted that the FSRDC at UC Berkeley experienced a slow initial take-up due to external events. In 1999, the IRS undertook its triennial safeguards review of the U.S. Census Bureau. In this process, the IRS found several deficiencies with its internal regulations on data confidentiality, leading to the temporary closure of the FSRDC network. Many projects active at UC Berkeley were paused until 2001, experiencing a substantial delay, as shown by Figure A2.

Figure A2: Impact of FSRDC Opening at the University of California, Berkeley



Note: This figure shows the direct impact of the FSRDC opened at the University of California, Berkeley in 1999. Panel A reports the count of articles written by Berkeley researchers using Census data. Panel B reports the count of articles written by Berkeley researchers citing past FSRDC work. The blue portion of each bar represents papers co-authored (Panel A) or citations made (Panel B) by theoretical researchers. See text for more details.

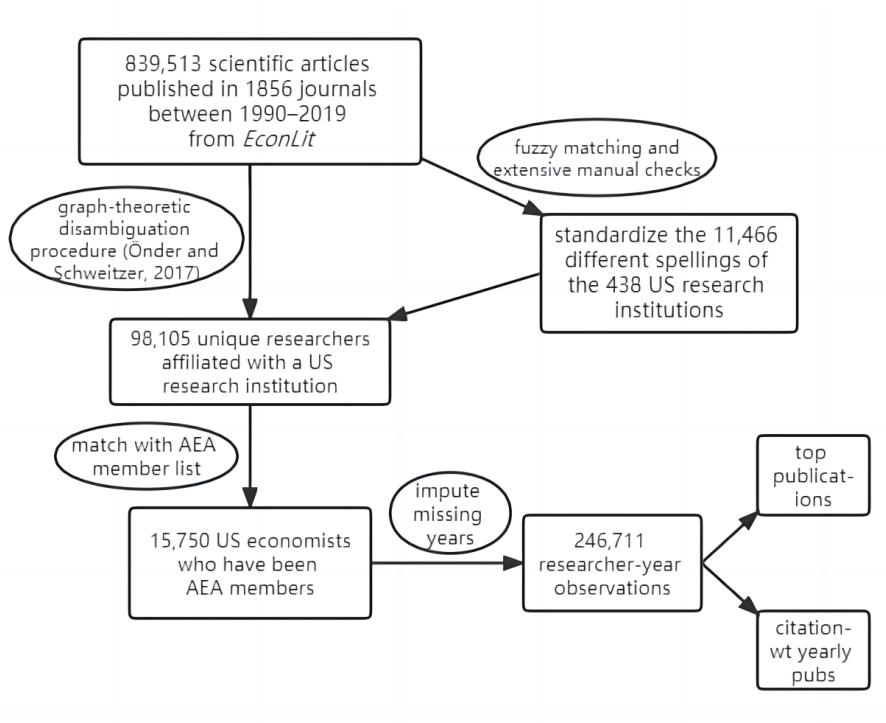
official press release of the opening, Nick Bloom was quoted saying that “having to go to Berkeley was an immense waste of time”²¹. The same point emerged in our interviews, with a researcher remarking about the long commute between Stanford and Berkeley (interview S94).

B Additional Data Details

The main 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 1856 journals between 1990 and 2019 inclusive. Unfortunately, EconLit lacks unique author identifiers, preventing from 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.

²¹Source: <https://news.stanford.edu/news/2010/february1/census-data-center-020210.html>

Figure B1: Steps Followed for Panel Construction



Note: This figure provides an illustration of 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 name 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 procedure 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 certain 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 misspelling in EconLit (e.g., “Tabellini, Gudio” instead of “Tabellini, Guido”) or to 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 consisting 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, as well as to 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.²² 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, and all the regional FED offices). The result is a list of 438 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 included 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.²³ The result is a list of 11,466 different spellings of the 438 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 fields but occasionally publishing in economics outlets. The result is an unbalanced panel of 15,750 publishing economists.

We use this set of authors to derive a panel of 246,711 researcher-year observations by imputing missing years between the first and the last year in which we see a researcher publishing. In practice, we interpolate researcher-year pairs that are missing 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

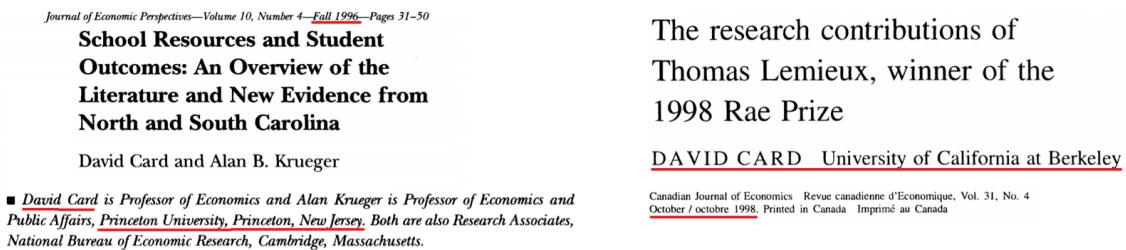
²²The Carnegie Classification of Institutions of Higher Education is available online at <https://carnegieclassifications.iu.edu/>.

²³The 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.

more common event than patenting, hence increasing the granularity of our panel and reducing the need of imputations.

Our research design takes advantage of the requirement that U.S. Census confidential data can only be analyzed in FSRDC facilities. In practice, we can approximate data access by measuring the distance between the location where the researcher works and 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



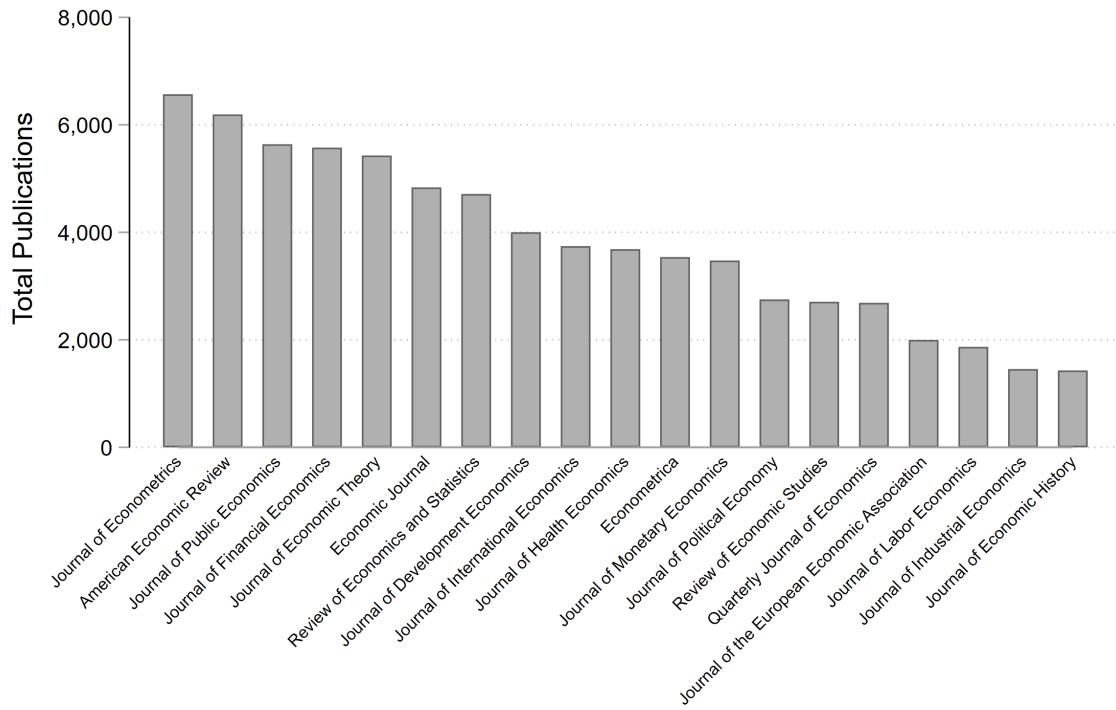
Note: This figure shows how bibliographic records can be leveraged to construct a panel of publishing economists. In particular, the affiliation details permit to detect mobility events over time and co-location to an FSRDC.

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 on 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 of the main outlets. Next, we augment EconLit by merging each article with the yearly citations 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 received by each article up to five years following publication.

Information on the ranking of economics departments is taken from Kalaitzidakis et al. (2003). We use this information to code seven tiers of universities that we use when estimating models that include a tier-specific time trend. We divide the 100 North American institutions reported by Kalaitzidakis et al. (2003) into five groups that have a roughly similar number of observations in our sample. The top tier include 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 forth 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

to foreign universities or to 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*. 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 8 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 of the other papers were correctly recorded in our data, suggesting that EconLit offers a 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 city 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 as a result of this type of measurement error.

However, downward bias could be especially 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. Overall, this implies that while junior scholars will enter our panel with some delay, the resulting attenuation bias in our estimates should be fairly small.²⁴

C FSRDC: Measuring Diffusion and Impact

This Appendix provides details on the diffusion of Federal Statistical Research Data Centers and the measurement of their research impact.

C.1 Access to FSRDCs Over Time

Table C1 shows the list of the 30 Federal Statistical Research Data Centers opened until 2019. The first was opened in Boston in the Census Bureau’s Boston Regional Office (Atrostic, 2007). After that, Carnegie Mellon University and the Center for Economic Studies pioneered a new institutional model where the data center would have been located and operated by an academic institution, with the Bureau just 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. 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. Appendix Figure E1 confirms qualitative evidence suggesting that the pattern of FSRDC openings did not explicitly prioritize higher-status universities.

Our main analysis explores the dynamics of individual productivity after gaining access to an FSRDC. We define as “treated” researchers working in a city with an active FSRDC. In the Appendix Table E8 we show the robustness of our main results with alternative definitions of the treatment, such as being affiliated with the same institution or being affiliated with an FSRDC consortium member. Alternatively, one could define the treatment status based on the actual distance between the institutional affiliation listed by researchers on

²⁴The mean time from PhD conferral to tenure is 7.13 years ($SD=2.65$), similar to the estimate of Heckman and Moktan (2020). Since the first publication appears on average one and a half years after the PhD, this validates our choice of using a cutoff of 7 years after the first publication to identify economists likely to be untenured.

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

Year opened	Institution of location	City	State
1994	NBER Cambridge	Boston	Massachusetts
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	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

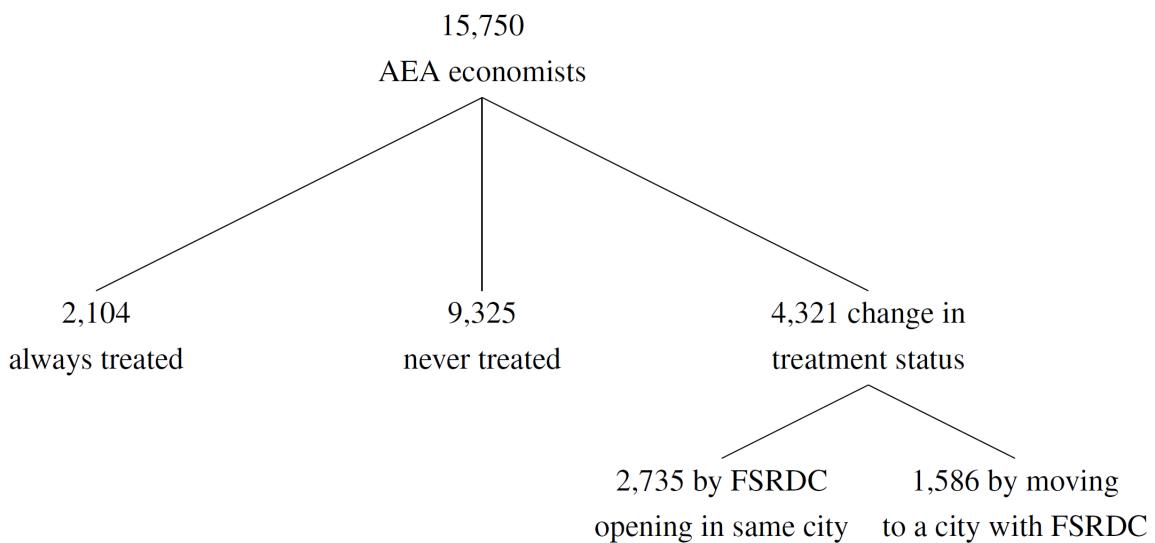
Note: The FSRDC at Carnegie Mellon University closed in 2004. Additional FSRDCs opened at the University of Utah (2020), University of Illinois Urbana-Champaign (2020), and University of Florida (2022) are outside of our sampling period.

their papers and the location of the nearest FSRDC. Appendix Figure E7 presents elasticities of our main results at various distances, showing results consistent with our main specifications.

Importantly, researchers can gain access to Census data in two ways: either when a new data center opens in the city of their current affiliation, or when they change employer and move to a new city 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 E2). We might worry that the individuals who advocated opening a local FSRDC might have an interest in using it, potentially inflating our estimates. Appendix Table E4 shows that our results are robust to the exclusion of the researchers that appeared as Principal Investigators (PIs) in the NSF grants that led to establishing the data centers. More in general, the institutions hosting the data center on their premises might be the ones that had the most to benefit from it, which would lead to upwardly biased estimates. Appendix Table E6 shows robustness of our results if we estimate our models only on the universities that gained vicarious access by being in the same city as the hosting institution.

An alternative concern is that researchers who gain access by moving to a city with an active data center might do so precisely because they want to work on confidential data. This endogenous geographical sorting could bias our estimates upward. To rule out this concern, we repeated the main analyses excluding all scholars who become co-located to a data center after changing affiliation. All the main results are robust to this test (Appendix Table E5). In addition, qualitative evidence from our interviews suggests that this concern is unlikely to hold in practice. While the presence of 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).

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



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

C.2 Articles using Census Data

We assembled a novel dataset of all articles that *directly* employ restricted-access microdata accessible only in U.S. Census' secure facilities. Since there is no official bibliographic record of research using Census data, we carefully searched for them using several complementary strategies.

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,²⁵ 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 Center for Economic Studies of the U.S. Census Bureau for online publication.²⁶ 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.²⁷ The high share of CES working papers that never get published suggests that output-based assessments of FSRDCs will underestimate the actual use of the network by academics (see Section C.3 for some alternative approaches).

Finally, we asked for access to all approved FSRDC projects via Freedom of Information Act (FOIA) request to the Census Bureau (FOIA ID No. DOC-CEN-2020-001640). Notably, the results of our request have been published online for the benefit of everyone interested in tracking the use of Census Bureau's administrative data.²⁸ 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 Census data consists of 587 articles written by 509 economics researchers.²⁹

C.3 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 Census 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 a FOIA request (FOIA ID No. DOC-CEN-2020-001640) that 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

²⁵The main limitations of JSTOR are that its coverage is unreliable for more recent years and that it does not encompass journals published by Elsevier.

²⁶The papers are available online at the following link: <https://ideas.repec.org/s/cen/wpaper.html>.

²⁷Published 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).

²⁸<https://www.census.gov/about/adrm/fsrdc/about/ongoing-projects.html>

²⁹During our data construction, we also found several articles from other disciplines, such as sociology, demography, and health policy. We excluded them because they are outside the scope of this analysis. However, it must be noted that a complete policy evaluation of the scientific impact of the FSRDC network would require expanding the focus outside economics sciences.

other agencies participating in the FSRDC program, all obtained with additional FOIA requests.³⁰ These data give us an upstream measure of data usage that is not dependent on the project’s outcomes.

Moreover, projects carried out using Census confidential data must be submitted as working papers to the Center for Economic Studies once completed. We use the count of CES working papers up to 2019 inclusive as an additional measure of data usage. Interestingly, only 48.4% of such 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). The large majority of 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).

C.4 Spillovers from Local Access to FSRDCs

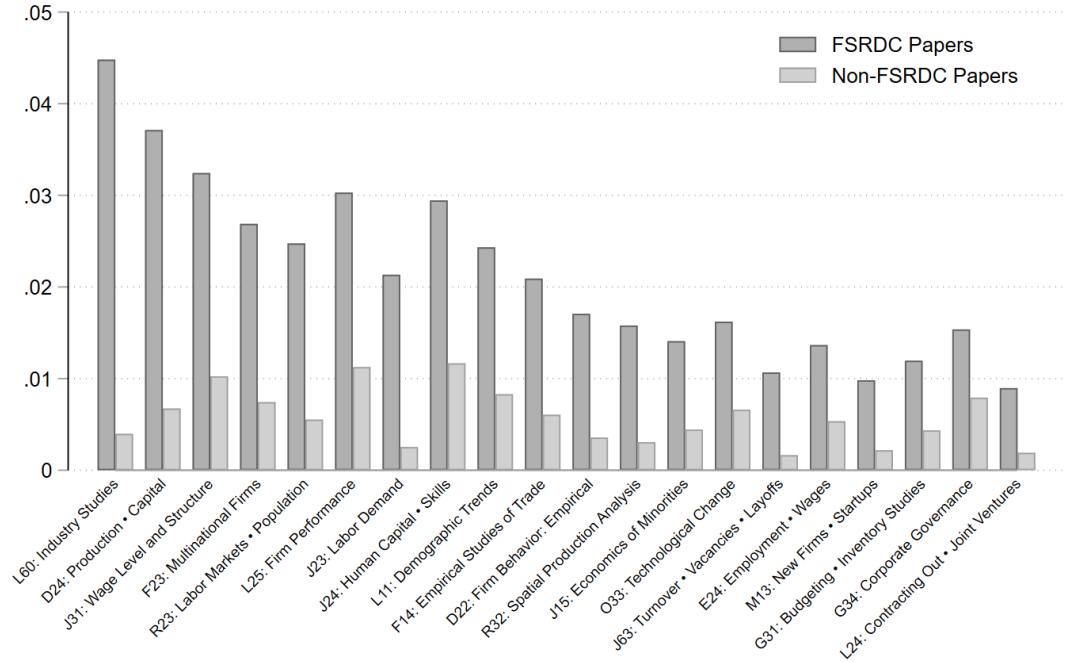
A. Articles citing FSRDC papers: Besides using confidential data available only in an FSRDC, we aim to capture how access to U.S. Census data might also affect research that does not directly use confidential Census data. Local access to Census data might be shaping the topics or questions researchers 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 Census data. This is done by taking the list of papers using Census data (identified with the procedure of Appendix Section C.2) and downloading their references from the Social Sciences Citation Index (SSCI) curated by Web of Science. We exclude papers directly using Census data 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 Census data. Researchers working in areas such as labor, productivity studies, trade, and environmental economics should benefit more from research exposure based on Census 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 research that is proximate in topical space.

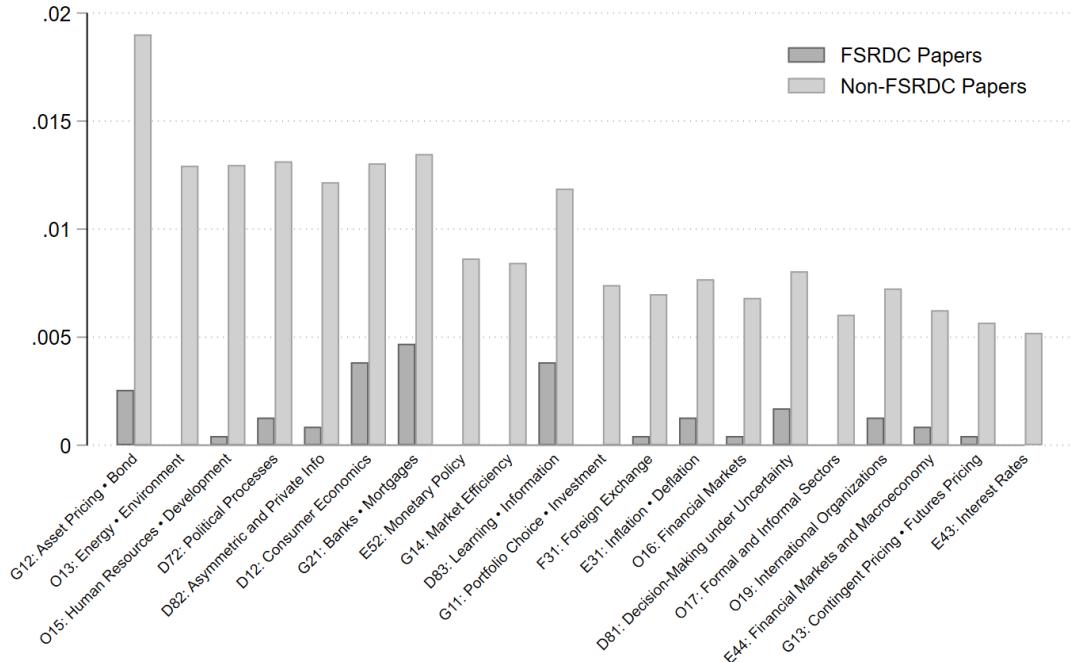
³⁰We obtained data from the AHRQ (FOIA ID No. 2021-00311-FOIA-PHS), the BEA (FOIA ID No. DOC-BEA-2021-000222), the BLS (FOIA ID No. 2021-F-00826) and the NCHS (FOIA ID No. 21-00102-FOIA).

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

(i) *JELs most representative of FSRDC research*



(ii) *JELs least representative of FSRDC research*



Note: This figure shows the JEL codes that are most distinctive of research carried out with and without confidential Census data. Each bar represents the share of papers that include that JEL code for each of the two groups of papers. Panel (a) reports the 20 JEL codes with the largest differences in frequency between papers written in an FSRDC and the rest of the sample. Panel (b) reports the 20 JEL codes with the largest negative differences in frequency between papers written in an FSRDC and the rest of the sample. JEL codes are sorted from largest to smallest difference in frequency. See text for more details.

We use the list of FSRDC papers to find the JEL codes most likely to be associated with confidential Census data. Figure C2 displays in Panel (a) the 20 JEL codes with the largest differences in frequency between papers written in an FSRDC and the rest of the sample. 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. Likewise, Panel (b) of Figure C2 shows that research carried out in Federal Statistical Research Data Centers is much less likely to encompass topics like financial markets, development, and monetary economics, as well as theoretical investigations of market efficiency and uncertainty. Appendix Table E11 shows the robustness of our main results to alternative definitions of which JEL codes are the most representative of research enabled by access to Census data.

C. Topical focus of research: The main shortcoming of relying on JEL codes is that they are assigned by the authors of the paper. Since there are no explicit rules on defining the proper JEL codes, differences in researchers' approach when choosing JEL codes might introduce measurement errors. We sidestep this issue by developing a data-driven categorization of research fields to characterize researchers' research trajectories (Furman and Teodoridis, 2020). Our approach consists in using topic modeling, a popular method for text analysis (Gentzkow et al., 2019). The main advantage of this choice is its ability to identify similarities between bodies of text without being influenced by authors' (potentially strategic) choice of JEL codes.

More in detail, we merge titles and abstracts of all the papers written by a researcher in a given year to generate researcher-year level bodies of text. Then, we follow common practices and pre-process the text using the “term frequency–inverse document frequency” (tf–idf) approach. In this way, we can downplay

Table C2: Latent Dirichlet Allocation (LDA) Topics and Their Most Salient Words

Topic Label	Most Salient Words					
topic 1 (innovation)	network	innovation	research	firm	social	
topic 2 (institutions)	patent	r&d	election	voter	vote	
topic 3 (trade)	trade	growth	export	firm	country	
topic 4 (energy)	energy	electricity	policy	cost	efficiency	
topic 5 (migrations)	quantile	immigrant	exit	immigration	alliance	
topic 6 (corporate governance)	governance	seller	buyer	culture	girl	
topic 7 (finance)	firm	bank	risk	stock	investor	
topic 8 (welfare)	food	jump	maternal	nutrition	childhood	
topic 9 (corporate finance)	merger	ambiguity	cultural	director	conservation	
topic 10 (geography)	friction	venture	gold	building	corruption	
topic 11 (econometrics)	model	forecast	method	estimator	test	
topic 12 (health/education)	health	child	household	woman	income	
topic 13 (natural resources)	renewable	sovereign	nuclear	screening	fine	
topic 14 (macroeconomics)	rate	monetary	shock	tax	price	
topic 15 (labor)	wage	worker	job	employment	labor	
topic 16 (statistics)	tail	dependence	nonparametric	bootstrap	covariate	
topic 17 (environmental)	emission	oil	carbon	climate	fuel	
topic 18 (microeconomics)	game	equilibrium	preference	agent	information	
topic 19 (accounting)	audit	dataset	reporting	auditor	transparency	
topic 20 (healthcare)	patient	hospital	ecosystem	rating	student	

Note: The table reports the five words most closely related to each LDA topic. After manual examination, we coded a plausible label to capture the content of each topic. See text for more details.

Table C3: Keywords Used to Tag Articles’ Research Design and Data

Category	Keywords (from Currie et al., 2020)
Administrative data	“administrative data” “admin data” “administrative-data” “admin-data” “administrative record” “admin record” “administrative regist” “admin regist” “registry data” “register data” “confidential data” “restricted data”
Survey data	“survey data” “phone survey” “survey administered” “household survey”
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, experimental methods. The list of keywords is taken from Currie et al. (2020).

the impact of common words while giving more weight to rare words when computing similarities between bodies of text. Next, we use an off-the-shelf algorithm known as latent Dirichlet allocation (LDA) to inductively derive 20 topics made of words that tend to co-appear in abstracts and titles. Each term is assigned a probability distribution over these clusters of words, with probabilities being higher when the term is more representative of that topic. Manual audits of the terms most representative of each data-driven cluster confirm that the procedure defines meaningful research topics (Table C2).

D. 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 n-grams we use is presented in Table C3 and directly taken from 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 C4 explores what keywords are more likely to be associated with papers using Census data. Confirming our priors, these papers are three times more likely to mention the use of confidential data or administrative records, while the same is not true for words denoting the use of traditional surveys compiled for research purposes. Further, Table C4 matches our intuition that research carried out in a Federal Statistical Research Data Center does not employ experimental research design (such as RCTs or behavioral laboratory experiments). The inclusion of year fixed effects helps alleviating the concern that these findings are just the

reflection of a wider “credibility revolution” in the years of FSRDC activity.

D Classification of Research by Methodological Orientation

In this Appendix, we explain the machine learning procedure followed in classifying 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.

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.³¹ Angrist et al. (2020) show that the algorithm 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.

³¹The algorithm is trained on a set of 5,469 hand-classified papers.

Table C4: Characteristics of Papers Using Census data

	Administrative data	Survey data	Quasi-exp. methods	Experimental methods
FSRDC Paper	0.0390*** (0.00848)	0.00841 (0.00574)	0.00975 (0.00904)	-0.0143*** (0.000530)
Year FE	YES	YES	YES	YES
N	93,694	93,694	93,694	93,694
Mean of the Dep Var	0.00475	0.0111	0.0366	0.0121

Note: This table presents estimates from OLS models evaluating the likelihood that papers using Census data mention the keywords listed in Table C3. *, **, *** denote significance at 5%, 1% and 0.1% level respectively. Robust standard errors reported.

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.

$$P(\text{Empirical}) = 96.95\%$$

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.

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.

$$P(\text{Theoretical}) = 83.6\%$$

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 has a remarkable performance also in cases where theoretical papers mention in their abstract or title keywords usually associated with empirical papers, such as “data” or “estimates”. We classify as empirical all papers with a predicted score larger than 0.5.³² 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 allow for a more precise classifications) and the fact that we consider in our analyses a broader set of journals.

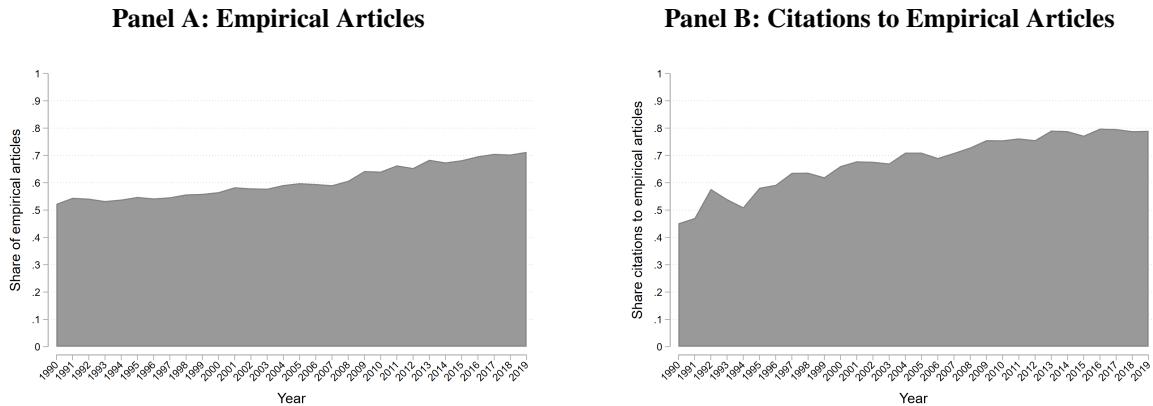
D.2 Robustness Checks

We define as empiricist all researchers with more than half of their publication output 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 73% 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 the results of our classification for the Editorial Boards of some high-profile economics

³²Unlike Angrist et al. (2020), we do not separately classify econometrics articles into an ad-hoc category.

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

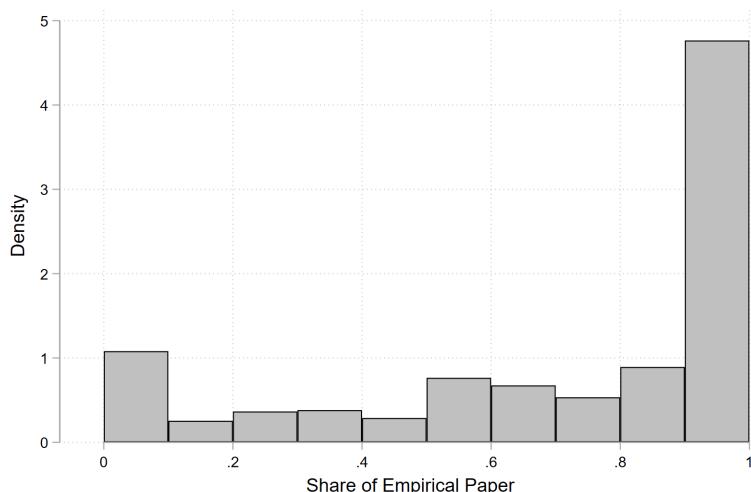


Note: This figure shows the share of empirical articles (Panel A) and citations received by empirical articles (Panel B) over time. See text for more details.

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 D1 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 98% 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 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

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



Note: The figure reports the distribution of the share of empirical published work for researchers in our sample. In our main analyses, we classified as empiricist all scholars with a share > 0.5. See text for more details.

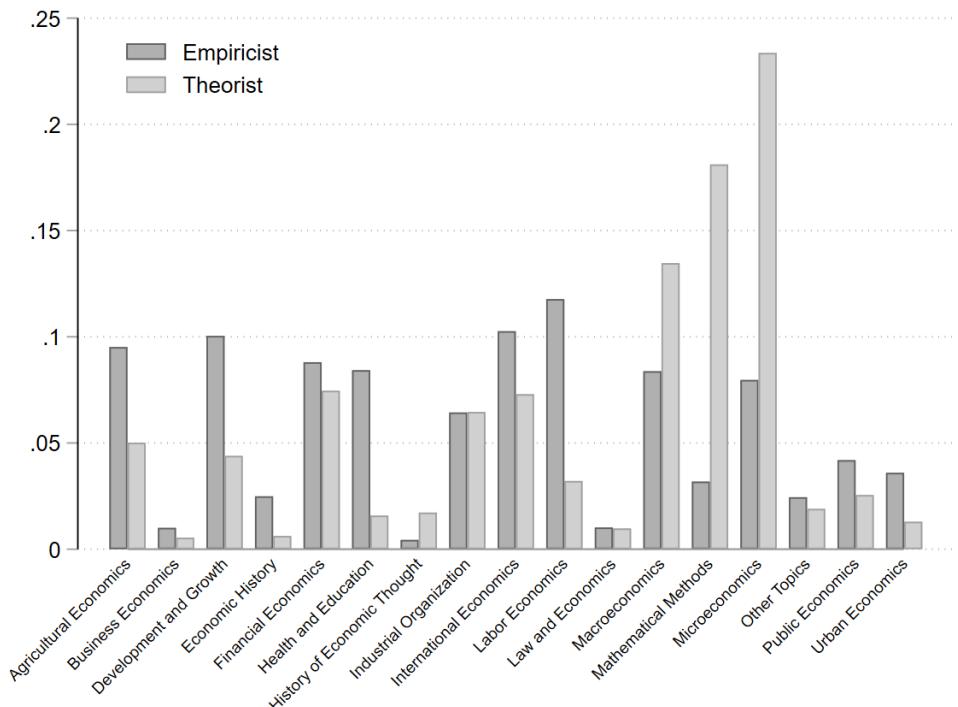
Table D1: Methodological Classification for the Editorial Board of Selected Economics Journals

Journal	Number of Editors	Ever U.S.-affiliated	Share Empirical
Journal of Economic Theory	62	47	2.13%
AEJ: Microeconomics	21	19	31.58%
AEJ: Macroeconomics	19	18	66.67%
AEJ: Applied Economics	36	33	96.97%
AEJ: Economic Policy	37	35	97.14%

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.

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, while the most popular fields for empiricists are Labor, Health, and Development Economics.

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. In total, we matched 5,209 economists, of which 3,876 classified as empiricists and 1,333 as theorists. See text for more details.

Our classification is based on the lifetime count of individual publications. As an alternative, we repeat it considering only articles published before being co-located to an FSRDC. The concordance between the two approaches is 92.8%, and grows to 95% when looking at researchers for whom we observe at least three

publications before they have access to a local FSRDC. This confirms our intuition that researchers rarely change methodological orientation during their career, and, if so, hardly because of gaining access to an FSRDC. 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 E8 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 monotonically growing when using more stringent thresholds to define empiricists.

E Appendix Tables and Figures

Table E1: Effect of FSRDC Access on Projects and Working Papers

	FSRDC Project Approval (0/1)		CES Working Paper (0/1)	
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.000996 (0.00052)	-0.00178 (0.00098)	-0.000883 (0.00052)	-0.000448 (0.00093)
Post-FSRDC × Empiricist	0.00457*** (0.00079)	0.00460*** (0.00080)	0.00200** (0.00074)	0.00182* (0.00078)
Researcher FE	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	No	Yes	No
University × Year FE	No	Yes	No	Yes
N	246532	245556	246532	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on additional measures of administrative data adoption. Columns (1) and (2) report results from OLS models, where the dependent variable is an indicator of whether the researcher has any new FSRDC project approved. Columns (3) and (4) report results from OLS 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 to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning based paper classifier. All models include individual fixed effects. Columns (1) and (3) further include year fixed effects interacted with university-tier dummies, and columns (2) and (4) further include year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively. See details in Appendix C.3.

Table E2: Effect of FSRDC Access on Research Output Using the Callaway-Sant'Anna Doubly-Robust DID Estimator.

	Use FSRDC	Cite FSRDC	Top Pubs	Cite-Pubs
	(1)	(2)	(3)	(4)
Post-FSRDC × Empiricist (ATT)	0.00598*** (0.001)	0.0209*** (0.004)	0.0259* (0.011)	2.235*** (0.460)
Researcher FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	230,348	230,348	230,348	230,348

Note: This table presents ATT estimates from models evaluating the impact of FSRDC access on measures of administrative data adoption and research output using the doubly-robust difference-indifference estimator developed by Callaway and Sant'Anna (2021). Column (1) reports result from models where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Columns (2) reports result from 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 Census data). Column (3) reports result from 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). Column (4) reports result from models where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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. Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E3: Effect of FSRDC Access for Universities

	Use FSRDC	Cite FSRDC	Top Pubs	Cite-Pubs
	(1)	(2)	(3)	(4)
Post-FSRDC	0.156*** (0.043)	0.931** (0.290)	1.927** (0.603)	105.2** (33.248)
University FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	9094	9094	9094	9094

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption and research output at the university level. Column (1) reports result from OLS models, where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Column (2) reports result 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 Census data). Column (3) reports result 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). Column (4) reports result from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-FSRDC equals one in all years after there is an operating FSRDC in the city of the university. All models include university fixed effects and year fixed effects. Standard errors are in parentheses, clustered at the university level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E4: Effect of FSRDC Access Excluding NSF Grant Applicants

	FSRDC Use (1)	FSRDC Cite (2)	Top Pubs (3)	Cite-Pubs (4)
Post-FSRDC	0.00224* (0.001)	-0.00196 (0.005)	-0.00422 (0.011)	-0.123 (0.436)
Post-FSRDC × Empiricist	0.00323** (0.001)	0.0161*** (0.004)	0.0316** (0.011)	1.664*** (0.445)
Researcher FE	Yes	Yes	Yes	Yes
University × Year FE	Yes	Yes	Yes	Yes
N	244850	244850	244850	244850

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption and research output excluding researchers who were the applicants of the NSF grant leading to a FSRDC center. Column (1) reports result from OLS models, where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Columns (2) reports result 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 Census data). Column (3) reports result 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). Column (4) reports result from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E5: Effect of FSRDC Access Excluding Authors Treated After Mobility Events

	FSRDC Use (1)	FSRDC Cite (2)	Top Pubs (3)	Cite-Pubs (4)
Post-FSRDC	0.00274 (0.001)	-0.00324 (0.005)	-0.0313* (0.015)	-0.928 (0.542)
Post-FSRDC × Empiricist	0.00347* (0.001)	0.0182*** (0.005)	0.0442*** (0.013)	2.141*** (0.560)
Researcher FE	Yes	Yes	Yes	Yes
University × Year FE	Yes	Yes	Yes	Yes
N	219941	219941	219941	219941

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption, research output, and research direction excluding researchers who got treated because of moving to an institution with an operating FSRDC. Column (1) reports result from OLS models, where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Columns (2) reports result 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 Census data). Column (3) reports result 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). Column (4) reports result from OLS models, where the dependent variable is the count of articles weighted by the number of citations received during the 5 years following their publication. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E6: Effect of FSRDC Access Excluding Authors in Institutions Hosting FSRDCs

	FSRDC Use (1)	FSRDC Cite (2)	Top Pubs (3)	Cite-Pubs (4)
Post-FSRDC	0.00249 (0.002)	-0.00400 (0.007)	-0.00573 (0.016)	-0.621 (0.636)
Post-FSRDC \times Empiricist	0.00325 (0.002)	0.0129* (0.006)	0.0315* (0.015)	2.024** (0.654)
Researcher FE	Yes	Yes	Yes	Yes
University \times Year FE	Yes	Yes	Yes	Yes
N	194110	194110	194110	194110

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption, research output, and research direction excluding researchers who affiliate with an institution with an operating FSRDC center. Column (1) reports result from OLS models, where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Columns (2) reports result 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 Census data). Column (3) reports result 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). Column (4) reports result from OLS models, where the dependent variable is the count of articles weighted by the number of citations received during the 5 years following their publication. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E7: Effect of FSRDC Access by Group

	Use FSRDC		Cite FSRDC		Top Pubs		Cite Pubs	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
City	0.00181 (0.00167)	0.00199 (0.00203)	0.00302 (0.00490)	0.00609 (0.00622)	-0.0155 (0.01689)	-0.0112 (0.01901)	-1.105 (0.75919)	-0.328 (0.89523)
Consortium	-0.000898 (0.00079)	-0.000122 (0.00162)	-0.0107* (0.00514)	-0.00629 (0.00716)	-0.000615 (0.01209)	0.0307 (0.01580)	-0.192 (0.41284)	0.0676 (0.63209)
City+Consortium	-0.0000178 (0.00133)	0.00222 (0.00200)	-0.0101 (0.00540)	-0.00579 (0.00746)	-0.00403 (0.01483)	0.0114 (0.01782)	-0.616 (0.62231)	-1.379 (0.86739)
University	-0.000956 (0.00103)	0.00403* (0.00177)	-0.0123** (0.00454)	0.00166 (0.00646)	-0.0195 (0.01284)	-0.00177 (0.01579)	-0.373 (0.47862)	1.391 (0.71213)
City × Empiricist	-0.00270 (0.00234)	-0.00307 (0.00241)	-0.00622 (0.00658)	-0.00837 (0.00686)	0.0254 (0.01898)	0.0218 (0.01953)	1.391 (0.91808)	1.190 (0.96321)
Consortium × Empiricist	0.00165 (0.00103)	0.00102 (0.00110)	0.0174** (0.00613)	0.0167** (0.00628)	0.00291 (0.01354)	0.00196 (0.01373)	0.629 (0.51243)	0.585 (0.54375)
City+Consortium × Empiricist	0.00322 (0.00204)	0.00244 (0.00212)	0.0212** (0.00685)	0.0185** (0.00716)	0.0176 (0.01683)	0.0177 (0.01684)	1.734* (0.79447)	1.857* (0.82103)
University × Empiricist	0.00741*** (0.00169)	0.00678*** (0.00171)	0.0298*** (0.00631)	0.0292*** (0.00654)	0.0540*** (0.01431)	0.0554*** (0.01449)	1.931** (0.62555)	1.932** (0.63957)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	No	Yes	No	Yes	No	Yes	No
University × Year FE	No	Yes	No	Yes	No	Yes	No	Yes
N	246532	245556	246532	245556	246532	245556	246532	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption and research output by type of researchers' affiliation. City equals one in all years after a researcher has been affiliated to a research institution located in a city (but not same university or consortium) with an operating FSRDC. Consortium equals one in all years after a researcher has been affiliated to a research institution part of an FSRDC consortium (but not in the same university or city). City+Consortium equals one in all years after a researcher has been affiliated to a research institution member of the FSRDC consortium and located in the same city of the FSRDC (but not same university). University equals one in all years after a researcher has been affiliated to a research institution with an operating FSRDC. The omitted category are researchers affiliated to a research institution without an FSRDC that is neither in the same city nor part of a consortium operating FSRDC. Columns (1) and (2) report results from OLS models, where the dependent variable is the count of articles that directly use confidential microdata 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 Census data). Columns (5) and (5) 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 is from Heckman and Moktan, 2020). Columns (7) and (8) report results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning based paper classifier. Columns (1), (3), (5), and (7) further include year fixed effects interacted with university-tier dummies and columns (2), (4), (6) and (8) further include year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E8: Alternative Methods to Measure FSRDC Exposure

	Top Pubs (1)	Cite-Pubs (2)	Top-Pubs (3)	Cite-Pubs (4)	Top-Pubs (5)	Cite-Pubs (6)
Post-FSRDC (Institution)	-0.00913 (0.016)	1.466 (0.778)				
Post-FSRDC (Institution) \times Empiricist	0.0525*** (0.014)	1.700** (0.645)				
Post-FSRDC (City)			-0.00712 (0.012)	-0.170 (0.435)		
Post-FSRDC (City) \times Empiricist				0.0352** (0.011)	1.737*** (0.445)	
Post-FSRDC (Consortium)					0.0127 (0.010)	0.00776 (0.462)
Post-FSRDC (Consortium) \times Empiricist					0.0246** (0.009)	1.395** (0.442)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University \times Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	245556	245556	245556	245556	245556	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research output using alternative measures of exposure. Columns (1), (3), and (5) 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 is from Heckman and Moktan, 2020). Columns (2), (4) and (6) report results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-FSRDC (Institution) equals one in all years after a researcher has been affiliated to a research institution hosting an operating FSRDC. Post-FSRDC (City) equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. And Post-FSRDC (Consortium) equals one in all years after a researcher has been affiliated to a research institution belonging to a consortium operating an FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E9: Effect of FSRDC Access Controlling for Broader Time Trends in Empirical Research

	Use FSRDC (1)	Cite FSRDC (2)	Top Pubs (3)	Cite-Pubs (4)
Post-FSRDC	0.00269* (0.001)	0.00653 (0.005)	0.00587 (0.012)	0.186 (0.433)
Post-FSRDC × Empiricist	0.00254* (0.001)	0.00408 (0.004)	0.0159 (0.011)	1.204** (0.449)
Researcher FE	Yes	Yes	Yes	Yes
University × Year FE	Yes	Yes	Yes	Yes
Empiricist × Year Bin FE	Yes	Yes	Yes	Yes
N	245556	245556	245556	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption and research output. Column (1) reports results from an OLS model, where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Column (2) reports results 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 Census data). Column (3) reports results 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 results from an OLS model, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning based paper classifier. All models include individual fixed effects, year fixed effects interacted with university dummies, and empiricist fixed effects interacted with time dummies (in bins of five years each). Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E10: Effect of FSRDC Access Excluding Direct FSRDC Users

	FSRDC Cite		Top Pubs		Cite-Pubs	
	(1)	(2)	(3)	(4)	(5)	(6)
Post-FSRDC	-0.00674** (0.003)	-0.00440 (0.004)	-0.0147 (0.009)	-0.00834 (0.011)	-0.510 (0.313)	-0.331 (0.392)
Post-FSRDC × Empiricist	0.0107*** (0.003)	0.0107** (0.003)	0.0341** (0.010)	0.0337** (0.010)	1.278*** (0.372)	1.318*** (0.391)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University Tier × Year FE	Yes	No	Yes	No	Yes	No
University × Year FE	No	No	No	No	No	No
N	238710	237747	238710	237747	238710	237747

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption and research output excluding researchers who ever used FSRDC. Columns (1) and (2) 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 Census data). Columns (3) and (4) report result 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 (5) and (6) report results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime publications are mostly empirical in nature, as assessed by our machine-learning based paper classifier. All models include individual fixed effects. Columns (1), (3) and (5) further include year fixed effects interacted with university-tier dummies, and columns (2), (4) and (6) further include year fixed effects interacted with university dummies. Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E11: Effect of FSRDC Access on Alternative Measures of Research Direction

Panel A: Papers Using JEL Codes (dummy)						
	Definition 1		Definition 2		Definition 3	
	FSRDC (1)	Non-FSRDC (2)	FSRDC (3)	Non-FSRDC (4)	FSRDC (5)	Non-FSRDC (6)
Post-FSRDC	-0.00332 (0.007)	-0.0125 (0.010)	-0.0107 (0.006)	-0.0105 (0.010)	-0.00817 (0.008)	-0.0127 (0.010)
Post-FSRDC \times Empiricist	0.0175** (0.006)	0.0278** (0.009)	0.0278*** (0.006)	0.0263** (0.009)	0.0355*** (0.007)	0.0294** (0.009)
Dependent Variable Mean	0.098	0.320	0.099	0.320	0.160	0.311
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University \times Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	245556	245556	245556	245556	245556	245556

Panel B: Papers Using JEL Codes (count)						
	Definition 1		Definition 2		Definition 3	
	FSRDC (1)	Non-FSRDC (2)	FSRDC (3)	Non-FSRDC (4)	FSRDC (5)	Non-FSRDC (6)
Post-FSRDC	-0.00837 (0.009)	-0.0180 (0.019)	-0.0188* (0.008)	-0.0159 (0.020)	-0.0135 (0.011)	-0.0160 (0.019)
Post-FSRDC \times Empiricist	0.0268** (0.008)	0.0668*** (0.018)	0.0412*** (0.008)	0.0646*** (0.018)	0.0576*** (0.011)	0.0676*** (0.018)
Dependent Variable Mean	0.116	0.450	0.116	0.451	0.197	0.434
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University \times Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	245556	245556	245556	245556	245556	245556

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of research directions. We experiment with alternative coding of what JEL codes are the most representative of FSRDC research. Definition 1: JEL codes with the largest frequency difference between FSRDC and non-FSRDC papers (this reproduces the main analyses). Definition 2: JEL codes that are among the twenty most frequent for FSRDC papers, but not among the twenty most frequent in non-FSRDC papers. Definition 3: JEL codes with the frequency difference between FSRDC and non-FSRDC papers larger than 0.03. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E12: Effect of FSRDC Access on Additional Proxies of Research Design

	Big Data (1)	Data (2)	Natural Exp. (3)
Post-FSRDC	-0.000457* (0.000)	0.0204* (0.009)	-0.00139 (0.001)
Post-FSRDC \times Empiricist	0.0000839 (0.000)	-0.00139 (0.008)	0.00335** (0.001)
Researcher FE	Yes	Yes	Yes
University \times Year FE	Yes	Yes	Yes
N	245556	245556	245556

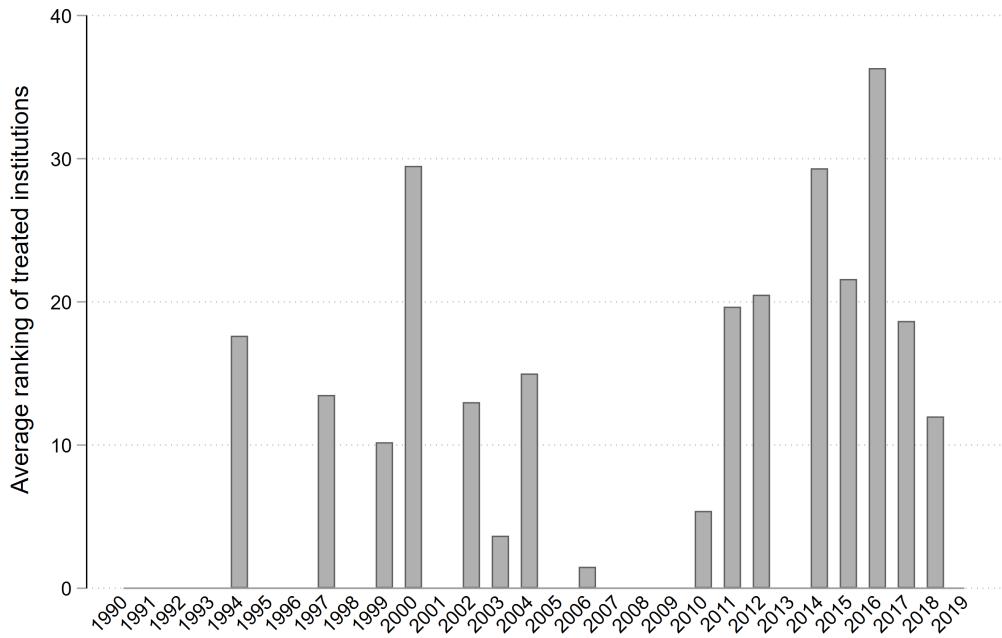
Note.: This table presents estimates from OLS models evaluating the impact of FSRDC access on additional text-based proxies of research design. Columns (1) and (2) report results from OLS models, where the dependent variable is the number of published articles that mention in their title or abstract keywords related to the use of big data or just data, respectively. Column (3) reports results from an OLS model, where the dependent variable is the number of published articles that mention in their title or abstract keywords related to the use of natural experiments. Post-FSRDC equals one in all years after a researcher has been affiliated to a research institution located in a city with an operating FSRDC. Empiricist equals one for those researchers whose lifetime 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 researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Table E13: Effect of FSRDC Access: Timing of NSF Grant

	FSRDC Use (1)	FSRDC Cite (2)	Top Pubs (3)	Cite-Pubs (4)
Post-Grant	0.000894 (0.001)	-0.00810* (0.003)	-0.0143 (0.011)	-0.558 (0.362)
Post-Grant × Empiricist	0.00152 (0.001)	0.0138*** (0.004)	0.0296** (0.011)	1.502*** (0.411)
Researcher FE	Yes	Yes	Yes	Yes
University Year × Year FE	Yes	Yes	Yes	Yes
N	229756	229756	229756	229756

Note: This table presents estimates from OLS models evaluating the impact of FSRDC access on measures of administrative data adoption and research output using alternative measures of exposure. Column (1) reports result from OLS models, where the dependent variable is the count of articles that directly use confidential microdata accessible only in an FSRDC. Columns (2) reports result 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 Census data). Columns (3) reports result 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). Column (4) reports result from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to 5 years following their publication. Post-Grant equals one in all years after a researcher has been affiliated to a research institution that received an NSF grant to establish a local FSRDC. Empiricist equals one for those researchers whose lifetime 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 fixed effects. Standard errors are in parentheses, clustered at the researcher level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

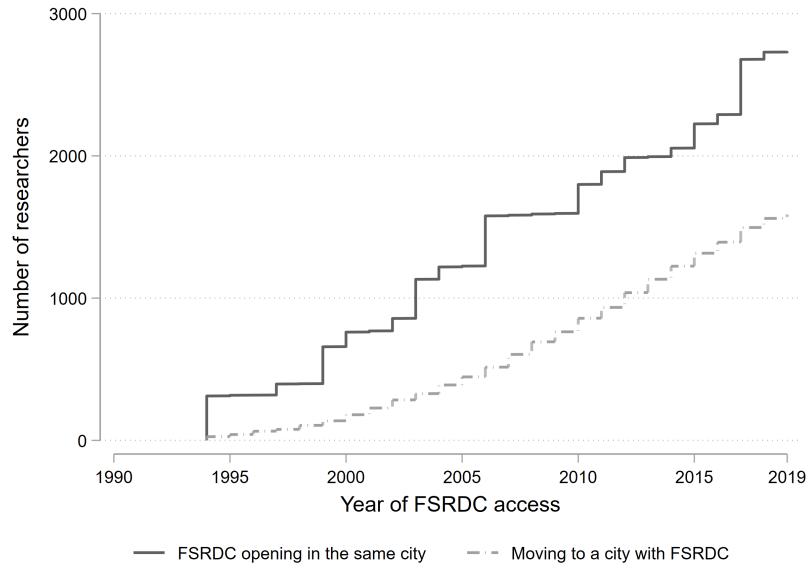
Figure E1: Rank of Universities Treated by FSRC Openings (average)



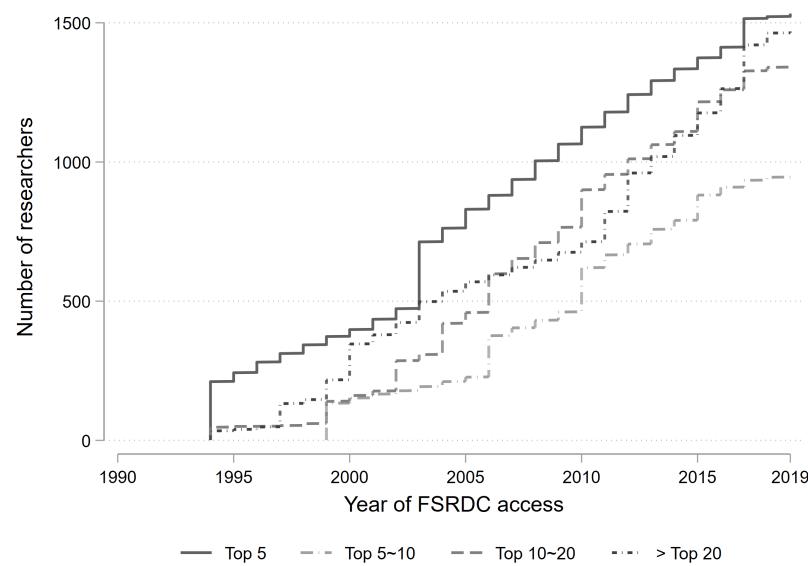
Note: The figure reports the average rank of universities treated by the opening of a new FSRDC in their city. Informations on the ranking of economics departments is taken from Kalaitzidakis et al. (2003). Lower numbers correspond to departments ranked higher. See text for more details.

Figure E2: Cumulative Number of Researchers Gaining FSRDC Access over Time

(i) *By Modality of Access*

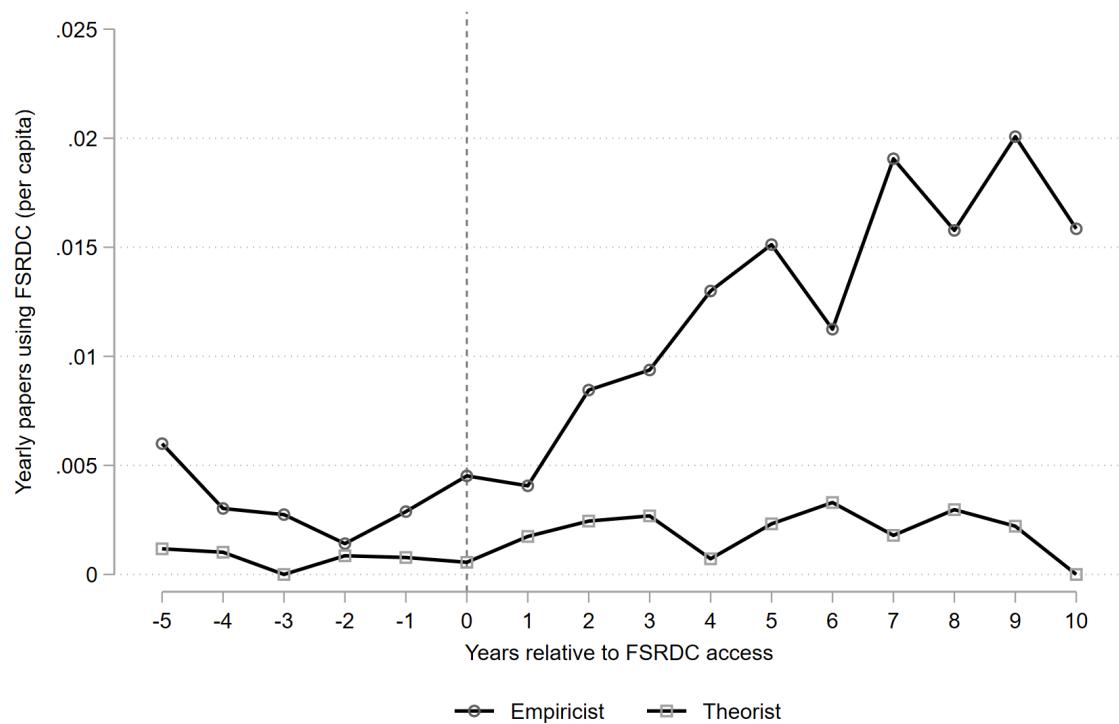


(ii) *By Ranking of Institution*



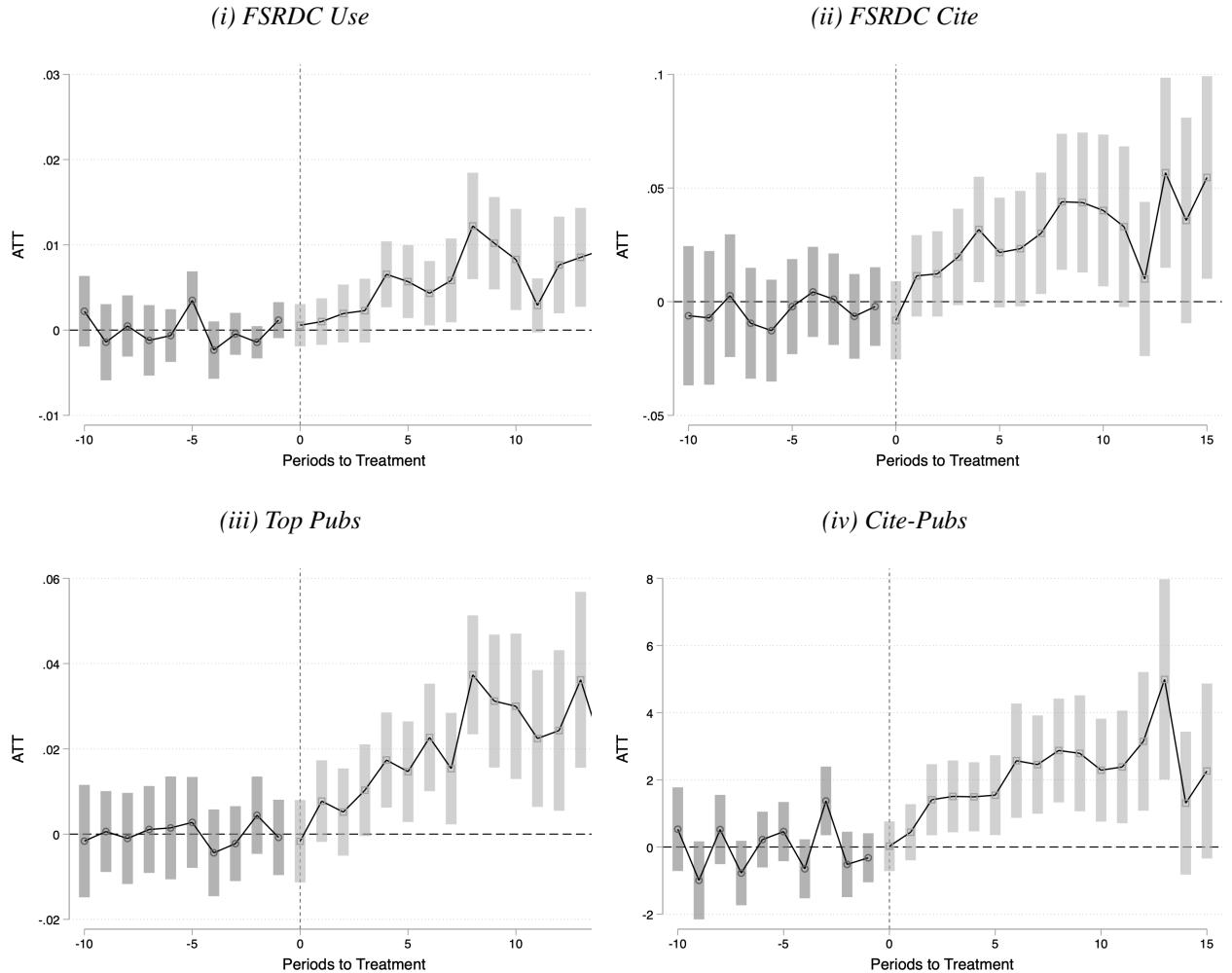
Note: This figure plots the cumulative number of researchers that gain access to an FSRDC in the city where they are located over time. The cumulative frequencies are split by the modality in which researchers become co-located to an FSRDC (panel (i)) and the ranking of their institution of affiliation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

Figure E3: Yearly Papers using Census Data (per capita averages)



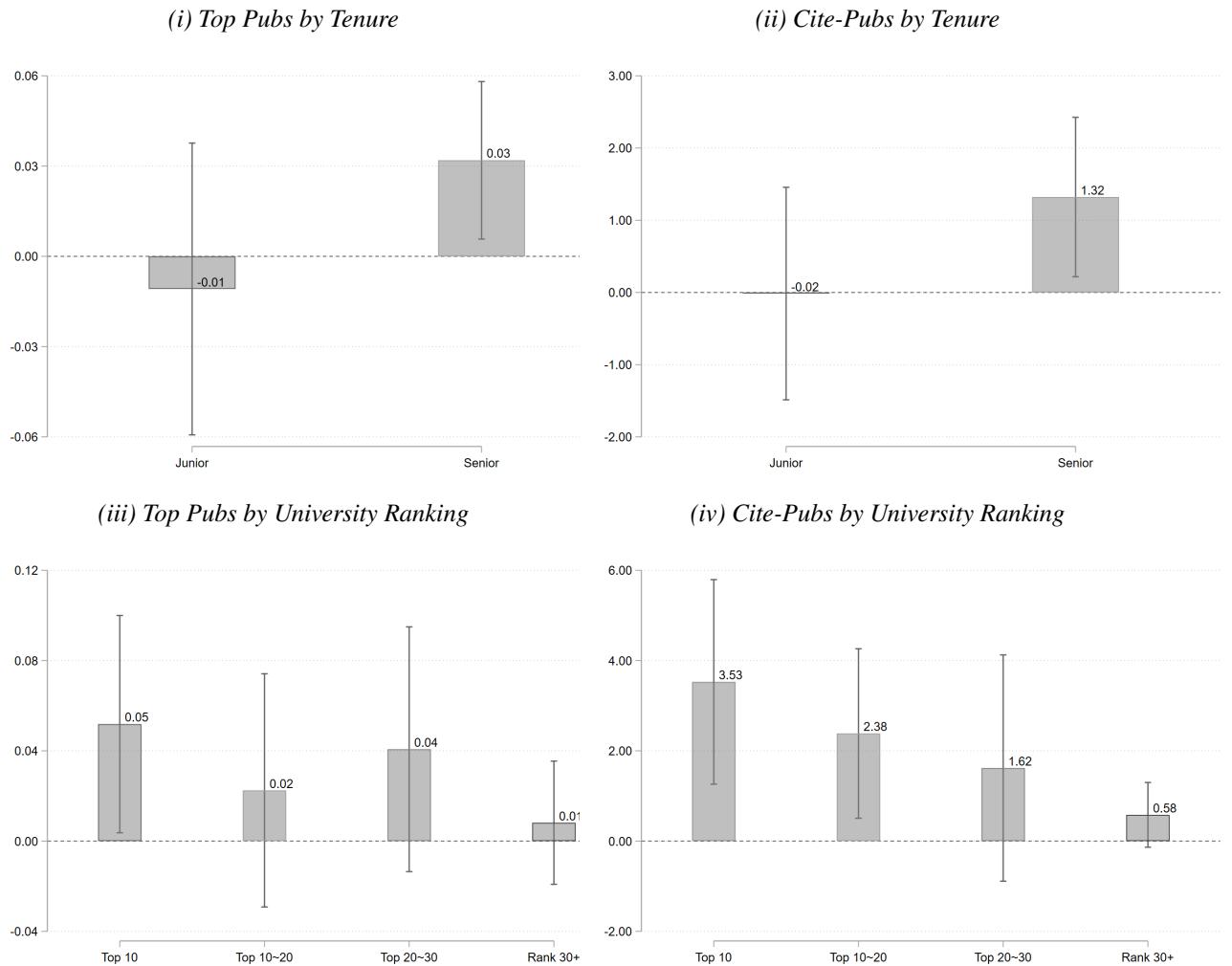
Note: The figure reports the average per capita number of yearly articles written using Census data. The yearly individual means are computed relative to the time of first access to a FSRDC. See text for more details.

Figure E4: Time-Varying Estimates of the Impact of FSRDCs on Research Output of Applied Scholars, Obtained with the Callaway-Sant'Anna Doubly-Robust DID Estimator.



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-indifference estimator developed by Callaway and Sant'Anna (2021). The main dependent variables are the number of papers written using Census data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), and the citation-weighted number of publications (Panel (iv)). The charts plot values of β for different lags before and after obtaining access to an FSRDC. Regressions include researcher and university \times year fixed effects. Standard errors are clustered at the researcher level.

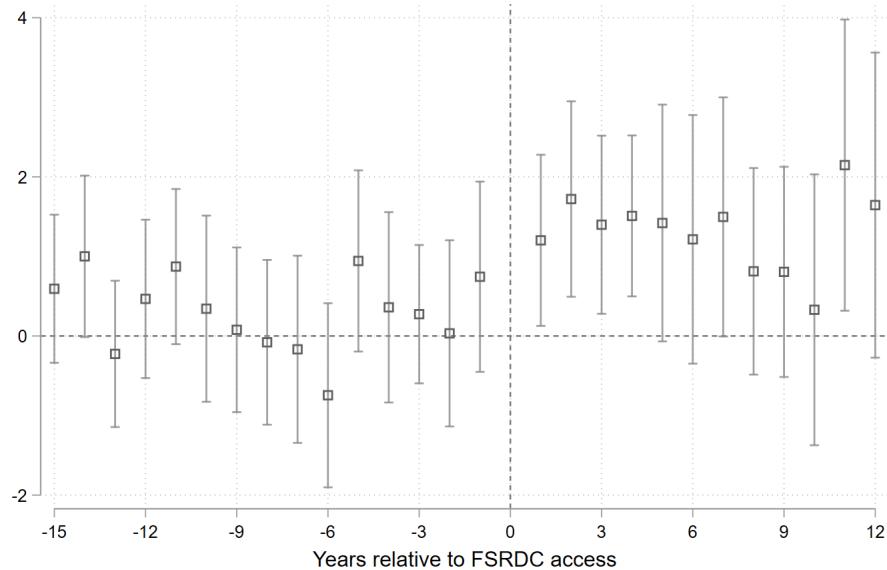
Figure E5: Heterogeneous Effect of FSRDC Access



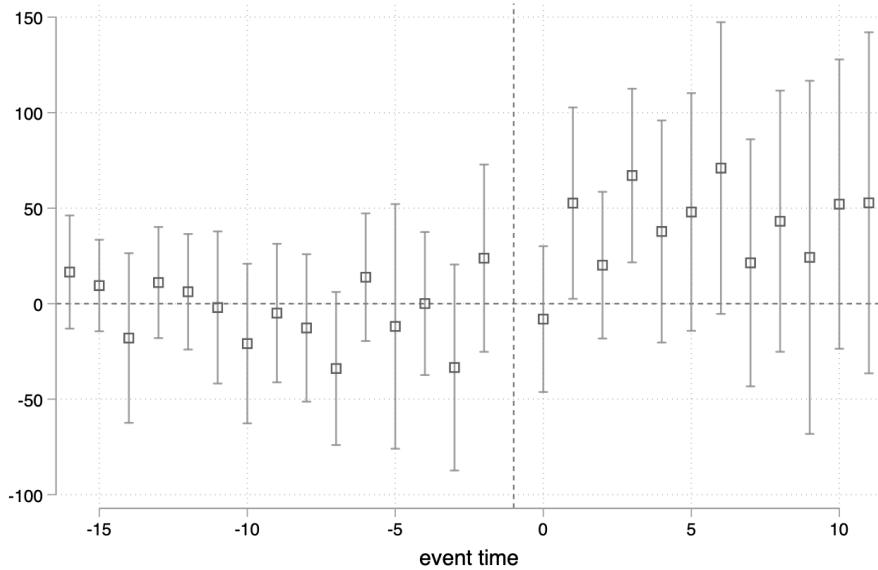
Note: This figure provides visual illustrations of the estimates from OLS models evaluating the impact of FSRDC access on research output. The main dependent variables are the number of top publications (Panel (i) and (iii)), and the citation-weighted number of publications (Panel (ii) and (iv)). The coefficients displayed are from split sample regressions. Junior: 0/1=1 for researchers that have started publishing less than seven years before. Information on the ranking of economics departments is taken from Kalaitzidakis et al. (2003). Regressions include researcher and university \times year fixed effects. Standard errors are clustered at the researcher level. See text for more details.

Figure E6: University-Level Time-Varying Estimates of the Impact of FSRDCs

(i) *Top Publications*



(ii) *Cite-weighted 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 count of articles published in the most prestigious economics journals (panel (i)) or the the count of articles weighted by the number of citations received up to 5 years following their publication (panel (ii)). The chart plots values of β for different lags before and after obtaining access to an FSRDC. All models include university fixed effects and year fixed effects. Standard errors are clustered at the university level. *, **, *** denote significance at 5%, 1% and 0.1% level respectively.

Figure E7: Effect of FSRDC Access by Distance

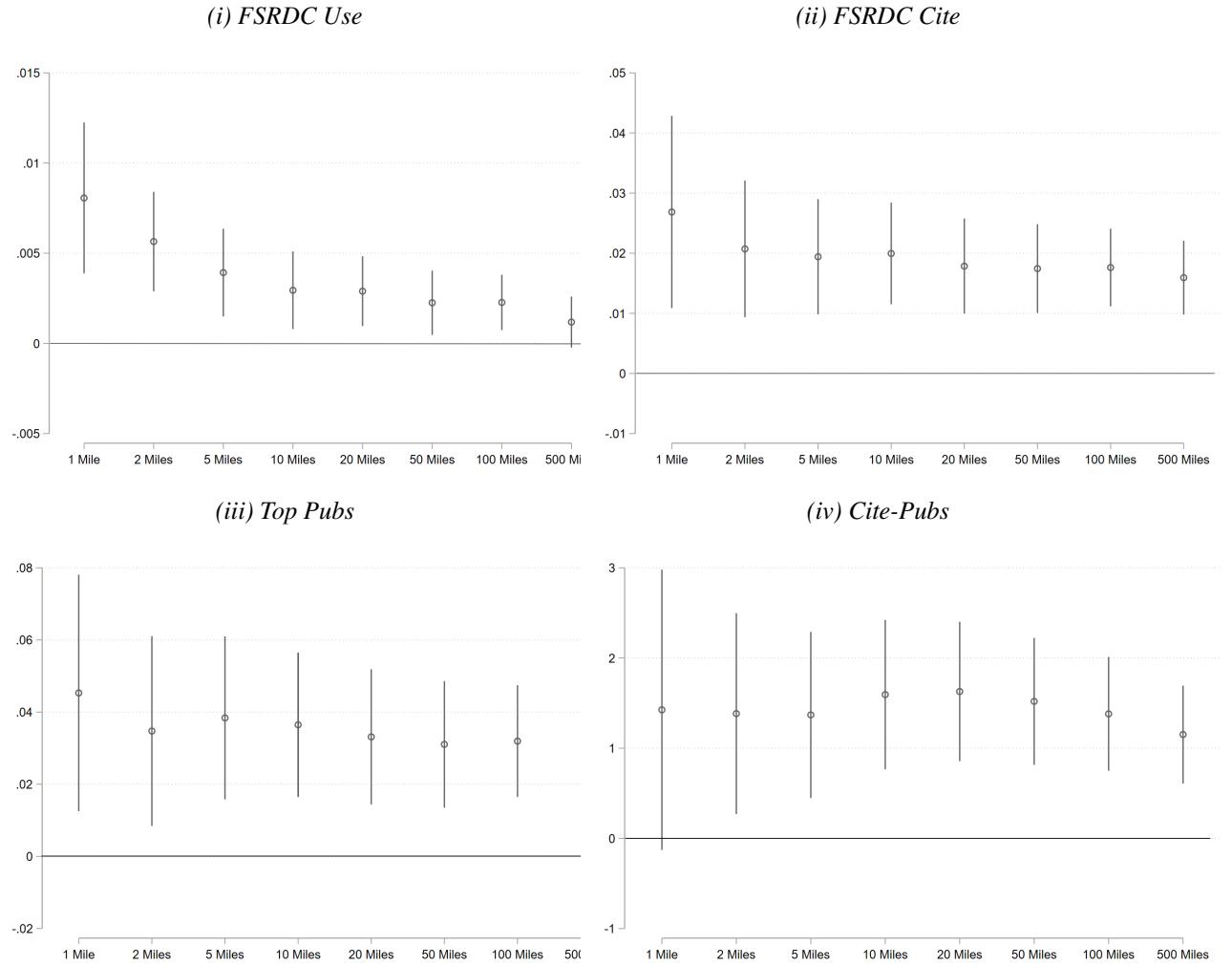
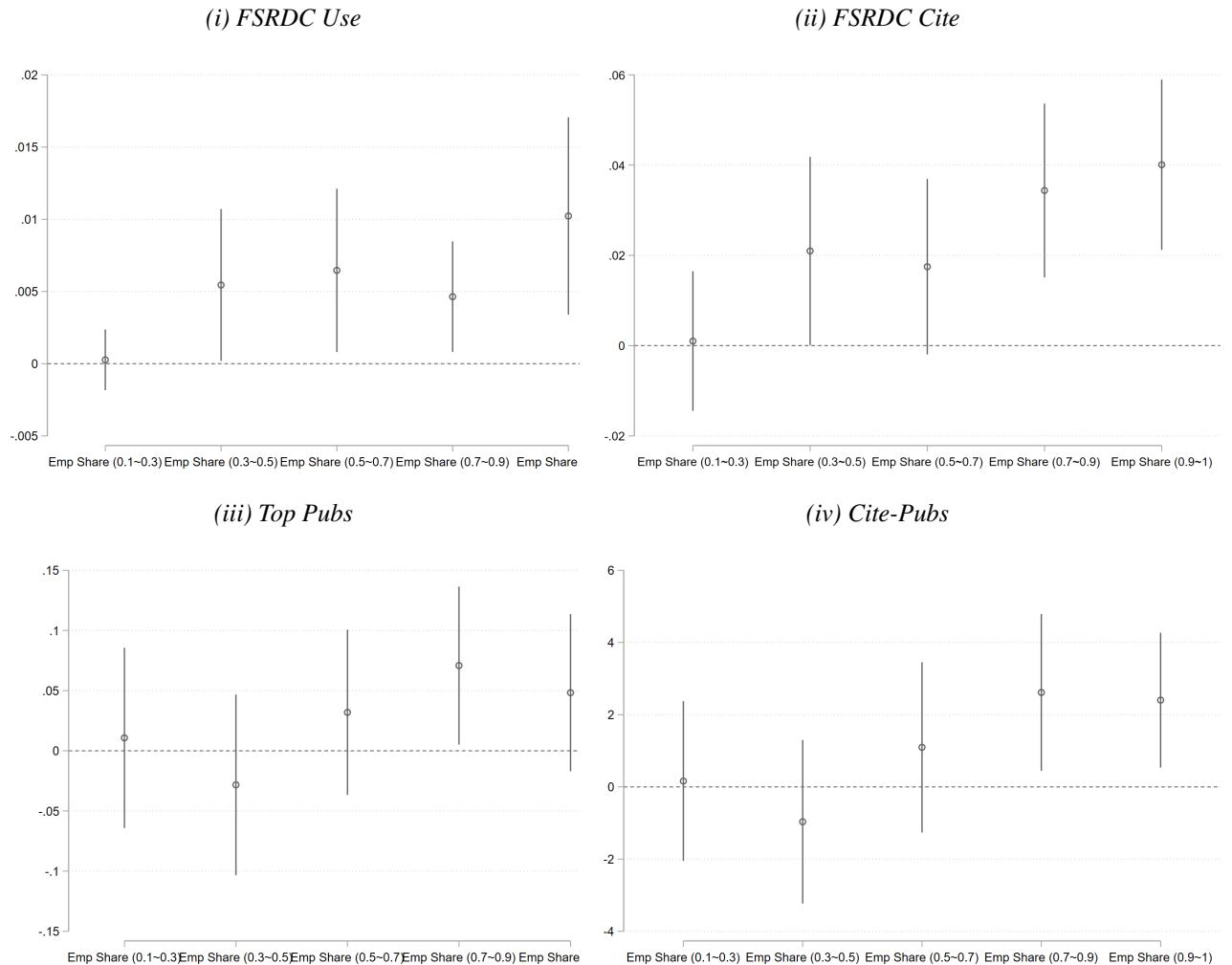
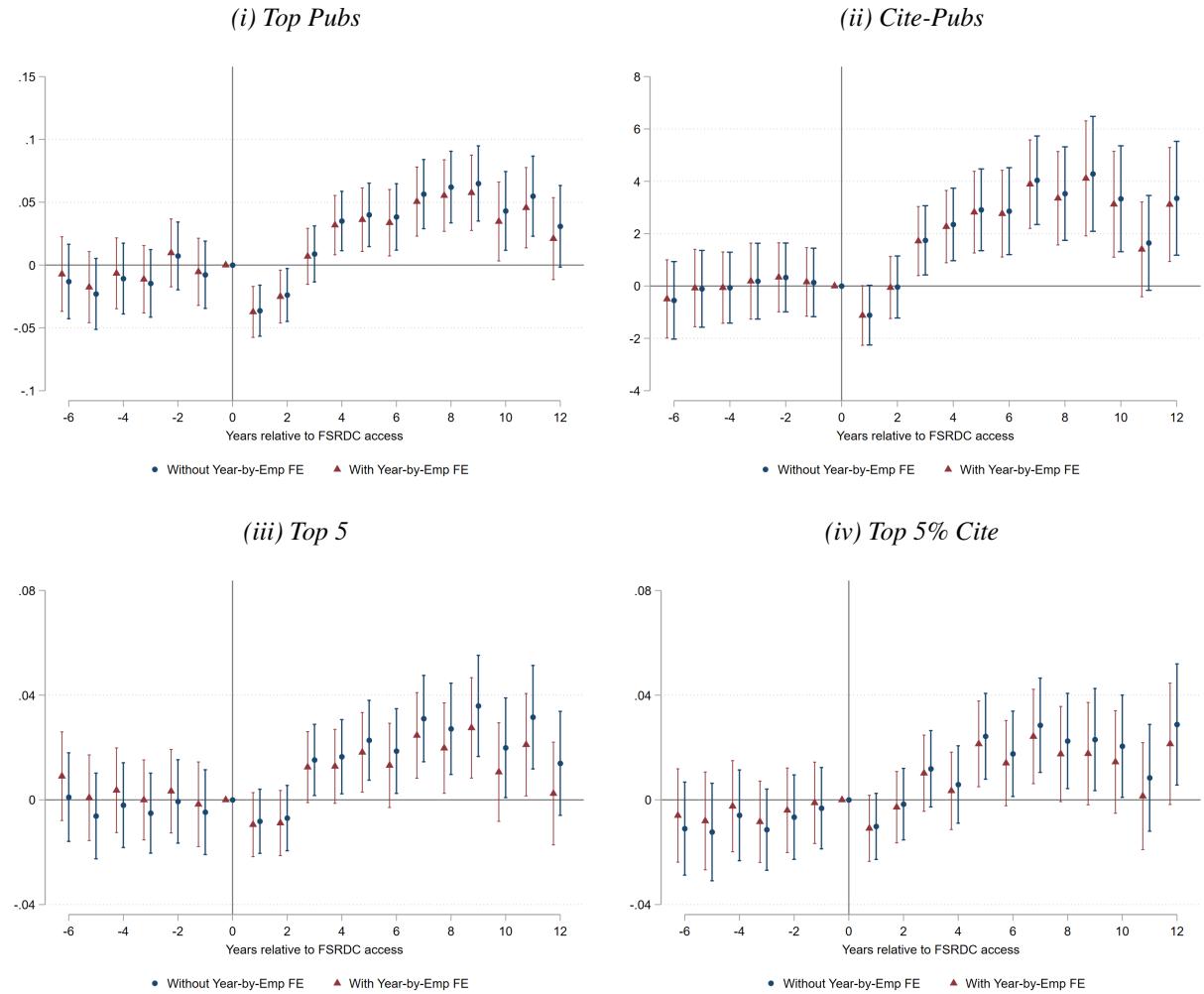


Figure E8: Effect of FSRDC Access by Share of Empirical Work



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 scholarship that is empirical in nature. The main dependent variables are the number of papers written using Census data (panel (i)), the number of papers that cite FSRDC papers (panel (ii)), the number of top publications (Panel (iii)), and the citation-weighted number of publications (Panel (iv)). The chart plots coefficients from a regression where researchers are classified into six mutually exclusive categories, depending on the share of their publications that is empirical in nature. The excluded category are researchers with no empirical papers. Standard errors are clustered at the researcher level.

Figure E9: Time-Varying Estimates of the Impact of FSRDCs on Research Output of Applied Scholars



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. 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 5 years following publication (panel (ii)), count of papers published in Top Five journals (panel (iii)), and count of papers that are in the top 5% of the citations distribution (panel (iv)). The charts plot values of β for different lags before and after obtaining access to an FSRDC. Regressions include researcher and university \times year fixed effects. Coefficients depicted with a triangle are from the same model estimated including additional time fixed effects for empirical researchers in bins of five years each. Standard errors are clustered at the researcher level.