NBER WORKING PAPER SERIES

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

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

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

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, Workshop on Big Data Analyses and New Developments in Research Data Centers at ZEW Mannheim, 2023 BITSS Annual Meeting, Columbia MAD 2023, and Center for Economic Studies seminar 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. Abhishek Nagaraj and Matteo Tranchero have no material interests to declare. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Does Data Access Shape Science? Evidence from the Impact of U.S. Census's Research Data Centers on Economics Research

Abhishek Nagaraj and Matteo Tranchero

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

ABSTRACT

This study examines the impact of access to confidential administrative data on the rate, direction, and policy relevance of economics research. To study this question, we exploit the progressive geographic expansion of the U.S. Census Bureau's Federal Statistical Research Data Centers (FSRDCs). FSRDCs boost data diffusion, help empirical researchers publish more articles in top outlets, and increase citation-weighted publications. Besides direct data usage, spillovers to non-adopters also drive this effect. Further, citations to exposed researchers in policy documents increase significantly. Our findings underscore the importance of data access for scientific progress and evidence-based policy formulation.

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

Matteo Tranchero
University of California, Berkeley
m.tranchero@berkeley.edu

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

1 Introduction

Modern science is largely empirical. In fields as diverse as astronomy, chemistry, and environmental sciences, researchers increasingly rely on large-scale, centralized datasets rather than on data curated for a single question (Hill and Stein, 2021; Locarnini et al., 2018; York et al., 2000). Economics is no exception (Backhouse and Cherrier, 2017). The share of theoretical papers published in top economics journals decreased from 50.7% in 1963 to 19.1% in 2011 (Hamermesh, 2013), while empirical work has surged during this period (Angrist et al., 2020). A key factor behind this surge has been the use of confidential administrative data from government sources. Around 20% of recent articles in the five most prestigious economics journals use such data (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). Academic economists have, therefore, urged for broader access to large-scale microdata from administrative sources claiming significant benefits to research and policy (Card et al., 2010). However, administrative data access is restricted and cumbersome due to significant privacy risks and the consequent need for tight security standards (Foster et al., 2009; Abowd and Lane, 2004). Understanding how and to what extent data access can benefit research and policy would help shed light on current debates, which are primarily centered around privacy protection (Abowd, 2018; Chetty and Friedman, 2019). Despite this urgent need, concrete evidence of the impacts of broadening data access remains thin.

That broadening data access will significantly benefit research and policy is not a given. First, data access must necessarily be accompanied by cumbersome security regulations, which might mean low levels of adoption by researchers. Even if these data do get adopted, access to the same pool of data could trigger racing dynamics that ultimately lower the quality of science (Hill and Stein, 2021). Lower quality could also result from asking more marginal questions to suit the "enormous bunch of data," rather than "a puzzle that needs to be explained" in the words of Bob Solow (Dizikes, 2019). Even if data access aids research, the magnitude of such benefits is essential to establish. Finally, it is important to understand whether benefits come from direct users or spillovers to non-adopters who build on their results and whether they remain confined to the ivory tower or diffuse more broadly to policy-makers.

We shed light on these questions by studying how access to confidential administrative data from the U.S. Census shapes the quantity, quality, and policy impact 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 was set up all over the country in a phased manner between 1994 and 2019. As multiple analyses will suggest, the timing and location of these openings were partly driven by factors governing geographic equity rather than by pre-existing trends in the use of administrative data or research output. Therefore, the focus on Census data 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 usage 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. Contrary to views relating data access to more marginal research, 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 in boosting the scientific 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, confirming qualitative accounts of the bureaucratic costs and uncertainties required to use confidential data.

Our results are robust to a host of potential concerns. First, detailed event studies, 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 (or assigning treatment based on a researcher's first institution), 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 data 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.

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 (Myers and Lanahan, 2022). Yet, these findings apply only to economists who either have colleagues using Census data or who cite Census-based papers. The effects disappear for departments without data adoption or authors not leveraging FSRDC-enabled research. These patterns suggest that local data access shapes academic output by exposing economists to research based on administrative data.

How does exposure spill over into greater productivity? The story does not seem to be one where researchers pursue pre-existing research topics and methods with better data. Instead, our results suggest that increased awareness about Census-based research might inspire a change in research direction or the adoption of research designs that lead to more impactful work. 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 new databases and robust methodologies.

Finally, we test whether access to administrative data increases the policy impact of economic research. Using novel data on citations of economic research in policy documents, we show that local access to an FSRDC leads applied economists to publish more policy-relevant research. Furthermore, the effect size is larger among U.S.-based policy sources, confirming the impact of the federal data infrastructure on evidence-based policymaking in the United States. This effect seems to be driven by the fact that administrative data lead to findings of higher scientific quality and, consequently, more policy relevance (Card et al., 2010; Chetty, 2012; Einav and Levin, 2014b).

Our work contributes to three different strands of research. New legislation in the U.S. (like the Evidence-

Based Policy Act and the Federal Data Strategy[1]) 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; Chetty et al., 2018, 2020). Yet, while some have hailed current data access programs as important, others have criticized them for being costly and cumbersome, with uncertain impact (Atrostic, 2007; Card et al., 2010; CES, 2017; Cole et al., 2020). Ironically, most of these debates have themselves been data-free. We contribute by providing, to our knowledge, the first systematic investigation of the effects of administrative data access on academic research, moving the debate beyond simply the privacy risks of data access. Further, by linking research to policy using novel data from policy documents, our work highlights how data access, even under restrictive conditions, can become a key driver of evidence-based policies (Hjort et al., 2021; Yin et al., 2022).

Second, we add to recent research that has investigated questions of relevance to the economics profession using bibliometric data. This includes past work documenting the empirical turn taken by economics (Backhouse and Cherrier, 2017; Brodeur et al., 2020; Currie et al., 2020; Einav and Levin, 2014a; Hamermesh, 2013). Our results show that increased access to high-quality data is an important factor driving the increase in the impact and credibility of empirical scholarship (Angrist et al., 2020; Brodeur et al., 2020). In addition, we 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; Feenberg 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 Teodoridis, 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, direction, and policy impact of scientific innovation (Hoelzemann et al., 2022; Nagaraj et al., 2020; Nagaraj, 2022; Williams, 2013). Further, researchers have recently looked at the drivers of topic selection among academics (Truffa and Wong, 2022; Myers, 2020). We show how research inputs, and in particular data, can shape topic choice. Finally, there is growing interest in examining how scientific progress can inform government and social policy (Hjort et al., 2021; Yin et al., 2021). Our work is among the first in the economics of innovation literature to link research inputs to policy-relevant research outcomes.

[1] https://strategy.data.gov

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 presents evidence on the policy impact of empirical research. Section 8 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 averages to 20% and gets to almost 70% when looking at studies on high-income 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 (Card, 2022; Heckman, 2001).

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, or similarly sensitive information 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 information 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, statistical agencies such as the U.S. Census Bureau do provide direct access to microdata but only through strong security barriers. This includes providing access exclusively to a set of vetted researchers for pre-approved projects (a process that can take over a year) 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. For this reason, 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. Nevertheless, the U.S. Census Bureau has been reluctant to move away from a more restrictive approach based on physical access.[2] Given the central importance that data policy plays in economic policymaking, our study aims to provide some empirical evidence on the effects of administrative data access under the current restrictive regime.

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

[2] In 2019, the Census Bureau began to provide remote virtual access for a select number of FSRDC researchers who do not work with records originating from the IRS. 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.

[3] The 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."

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, including the LBD and the LEHD (see Appendix F for details on the use of specific datasets). 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 growth of the FSRDC network has led other federal agencies to make their confidential data available through the same infrastructure. The agencies partnering with the FSRDC network include the Agency for Healthcare Research and Quality, Bureau of Economic Analysis, Bureau of Justice Statistics, BLS, NCHS, and National Center for Science and Engineering Statistics. Given this diversity, while the CES was originally established to give access to data for the manufacturing sector, FSRDCs now permit the investigation of a large variety of economic phenomena (McGuckin et al., 1993).[5]

Several accounts suggest 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-

[4]The 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 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.

[5]We 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 cannot precisely track this expansion in our study. In our estimates, we rely on a tight set of time fixed effects to control for the introduction of new datasets. In Appendix F, we provide additional details.

edge empirical research (Card et al., 2010). Moreover, research shows that restrictive access to data is disproportionally 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 and policymaking to shape the design of the federal data infrastructure.

3 Research Design

There are two key challenges in empirically assessing the impact of data access on economics research. First, it is 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 upward biased. 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. 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.

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 the FSRDC at these two institutions opens. As is clear

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

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 cutting-edge economics departments. Note also that by the time the Stanford FSRDC opened in 2010, confidential Census data were already commonly used in research, and yet almost no Stanford researcher was using them. This example suggests that opening 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 host 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 when they 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 primarily 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 to 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.[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 host the data center was often the result of a compromise. Taken together, this qualitative evidence suggests that both the timing and the locations of FSRDCs were strongly influenced by idiosyncratic factors unrelated to underlying trends in research productivity, 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 each researcher's methodological orientation, (d) measuring the diffusion of FSRDCs and the adoption of confidential Census data and (e) measuring the policy impact of a given publication.

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

[7]One 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.

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 researchers' names using a graph-theoretic disambiguation procedure (Önder and Schweitzer, 2017). This approach assumes that the combination of first, middle, and last names uniquely identifies each economist (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., 2022).

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 citation 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

[8] We 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.

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

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 (see Appendix D for more). 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 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 A2). 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 using 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.

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

[10] Unfortunately, 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.

[11] We 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.

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

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

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

[13] Appendix D provides additional details and shows the robustness of our classification of empirical economists.

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 considers 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 G1 and G2 show that the average ranking position of treated institutions and treated researchers are relatively constant over time. The remarkably balanced expansion of the FSRDC network further excludes explicit prioritization of high-status universities.

4.5 Measuring Policy Impact

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

4.6 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 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, 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.

Table G15 in the Appendix presents the descriptive statistics for policy citations. Around 47.5% of economists in our sample have written at least one paper mentioned in a policy source. The fraction of individual papers mentioned at least once in policy-related documents is 14.96%, but it grows substantially for articles appearing in more prestigious or explicitly policy-oriented journals (Appendix Figure E3).

5 Results

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

$$y_{i,j,t} = \alpha + \beta_1 PostFSRDC_{j,t} + \beta_2 PostFSRDC_{j,t} \times Empiricist_i + \mu_i + \delta_t \times \omega_j + \epsilon_{i,j,t}, \quad (1)$$

where the dependent variable $y_{i,j,t}$ is publications at the researcher-year level. The main independent variable $PostFSRDC_{j,t}$ is a time-varying dummy that takes a value of one after a researcher has gained access to an FSRDC facility located in the same city of her university 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 whom 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.[14]

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_2 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.[15] To provide a more meaningful estimate of β_1 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 (Kalaitzidakis et al., 2003).[16]

5.1 Impact of FSRDC Openings on Data Use

The 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 with an FSRDC, it was always possible to collaborate with a Census Bureau researcher with access to the data or to commute to a center in another city to access the data; therefore, it is not obvious that being co-located with an FSRDC will lead to a substantial diffusion of confidential 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 G3). 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 G1). Interestingly, we also find a large impact on the likelihood of citing past work (from any university) based on Census data (Columns 3

[14] This specification differs from a triple-difference estimate because the term "Post" is not defined for never treated units due to the staggered rollout of FSRDCs. In our case, we can sidestep the need for a matching estimator by exploiting within-university variation between theorists and empiricists. All other time-invariant terms are absorbed by the researcher fixed effects.

[15] Said 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. Indeed, β_1 cannot be identified once we drop researchers losing local access after a mobility event (Appendix Table G6).

[16] We 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.

and 4). While the increase on the sample mean is much smaller (around 53%–57%), this result is even more remarkable since geography should not be an impediment to learning about papers published in academic journals.[17] This suggests that the opening of an FSRDC is enculturating local economists into research using administrative data, potentially through interactions with others directly dealing with such data in their own work.

The validity of our regressions hinges on the assumption of a parallel trend between treated and control units in the absence of treatment. We empirically assess this by estimating the event study version of the results reported in Table 2. Specifically, we estimate $y_{i,j,t} = \alpha + \Sigma_z \, \beta_t \times 1(z) + \mu_i + \delta_t \times \omega_j + \epsilon_{i,j,t}$, where μ_i and $\delta_t \times \omega_j$ represent researcher and university $\times$ 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 generalist 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.

Panel A of 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

[17] Note 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.

and meaningful, but they are against a small denominator and so do not represent implausible shifts in 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 Panel B of Table 3, the results are larger, supporting anecdotal evidence that administrative data lead to breakthrough findings.

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 G2 and Appendix Figure G4 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. Appendix Figure G5 confirms that our results carry through if we exploit only variation in treatment time across universities using split-sample DID regressions between theorists and empiricists. In sum, 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 G6. 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 (see Appendix A). Our conversations with FSRDC users pointed out that applying to use confidential data is especially risky for researchers on a tenure clock due to bureaucratic uncertainties, which helps explain our findings.

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 G3). Appendix Figure G7 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 by excluding from the sample either the researchers or the institution directly involved in bringing an FSRDC to a given location. We code the recipients of NSF grants establishing each FSRDC, and Appendix Table G4 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 G7). One might also be worried that researcher mobility events are driving our results, especially since we impute the year of the move based on publication data. Appendix Tables G5 and G6 show that the results are robust to excluding researchers that gain or lose access due to mobility events.[18]

Third, we experiment with alternative specifications of "exposure" to administrative data in geographical and intellectual space. Appendix Figure G8 shows that an economist's likelihood of using confidential Census data rapidly decays with distance. 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 G8 and G9). Similarly, the effects of access monotonically grow when we use increasingly stringent definitions of empirical researchers (Figure G9). This confirms our research design and reassures us 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. For instance, changes in editorial preferences in top journals might explain some of our results. 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 G10 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 G10), but the results remain robust and significant when weighting publications by their impact.

[18] As an additional test, we repeated our main results after artificially suppressing all variation due to mobility events in our data. In practice, we code each researcher as if they spent their entire career in the institution of their first placement. Appendix Table G17 shows that results are attenuated but still significant.

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 G11 reports all of our main results excluding from the analyses researchers whom 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 data usage, there are significant spillovers from data access even among those who do not directly adopt (Myers and Lanahan, 2022).

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 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 compared to those who do not. We find that the positive effects on productivity do not extend to all empirical researchers—instead, they are limited to researchers who cite past work based on FSRDCs. 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 the effects for those empirical researchers with and without colleagues using confidential Census data in their work. We find no positive spillovers for researchers affiliated with departments where nobody directly uses Census data. Both results support the conclusion that spillovers from co-localization operate by making researchers aware of research based on confidential administrative Census data.

We 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 4 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%.[19] 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 5 shows that treated researchers increase mentions of quasi-experimental methods, such as DID or synthetic controls. The absence of a similar effect on laboratory experiments or randomized control trials implies that this result is not a byproduct of a wider "credibility revolution."[20] 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 G13).

7 Do Administrative Data Increase the Policy-relevance of Research?

Our results highlight the crucial role of data access in shaping scientific advances in economic research. However, it is not guaranteed that improvements in scientific quality will translate into higher policy relevance. It might well be that a focus on advancing scholarly knowledge comes at the cost of research directly applicable in the policy realm (Landry et al., 2003). Academic researchers often lack incentives to disseminate policy insights from their work, despite evidence showing strong policy responsiveness to scientific evidence (Hjort et al., 2021; Yin et al., 2022). Focusing on research using administrative data specifically, evidence of its policy relevance is primarily anecdotal and confined to the impact of few high-

[19] This result is robust to alternative ways to select which JEL codes most represent FSRDC-based research; see Appendix Table G12.

[20] As 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 G13).

profile studies (Card et al., 2010; CES, 2017; Cole et al., 2020; Einav and Levin, 2014b). Whether access to FSRDC data leads to research with higher policy relevance remains an empirical question.

We first use our data on policy citation by correlating paper-level attributes with their citations in policy sources. Evidence in Appendix Table E1 suggests that empirical articles are generally more cited by policy sources than comparable theory contributions that appear in the same journal and year. Interestingly, the effect is much more extensive for papers directly using FSRDC data. Consistent with the intuition that FSRDC data help shed light on economic and social trends specific to the United States, the impact of such papers is almost three times larger among U.S. policy sources. Furthermore, Appendix Table E1 shows that papers with a better research design generally achieve larger interest from policymakers, echoing some of the findings by Hjort et al. (2021). We also document large differences in policy consumption across fields of economic research (Appendix Figure E4).

Next, we estimate the causal impact of FSRDC access on the policy impact of economics research. We estimate the same specification as in Eq. (1), but where the dependent variables are measures of policy use of academic science. Table 6 presents the results. Columns 1 and 4 show that the work of applied researchers receives a larger number of policy citations and is more likely to be referenced by policy sources after becoming exposed to FSRDC data. The percentage increase over the sample mean is slightly bigger among U.S.-based policy sources, confirming the impact of the federal data infrastructure on evidence-based policy-making in the United States. Event study specifications in Appendix Figure G11 help rule out that the effects are due to pre-trends in the dependent variables.

Interestingly, our results do not seem to be driven by a dramatic shift in research topics toward more policy-relevant fields. If this were the case, we should see the use of language that emphasizes the policy implications of research or a direct shift to policy topics. However, we do not find either of these effects when directly looking at the language used in the abstracts or the JEL codes indicated by the authors (Table G16). Instead, our results could be consistent with an increase in policy relevance due to the higher scientific quality of applied research caused by access to better data.

Combined, our analysis can paint a picture of *how* FSRDC openings drive improved scientific productivity and policy impact among applied researchers. First, a cadre of researchers can now access these data at a lower cost and use them in their own work to great effect. However, given the high costs of direct access, this group is only partially responsible for driving the overall effects. Spillovers to non-adopters are also important: FSRDC openings tend 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 research focus to become more explorative, working on topics typically studied with administrative data, and adopting other types of administrative data and quasi-experimental research designs in their work.

In turn, the improvement in scientific quality is accompanied by a growing number of citations from policy documents. We also find suggestive evidence that the increased policy impact of economic research is due to the increased quality of the empirical evidence provided and not to a crowding out of scientifically important questions by more policy-relevant topics.

8 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 assess 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. Finally, we show that increased scientific quality translates into additional policy impact, leading to higher consumption of empirical evidence in policy documents.

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 show how evaluations of the impact of data access programs need to extend beyond those directly using the data. Data access can shape research by changing the topical and methodological focus of 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 demography. 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 consider 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 significant and

directly leads to the production of higher-quality scientific output with an extensive policy impact.

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 entails 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 considerable 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. (2018): "The US Census Bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2867–2867.

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.

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.

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

——— (2022): "Gender Differences in Peer Recognition by Economists," *Econometrica*.

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

Chetty, R., J. N. Friedman, N. Hendren, M. Stepner, et al. (2020): "The economic impacts of COVID-19: Evidence from a new public database built using private sector data," Tech. rep., national Bureau of economic research.

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.

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

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.

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

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.

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

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

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

Hill, R. and C. Stein (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. Manso, 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," CES working paper.

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 (2022): "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.

Landry, R., M. Lamari, and N. Amara (2003): "The extent and determinants of the utilization of university research in government agencies," *Public Administration Review*, 63, 192–205.

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.

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

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

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

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

Truffa, F. and A. Wong (2022): "Undergraduate gender diversity and direction of scientific research," Stanford University.

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

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

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

9 Tables and Figures

Table 1: Summary Statistics

Panel A: Researcher Level

	N	Mean	Std. Dev.	Median	Min	Max
Ever Had 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)	230961	0.348	0.48	0	0	1
New LDA Topics (0/1)	230961	0.249	0.43	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 = 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 95^{th} 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 95^{th} 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 (this variable is not defined for the first year of each researcher). 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 (this variable is not defined for the first year of each researcher). 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	0.00216	-0.00884**	-0.00235
	(0.00069)	(0.00114)	(0.00333)	(0.00463)
Post-FSRDC × Empiricist	0.00392***	0.00332**	0.0189***	0.0174***
	(0.00107)	(0.00113)	(0.00425)	(0.00444)
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 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. The number of observations changes slightly due to the inclusion of different fixed effects. 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

Panel A: Research Impact

	Top Publications		Cite-weighted Publications	
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.0139	-0.00712	-0.676*	-0.170
	(0.010)	(0.012)	(0.336)	(0.435)
Post-FSRDC × Empiricist	0.0353***	0.0352**	1.714***	1.737***
	(0.011)	(0.011)	(0.425)	(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

Panel B: Right Tail of Research Impact

	Top Five Pubs		Top 5% Cite	
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.0131*	-0.00994	-0.0193***	-0.0147*
	(0.006)	(0.007)	(0.005)	(0.006)
Post-FSRDC × Empiricist	0.0218**	0.0231***	0.0285***	0.0287***
	(0.007)	(0.007)	(0.006)	(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 measures of research output. Columns (1) and (2) of Panel A 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) of Panel A report results from OLS models, where the dependent variable is the count of articles weighted by the number of citations received up to five years following their publication. Columns (1) and (2) of Panel B 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) of Panel B 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. The number of observations changes slightly due to the inclusion of different fixed effects. 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 Direction of Research

	New JEL Code (0/1)	New LDA Topic (0/1)	FSRDC JELs (0/1)	Non-FSRDC JELs (0/1)
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.00686	-0.00444	-0.00332	-0.0125
	(0.011)	(0.011)	(0.007)	(0.010)
Post-FSRDC × Empiricist	0.0475***	0.0339***	0.0175**	0.0278**
	(0.010)	(0.010)	(0.006)	(0.009)
Researcher FE	Yes	Yes	Yes	Yes
University × 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. 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. The dependent variables in Columns (1) and (2) are not defined for the first year of each researcher, hence the lower number of observations. 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. The number of observations changes slightly due to the inclusion of different fixed effects. 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 Design of Empirical Research

	Data		Research Methods	
	Admin	Survey	Quasi-Exp.	Experiment
	(1)	(2)	(3)	(4)
Post-FSRDC	-0.00243*	0.00220	0.000395	0.00372
	(0.001)	(0.001)	(0.003)	(0.005)
Post-FSRDC × Empiricist	0.00328***	0.000364	0.00786**	0.000294
	(0.001)	(0.001)	(0.003)	(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. The number of observations changes slightly due to the inclusion of different fixed effects. Standard errors are in parentheses, clustered at the researcher level. *, **,*** denote significance at the 5%, 1%, and 0.1% level, respectively.

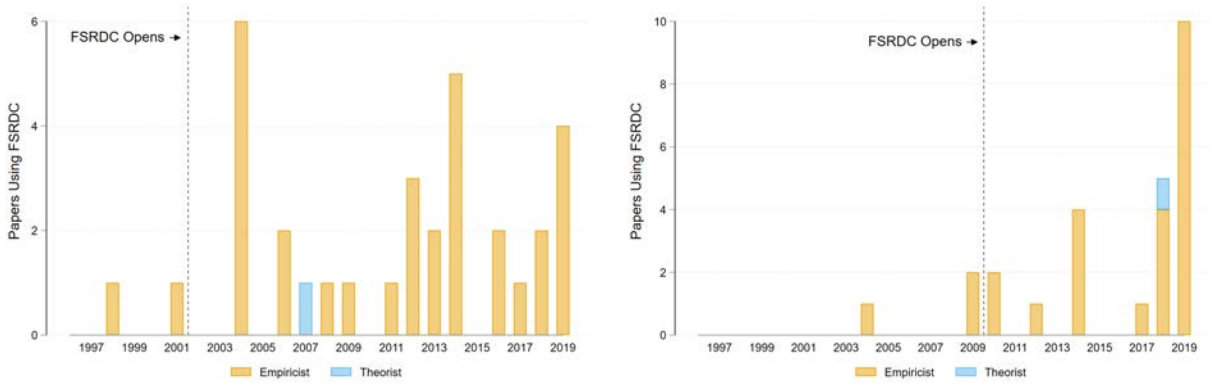
Table 6: Effect of FSRDC Access on the Policy Impact of Economic Research

	Count Policy Cites			Count Papers Cited by Policy		
	All	US Only	Non-US Only	All	US Only	Non-US Only
	(1)	(2)	(3)	(4)	(5)	(6)
Post-FSRDC	-0.0611	-0.0263	-0.0348	-0.0187*	-0.0157**	-0.0112
	(0.055)	(0.027)	(0.032)	(0.008)	(0.006)	(0.007)
Post-FSRDC × Empiricist	0.203***	0.100***	0.103**	0.0486***	0.0392***	0.0319***
	(0.058)	(0.028)	(0.034)	(0.008)	(0.006)	(0.006)
Researcher FE	Yes	Yes	Yes	Yes	Yes	Yes
University × 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 the policy relevance of economic research. Column (1) report results from OLS models, where the dependent variable is the number of citations from policy sources received by the articles published in a given year. Columns (2) and (3) report results from the same OLS models, where the dependent variable is splitted between citations from U.S. and non-U.S. policy sources, respectively. Column (4) report results from OLS models, where the dependent variable is the number of articles published in a given year that received at least one citation from policy sources. Columns (5) and (6) report results from the same OLS models, where the dependent variable is splitted between papers cited by U.S. and non-U.S. policy sources, 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. The number of observations changes slightly due to the inclusion of different fixed effects. 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.

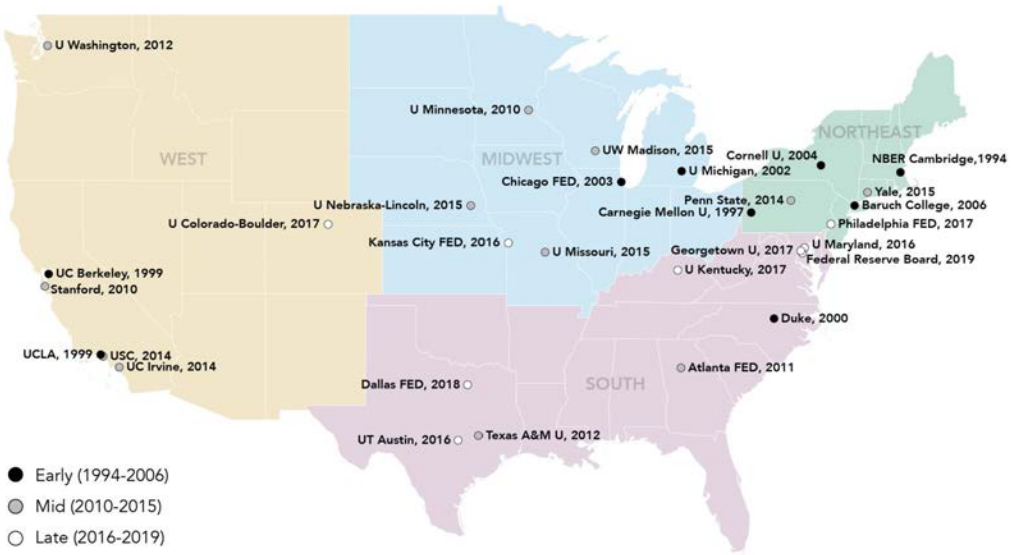
(i) University of Michigan-Ann Arbor *(ii) Stanford University*



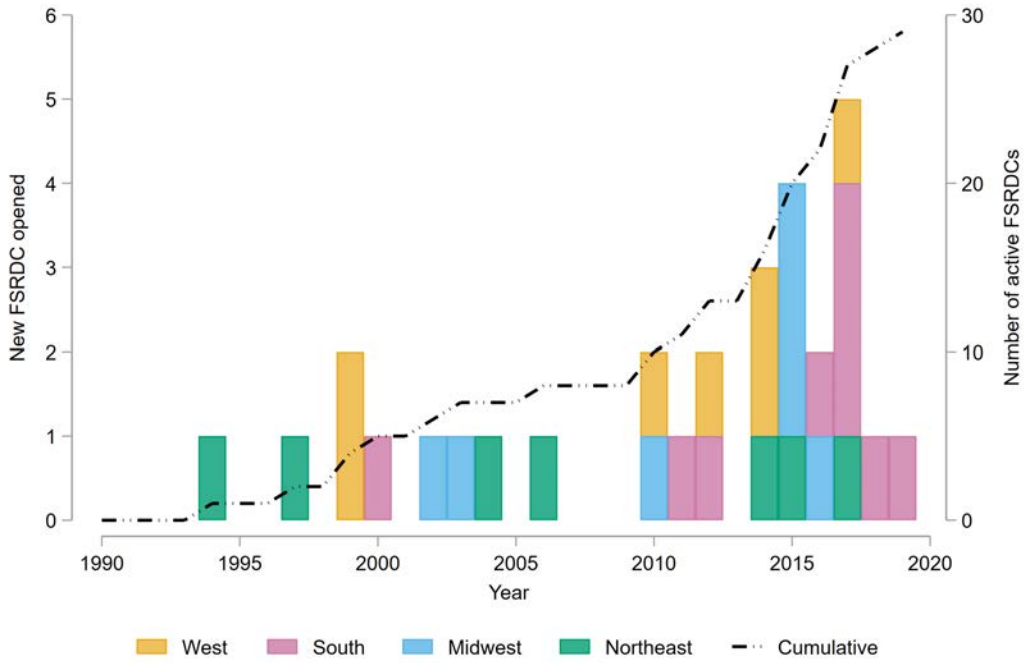
Note: This figure shows the number of papers using Census data published by researchers at the University of Michigan-Ann Arbor and Stanford University, respectively. The blue portion of each bar represents papers co-authored by at least one theoretical researcher. See text for more details.

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

Panel A: Geographic Expansion of FSRDCs



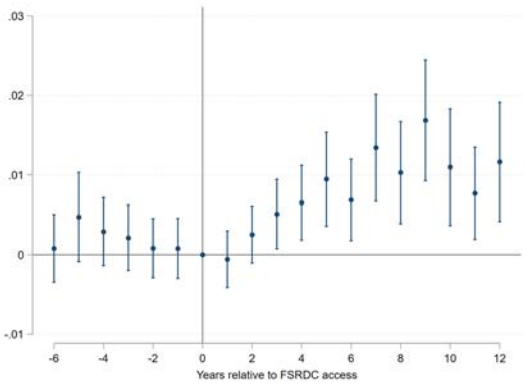
Panel B: Time Expansion of FSRDCs



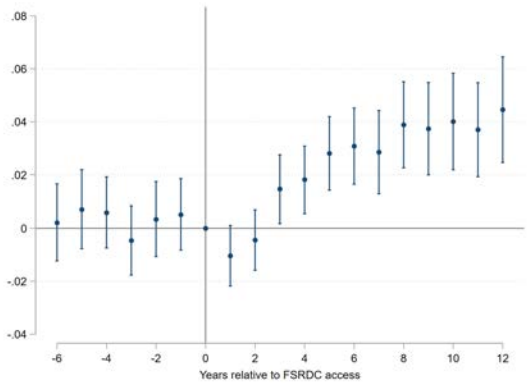
Note: This figure provides an illustration of the timing and geographic scope of the expansion of the FSRDC network. See text for more details.

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

(i) Papers Using FSRDC

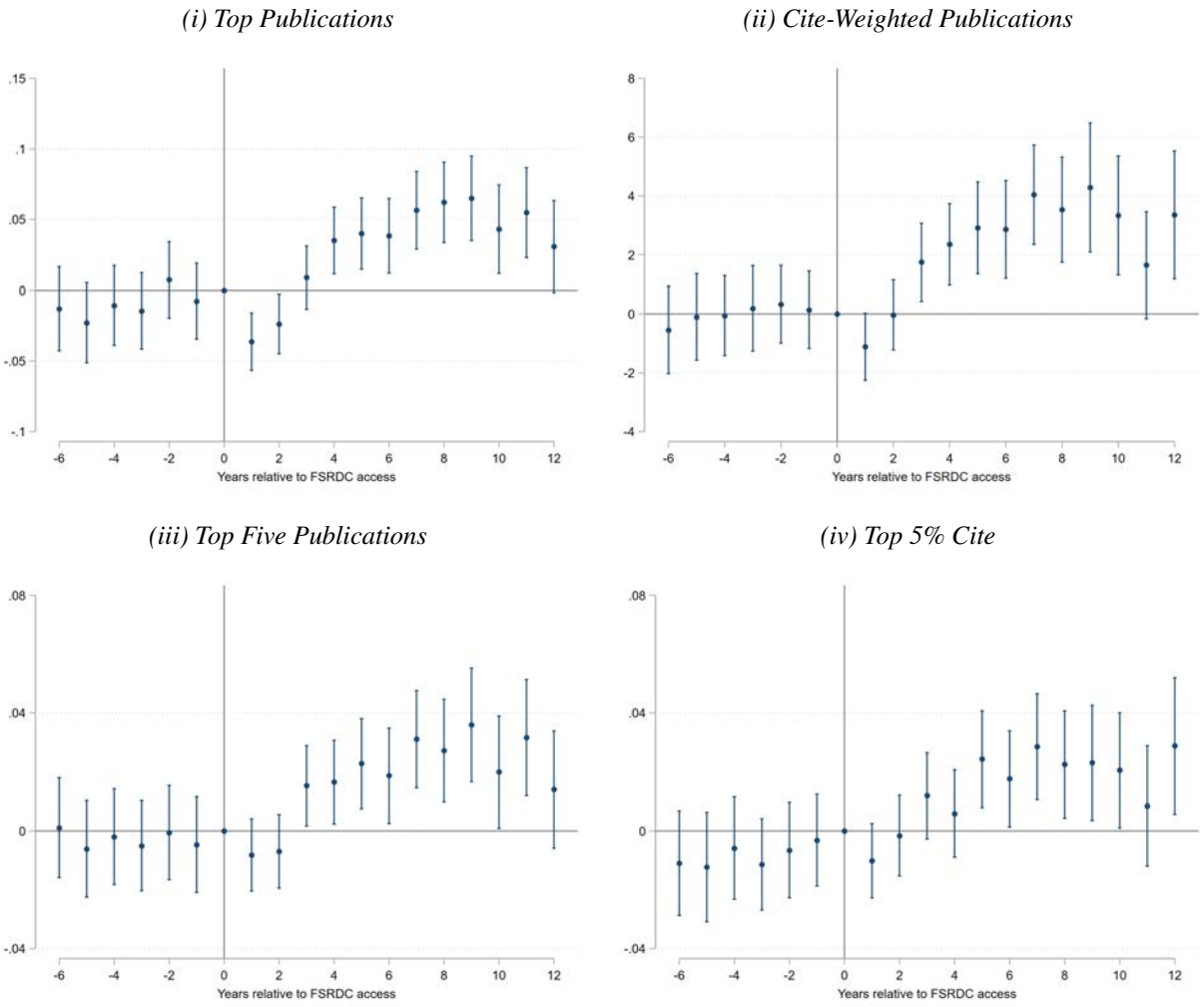


(ii) Papers Citing FSRDC



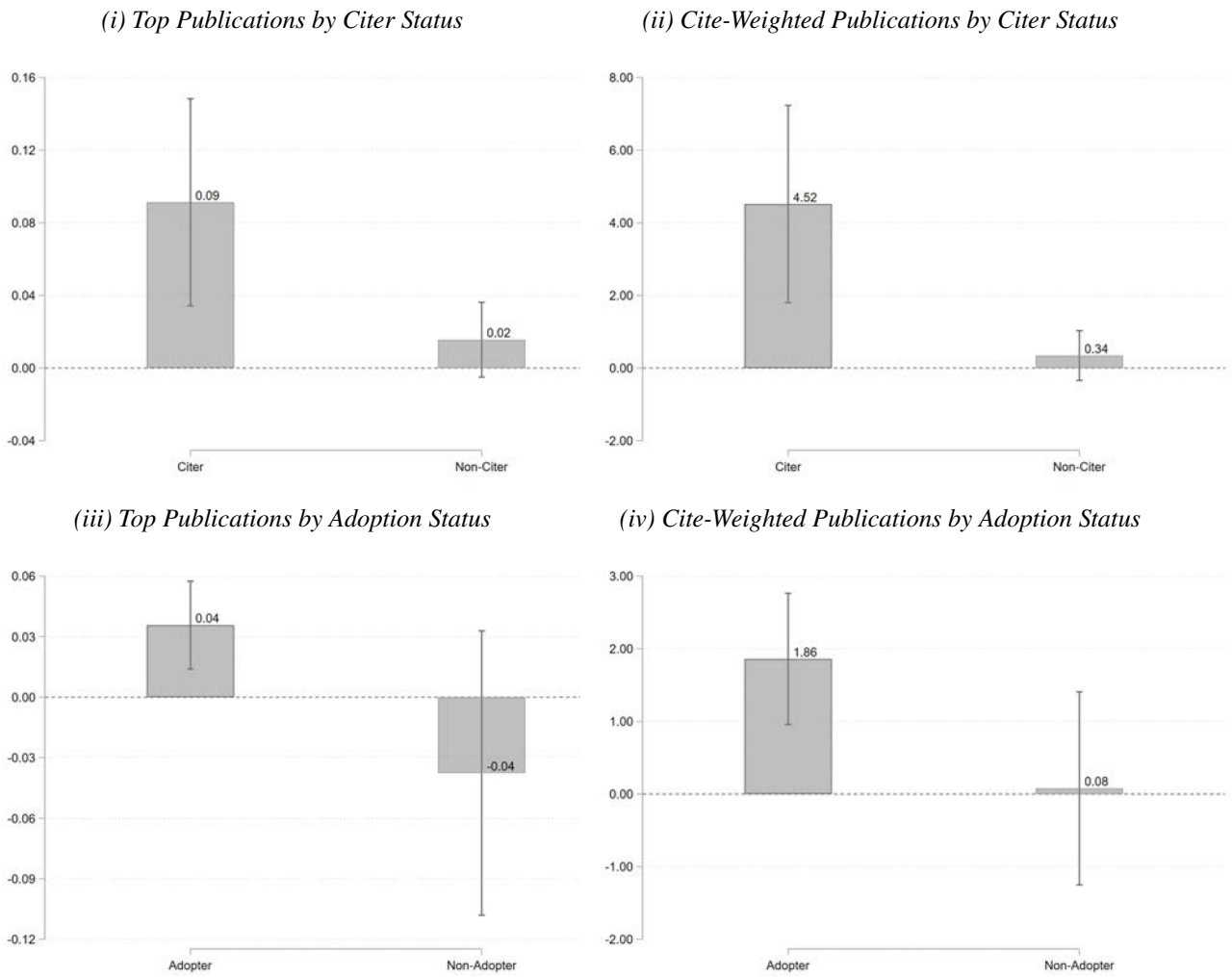
Note: This figure provides visual illustrations of the event study version of the main regression on administrative data adoption. The main dependent variables are the number of papers written using Census data (panel (i)) or the number of papers that cite papers using Census data (panel (ii)). The chart plots values of β for different lags before and after obtaining access to an FSRDC with 95% confidence intervals. 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 with 95% confidence intervals. 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.