HDSR
HARVARD DATA SCIENCE REVIEW

Just Accepted
DOI: 10.1162/99608f92.19a0e1c7
ISSN: 2644-2353

Restricted-Access Data in Economics: Adoption, Diffusion, and Impact of U.S. Census Bureau's Microdata

Abhishek Nagaraj[†], Fernando Stipanicic[‡], Matteo Tranchero[◇,*]

[†] UC Berkeley-Haas

[‡] University of Oslo

[◇] The Wharton School

Abstract. Microdata from government agencies is believed to be valuable for economics research, and yet access to this data is highly restricted due to concerns about privacy and security. We provide an empirical assessment of the use and impact of restricted-access data that researchers can analyze at the U.S. Census Bureau's secure facilities. Our findings show that the use of Census Bureau's confidential data is growing and the publications employing it have a higher impact on the scientific and policy debate. However, adoption remains largely limited to established researchers from prestigious institutions. Our results and discussion inform the design of policies that balance privacy protection with accessibility to confidential microdata.

Keywords: Confidential microdata, U.S. Census Bureau, FSRDCs, data access, economics research, inequality in science

1. Introduction

In recent decades, economics has taken a sharp turn towards becoming a more empirical science (J. Angrist et al., 2020; Hamermesh, 2013). While there are many potential drivers for this shift, the increasing availability of high-quality, large-scale, longitudinal microdata from government agencies is a major factor (Currie et al., 2020; Einav & Levin, 2014). Important empirical and theoretical breakthroughs have come from data that were originally collected for administrative or statistical needs (Lane & Potok, 2024). For example, plant-level records from the U.S. Census revealed numerous descriptive facts about exporters (Bernard & Jensen, 1999) that inspired the analysis of Melitz (2003), perhaps *the* workhorse model in international trade.[1] Similar examples can be found

[1] At the moment of writing, Melitz (2003) appears as the 17th most cited article of all times in Ideas: https://ideas.repec.org/top/top.item.nbcites.html.

in many other fields of economic research.[2]

Despite their scientific potential, the use and potential release of microdata from government sources could pose privacy and security risks because they contain sensitive information. In many jurisdictions, these confidentiality concerns have translated into a system prioritizing strong guardrails (Cole et al., 2020). In other locations, such data are more easily available to academic researchers (Card et al., 2010). Notwithstanding the growing importance of microdata for economics and enormous variation in access rules, the question of how one might balance privacy and scientific progress has ironically been data-free to date. A few papers have pointed to the growing diffusion of restricted-access data in general (Abraham et al., 2022; Chetty, 2012; Currie et al., 2020; Einav & Levin, 2014), but evidence on the impact of such data remains mostly anecdotal (Atrostic, 2007; CES, 2017; Davis & Holly, 2006).

Several questions require more attention. How does research stemming from restricted-access data from government sources shape scientific and policy debates in the United States? What types of researchers are using the data and what questions are they exploring? Are some types of researchers systematically less likely to use these data? While balancing the trade-off between access and confidentiality is a function of social preferences (Abowd & Schmutte, 2019; Fobia et al., 2020), data-driven answers to these questions could help policymakers strike a better balance between protecting privacy and enabling scientific innovation.

In this paper, we contribute to answering these questions by examining the use and impact of confidential data curated by the U.S. Census Bureau on economics research.[3] Census-curated data are perhaps the most pre-eminent sources of confidential data and have tight access restrictions, making them a prime setting for study. For example, data can only be accessed via physical enclaves, projects need to be approved in advance, and the disclosure of results is intensely scrutinized (Abowd & Schmutte, 2019; Cole et al., 2020; Foster et al., 2009; Nagaraj & Tranchero, 2024). However, the current organization of the U.S. data infrastructure has been criticized for being restrictive and dysfunctional (Lane, 2021). Some leading economists indicated that barriers in data access could penalize U.S.-focused research (Card et al., 2010), posing the risk of shifting scientific attention toward topics with relatively more accessible data (Hoelzemann et al., 2024). By bringing new evidence on the use of Census-curated confidential data, we contribute to the debate on how the access regime shapes data adoption and impact in the United States.

[2]Including in employment (Haltiwanger et al., 2013) and wages (Abowd et al., 2018), racial disparities (Chetty et al., 2020), innovation (Jaravel et al., 2018), firm productivity (Olley & Pakes, 1996), health care provision (Finkelstein et al., 2021) and mortgage markets (Beraja et al., 2019).

[3]Technically, the U.S. Census Bureau collects *statistical data* mostly through surveys (e.g., the Economic Census). The protection of these data is regulated by Title 13 of the U.S. Code. The Bureau also uses *administrative records* from other agencies that provide services to the public. Different legal provisions often protect such data (e.g., IRS tax returns fall under Title 26 of the U.S. Code). For our purposes, this distinction is blurred since many Census-curated datasets stem from both types of data (e.g., the LEHD, see Abowd et al. 2009) and because our focus is on the confidential nature of the data, and not their precise legal status or mode of collection.

We document new facts on the use and impact of restricted-access data using comprehensive records of publications in economics. The primary source of the data is EconLit, a proprietary database of economic scholarship curated by the American Economic Association. In our analyses, we consider over 90,000 articles published in 158 peer-reviewed journals by researchers based in the United States. These data do not possess unique identifiers for researchers or institutions. We engage in a painstaking disambiguation effort, which allows us to measure outcomes for 17,820 researchers affiliated with 342 North American institutions between 1991 and 2019. Through various methods, including natural language processing and matching titles with information on approved projects, we can track the adoption of confidential data distributed by the U.S. Census Bureau at the paper level. Finally, we augment these paper-level data with information on authors' characteristics, paper-to-paper citations, and policy document-to-paper citations.

The analysis of our publication data shows that the incidence of articles using restricted-access data curated by the U.S. Census Bureau has steadily increased over time, from 0.21% of all published papers in 1991 to 1.27% in 2019. Almost 6% of all papers in economics cite at least one paper using such restricted-access data. Scientific papers employing restricted-access data are highly impactful, being 50% more likely to be published in the so-called "top five" economics journals[4], receiving 28% more citations from other papers, and 80% more citations from policy documents. Although descriptive in nature, the results are robust to restricting the comparison to papers that use microdata from the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS). Confidential data from the Census Bureau are predominantly used in labor, applied microeconomics, industrial organization, and international economics research. However, papers using such data are more likely to include authors who have previously published in a top five journal or are affiliated with higher-status U.S. universities. Taken together, our descriptive analysis suggests that while restricted-access microdata are valuable in producing impactful research, their adoption remains relatively limited and restricted to established researchers from prestigious institutions.

We organize the article as follows. First, we provide an overview of the restricted-access data made available by the U.S. Census. Then, we describe the dataset we have built to assess the adoption and impact of these data in the economics profession. We then overview key facts emerging from our empirical analysis. Finally, we discuss the implications of our work.

2. Background

2.1. **Confidential Microdata in Economics Research.**

Economic scholarship has changed from being primarily theory-driven into a more data-intensive discipline (Backhouse & Cherrier, 2017). The share of theoretical papers published in top economics journals has decreased from 50.7% in 1963 to 19.1% in 2011 (Hamermesh, 2013), while empirical work has surged both in incidence and impact (J. Angrist et al., 2020; J. D. Angrist & Pischke, 2010). The increased availability of government microdata, most notably administrative records and statistical surveys, has played an important role (Cole et al., 2020; Groves, 2011). Around a fifth of the articles recently published in the five most prestigious economics journals employ data derived from administrative sources (Currie et al., 2020). This share increases up to 70% when focusing on studies about high-income

[4]These journals traditionally encompass the *American Economic Review*, the *Quarterly Journal of Economics*, the *Journal of Political Economy*, *Econometrica*, and the *Review of Economic Studies*.

countries (Chetty, 2012).

Even if administrative records and statistical surveys are not originally collected for research purposes, they offer tremendous advantages for economic research (Cole et al., 2020). These data are usually granular, highly disaggregated, and display a longitudinal structure that naturally allows tracking of the same individuals, or companies, over time and before and after certain interventions. Such microdata are often comprehensive in coverage and can include large samples that allow precise estimation for subgroup heterogeneity analyses (Abowd et al., 2009; Heckman, 2001). Anecdotal evidence suggests that microdata from government sources can open up entirely new fields of inquiry based on new questions (Einav & Levin, 2014). For instance, creative combinations of administrative records from multiple sources have been instrumental during the COVID-19 pandemic in following real-time developments and targeted assistance programs (Vavra, 2021).

The unique value of microdata for policy-relevant research derives from the level of detail that it affords. However, this very characteristic might put the privacy of subjects and respondents at risk. Information on an individual's employment, earnings, tax identification numbers, etc., must be kept private to prevent identity fraud and harmful targeting. The leakage of granular sales and employment information at the business level could also have negative competitive implications for firms. The risk of privacy loss does not just stem from the release of direct identifiers, but also from indirect identification through any combination of data records (Sweeney, 2000). Even beyond these economic harms, moral and legal frameworks in most contexts, and especially in the U.S., support privacy as a fundamental right that must be preserved. For example, Title 13 of the U.S. Code makes it illegal for the U.S. Census to disclose any private information that identifies (directly or indirectly) an individual or business.[5] Government agencies thus face a trade-off between granting access to their microdata, which are potentially very valuable for research purposes, and their duty to protect the confidentiality of the information entrusted to them (Foster et al., 2009).

The trade-off between limiting access to sensitive data and their value for research is further complicated because access to confidential data faces problems similar to public goods provision (Lane et al., 2024). The benefits from realized projects accrue to the whole society while costs are borne by the data provider (Hoelzemann et al., 2024; Ritchie & Welpton, 2011). Therefore, it is reasonable to expect that from a welfare perspective, access to confidential data could be under-provided. Speaking specifically about restricted-access data from U.S. government sources (the focus of this paper), informed commentators have argued that current access policies prioritize security and privacy risks without fully considering its impacts on value creation (Card et al., 2010; Lane, 2021).

2.2. **Balancing Access and Security at the U.S. Census Bureau.** Over the years, government agencies have experimented with several solutions to disseminate their data while protecting individual privacy. Some of them include releasing anonymized samples for public use or the development of synthetic data (Abowd & Lane, 2004; Kinney et al., 2011), which has become an increasingly popular way of disclosure control (Reiter, 2023). Yet, depending on the research application, these solutions can turn out to be only second-best replacements for the possibility of

[5]See the U.S. Census Bureau webpage on Data Stewardship for more details about Title 13 of the U.S. Code: https://www.census.gov/about/policies/privacy/data_stewardship.html

working with the universe of respondent-level microdata.

To address these limitations, the U.S. Census Bureau has pioneered an alternative approach. In 1982, it established the Center for Economic Studies (CES) to enable non-employee access to confidential microdata (Atrostic, 2007). The objective of the CES was to develop longitudinal databases and to host qualified academic researchers who could analyze confidential data directly onsite (Foster et al., 2009). The Census Bureau enforces tight restrictions on the types of projects and people who can work with confidential data. Researchers must write a detailed proposal showing how the proposed research benefits the Census Bureau, justifying the project's feasibility and requirement for non-public data, and proving that the project does not pose a risk of unauthorized disclosure. Researchers must also apply for "special sworn status" (SSS) with the U.S. Census Bureau, which involves passing a background check, among other legal requirements. Further, the proposal must specify which variables and datasets the researchers will use. Access is then provided to only the data requested and approved.

Regarding how the data is accessed, the U.S. Census Bureau has adopted the approach of allowing only in-situ analysis in a data enclave to minimize the risk of privacy breaches. Since traveling to the CES' headquarters is not practical for most researchers, over time the Bureau has opened multiple secure facilities across the country as part of a program known as the Federal Statistical Research Data Centers (FSRDCs).[6] FSRDCs are operated by Census staff in partnership with local universities or research institutions. Each branch meets the same physical security standards as the CES. Researchers are monitored closely when accessing data in these enclaves, and no data or outputs can leave the secure facilities without a detailed disclosure review from Census officials. Through these multiple steps, the Census ensures that the privacy and security risks of sharing sensitive data are minimized while simultaneously allowing access for academic research purposes. Gradually, other Federal Agencies started making their data available to researchers through the FSRDC network, including the Agency for Healthcare Research and Quality, the Bureau of Economic Analysis, the Bureau of Justice Statistics, the Bureau of Labor Statistics, the National Center for Health Statistics, and the National Center for Science and Engineering Statistics.

Despite the strict limits on access, anecdotal accounts suggest that the FSRDC program helped foster data diffusion and provided benefits for the Census Bureau and the other agencies involved (CES, 2017; Davis & Holly, 2006). Several policy-relevant findings were enabled by the granularity of confidential data uniquely accessible in the data centers (Card et al., 2010; Einav & Levin, 2014). Nagaraj and Tranchero (2024) show that the expansion of the FSRDC network alleviated access constraints, leading to a large increase in the use of microdata by researchers in nearby institutions. Researchers near an FSRDC improve the quality of their publications, reflecting both the direct effect of data access and the spillover effects of being exposed to high-impact research using restricted-access data. However, understanding the potential trade-offs entailed by the tight confidentiality regime requires a better characterization of the research carried out using restricted-access data. Information on the type, impact, and topical focus of research based on confidential data is important to inform the design of data access policies.

[6]See Appendix A.1 for additional details on the history and diffusion of FSRDCs.

3. Data

We assemble a database of economic scholarship using article-level data from EconLit, a proprietary bibliographic database curated by the American Economic Association. Compared to other popular databases of scientific publications, EconLit has a wider coverage of economics journals and includes JEL codes that allow classifying papers into research fields (J. Angrist et al., 2020; Nagaraj & Tranchero, 2024). Our analysis required intense data cleaning and the combination of various data sources. In what follows, we provide a short description of the main steps we followed, while further details are reported in Appendix B.

We consider in our analyses all the articles published in peer-reviewed journals between 1991 and 2019 by researchers based in the United States. These data do not possess unique identifiers for researchers or institutions. We engage in a painstaking disambiguation effort using the disambiguation algorithm developed by Önder and Schweitzer (2017).[7] We also disambiguate institutions, taking the institution of each author to be the one reported in the majority of articles published in that year.[8] These procedures allow us to measure outcomes for 17,820 researchers affiliated with 344 North American institutions. We augment these paper-level data with information about authors' characteristics and paper-to-paper citations from the Web of Science database.

We obtain the count of citations from policy documents from Overton. Policy documents indexed by Overton range from white papers of international development organizations to parliamentary transcripts and think tanks' reports (Appendix Figure B1). We exclude citations from NBER, CEPR, and IZA sources since those are likely working paper versions of academic articles rather than policy documents. In Appendix B.2.3 we provide additional details on Overton's coverage of policy documents.

Information on the prestige of economics departments is taken 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 in the five years from 1995 to 1999. This ranking fits well for our purposes because it is based on research intensity (and not on alumni surveys or brand recognition) and focused specifically on the economics department (instead of being a university-wide assessment). An additional advantage is that it considers publications predating the establishment of most FSRDCs. When comparing it with other common rankings such as Times Higher Education, we noticed very marginal changes over time, suggesting that the level of research intensity is fairly constant over the years considered in our analysis.

We collected new information to measure the adoption of confidential data available only at the Census Bureau or through the FSRDC network. We carefully sifted published records with several complementary strategies. We exploited the fact that projects using confidential microdata

[7]The disambiguation approach is similar to Card et al. (2022). Cases where authors could not be clearly disambiguated have been kept as a singleton in the raw data. While the disambiguation keeps such single-paper authors, they would later be dropped in the regressions due to author fixed effect.

[8]While this method to infer affiliations may introduce measurement error if publication delays mean that articles were completed while still being at the previous institution, we are unfortunately unable to fully account for it.

distributed by the Census Bureau are expected to indicate it clearly in the acknowledgment of the published version of the paper. Using natural language processing, we searched for the most commonly used acknowledgment formulas in databases such as Web of Science, Scopus, JSTOR, Google Scholar, IDEAS RePEc, and the NBER website. However, we noticed that several publications omit the disclaimer. As a solution, we also gathered information on what projects had been approved for in-situ analysis by the U.S. Census Bureau. We then manually matched projects with their resulting output in EconLit and used this information to supplement the information obtained from paper acknowledgments. Finally, we merged our data with the list of publications using two popular micro-level databases not hosted by the Census Bureau: the Panel Study of Income Dynamics (PSID) and the National Longitudinal Surveys (NLS) (details in Appendix B). These surveys were the gold standard microdata before the rollout of FSRDCs and have minimal access restrictions, constituting an interesting comparison group for articles using FSRDC data.[9]

Table 1. Descriptive statistics of publication data for U.S.-based economics researchers

Panel A: Publications by all U.S.-based researchers

Type of microdata used	N. Articles	N. Authors	N. Institutions	Avg. yearly citations	% Articles in Top 5 journal	Avg. yearly cit. Top 5 journal	Avg. yearly policy cit.
Curated by Census	589	525	122	7.26	14.60	17.25	2.36
From NLS/PSID	1,586	1,315	194	4.67	13.93	9.97	0.94
Other or None	89,405	17,694	342	3.99	7.53	9.99	0.57
Total	91,570	17,820	342	4.02	7.69	10.08	0.59

Panel B: Publications by users of Census-curated confidential data

Type of microdata used	N. Articles	Avg. yearly citations	% Articles in Top 5 journal	Avg. yearly cit. Top 5 journal	Avg. yearly policy cit.
Curated by Census	589	7.26	14.60	17.25	2.36
From NLS/PSID	264	5.84	14.39	10.41	1.34
Other or None	5,463	7.03	15.47	15.72	1.32

Note: See Appendix B for more details on the sources and data constructions.

Descriptive statistics of our article-level sample are presented in Table 1. In total, we have 589 papers that used confidential data curated by the Census Bureau written by 525 U.S.-based economists in the period 1991-2019. Panel A shows that papers using confidential data from the Census Bureau are more impactful on average and more likely to appear in high-profile economics outlets. When compared with paper employing microdata from PSID or NLS, we see that the gap in impact shrinks, but does not disappear. The table also suggests that confidential microdata are within the purview of a restricted set of institutions. Panel B shows the same summary statistics, limiting the sample to publications authored by users of confidential data curated by the Census

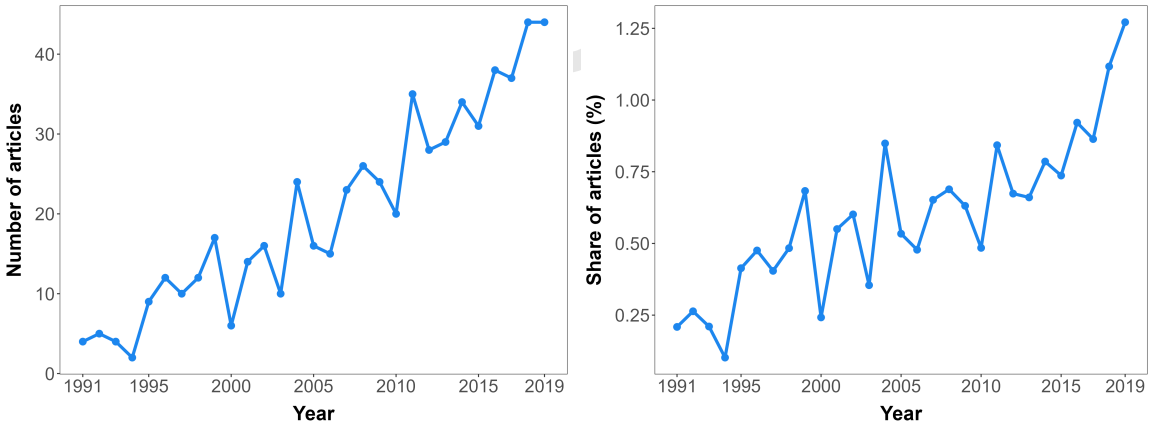
[9]A small part of these surveys is confidential and only provided through restricted-use agreements to approved researchers. We are unable to distinguish which articles use the restricted or unrestricted parts of the PSID and NLS. It would be reasonable to assume that most of such papers use the unrestricted parts of the surveys.

Bureau. In general, users of Census' confidential data appear positively selected on proxies of research quality. But even among articles written by confidential data users, we see that papers using data available only in FSRDCs have a larger impact, especially in the policy realm.[10]

4. Stylized Facts on Research Using Restricted-Access Data

We present descriptive findings emerging from our analysis of research using restricted-access microdata made available at the U.S. Census Bureau. To the best of our knowledge, we are the first to provide an empirical assessment of the use and impact of these data in the economic profession. Our results can be summarized in the following six facts.

Fact 1: The scientific impact of confidential Census Bureau's microdata is increasing over time. The number of peer-reviewed publications employing restricted-access microdata from administrative sources or statistical surveys accessed through FSRDCs has steadily increased during our sample period. In relative terms, these papers went from 0.21% to 1.27% of economics scholarship published in the last three decades by U.S.-based authors. The increase is confirmed by the significant coefficient of a trend line over the yearly count of FSRDC papers (Appendix Figure C1). Papers using restricted-access data are often published in the most prestigious outlets. The share of papers using restricted-access microdata that appear in a top five journal is around 15% in our sample period (Appendix Figure C2). Despite some yearly fluctuations, there is no evidence of a decline in the prestige of journals where these papers appear.



(A) Articles using restricted-access data (B) Share of articles using restricted-access data

Figure 1. Use of restricted-access microdata in economics research

Note: The figures show the count and percentage of yearly publications in peer-reviewed economics journals that employ restricted-access data available for analysis at the U.S. Census Bureau facilities.

[10]Articles using restricted-access Census data would seem slightly less likely to be published on a top five outlet relative to articles by the same authors that do not use them. The following section shows that this pattern flips once taking into account author and year fixed effects in a regression framework.

Fact 2: Articles using confidential Census Bureau's microdata are more scientifically impactful. Despite constituting a fraction of all published economic scholarship, papers using confidential microdata hosted at the Census Bureau are disproportionally impactful. These papers represent 0.64% of all papers published in 1991-2019, but they received 1.16% of all citations and constitute 1.22% of the papers appearing in a top five journal. Results of regression analysis presented in Columns (1), (2), and (3) of Table 2 confirm these patterns. Even when restricting the comparison to papers of the same author, papers written in an FSRDC are 50% more likely to be published in the top five journals. The result is not driven by the prestige of the journal: these papers receive, on average, 28% more citations than other papers of the same author appearing in the same journal.[11] In the second part of the table, we repeat the same analyses restricting the analysis to articles that use other microdata available in the U.S., namely PSID and NLS. The results are consistent and suggest that our findings are not driven by systematic differences in methods or topic choices among papers using confidential microdata. As a further measure of scientific quality, we collected data on papers that won best paper prizes (e.g., the *AEJ Best Paper Award* or the *Frisch Medal*). We found that the incidence of paper prizes is around 2.5 times higher in the sample of papers using restricted-access data hosted in FSRDCs. However, we point out that our results are descriptive, and the larger impact of articles using restricted-access microdata could be driven partially by project-level selection. Since applying for data access requires substantial effort, we may expect authors to be more likely to apply if they deem their project as high-potential. While we cannot rule out this possibility, the selection effect would have to be substantially large to explain the differences that we find. Additionally, given the fixed effects that we employ, the selection would have to take place across articles of the same authors.

[11] Columns (2) and (3) of Table 2 use the inverse hyperbolic sine (*asinh*) of citations as the dependent variable. The *asinh* function closely follows the natural logarithm function. Given that the right-hand side variable is a dummy, coefficients should be interpreted as being close to the semi-elasticity of citations when using confidential microdata. The percentage increase in citations due to a paper using restricted-access microdata is $\exp(\beta) - 1$, which is approximately 28% using the coefficient of column (3). Estimations through Poisson Pseudo Maximum Likelihood give a similar result.

Table 2. Economics research using restricted-access microdata has higher scientific impact

Dependent Variables:	Top Five publication	asinh(citations received)		Top Five publication	asinh(citations received)	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Uses restricted-access data (0/1)	0.056***	0.398***	0.250***	0.058*	0.573***	0.509***
	(0.014)	(0.043)	(0.038)	(0.034)	(0.122)	(0.112)
Fixed-effects						
Author	Yes	Yes	Yes	Yes	Yes	Yes
Year-field	Yes	Yes	Yes	Yes	Yes	Yes
Journal			Yes			Yes
Fit statistics						
Observation level	paper-author	paper-author	paper-author	paper-author	paper-author	paper-author
Comparison group	All	All	All	PSID/NLS	PSID/NLS	PSID/NLS
Standard errors	Author	Author	Author	Author	Author	Author
Observations	132,259	132,259	132,259	2,764	2,764	2,761
Mean dependent variable	0.09	55	55	0.161	75	75

Clustered (Author) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: The sample in columns (1) to (3) includes all papers, while in columns (4) to (6) it includes only papers that used FSRDC, NLS, or PSID data. Results of estimating by OLS: $y_{iajft} = UsesRestrictedAccessData_i + FE_{..} + \epsilon_{iajft}$ for paper i, author a, journal j, field f and year of publication t. In columns (1) and (4) the outcome variable y_{iajft} is whether the paper was published in a Top Five journal (i.e., American Economic Review, Econometrica, Journal of Political Economy, Review of Economic Studies, The Quarterly Journal of Economics). In columns (2) to (3) and (5) to (6) the outcome variable y_{iajft} is the inverse hyperbolic sine of the amount of citations that the paper i has received up to 2019 inclusive. In columns (2), (3), (5) and (6) the mean of dependent variable presented is the mean of citations received (rather than mean of asinh(citations received)). Each paper is classified into only one 16 economics fields using the algorithm of Angrist et al. (2020). In order to identify the author fixed effect it is required that the author publishes more than one paper in a given field-year. We keep the sample of papers constant across regressions at the paper-author level by dropping single-paper authors. The amount of observations reported is the effective sample size.

Fact 3: Articles using confidential Census Bureau's microdata receive more citations from policy documents. While 44% of articles in our sample have at least one policy citation, the share is 75% for papers using restricted-access data from U.S. administrative sources or statistical surveys accessed through FSRDCs. This share grows to 89% for articles published in the top five journals, compared to a mean of 65% for the other articles in such journals. The regression analysis presented in Table 3 shows that the difference persists after controlling for field-year, author, and journal fixed effects: articles that use confidential data are 24% more likely to receive policy citations and receive 80% more policy citations. Finally, the last three columns of Table 3 restrict once more the comparison to papers using U.S. microdata from alternative sources. The coefficients are consistent albeit slightly smaller in magnitude, suggesting that papers employing PSID and NLS data tend to be referenced more prominently in the policy debate than the average publication, but less than those using restricted-access microdata.

Table 3. Economics research using restricted-access microdata has higher policy impact

Dependent Variables:	Any policy cites (0/1)	asinh(count policy cites)		Any policy cites (0/1)	asinh(count policy cites)	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Uses restricted-access data (0/1)	0.116***	0.701***	0.587***	0.121**	0.594***	0.468**
	(0.017)	(0.062)	(0.055)	(0.052)	(0.196)	(0.193)
Fixed-effects						
Author	Yes	Yes	Yes	Yes	Yes	Yes
Year-field	Yes	Yes	Yes	Yes	Yes	Yes
Journal			Yes			Yes
Fit statistics						
Observation level	paper-author	paper-author	paper-author	paper-author	paper-author	paper-author
Comparison group	All	All	All	PSID/NLS	PSID/NLS	
Standard errors	Author	Author	Author	Author	Author	Author
Observations	132,259	132,259	132,259	2,764	2,764	2,761
Mean dependent variable	0.469	7	7	0.707	16	16

Clustered (Author) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: In columns (1) to (3) the sample includes all papers, while in columns (4) to (6) the sample includes only papers that used FSRDC, NLS, or PSID data. Results of estimating by OLS: $y_{iajft} = UsesRestrictedAccessData_i + FE_{...} + \epsilon_{iajft}$ for paper i, author a, journal j, field f and year of publication t. In columns (1) and (4) the outcome variable y_{iajft} is whether the paper received at least one citation from a policy source. In columns (2) to (3) and (5) to (6) the outcome variable y_{iajft} is the inverse hyperbolic sine of the count of policy citations that the paper i has received up to 2023 inclusive. Each paper is classified into only one of 16 economics fields using the algorithm of Angrist et al. (2020). In order to identify the author fixed effect it is required that the author publishes more than one paper in a year-field. We keep the sample of papers constant across regressions at the paper-author level by dropping single-paper authors. The amount of observations reported is the effective sample size.

Fact 4: Confidential Census Bureau's microdata are mostly used in labor economics and applied microeconomics. 30% of articles using confidential microdata can be classified as labor economics and an additional 28% as applied microeconomics. Notably, the incidence of these two fields in the set of papers using confidential data available in FSRDCs is more than double the incidence of these fields in the rest of the economic scholarship. These changes are a likely reflection of the increased emphasis on the identification of causal effects in applied economic research, which often requires microdata at the level of individual units. This "credibility revolution" was started in labor economics (J. D. Angrist & Pischke, 2010), hence leading to faster adoption of administrative and confidential microdata in this field.[12] Confirming our priors, articles based on confidential microdata are substantially less focused on macroeconomics or econometric methods (5% and 2% respectively) relative to the other articles (11% and 9% respectively). Within the body of papers using restricted-access data, we noticed a decrease in papers in industrial organization from 18.5% in the first decade to 5.6% in the last decade of our sample. Meanwhile, there has been a constant growth of reduced-form applied micro papers (Appendix Figures C3 and C4), consistent with the broader trends documented by Hamermesh (2013) and J. Angrist et al. (2020).

Fact 5: Articles using confidential Census Bureau's microdata have larger and more age-diverse co-authorship teams. Papers employing confidential data have an average of 2.1

[12]We thank an anonymous referee for this point.

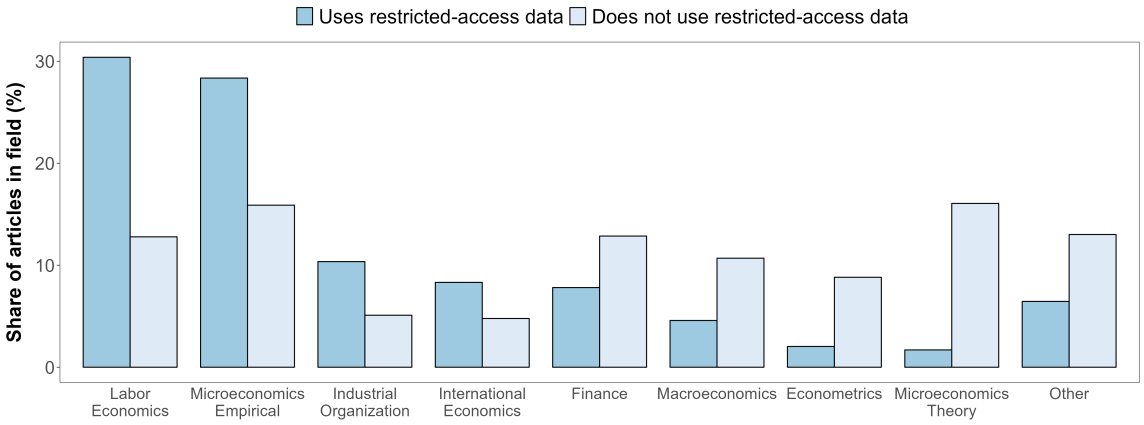


Figure 2. Share of articles using restricted-access microdata by research field

Note: The figure shows the share of articles by research field, separately for papers that employ restricted-access data curated by the U.S. Census Bureau and for all other papers.

authors, while all other economics papers have an average of 1.5 coauthors. Compared to the rest of the sample, the within-paper minimum seniority of coauthors is lower for restricted-access data papers, while the maximum seniority is higher (Appendix Table C1).[13] The result is that papers using administrative and statistical survey microdata are more likely to combine experienced and inexperienced researchers. Figure 3 plots the within-paper difference in seniority between coauthors, which is a synthetic metric capturing the age composition of authorship teams. We notice that the distribution is shifted to the right for Census papers, suggesting that they tend to include senior authors working with junior colleagues. This fits well with oral accounts of the division of labor in those projects, whereby it is usually the junior team member who physically accesses the data enclave to carry out empirical analyses.

Fact 6: Articles using confidential Census Bureau's microdata are more likely to include established researchers affiliated with high-status institutions. Authors of research based on confidential microdata are more established in the discipline: 59% of Census papers have authors who previously published in a top five journal, while this is true only for 38% of the papers in our sample. Regressions show that the difference remains after controlling for field, year, number of coauthors, average seniority, and maximum seniority difference between coauthors. Similarly, the share of papers that have at least one author affiliated with a top 10 department is 42% for papers using restricted-access microdata while it is 35% for the other papers. Using the best-ranked institution among coauthors in an article from Kalaitzidakis et al. (2003), we find that the difference persists after controlling for field-year effects and number of coauthors in the article (Appendix Table C2). While this result is just descriptive in nature, it suggests that the current regime of confidentiality might pose larger challenges to early-career scholars, and risks slowing down scientific progress by constraining access to a crucial research input (Card et al., 2010). Furthermore,

[13]Age is computed as the year of publication minus the first year in which the author published an article. Within-paper minimum seniority of coauthors is 6.8 years for confidential data papers vs. 9.7 years for non-confidential data papers; the average maximum seniority is 18.1 years for confidential data papers and 15.8 years for non-confidential data papers.

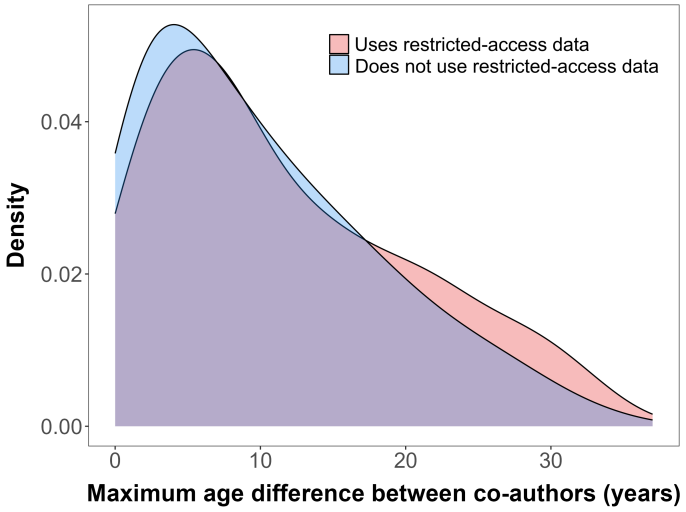


Figure 3. Maximum age difference between coauthors of articles using restricted-access microdata

Note: The figure shows the distributions of the age difference between coauthors of papers that employ restricted-access data curated by the U.S. Census Bureau and for all other papers. Age is computed as the year of publication minus the first year in which the author published an article.

the status quo is potentially contributing to making the profession more hierarchical and unequal, entailing distributional consequences in terms of scientific careers (Nagaraj & Tranchero, 2024).

5. Discussion

This article provides new empirical estimates on the value of restricted-access microdata for economics research. Our analysis shows the dual impact of the federal data infrastructure on scientific production and evidence-based policy-making by focusing on the use of restricted-access data curated by the U.S. Census Bureau. Articles using these data are 50% more likely to be published in the best economics journals, receive 28% more citations from other academic articles, and 80% more citations from policy documents. However, our analysis also shows that the use of these data remains fairly limited, and researchers using it are more likely to be established economists from prestigious institutions. Our article highlights that there are benefits stemming from restricted-access data, suggesting that a further expansion of access may be warranted.[14]

Of course, expanding access would likely entail direct monetary costs for the U.S. Census Bureau and partner universities. We hope that the evidence of this paper, together with cost-benefit calculations from Nagaraj and Tranchero (2024), can help policymakers better weigh the monetary costs with the scientific benefits. More broadly, the creation and distribution of restricted-access microdata share similarities with the provision of a public good (Hoelzemann et al., 2024). While the benefit of using such data for research accrues to society as a whole, the privacy and security

[14]While we cannot rule out the possibility that part of the higher impact could be due to selection of which projects authors apply for data access, the selection effect would have to be very large to completely explain the results.

risks are borne by the data provider (Lane et al., 2024; Ritchie & Welpton, 2011). Data providers thus face a trade-off between granting access to their confidential records and their duty to protect the confidentiality of the information entrusted to them (Foster et al., 2009). Agencies of different countries vary in terms of their policies for (a) reviewing and approving projects and researchers, (b) providing ongoing access to microdata, and (c) reviewing public disclosure of data and results.

In the case of the U.S. Census Bureau, access to confidential data is tightly circumscribed by the U.S. Code and CIPSEA. Within these limits, the Census Bureau has provided access to researchers by setting up physical enclaves with strict security controls. Nagaraj and Tranchero (2024) show that the expansion and opening of new physical enclaves led to an increase in the use of confidential microdata by researchers in nearby institutions. In spite of this, several commentators have worried that the use of administrative data in economics remains far below its true potential, even holding the current regulatory framework constant (Card et al., 2010; Lane, 2021). Promisingly, the Census Bureau has recently taken steps toward providing remote access to confidential data. This process is in its very early stages, and future work should investigate whether this policy change is effective in democratizing and streamlining access to data without significantly compromising privacy or security (Frieder, 2024).

To conclude, this article provides evidence about the importance and growing role of confidential microdata in economics and a discussion of the challenges associated with making these data more accessible. However, the current analysis has limitations. Our findings are descriptive in nature and limited to a single provider of confidential data in the United States. Furthermore, our work is limited to the field of economics, which has recently witnessed a surge in interest in microdata as a result of a "credibility revolution" (J. D. Angrist & Pischke, 2010). Documenting if our findings hold in other fields is an interesting extension. Future work should also examine the career implications of gaining access to such data, and how considerations regarding data access affect the location choices of academics. These limitations notwithstanding, this paper constitutes an important first step in informing the discussions about how to best provide access to confidential microdata for academic scholarship (Lane & Potok, 2024).

Data and Replication Package. Data and code to replicate the analyses of this paper are available in the Harvard Dataverse at the following link:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IJ2ZMM

Disclosure Statement. The authors have no conflicts of interest to declare.

Acknowledgments. We thank Cecil-Francis Brenninkmeijer and Randol Yao for excellent research assistance and participants at the NBER conference "Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences" in May 2024 for their feedback. We gratefully acknowledge the financial support of the Alfred P. Sloan Foundation (Grant Number: G-2021-16965). Any opinions and conclusions expressed herein are those of the authors only, and any errors are our own.

References

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

Abowd, J. M., McKinney, K. L., & Zhao, N. L. (2018). Earnings inequality and mobility trends in the united states: Nationally representative estimates from longitudinally linked employer-employee data. *Journal of Labor Economics*, *36*(S1), S183–S300.

Abowd, J. M., & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, *109*(1), 171–202.

Abowd, J. M., Stephens, B. E., Vilhuber, L., Andersson, F., McKinney, K. L., Roemer, M., & Woodcock, S. (2009). The LEHD infrastructure files and the creation of the quarterly workforce indicators. In *Producer dynamics: New evidence from micro data* (pp. 149–230). University of Chicago Press.

Abraham, K. G., Jarmin, R. S., Moyer, B., & Shapiro, M. D. (2022). *Big data for 21st century economic statistics*. NBER Book Series Studies in Income/Wealth.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., & Lu, S. F. (2020). Inside job or deep impact? extramural citations and the influence of economic scholarship. *Journal of Economic Literature*, *58*(1), 3–52.

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, *24*(2), 3–30.

Atrostic, B. (2007). The center for economic studies 1982-2007: A brief history. *CES working paper*.

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

Beraja, M., Fuster, A., Hurst, E., & Vavra, J. (2019). Regional heterogeneity and the refinancing channel of monetary policy. *The Quarterly Journal of Economics*, *134*(1), 109–183.

Bernard, A. B., & Jensen, J. B. (1999). Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics*, *47*(1), 1–25.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the united states. *American Economic Association, ten years and beyond: Economists answer NSF's call for long-term research agendas*.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2022). Gender differences in peer recognition by economists. *Econometrica*, *90*(5), 1937–1971.

CES. (2017). Center for economic studies and research data centers research report: 2016. *Available on the US 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., Hendren, N., Jones, M. R., & Porter, S. R. (2020). Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, *135*(2), 711–783.

Cole, S., Dhaliwal, I., Sautmann, A., Vilhuber, L., et al. (2020). Handbook on using administrative data for research and evidence-based policy. *https://admindatahandbook. mit. edu/book/v1. 0-rc5/index. html*.

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

Davis, J. C., & Holly, B. P. (2006). Regional analysis using census bureau microdata at the center for economic studies. *International Regional Science Review*, *29*(3), 278–296.

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

Finkelstein, A., Gentzkow, M., & Williams, H. (2021). Place-based drivers of mortality: Evidence from migration. *American Economic Review*, *111*(8), 2697–2735.

Fobia, A. C., Childs, J. H., & Eggleston, C. (2020). Attitudes toward data linkage: Privacy, ethics, and the potential for harm. *Big Data Meets Survey Science: A Collection of Innovative Methods*, 683–712.

Foster, L., Jarmin, R., & Riggs, L. (2009). Resolving the tension between access and confidentiality: Past experience and future plans at the us census bureau. *Statistical Journal of the IAOS*, *26*(3, 4), 113–122.

Frieder, O. (2024). On democratizing data: Diminishing disparity and increasing scientific productivity. *Harvard Data Science Review*.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861–871.

Haltiwanger, J., Jarmin, R. S., & Miranda, J. (2013). Who creates jobs? small versus large versus young. *Review of Economics and Statistics*, *95*(2), 347–361.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, *51*(1), 162–72.

Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of Political Economy*, *109*(4), 673–748.

Hoelzemann, J., Manso, G., Nagaraj, A., & Tranchero, M. (2024). The streetlight effect in data-driven exploration. *NBER Working Paper 32401*.

Jaravel, X., Petkova, N., & Bell, A. (2018). Team-specific capital and innovation. *American Economic Review*, *108*(4-5), 1034–73.

Kalaitzidakis, P., Mamuneas, T. P., & Stengos, T. (2003). Rankings of academic journals and institutions in economics. *Journal of the European Economic Association*, *1*(6), 1346–1366.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, *79*(3), 362–384.

Lane, J. (2021). *Democratizing our data: A manifesto*. MIT Press.

Lane, J., & Potok, N. (2024). Democratizing data: Our vision. *Harvard Data Science Review*.

Lane, J., Spector, A., & Stebbins, M. (2024). An invisible hand for creating public value from data. *Harvard Data Science Review*.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, *71*(6), 1695–1725.

Nagaraj, A., & Tranchero, M. (2024). How does data access shape science? the impact of federal statistical research data centers on economics research. *NBER Working Paper 31372*.

Olley, G. S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, *64*(6), 1263–1297.

Önder, A. S., & Schweitzer, S. (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*(3), 1297–1331.

Reiter, J. P. (2023). Synthetic data: A look back and a look forward. *Transactions on Data Privacy*, *16*(1), 15–24.

Ritchie, F., & Welpton, R. (2011). Sharing risks, sharing benefits: Data as a public good. *Work Session on Statistical Data Confidentiality, Eurostat*.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Carnegie Mellon University*.

Vavra, J. (2021). Tracking the pandemic in real time: Administrative micro data in business cycles enters the spotlight. *Journal of Economic Perspectives*, *35*(3), 47–66.

Appendix A. Short History of the FSRDC Network

A.1. **The Objectives of the FSRDCs Program.** The U.S. Census Bureau constantly collects individual and business information via statistical surveys and enumerations as part of its mandate to assemble timely data about the nation's demographic and economic trends. A crucial challenge in achieving its mission involves earning the trust of households and businesses to elicit complete and truthful answers (Foster et al., 2009). For this reason, the data collected by the U.S. Census Bureau is tightly regulated by the U.S. Code and federal laws like the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[15] However, the U.S. Census Bureau was well aware of the immense research potential of its microdata. In 1982, it established the CES to make these resources accessible to researchers in economics (CES, 2017). From a legal standpoint, the Census Bureau is authorized to give access to its confidential information only as long as it derives a clear benefit for its data programs (McGuckin et al., 1993). The objective of the FSRDC network is to maximize the benefit to Census' programs subject to its confidentiality constraints, and not necessarily maximizing confidential data diffusion (Foster et al., 2009). Therefore, it must be noted that the intention of this study is not to provide a complete policy evaluation of the FSRDC network. First, we would need to expand our focus to other disciplines that benefit from FSRDC data, including demography and especially health policy, which are outside of the scope of our dataset. Second, we would need to consider all benefits occurring to the Census' data programs beyond those encapsulated in peer-reviewed publications. Significant examples include improvements to existing datasets, refining survey questionnaires, or even designing de novo surveys.

A.2. **The Expansion of the FSRDCs Network.** A network of 31 FSRDCs was set up all over the country in a phased manner between 1994 and 2019. After the COVID-19 pandemic, a greater emphasis on remote access has enabled more opportunities to access the FSRDC network.

A.2.1. *Physical Locations.* The first FSRDC was opened in the Census Bureau's Boston Regional Office (Atrostic, 2007). After that, Carnegie Mellon University and the CES pioneered a new institutional model where the data center would have been located and operated by a research institution, with the Census Bureau keeping an oversight role. This institutional arrangement became the standard model followed by all subsequent FSRDCs.[16] More recently, new FSRDCs tend to be sponsored by consortia of local institutions partnering up to split the costs of running the data center. Such institutions include universities, regional branches of the Federal Reserve System, and other research centers active in the social sciences, such as the RAND Corporation or the Russell Sage Foundation.

[15]The legal framework regulating the acquisition, use, and protection of statistical surveys by the U.S. Census Bureau data is outlined in Title 13 of the U.S. Code. Title 13 also requires that anyone accessing data hosted by the Census Bureau be sworn to observe their confidentiality for life or face severe penalties, precisely as if they were Census employees. This is why external researchers are required to undergo background checks and face an application procedure that necessarily involves a certain degree of bureaucratic hurdles (and several months of wait). IRS data are governed by the confidentiality provisions outlined in Title 26 of the U.S. Code. For most other statistical data, the legal framework is given by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).

[16]Interestingly enough, the FSRDC at CMU is also the only one that later closed because of low usage. The other exception is the RTI International FSRDC, which was closed as part of a reorganization of the Triangle Research Data Center with Duke University and the University of North Carolina at Chapel Hill.

Appendix Figure A1 shows the list of the 31 Federal Statistical Research Data Centers opened until 2019. In Nagaraj and Tranchero (2024), we investigate the drivers of this expansion thoroughly. Opening a new FSRDC requires universities to submit a competitive application to the National Science Foundation (NSF), which is then jointly evaluated by the NSF and the Census Bureau. Through a series of interviews and archival research, we found that the Census Bureau and especially the NSF were trying to balance researchers' demand with equitable geographical coverage across the United States. The presence of a nearby data center prevented even top-tier universities with large user bases from obtaining an FSRDC for many years, thus allaying concerns about sorting. Coupled with inevitable idiosyncratic factors behind the establishment of each specific FSRDCs, this means that the timing and location of these openings were partly driven by factors governing geographic equity rather than by pre-existing trends in the use of microdata or research productivity.

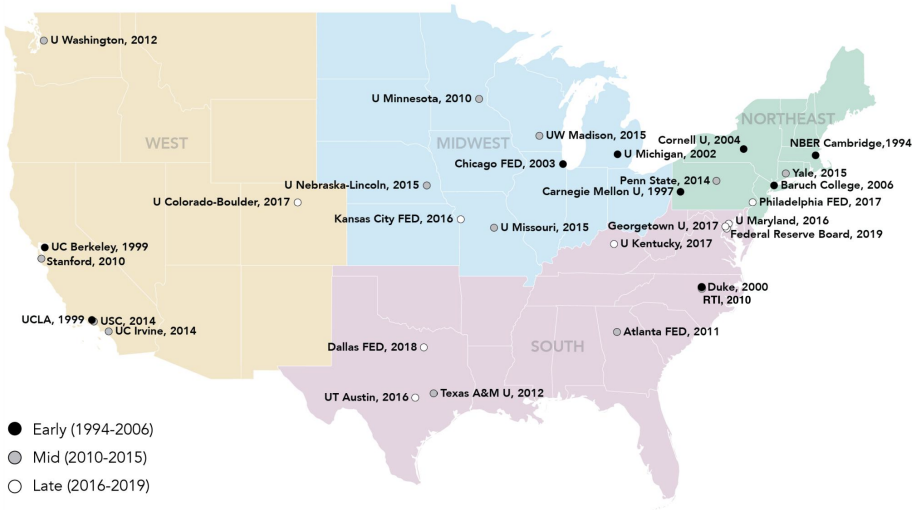


Figure A1. Geographic expansion of FSRDCs over time
Note: Figure reproduced with permission from Figure 1 of Nagaraj and Tranchero (2024)

A.2.2. *Remote Access.* In 2019, the Census Bureau began to provide remote virtual access to a select number of FSRDC researchers working only with Title 13 data. The pilot has been scaled up during the COVID-19 pandemic, resulting in 83 approved projects by mid-2021. However, the possibility of remote access is not extended to data originating from the IRS – which includes several of the most popular databases, such as the LEHD. More details on these exciting developments are available at: https://www.census.gov/about/adrm/fsrdc/about/secure-remote-access.html. While the expansion of remote access is an exciting development, our sample period ends before it was rolled out, thus making it outside the scope of the current paper. Future work should evaluate its impact on democratizing the diffusion of restricted-access microdata.

A.3. **Restricted-access Data Accessible in FSRDCs.** The initial focus of the CES was the creation of data resources on the manufacturing sector. The first matched dataset, developed in 1984, was called the Longitudinal Establishment Database (LED) and consisted of pooling the manufacturing data from 1972 to 1981. Following this first step, CES staff members continued

expanding the LED with data from the Economic Censuses and the Annual Survey of Manufactures, eventually creating what is known as the Longitudinal Research Database (LRD). The LRD became popular among academic researchers, allowing them to conduct pathbreaking empirical research on business dynamics and business demographics (Davis et al., 1998). In the late 1990s, CES began developing a new database, later called LBD (Jarmin & Miranda, 2002). The LBD was created by merging the Standard Statistical Establishment List with the Economic Census, thus creating a source that contains basic information on the universe of all U.S. business establishments. This impressive data effort was joined by the concurrent development of the LEHD data (Abowd et al., 2004). The LEHD merges worker and employer records from Census Bureau surveys with state unemployment insurance claims to create matched employer-employee panel data. Together, the LBD and LEHD constitute unique research tools to investigate the dynamics of the U.S. economy.

Over time, the range of confidential datasets FSRDC offers has expanded. Several other statistical agencies have started making their microdata available through the FSRDC network, most notably the National Center for Health Statistics (NCHS), the Bureau of Labor Statistics (BLS), and the Agency for Healthcare Research and Quality (AHRQ). The advantages of doing so are clear since they can leverage the existing system of data centers without having to muster the resources for creating a similar infrastructure (CES, 2017). Moreover, one farsighted feature of the FSRDC is the ability to allow approved users to merge existing microdata and collaborate in developing new databases. For example, the recent development of the Management and Organizational Practices Survey (MOPS) has been spearheaded by academic researchers, who have collaborated with the U.S. Census Bureau to design a novel survey instrument (Bloom et al., 2019). As a result, statistical surveys and administrative microdata available in FSRDC are increasingly powerful in investigating a broad swath of economic, social, health, and policy questions.

Appendix B. Additional Data Details

B.1. **Bibliographic Records.** The primary data source we leverage is EconLit, a proprietary database of publications in economics curated by the American Economic Association. Our version of EconLit includes 839,513 articles published in 1,856 journals between 1990 and 2019 inclusive. For more details on the data construction, the reader is referred to Nagaraj and Tranchero (2024).

B.1.1. *Disambiguation of authors.* We disambiguate the names of researchers appearing in EconLit following three steps. First, we replace all non-English characters (e.g., "ó" is replaced by "o") and transform the most common name abbreviations into standardized names (e.g., "Ted" is replaced by "Edward"). Second, we apply the disambiguation procedure developed by Önder and Schweitzer (2017). The method employs a graph-theoretic approach that follows a hierarchical process. After identifying the set of name entries with identical surnames, the algorithm constructs a graph of the relationships of all the corresponding first names to each other. Names associated with a particular surname can be identical, different, or subsets of each other. For example, the first name "Michael J." is identical to "Michael J.", it is different from "Tom", and it is a subset of "Michael". The algorithm classifies "Michael J." and "Michael" as the same person if no other "Michael x.", with x. different from J., appears in the data. This approach is equivalent to assuming that the combination of first, middle, and last names uniquely identifies each economist while being conservative in assigning

ambiguous names that lack a clear middle name (Card et al., 2022). Third, we performed extensive manual checks and corrected several misclassifications due to either misspellings in EconLit (e.g., "Tabellini, Gudio" instead of "Tabellini, Guido") or ambiguous names (e.g., "David Levine", that could refer to David I. Levine or David K. Levine). This three-step procedure results in a database of 434,938 unique researchers from the 552,570 names originally appearing in EconLit.[17]

B.1.2. *Disambiguation of institutions.* We standardize the names of the 178,798 affiliations appearing in EconLit. We begin with a list of all research universities in the United States taken from the 2018 Carnegie Classification of Institutions of Higher Education.[18] In particular, we consider all doctoral universities (corresponding to the codes 15, 16, and 17 in the Carnegie Classification), to which we add the main institutions active in economic research (such as the IMF, RAND Corporation, World Bank, RTI, and all the regional FED offices). The result is a list of 439 universities and research centers, which we merge with the EconLit record via fuzzy string matching. This is done by employing string partial ratio similarity to consider the information in the affiliation word ordering (e.g., to distinguish "University of Washington" from "Washington University"). We retain all matches achieving a partial ratio similarity score equal to or greater than 90 over 100, and we manually check them.[19] Together with the previous name disambiguation efforts, this results in a sample of 91,918 peer-reviewed articles by 17,820 researchers affiliated with 344 North American institutions.

B.1.3. *Data quality checks.* A possible concern of our data construction is that it might lead to measurement errors in our dependent variables if the bibliographic records employed are incomplete. To ensure that EconLit provides good coverage of published research, we randomly drew 15 researchers from our panel and manually searched their publication records. For 14 of them, we could find a PDF version of their CV with a list of their scholarly work. We recorded all the publications listed in their CV as the "ground truth" against which we assessed the coverage of EconLit. These researchers collectively published eight articles in a top five journal and 21 articles in a top field journal during our sample period. Except for one article appearing in the *Journal of Economic History*, all other papers were correctly recorded in our data, suggesting that EconLit offers reliable coverage.

B.2. **Research Impact Metrics.** We compute a number of metrics to proxy for the research impact of each article in EconLit.

B.2.1. *Top Five Journals.* We consider as "top publications" all papers appearing in the five economics journals traditionally considered the most prestigious in the field. These journals encompass the *American Economic Review*, the *Quarterly Journal of Economics*, the *Journal of Political Economy*, *Econometrica*, and the *Review of Economic Studies*.

[17]Note that our disambiguation procedure does not drop any data. In cases where authors could not be clearly disambiguated, they have been kept as a singleton in the raw data (even though they would be later dropped in regressions with author fixed effect).

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

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

B.2.2. *Scientific citations.* We augment EconLit by merging each article with its citation count extracted from SSCI/Web of Science. We can match 97.2% of our data with the corresponding citation records to construct citation-weighted metrics of research impact. We use this information to measure the number of citations each article receives up to five years following publication.

B.2.3. *Policy citations.* We merge the articles in our data with the list of mentions in policy documents assembled by Overton, a company that maintains a large database of policy documents linked to the academic research they cite. In particular, we leverage the fact that Overton indexes over 10 million policy documents from 1,700 sources in over 180 countries.[20] The company parses data within policy documents going as far back as the 1920s, thus capturing impact that might take several years to materialize. The type of documents tracked range from white papers of international development organizations to parliamentary transcripts and think tanks' reports. We merge our publication-level dataset to the Overton data using the articles' DOI. In total, we find 750,842 citations from 129,042 policy documents to 70,895 economics articles (38% of our sample). We exclude from the Overton data citations from NBER, CEPR, and IZA sources since they are the working paper versions of academic articles and not policy documents. Figure B1 shows the type of documents citing economics articles.

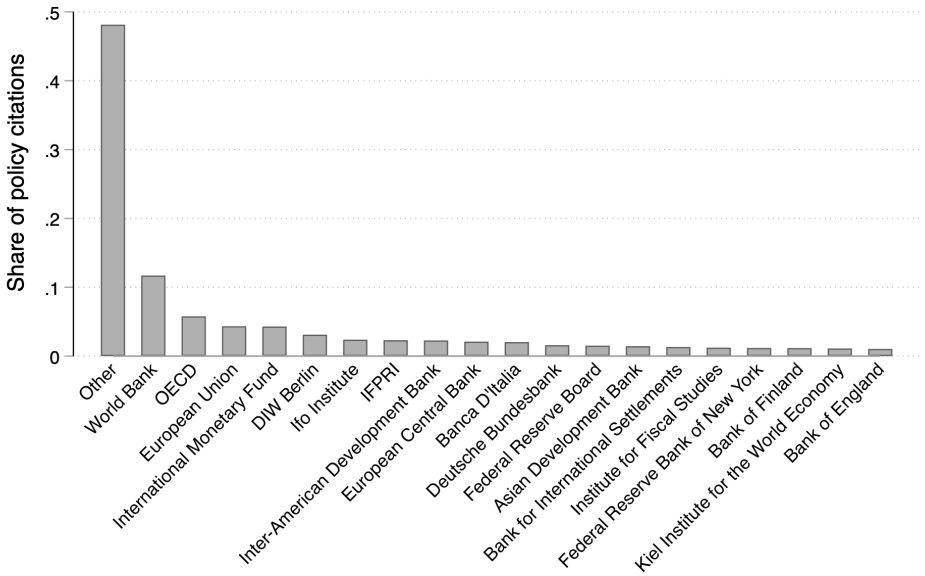


Figure B1. Origin of policy citations to economic research
Note: Figure reproduced with permission from Appendix E of Nagaraj and Tranchero (2024)

B.3. **Articles Using Microdata.** We assembled a novel dataset of articles that *directly* employ specific microdata made available only under access restrictions.

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

B.3.1. *Articles using restricted-access data available in FSRDCs.* When we started our project, there was no official bibliographic record of research using FSRDC data, so we constructed one by carefully searching the bibliome using several complementary strategies. As a starting point, we exploited the fact that papers using restricted-access data are expected to indicate it clearly in the acknowledgment of the published version of the paper. We started by collecting all the most commonly used sentences appearing in the acknowledgments of a sample of FSRDC papers (such as "Census Research Data Center", "do not reflect the views of the Census Bureau", and "Special Sworn Status researchers"). Then we searched for them in the main databases of published research, Web of Science and Scopus, which recently started collecting the acknowledgment sections of journal articles. However, we found that many papers do not report the standard disclaimers required by the Census Bureau. We tried to overcome this limitation with additional searches in databases that allow full-text searches, such as JSTOR,[21] Google Scholar and the NBER working paper repository. We further expanded our search by exploiting the fact that projects approved by the FSRDC are expected to submit a final working paper to the CES for online publication.[22] We collected the metadata of 1,081 working papers and matched them to EconLit through a combination of fuzzy title matching and extensive manual checks. Overall, just about half of these papers were ever published, and only 455 papers could be linked to the corresponding EconLit record.[23] Finally, we leveraged the list of approved FSRDC projects that the U.S. Census Bureau publishes online. These data are updated regularly for everyone interested in tracking the use of FSRDC data and available here: https://www.census.gov/about/adrm/fsrdc/about/ongoing-projects.html. We manually searched the publication records of each researcher whose projects were approved to be carried out in an FSRDC. In total, we were able to find a total of 861 papers published in peer-reviewed journals that could be matched with EconLit. Once we restrict our sample to U.S. researchers affiliated with the AEA, the final sample of papers employing FSRDC data consists of 589 articles written by 525 economics researchers.[24]

B.3.2. *Articles using non-restricted microdata.* In addition to papers using data accessible in FSRDCs, we collected information about articles using microdata that have lower restrictions to use. We chose to focus on two databases popular among U.S.-based researchers. The first is the Panel Study of Income Dynamics (PSID), a longitudinal panel survey of American households conducted by the Survey Research Center at the University of Michigan. Importantly, for our purposes, PSID's website provides an updated bibliography of published work that employs its data.[25] We match this information with our publication database using a fuzzy string matching based on the title

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

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

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

[24]We cannot separately code papers stemming from internal Census projects that are not subjected to the FSRDC application procedure. During our data construction, we found articles from other disciplines, such as sociology, demography, and especially health policy. We excluded them because they are outside the scope of this analysis.

[25]See: https://psidonline.isr.umich.edu/publications/bibliography/Search.aspx. Note that some of PSID's data are provided only under conditions of a restricted-use contract between approved researchers and the University of Michigan. We are unable to track which articles employ their restricted data, but in general, it is safe to assume that restrictions to accessing PSID data are lower than for FSRDC ones.

and year of the paper. We were able to recover 1,023 papers using PSID. Next, we do the same for the National Longitudinal Surveys (NLS) sponsored by the Bureau of Labor Statistics (BLS). These survey data track the labor market experiences and other significant life events of cohorts of Americans over the span of several decades. Once more, we merge our publication data with the curated NLS bibliography using fuzzy matching based on title and years.[26] In total, we are able to recover 634 papers using NLS. Note that the bibliography of PSID and NLS list a larger number of articles, but we can match only a portion of them because of our focus on a specific time frame (1990-2019) and only on U.S.-based researchers.

Appendix C. Appendix Tables and Figures

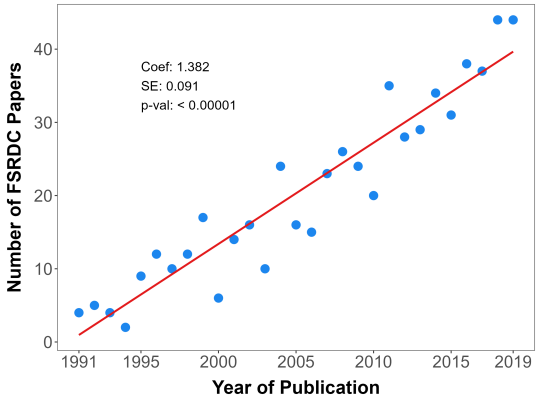


Figure C1. Yearly publications employing confidential Census Bureau's microdata

Note: This figure shows a scatterplot of the yearly number of papers using confidential microdata distributed by the U.S. Census Bureau. We also display the coefficient estimate, standard errors, and fitted line from a regression of the yearly number of publications on the year of publication.

[26] See: https://nlsinfo.org/bibliography-start. Once again, we are unable to determine which articles used NLS datasets with higher confidentiality restrictions.

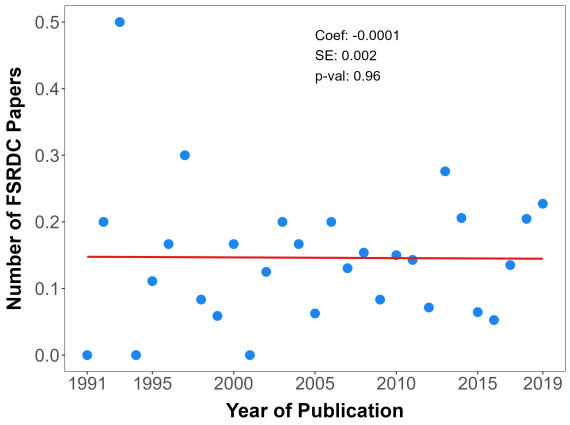


Figure C2. Likelihood that paper employing confidential Census Bureau's microdata appear in a top five journal

Note: This figure shows a scatterplot of the likelihood that an article appears in a top five economics journal over time for papers using confidential microdata distributed by the U.S. Census Bureau. We also display the coefficient estimate, standard errors, and fitted line from a regression of the likelihood of top five publications on the year of publication.

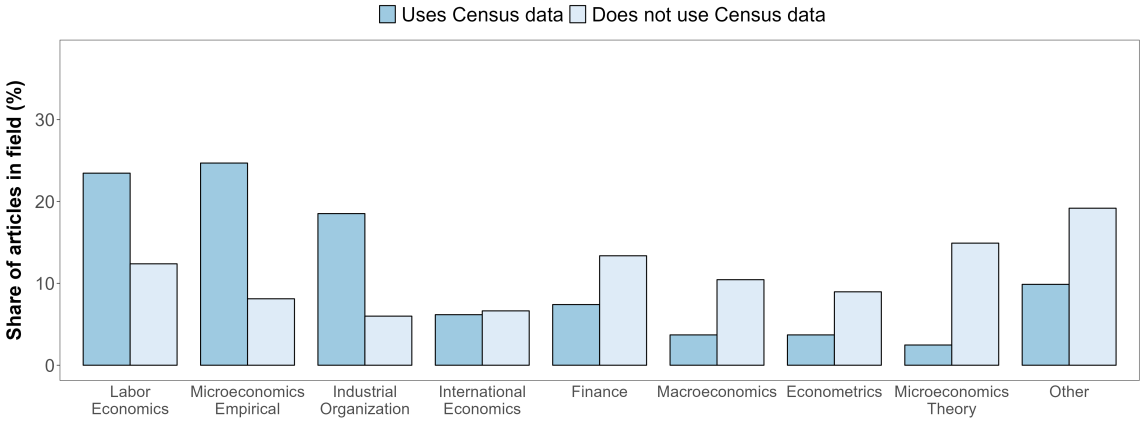


Figure C3. Share of articles using confidential Census Bureau's microdata by research field 1991-2000

Note: The figure shows for the period 1991-2000 the share of articles by research field, separately for papers that employ restricted-access data curated by the U.S. Census Bureau and for all other papers.

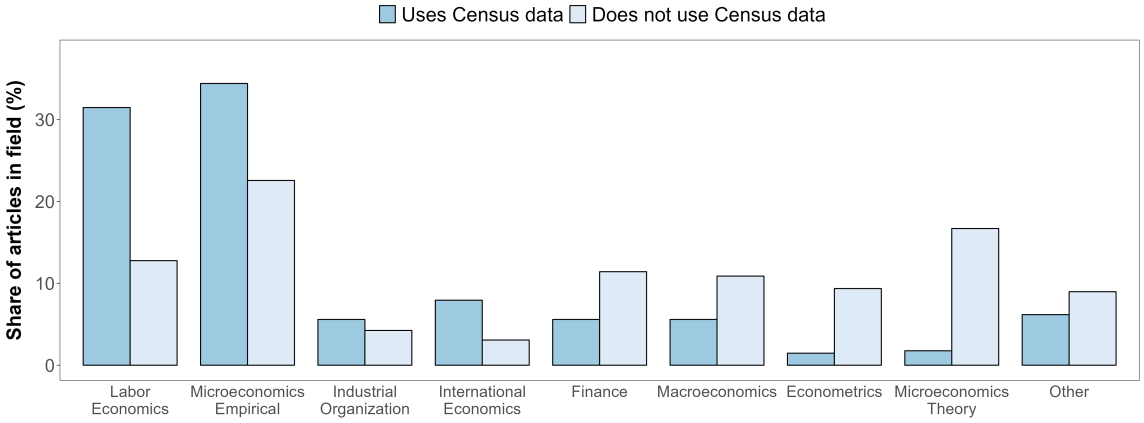


Figure C4. Share of articles using confidential Census Bureau's microdata by research field 2010-2019

Note: The figure shows for the period 2010-2019 the share of articles by research field, separately for papers that employ restricted-access data curated by the U.S. Census Bureau and for all other papers.

Table C1. Articles using confidential Census Bureau's microdata are more likely to include a mix of senior and junior authors.

Dependent Variable:	Maximum age difference		
Model:	(1)	(2)	(3)
Variables			
Data used = FSRDC	1.17**	1.09**	1.28**
	(0.482)	(0.476)	(0.597)
Data used = NLS	0.869	0.909	0.935*
	(0.570)	(0.577)	(0.538)
Data used = Others	0.149	0.035	0.132
	(0.394)	(0.401)	(0.386)
Fixed-effects			
Year	Yes	Yes	
Field		Yes	
Year-field			Yes
Fit statistics			
Observations	35,558	35,558	35,558
Dependent variable mean	10.047	10.047	10.047

Robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: Results of estimating by OLS: $y_{ift} = \beta \times$ Microdata used$_i + FE_{...} + \epsilon_{ift}$ for article i, field f and year of publication t. The outcome variable y_{ift} is the maximum age difference between co-authors of the article. PSID is the reference category. Only papers with more than one coauthor are included in the sample. The amount of observations reported is the effective sample size.

Table C2. Articles using confidential Census Bureau's microdata are more likely to include authors from prestigious institutions

Dependent Variable:	Uses restricted-access data (0/1)	
Model:	(1)	(2)
Variables		
log(min university rank)	-0.002***	-0.001***
	(0.0004)	(0.0004)
Fixed-effects		
Year-Field	Yes	Yes
Number of authors		Yes
Fit statistics		
Observations	57,900	57,900
Dependent variable mean	0.00813	0.00813

Clustered (year-field) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: The dependent variable is a dummy for whether the paper uses FSRDC data. The independent variable is the logarithm of the minimum institution ranking across coauthors (within a paper). The minimum rank possible is 1 and is the highest quality institution isong the ranking by Kalaitzidakis et al. (2003). An observation is a paper. Only papers for which at least one coauthor's institution's ranking is coded are included in the sample. The amount of observations reported is the effective sample size.

Table C3. Articles using confidential Census Bureau's microdata are more likely to include authors with previous Top Five publications

Dependent Variable:	Uses restricted-access data 0/1				
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
Coauthor with top five publication	0.006***	0.003***	0.004***	0.004***	0.004***
	(0.0008)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Fixed-effects					
Year-Field	Yes	Yes	Yes	Yes	Yes
Number of coauthors		Yes	Yes	Yes	Yes
Controls					
Average coauthor age			Yes	Yes	Yes
Max age difference				Yes	Yes
Max age coauthor					Yes
Fit statistics					
Observations	91,570	91,570	91,570	91,570	91,570
Dependent variable mean	0.00643	0.00643	0.00643	0.00643	0.00643

Clustered (Year-Field) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Note: The dependent variable is a dummy for whether the paper uses FSRDC data. The independent variable is whether the paper's coauthorship team had at least one author with a paper published in a top five journal previous to the publication of the FSRDC paper in question. Controls for average age of coauthors in team and maximum age difference within team are added. An observation is a paper. All papers are included in the sample.

Appendix References.

Abowd, J. J., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the united states. *American Economic Review*, *94*(2), 224–229.

Atrostic, B. (2007). The center for economic studies 1982-2007: A brief history. *CES working paper*.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I., & Van Reenen, J. (2019). What drives differences in management practices? *American Economic Review*, *109*(5), 1648–83.

Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2022). Gender differences in peer recognition by economists. *Econometrica*, *90*(5), 1937–1971.

CES. (2017). Center for economic studies and research data centers research report: 2016. *Available on the US Census Bureau's website*.

Davis, S. J., Haltiwanger, J. C., & Schuh, S. (1998). Job creation and destruction. *MIT Press Books*.

Foster, L., Jarmin, R., & Riggs, L. (2009). Resolving the tension between access and confidentiality: Past experience and future plans at the us census bureau. *Statistical Journal of the IAOS*, *26*(3, 4), 113–122.

Jarmin, R. S., & Miranda, J. (2002). The longitudinal business database. *CES working paper*.

Kalaitzidakis, P., Mamuneas, T. P., & Stengos, T. (2003). Rankings of academic journals and institutions in economics. *Journal of the European Economic Association*, *1*(6), 1346–1366.

McGuckin, R. H., McGukin, R. H., & Reznek, A. P. (1993). The statistics corner: Research with economic microdata: The census bureau's center for economic studies. *Business Economics*, 52–58.

Nagaraj, A., & Tranchero, M. (2024). How does data access shape science? the impact of federal statistical research data centers on economics research. *NBER Working Paper 31372*.

Önder, A. S., & Schweitzer, S. (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*(3), 1297–1331.