SAP Template for FSRDC Projects that Request Census Bureau Data

To facilitate proposal development, fill out this proposal template and discuss with your local RDC administrator. The template is intended to be used with FSRDC Research Proposal Guidelines for Projects Requesting Access to U.S. Census Bureau Data, available at https://www2.census.gov/adrm/FSRDC/Apply_For_Access/Research_Proposal_Guidelines.pdf. Once you and the RDC administrator agree that the proposal is ready to submit for Census review, copy the responses from this template to the SAP portal at Research Data Gov.

Datasets

List the datasets that you will request in SAP before starting the application.

Researcher Information

PI/Lead Researcher

Name of principal investigator or lead researcher. Character limit: 200

Institutional Affiliation

Name of employer or affiliate. Character limit: 200

Title

Primary position type.

Choose an item.

Text entry if Other is chosen:

Email

Email address. Once the application is *submitted*, this address will receive all communications about the application and the individual with this address will be the primary person to act upon the application using his/her account. Character limit: 200

Phone number

Phone number. Character limit: 15

Citizenship

Is this person a U.S. citizen?

☐ No ☐ Yes

Residency (item appears if No is selected for Citizenship)

Has this person lived in the United States for at least three of the past five years?

☐ No ☐ Yes

Special Sworn Status

Does this person have active special sworn status (SSS) with the U.S. Census Bureau?

☐ Don't Know ☐ No ☐ Yes

Data Access

Will this person access the data listed under "Data Requested" in the sidebar of this application? Note: Some options may require follow up with the data provider.

Select **Yes** if this person needs a security clearance to obtain an enclave seat or access point to do statistical work **or** needs a security clearance to participate in discussions involving unvetted output in a secure location (this option is only available for some data providers).

Select **No** if this person does not need a security clearance and is only supporting the research project in ways that involve access to information that has been cleared for disclosure.

☐ No ☐ Yes

Add Co-Principal Investigator (fill out and copy this section as needed)

Co-PI/Lead Researcher

Name of co-principal investigator or lead researcher. Character limit: 200

Institutional Affiliation

Name of employer or affiliate. Character limit: 200

Title

Primary position type.

Choose an item.

Text entry if Other is chosen:

Email

Email address. Once the application is submitted, this address will receive all communications about the application and the individual with this address will be the primary person to act upon the application using his/her account. Character limit: 200

Phone number

Phone number. Character limit: 15

Citizenship

Is this person a U.S. citizen?

☐ No ☐ Yes

Residency (item appears if No is selected for Citizenship)

Has this person lived in the United States for at least three of the past five years?

☐ No ☐ Yes

Special Sworn Status

Does this person have active special sworn status (SSS) with the U.S. Census Bureau?

☐ Don't Know ☐ No ☐ Yes

Data Access

Will this person access the data listed under "Data Requested" in the sidebar of this application? Note: Some options may require follow up with the data provider.

Select **Yes** if this person needs a security clearance to obtain an enclave seat or access point to do statistical work **or** needs a security clearance to participate in discussions involving unvetted output in a secure location (this option is only available for some data providers).

Select **No** if this person does not need a security clearance and is only supporting the research project in ways that involve access to information that has been cleared for disclosure.

☐ No ☐ Yes

Add Researcher (fill out and copy this section as needed)

Researcher

Name of researcher. Character limit: 200

Institutional Affiliation

Name of employer or affiliate. Character limit: 200

Title

Primary position type.

Choose an item.

Text entry if Other is chosen:

Email

Email address. Once the application is submitted, this address will receive all communications about the application and the individual with this address will be the primary person to act upon the application using his/her account. Character limit: 200

Phone number

Phone number. Character limit: 15

Citizenship

Is this person a U.S. citizen?

☐ No ☐ Yes

Residency (item appears if No is selected for Citizenship)

Has this person lived in the United States for at least three of the past five years?

☐ No ☐ Yes

Special Sworn Status

Does this person have active special sworn status (SSS) with the U.S. Census Bureau?

☐ Don't Know ☐ No ☐ Yes

Data Access

Will this person access the data listed under "Data Requested" in the sidebar of this application? Note: Some options may require follow up with the data provider.

Select **Yes** if this person needs a security clearance to obtain an enclave seat or access point to do statistical work **or** needs a security clearance to participate in discussions involving unvetted output in a secure location (this option is only available for some data providers).

Select **No** if this person does not need a security clearance and is only supporting the research project in ways that involve access to information that has been cleared for disclosure.

☐ No ☐ Yes

Researcher role on the project

Role on project. Character limit: 200

Research Description

Project Title

Title of project. Character limit: 200

Project Duration

How long is your proposed project **in months**? Make sure there is sufficient time to achieve project objectives and that the duration is not greater than the data provider's maximum duration. Requests for extensions beyond the initial proposed duration depend on the practices of the providing agency. Refer to the document below for the maximum length allowed by the agency from whom you are requesting data. Character limit: 2

Five years is the maximum duration for Census projects. We recommend requesting the full 60 months.

Funding

Please list any grants, FSRDC funding, university funding, and any other sources of funding for this project. If none, enter "None." Character limit: 3800

The University of Michigan and Michigan State University provide FSRDC funding for all faculty and students.

Timeline

What is the timeline for completing project tasks? Refer to the FSRDC Research Proposal Guidelines for Projects Requesting Access to U.S. Census Bureau Data for an example.

Projects that use data from the U.S. Census Bureau are required under Title 13, U.S.C. to contribute to the Census Bureau's mission. These contributions are called "benefits." The development of the timeline requires linking Census benefits to each of the project activities. Assistance from an FSRDC administrator will greatly reduce the need for future revisions to the timeline after submission. To find an FSRDC administrator near you, please check the list of FSRDC contacts. If you are unsure which FSRDC to contact or have a general question, please reach out to the Census Bureau at ced.fsrdc.info@census.gov.

Upload as .docx or .pdf file.

Task	Year				
	1	2	3	4	5

For projects with multiple benefits, Census and IRS like to see how each row in the timeline ties back to benefits to the Census Bureau.

Examples of timelines for demographic, economic, and mixed projects are available in Addendum 1: Project Timelines.

Research Question

What is the proposed research question? Character limit: 3800 - *Note: Spaces are counted towards the character limit.*

Research question examples for demographic, economic, and mixed projects are available in Addendum 2: Research Questions.

Demonstrated Need

Explain why the research question can only be addressed using the requested restricted-use microdata. Be as specific as possible, including listing key variables or methodological advantages of the restricted file compared to a public-use file (where available). Character limit: 10000

Demonstrated need examples for demographic, economic, and mixed projects are available in Addendum 3: Demonstrated Need. It is worth noting that most sections of the SAP application should demonstrate a need for the restricted data.

Study Population

Briefly describe the study population or universe and how it relates to the research question. Character limit: 10000

This section is new to the SAP, so we do not yet have examples to share.

Project Abstract

Provide a project abstract of approximately the length that would be published for a journal article. The abstract should broadly describe the purpose of the research, the type of data to be used, and the hypotheses to be tested. Character limit: 3800

Examples of abstracts for demographic, economic, and mixed projects are available in Addendum 4: Project Abstracts.

Time, Geographic, and Other Units Requested

Some datasets are made available only for the specific years or states that you need for your project. For each dataset you're requesting, look at the "Application-related" tab on the main dataset page. If there are "Provisioned by…" fields, please list the years, states, or other units you are requesting for each dataset, indicating why these specific data are necessary for your research. If you have multiple datasets provisioned this way, list the dataset name and other information for each. If the "Application-related" tab shows no "Provisioned by" fields, you may skip this question. Character limit: 10000

This section is not usually required for Census projects. However, requests for TANF, SNAP, and LEHD data or any other data with "Provisioned by…" fields, will require completion of this section.

Work Location

Where will the data be accessed? If you and, if applicable, other members of the research team plan on using data in multiple places, please mark all that apply. Note that there are often fees for accessing data

in a Federal Statistical Research Data Center (FSRDC), please contact the FSRDC location you intend to use for more information. Other data may have fees for access as noted in the "Fees" field in the "Data Access" section of the information about the asset. Select the appropriate FSRDC location(s).

- ☐ Boston
- ☐ New York - Baruch
- ☐ New York - Cornell
- ☐ Penn State
- ☐ Philadelphia
- ☐ Yale
- ☐ Central Plains
- ☐ Chicago
- ☐ Kansas City
- ☐ Michigan
- ☐ Minnesota
- ☐ Missouri
- ☐ U. of Illinois Urbana-Champaign
- ☐ Wisconsin
- ☐ Atlanta
- ☐ Dallas-Fort Worth
- ☐ Federal Reserve Board
- ☐ Florida
- ☐ Georgetown
- ☐ Maryland
- ☐ Kentucky
- ☐ Texas
- ☐ Texas-UT Austin
- ☐ Triangle-Duke
- ☐ California-Berkeley
- ☐ California-Irvine
- ☐ California-Stanford
- ☐ California-UCLA
- ☐ California-USC
- ☐ Northwest
- ☐ Rocky Mountain
- ☐ Wasatch Front

Data Linkages

Discuss data linkages planned for the research, if any. Please specify datasets to be linked, whether linkages are at the record level (e.g., person, household, business), the purpose of the linkage (e.g., geographic/industry context), and provide basic information on how the linkages are to be performed. Expected length, for projects including linkages only: 2-5 pages. Upload a .docx or .pdf file.

Examples of the proposed data linkages for demographic, economic, and mixed projects are available in Addendum 5: Data Linkages.

User-provided Data

If you are planning to provide other data for use in this project, please describe those data below. Enter information about one dataset at a time, using the "Add User-provided Data" button to include additional sources.

For each user-provided dataset:

Name and Description of Data

Provide the name of the dataset, a brief description of its contents, and the approximate size of the file. Character limit: 200

Ownership

Please indicate whether these data are publicly available or proprietary and provide the source (including URL) from whom you obtained the data. Character limit: 200

Anonymized

Will personal identifiers be removed -- that is, will the data be anonymized?

☐ Don't Know ☐ No ☐ Yes

Linkage

Will you be linking these to other data at the record (e.g., person, household, business) level?

☐ No ☐ Yes

Protected Identification Keys (PIKs)

Will you require Census protected identification keys (PIKs) to be applied to these data in order to link to other Census Bureau data?

☐ Don't Know ☐ No ☐ Yes

Examples of the user provided data sections for demographic, economic, and mixed projects are available in Addendum 6: User Provided Data.

Software Requirements

If your project requires statistical software other than that which is currently available in the location in which you will access the data, please indicate that here. Character limit: 200

This section is not usually required for Census projects.

Methodology

Explain the methodology that will be used for the project. The methodology should be clearly stated and appropriate for the research questions. The metadata catalog, agency publications and statistical products, agency webpages describing the restricted access data, and agency contacts are valuable resources for background information for drafting a strong methodology. Expected length: 5-10 pages.

Your methodology may include, but is not limited to, the following information, as appropriate

- How each requested data set will be used
- Model equations to be estimated
- Estimation methods
- How previous research supports the feasibility of the methodology of the project
- How model variables will be constructed
- Strategies for addressing data quality issues
- Expected sample size and subsamples
- Unit of analysis including level of geography
- Ability to link datasets
- Availability of the study population in the data
- Use of sample weights, design variables, and adjustments for use of complex survey design
- Expected outcomes

Projects that use data from the U.S. Census Bureau are required under Title 13, U.S.C. to contribute to the Census Bureau's mission. These contributions are called "benefits." Your proposed methodology must demonstrate that the use of the data can achieve the required Census benefits. Assistance from an FSRDC administrator will greatly reduce the need for future revisions to your methodology after submission. To find an FSRDC administrator near you, please check the list of FSRDC contacts. If you are unsure which FSRDC to contact or have a general question, please reach out to the Census Bureau at ced.fsrdc.info@census.gov. Upload as .docx or .pdf file.

Examples of the user provided data sections for demographic, economic, and mixed projects are available in Addendum 7: Methodology.

List of References

List any publications referenced in this application as well as any other works that are of importance to this project. Expected length: 2-5 pages. Upload as .docx or .pdf file.

Examples of the List of References for demographic, economic, and mixed projects are available in Addendum 8: List of References.

Project Products

What are the anticipated journal articles, books, working papers, conference presentations, technical memoranda, dissertations, government reports, or other products for this project? Please include the names of journals. Character limit: 3800

This section is new to the SAP, so we do not yet have examples to share.

Requested Output

Describe the anticipated output for this project, including regression/modeling output, summary statistics, and any other output you intend to submit for disclosure review as well as anticipated methods to meet disclosure requirements (e.g., noise infusion). Please check with the agency contact if you are unsure of the agency's output and disclosure requirements.

- For modeling output include descriptions of the samples you anticipate using and variables you plan on reporting results on, including descriptions of categorical variables.
- For tabular output, describe the output needed for the project products in detail, including examples when applicable (i.e., state groupings, levels of output and how you will display restricted-use data, analytic methods to reduce disclosure, etc.).

Each agency will assess this section based on its disclosure requirements. Some agencies' disclosure requirements only allow for projects that emphasize regression/modeling output and a limited number of summary statistics that support this output, while others require table shells of requested output be included in the application.

Expected length: 2-15 pages. Upload as .docx or .pdf file.

Examples of the Requested Output for demographic, economic, and mixed projects are available in Addendum 9: Requested Output.

Census Benefits

Projects that use data from the U.S. Census Bureau are required under Title 13, U.S.C. to contribute to the Census Bureau's mission. These contributions are called "benefits." List each proposed criteria, explaining how the benefit will be achieved. The development of Census benefits that will be accepted during the application review process can be challenging for many researchers. Assistance from an FSRDC administrator will greatly reduce the need for future revisions to your benefit statements after submission. To find an FSRDC administrator near you, please check the list of FSRDC contacts. If you are unsure which FSRDC to contact or have a general question, please reach out to the Census Bureau at ced.fsrdc.info@census.gov.

Refer to criterion details for more information

Choose an item.

Character limit: 10000 (per benefit)

Copy and repeat the criteria drop-down and description as needed.

Examples of the Census Benefits for demographic, economic, and mixed projects are available in Addendum 10: Census Benefits.

Addendum 1: Project Timelines

Exemplar Timelines from Approved Census Proposals

Demographic Proposal – Project Timelines

Task	**Year**				
	1	2	3	4	5
Link core datasets (ACS, CPS, and SIPP) to NUMIDENT and datasets with adult outcomes (e.g., CJARS and Tenant Rental Assistance).	x				
Add PI-supplied data to the linked restricted data	x				
Clean data as necessary to prepare for analyses	x	x			
Conduct the following data analyses to produce deliverables for the Census: (1) assess sensitivity when long-run outcomes of individuals exposed to ECE are computed using point-in-time outcome data and longitudinal outcome data (Criterion 9); (2) demonstrate the inadequacy of existing data for connecting childhood and family characteristics to long-run outcomes and propose solutions (Criterion 9).	x	x	x	x	
Conduct data analyses as described in project proposal to answer research questions on the long-run outcomes (educational attainment, employment and earnings, receipt of government benefits, criminal behavior, etc.) of children exposed to public ECE (Criterion 11).	x	x	x	x	
Prepare and circulate paper(s) on early childhood education and long-run outcomes (Criterion 11).		x	x		
Revise paper(s) on early childhood education and long-run outcomes (Criterion 11).			x	x	
Prepare and submit technical paper on early childhood education and long-run outcomes to the CES working paper series (Criterion 11).				x	x
Prepare and submit: (1) Post Project Certification document and (2) technical memorandum on findings to Census Bureau (Criterion 9).					x
Submit annual Census benefits report	x	x	x	x	x

Economic Proposal – Project Timelines

Task	**Year**				
	1	2	3	4	5
Construct QFRCEN sample	x				
Link QFRCEN sample to LBD, CMF/ASM, and Compustat in support of Criterion 5.	x				
Prepare an internal technical memorandum (for Census Bureau only) documenting the match between QFRCEN, LBD, CMF/ASM, and Compustat in support of Criterion 5.	x				
Submit annual Census benefits report	x				
Link CFS to LBD, iLBD, CMF/ASM, and Compustat in support of Criterion 5.		x			
Prepare an internal technical memorandum (for Census Bureau only) documenting the match between CFS, LBD, CMF/ASM, and Compustat in support of Criterion 5.		x			
Link the Factset Revere data to LBD, iLBD and CMF/ASM in support of Criterion 5.		x			
Start preparing estimates of the population and characteristics of the population prepared for part #1 in support of Criterion 11.		x			
Submit annual Census benefits report		x			
Link the Factset Revere data to CFS in support of Criterion 5.			x		
Finalize the models and the empirical specifications, and Perform data analyses as a part of part #1 in support of Criterion 11.			x		
Submit annual Census benefits report			x		
Start preparing estimates of the population and characteristics of the population prepared for part #2 in support of Criterion 11.				x	
For the part #1, continue data analyses, revisions, and refinements, and improve the previous work based on the comments and suggestions from the Census Bureau and external referees.				x	
Submit annual Census benefits report				x	
Finalize the models and the empirical specifications, and Perform data analyses as a part of part #2 in support of Criterion 11.					x
Continue data analyses, revisions, and refinements, and improve the previous work based on the comments and suggestions from the Census Bureau and external referees.					x
Work on revisions and improvements of the previous work on Criteria 5 and 11.					x
Submit Post Project Certification					x

Mixed Proposal – Project Timelines

Task	Year 1	Year 2	Year 3	Year 4	Year 5
Clean data	X				
Finalize empirical specifications and theoretical underpinnings	X				
Compare the LEHD with the IRS Form 990 data (Criteria 5 and 7) e.g. comparing organizational classifications and number of employees vs number of employees + volunteers	X				
Summarize findings comparing the LEHD to IRS Form 990 data in a technical memorandum to the Census (Criteria 5 and 7) and submit	X				
Submit annual report	X				
Document basic facts about charity employees as described in Criterion 11		X			
Submit annual report		X			
Estimate regression equations related to the effect of shocks to a nonprofit's finances on charity employees			X		
Submit annual report			X		
Estimate regression equations related to the effect of the PSLF Program				X	
Submit annual report				X	
Prepare and submit Post Project Certification document and technical memorandum on findings related to (Criterion 11) to Census Bureau					X

Addendum 2: Research Questions

Research Questions from Approved Census Proposals – Character limit 3800

Demographic Proposal – Research Questions

The proposed research questions are: 1. How does early childhood education (ECE) impact long- run adult outcomes (such as educational attainment, employment, criminality, etc.)? How do those impacts vary by race/ethnicity, sex, and birth county characteristics? 2. What are the impacts of early childhood education on the short- and medium-run outcomes that mediate long- run outcomes (i.e., parental employment, family income, etc.)?

Economic Proposal – Research Questions

The purpose of this project is to conduct a micro-level analysis that documents how shocks that disrupt supply chains affect customers' investment, employment, and asset redeployment decisions. The project will conduct this analysis in two parts that complement each other: First, we conduct an "ex post" analysis by examining how realized disruptions of suppliers' production affect customers' investment, employment, and asset redeployment decisions as well as their productivity. Second, we complement the "ex post" analysis in the first part with an "ex ante" analysis. We construct a new measure of supply chain risk faced by firms and analyze what firms do to manage the risk that suppliers will not be able to deliver the inputs required from them.

In part #1, we focus on realized disruptions to suppliers' operations due to natural disasters. Barrot and Sauvagnat (2016) and Hines et al. (2008) report that natural disasters including blizzards, earthquakes, floods, and hurricanes disrupt production by causing power outages or damaging machines and buildings. Barrot and Sauvagnat (2016) document that disruptions to suppliers' operations cause their customers to experience declines in sales. However, firms may limit these damages by reorganizing their operations and investing in other relationships to prevent a permanent loss of economic activity. The goal of this part is to understand how firms respond to shocks to their suppliers' operations. We will generate estimates of financing, productivity, profitability, investment, employment, wages, and plant closure and sale for firms whose suppliers were affected by natural disasters over the years 1976 to 2026. We will provide a unique analysis of how natural disasters to suppliers affect asset redeployment decisions of customer firms. Specifically, we will seek to understand which types of assets or plants are sold, bought, closed, and allocated more resources to following natural disasters to suppliers.

In part #2, rather than focusing on realized disruptions, we focus on the risk of disruption. In line with this, we use textual analysis of quarterly earnings conference-call transcripts and SEC current report filings (i.e., 8-Ks) to construct firm-level measures of the extent and type of supply chain risk faced by public firms in the U.S. between 1992 and 2026. We will quantify the supply chain risk faced by a firm based on the share of conversations on conference calls and regulatory filings that centers on risks associated with supply chains (Hassan et al. 2019). Elliott, Golub, and Leduc (2020) argue that a decrease in the probability that suppliers are able to deliver the inputs required from them can lead firms to increase their capital investments. However, there is limited understanding of what kinds of investments firms are making when faced with supply chain risks. For this reason, we will approach this question from both the intensive and the extensive margin: First, on the intensive margin, do firms increase/decrease capital investments and employment at their existing plants when faced with supply chain risks? Second, on the extensive margin, do they buy, sell, or close plants? Finally, what type of assets or plants are bought, sold, or closed due to supply chain risks?

Mixed Proposal – Research Questions

Despite the fact that the non-profit sector employs roughly 10% of the American workforce, making it the third largest workforce behind retail and manufacturing (The Independent Sector, 2020), relatively little is known about its employees. Indeed, very few studies conduct a detailed analysis of a particular sector. Those that do consider a narrower subset of workers, are mainly concerned with public sector workers, and do not focus exclusively on charity employees. This project will be the first to use administrative data to paint a picture of the U.S. non-profit sector, and answer important questions such as who works in the non-profit sector, and how do shocks and government policies affect its employees. We are requesting access to the Employer Characteristics File, the Employment History File, the Individual Characteristics File, the Unit to Worker File, and the restricted American Community Survey.

Up until now, researchers have used survey data to focus on the motivation of non-profit workers, and how their behavior differs from for-profit workers. Using British panel data, Gregg et al. (2011) show that non-profit workers are more likely to donate their labor (as measured by unpaid overtime), than their for-profit counterparts. In the U.S. context, Houston (2000, 2006) finds that public employees place a higher value on intrinsic reward, and government employees are more likely to volunteer for charities and donate blood than for-profit employees. However, they found no difference among public service and private employees in terms of individual philanthropy. This is in contrast to Buurman et al (2012), who conclude from a Dutch survey that public sector employees contribute less to charity because they feel that they contribute enough to society at work for too little pay.

Therefore, although we know something about the motivations of charity employees, we know very little about their basic demographics, how they interact with the charity labor market, and how they move between the for- and non-profit sectors. Furthermore, understanding how organization-level shocks affect charity employees, and how government policies affect non-profit employment decisions will add to our knowledge of the charity labor market and contribute to the policy debate over the benefits of public versus private provision of services.

The use of administrative data is crucial in our setting; the Longitudinal Employer Household Dynamics (LEHD) is the only data source that allows us to focus on the employer-employee relationship in the United States. Using the LEHD also provides several benefits over using survey data, such as larger sample sizes and avoiding concerns about non-random measurement error. Indeed, previous studies have found the size of the non-profit sector to be misrepresented in surveys (Millard and Machin, 2007).

The predominant purpose of this study is to increase the utility of Title 13, Chapter 5 data of the Census Bureau by meeting the criteria listed below and as described. We plan to meet Criterion 7 by producing crosswalks between the set of nonprofit organizations included in the Employer Characteristics File of the LEHD Data and the IRS Form 990 Filings. These filings come from two sources: the National Center for Charitable Statistics provides data from Form 990 on a sample of 501(c)(3) nonprofit organizations from 1989 through 2015, and the IRS directly provides data from Form 990 on the universe of nonprofit organizations which file this form electronically, for all years between 2009 and 2018. This crosswalk will greatly expand the set of characteristics available for these employers, including their revenues and expenditures by source, their use of tax preparers and lobbyists, their total assets and liabilities by the end of each fiscal year, and more. Notably, the Form 990 includes a measure of the number of employees working at each nonprofit. Comparison of this figure to its analogue in the LEHD data would allow researchers to assess the quality of the data produced through the LEHD, thus meeting Criterion 5. We also plan to meet Criterion 11 by characterizing the population of nonprofit employees.

Addendum 3: Demonstrated Need

Demonstrated Need from Approved Census Proposals – Character limit 10,000

Demographic Proposal – Demonstrated Need

The core datasets I require are the ACS, CPS, and SIPP. These will provide the sample of individuals whose long-run outcomes I will examine. There are two primary reasons I need access to the restricted versions of these datasets. The first is that the sample size would only be large enough to conduct subgroup analyses when the survey samples are combined—and one of the key contributions of my paper will be estimating treatment effect heterogeneity by subgroup. The restricted version of the ACS contains a larger sample size, and as I will discuss next, I cannot use the CPS or the SIPP at all without linking them to the NUMIDENT file. Second, the restricted ACS, CPS, and SIPP are necessary for my project because they can be linked to the NUMIDENT file.

One reason I need the NUMIDENT file is to identify individuals' county of birth, which is crucial for three reasons. First, having county of birth would allow me to use the CPS and SIPP, which is key for achieving a sufficient sample size and because they contain variables not in the ACS. The CPS and SIPP would be unusable to me without the NUMIDENT linkage because they do not contain the information on birthplace required to define ECE exposure. Second, with only public data, I would have extremely limited knowledge of individuals' pre-treatment characteristics. I would only observe sex, race/ethnicity, and (when available) quarter of birth.

With county of birth, I could merge on county-level characteristics (rural/urban, median income, school district spending per pupil, etc.) and control for them in my analyses. I could also account for time-invariant county of birth characteristics in the DiD analysis by including county of birth fixed effects. Third, without county-level characteristics, I could not sufficiently assess the balance between treatment and control groups. If I find that the two groups are unbalanced on important dimensions, I might employ a re-weighting strategy or construct a control group sample matched on individual- and county-level characteristics. Such strategies would greatly improve the credibility of my results. I also need the NUMIDENT file for exact birthdate. For

the long-run analyses, having exact birthdate will allow me to accurately categorize people into pre-K cohorts. Although quarter of birth is available in some public datasets, this level of detail is inadequate for categorizing individuals on the margin of two pre-K cohorts. Using quarter of birth would result in some cohort misclassification, which would attenuate my results. Having exact birthdates would allow me to overcome this shortcoming of the public data and obtain unbiased results. For the short- and medium-run analyses (i.e., estimating first stage effects and mechanisms), having exact birthdate is essential for performing RD analyses since it is the running variable.

The CPS School Enrollment Supplement is necessary for understanding ECE enrollment in treatment and control group states. The monthly CPS files contain educational attainment, but not enrollment. Using the enrollment variable in the School Enrollment Supplement, I can measure the change in public ECE enrollment across cohorts. This information is essential for converting intent-to-treat effects into treatment effects on the treated. More generally, understanding ECE enrollment patterns is essential for giving context to long-run outcomes. I need the restricted version of this dataset for all the reasons mentioned above for the monthly CPS.

The CJARS data is necessary for estimating ECE's impact on crime. Without CJARS, the best I could do is infer incarceration based on residence in group quarters, which has substantial measurement error. Moreover, incarceration is an extreme event; there are several types of criminal activities I would miss if I could only examine incarceration. The CJARS data would overcome this major limitation by providing data on more minor criminal incidents.

The SSA Summary Earnings Record (SER) and Detailed Earnings Record (DER) files are necessary for estimating long-run impacts because they provide longitudinal accounts of the labor market outcomes I plan to examine. With only survey data, I would be unable to track individuals across time, and I would be forced to rely on self-reported measures of income. Because I intend to estimate long-run effects, it is important that I track individuals for as long as possible and observe year-to-year fluctuations in economic well-being. Moreover, longitudinal data is necessary for capturing individuals who were too young to have labor market outcomes when they were surveyed by the ACS, CPS, or SIPP. Finally, for the mechanism analysis, the SSA data are the only source of information on family earnings in the years following ECE.

Similarly, the Tenant Rental Assistance file is necessary because it provides a longitudinal account of housing assistance. My intent is to measure the impact of ECE on cumulative government benefit receipt as an adult, not just at a point in time. Public data would require that I exclude anyone surveyed at too young an age and it would force me to estimate point-in-time benefit receipt.

The CPS March Supplement and Food Security Supplement are necessary because they contain variables of high interest that are not available in the monthly CPS files. For instance, with only the monthly files, I could not examine receipt of government benefit programs such as the Supplemental Nutrition Assistance Program (SNAP), the Supplementary Nutrition Program for Women, Infants, and Children (WIC), Temporary Assistance for Needy Families (TANF), etc.

The CPS Fertility Supplement is necessary for measuring teen pregnancy and other fertility outcomes not available in the monthly CPS files.

Finally, the CPS and SIPP Supplemental Security (SSR) Extracts are necessary for measuring participation in the Supplemental Security Income program. I need the restricted version of these datasets for all the reasons mentioned above for the regular SIPP and monthly CPS. As much as possible, I will link the adults in my sample to their parents, which is not possible in the public data. This linking is important to my project for three reasons: 1) it would provide an expanded range of childhood controls, 2) it would allow me to estimate impacts on educational and economic mobility across generations, and 3) it would allow me to estimate treatment effect heterogeneity based on characteristics like parental education or family income.

To make these links, I need access to the Master Address Files, the Household Composition Key, and the Decennial Censuses. These datasets contain information on shared residency between parents and their children that can facilitate links. Moreover, for individuals whose parents aren't surveyed in other Census products, the Decennial Census may be the only source available with information on parental demographic and economic characteristics. My analysis also requires the Master Address File for analyzing short- and medium-run mechanisms that mediate ECE's long- run impacts. To estimate how the effects of ECE vary with local context, I must first assign households to local areas. As an example, estimating how take-up of public ECE varies with county-level childcare prices requires that I know which county a household resides in. Public datasets do not have residential location with this degree of specificity, and the Master Address File would overcome this shortcoming. Finally, the Decennial Census will also allow me to calculate county-level statistics to be used as controls and for checking balance between treatment and control groups (as discussed earlier).

Economic Proposal – Demonstrated Need

Access to non-public Census data is essential for the completion of this research project. This access is vital for both the firm- and plant-level analyses.

Our empirical approach requires firm- and plant-level measures of investment, employment, payroll, plant closure, and plant ownership changes. No publicly available source of data (e.g., Compustat) reports these variables at the firm, let alone plant level. Thus, the only way to construct these outcome variables and conduct our empirical analysis is to calculate them using the LBD, CMF, and ASM. For the plant-level analysis, the the CMF and the ASM are the highest quality data available for this project. Access to the CMF and the ASM will allow us to show how realized disruptions to suppliers' operations or increasing supply-chain related risks influence resource allocation within firms. Another advantage of the CMF and the ASM is that we can see both the location and the industry of the plant, which will allow us to observe all plants to a firm that differ across several dimensions (location, focus, input-output), and see which plants experience investment and employment cutbacks.

The plant sales and closures data constitutes another advantage of the Census data for the completion of this project. Maksimovic and Phillips (2001) argue that mergers comprise only half of the total number of asssets traded. Partial-firm assets sales including plant sales as well as plant closures are variables that will help us see how firms reorganize their production when faced with supply chain disruptions or increasing supply-chain risks. No publicly available source of data reports these variables.

The CFS, LFTTD, M3, and QPC will be important to identify suppliers and disruptions to their operations. The CFS provides detailed information about suppliers, their shipment amounts, and their shipment locations. Data at this granularity will be important for us to see suppliers' operations. Data from the CFS will be merged with shipments and capacity utilization data from M3 and the QPC, respectively, to identify whether suppliers experienced disruptions in their operations. No publicly available dataset provides capacity utilization level at the plant-level, which makes it criticial to our project. In addition to domestic suppliers, data from LFTTD will be used to see which firms' global suppliers experienced disruptions in their operations.

The QFRCEN will be essential to analyze firms' financial variables. Existing publicly-available datasets do not provide sufficient information for a representative sample of firms; notably, they exclude private firms. The QFRCEN will first help us construct a measure of trade credit for both public and private firms, which will be important in our analysis of how supply chain disruptions affect the provision of trade credit.

Mixed Proposal – Demonstrated Need

Our project requires restricted data so that we can match employees to their employers. Our primary research question, listed above as 2(b), seeks to discover the extent to which organizational-level shocks are transmitted to individual workers within the firm. We know of no other dataset that would allow us to match U.S. workers to firms in a way that would enable us to observe workers' incomes and transition rates simultaneously with the EINs of the firms for which they work. Furthermore, it is crucial that these data offer comprehensive coverage of employment relationships within the states represented in the LEHD data. The comprehensive nature of this dataset will ensure that we can match nearly all of the nonprofit organizations in these states which were subject to these positive and negative shocks, and thus have sufficient power to answer our research question in a satisfactory way.

Addendum 4: Project Abstracts

Abstracts from Approved Census Proposals

Demographic Proposal – Project Abstracts

Public early childhood education (ECE) expanded rapidly beginning in the 1980s. Between 1980 and 2000, the number of states funding preschool rose from 4 to 30. Importantly, these programs differed on a number of dimensions from the smaller, much-studied programs of the 1960s and 1970s. Enough time has now passed that we can begin to examine the long-run outcomes of the participants of public ECE in the 1980s, 1990s, and 2000s. We hypothesize that participants will have better long-run outcomes (educational attainment, employment, earnings, etc.) than non- participants, although the differences may be smaller than those between individuals born in the 1960s and 1970s. To understand any observed differences, we will also examine the short- and medium-run mechanisms (parental employment, family earnings, etc.) that mediate children's long-run outcomes. Our analyses will use a difference-in-differences (DiD) framework, the synthetic control method (SCM), and regression discontinuity (RD) designs, as applicable. The project uses restricted-use CPS, ACS, and SIPP data, as well as administrative data on criminality, earnings, and government assistance, which grant us greater precision for conducting subgroup analyses, county of birth information for linking adults to their childhood circumstances, exact date of birth for conducting RD analyses, and longitudinal data for examining long-run outcomes—none of which would be possible with public data alone.

Economic Proposal – Project Abstracts

The modern economy is characterized by a complex network of customer and supplier relationships. Idiosyncratic shocks affecting one firm are known to propagate upstream and downstream over supply chains. Direct and indirect customers of firms hit by idiosyncratic shocks, such as natural disasters, experience significant declines in sales growth and profitability. However, while the existing literature documents the risk of propagation, little is known on how firms react when negative shocks affect their customers and suppliers. The purpose of this project to conduct a micro-level analysis that documents how shocks that disrupt supply chains affect customers' investment, employment, and asset redeployment decisions. The project will conduct this analysis in two parts that complement each other: First, we conduct an "ex post" analysis by examining how realized disruptions of suppliers' production affect customers' investment, employment, and asset redeployment decisions as well as their productivity. Second, we complement the "ex post" analysis in the first part with an "ex ante" analysis. We construct a new measure of supply chain risk faced by firms and analyze what firms do to manage the risk that suppliers will not be able to deliver the inputs required from them. The Census of Manufacturers and Annual Survey of Manufacturers, the Longitudinal Business Database, the Longitudinal Employer-Household Dynamics data, the Quarterly Financial Report Census Years, the Commodity Flow Survey, the Manufacturers' Shipments, Inventories, and Orders, the Quarterly Survey of Plant Capacity Utilization, the Survey of Industrial Research and Development, the Business Research & Development and Innovation Survey, and the Census of Auxiliary Establishments and Standard Statistical Establishment List will be used to quantify the effect of supply chain risks on firm behavior and performance.

Mixed Proposal – Project Abstracts

Although the non-profit sector is one of the largest employers in the US, very little is known about its employees. This project attempts to contribute to our understanding of the charity labor market. In particular, the researchers will first investigate the consistency between the LEHD and IRS Form 990 data, and document some basic facts about the nonprofit workforce. Having described worker characteristics (such as age, tenure and education), wage differentials and transition rates within and across sectors, we will proceed by examining the effect of shocks to a nonprofit's finances (or financial prospects) on wage growth and the probability of separation. Finally, we ask how government programs such as the Public Service Loan Forgiveness plan of 2007 affect the decision to work in the non-profit sector and the wages earned.

Addendum 5: Data Linkages

Data Linkages from Approved Census Proposals

Demographic Proposal – Data Linkages

The basis of my sample will consist of the American Community Survey (ACS), monthly Current Population Survey (CPS), and Survey of Income and Program Participation (SIPP). Henceforth, I will refer to these three datasets as my "base sample."

The first linkage I will make is between the base sample and the NUMIDENT file, at the person level. The purpose of this linkage is to attach county of birth and exact birthdate to each observation in the base sample. Exact date of birth will allow me to categorize people accurately into their school cohorts, and it will serve as the running variable in regression discontinuity (RD) analyses. In my regression analyses, I will include fixed effects for counties of birth to account for many of the unobservable characteristics of individual's childhood circumstances. I will also use county of birth to merge on other county-level characteristics, as I will describe next.

The next linkages I will make will be between the base sample and several user-provided county- level datasets. The county of birth variable (merged on from the NUMIDENT) will facilitate these merges. The datasets I will merge include childcare prices from the Department of Labor; treatment indicators for preschool expansion in North Carolina from Anders et al. (2023); replication data from Bailey and Goodman-Bacon (2015); rollout of Food Stamps from Hoynes and Schanzenbach (2009); annual estimates of payroll and employment from the Bureau of Economic Analysis Region (BEAR) Economic Accounts; annual population estimates from Surveillance, Epidemiology, and End Results (SEER); quarterly counts of employment and wages from the Quarterly Census of Employment and Wages (QCEW); estimates of payroll and hours from the Local Area Unemployment Statistics (LAUS); local government expenditures from the Annual Survey of State and Local Government Finances; estimates of population size and demographic/economic characteristics from the 1980, 1990, and 2000 Census Summary Tape files; revenues and expenditures of public school systems from the Annual Survey of School System Finances (F-33); and arrest and offense data from the Uniform Crime Reporting (UCR) Program.

Next, I will link the base sample to a set of user-provided datasets that are at the state level. First I will use county of birth information to assign each observation a state of birth, then I will use state of birth to facilitate the merge. The datasets I will merge include treatment indicators for 1980s Medicaid expansion from East et al. (2022); treatment indicators for ACA Medicaid expansion from Miller et al. (2019); treatment indicators for school finance reforms from Jackson, Johnson, and Persico (2016); AFDC participation rates from Goodman-Bacon (2021) (1958 to 1997); and Head Start enrollment from the Kids Count Data Center.

I will also perform several linkages using Census-provided datasets to obtain more information at the individual level. The SSA summary and detailed earnings records are already linked to the CPS and SIPP, so I will not need to perform linkages there. I will use the ACS PIK crosswalk, the CPS PIK crosswalk, and the SIPP PIK crosswalk to link the base sample to the CJARS dataset to obtain long-run criminality outcomes and to the Tenant Rental Assistance Certification system to obtain long-run rental assistance outcomes. I will also use CPS crosswalks to link the

base sample to the CPS supplement files (ASEC, School Enrollment, Food Security, Fertility, and Supplemental Security Record).

To examine short- and medium-run outcomes, I will merge on characteristics of individual's current counties of residence. Note that the county-level linkages I discussed before were for birth county characteristics; this set of linkages is for current county characteristics. First, I will use the county of residence information available in the base sample surveys. Then, I will fill in missing counties using the Master Address File. Once I have individuals' current county of residence, I will merge on treatment indicators for Transitional Kindergarten programs in CA, MI, and WA; measures of childcare supply from the Center for American Progress; childcare prices from the U.S. Department of Labor; annual estimates of payroll and employment from the Bureau of Economic Analysis Region (BEAR) Economic Accounts; annual population estimates from Surveillance, Epidemiology, and End Results (SEER); quarterly counts of employment and wages from the Quarterly Census of Employment and Wages (QCEW); estimates of payroll and hours from the Local Area Unemployment Statistics (LAUS); local government expenditures from the Annual Survey of State and Local Government Finances; estimates of population size and demographic/economic characteristics from the 1980, 1990, and 2000 Census Summary Tape files; revenues and expenditures of public school systems from the Annual Survey of School System Finances (F-33); and arrest and offense data from the Uniform Crime Reporting (UCR) Program.

Economic Proposal – Data Linkages

Census Bureau Data

- Annual Survey of Manufacturers (ASM), 1976-2019, and if available 2020-2026. This dataset will provide information for a subset of CMF plants (those with greater than 250 employees and a randomly selected subset of smaller plants) in non-Census years. This will include key plant-level information such as capital expenditures, total assets, value of shipments, employment, industry, and location.
- Census of Manufacturers (CMF), 1977, 1982, 1987, 1992, 1997, 2002, 2007, 2012, 2017, and 2022 if and when available. This provides information for all U.S. manufacturing plants with at least one paid employee and is conducted every five years in "Census years" (years ending with either 2 or 7). This will include information such as capital expenditures, assets, value of shipments, employment, industry, and location.
- Longitudinal Business Database (LBD), 1976-2019, and if available 2020-2026. We will use longitudinal establishment identifiers in the LBD to construct linkages between the CMF and ASM. Once the data are linked at the plant level they will then be aggregated to the firm level to construct our panel data set.
- Integrated Longitudinal Business Database (iLBD), 1976-2018, and if available 2019-2026. We will use iLBD to identify non-employer firms on the supply chain. More specifically, we will analyze how shocks affect the closure or birth of non-employer firms on the supply chain.
- Quarterly Financial Report (QFRCEN), 1977, 1982, 1987, 1992, 1997, 2002, 2007, 2012, 2017, and 2022 if available. QFRCEN will be necessary for our analysis of the financial condition of firms whose suppliers' operations were disrupted. We will use these data to construct measures of trade credit supplied to customers and trade credit received from suppliers.
- Commodity Flow Survey (CFS), 1993, 1997, 2002, 2007, 2012, 2017, and 2022 if available. These data will be merged with Compustat and the Factset Revere databases to identify a firm's suppliers.
- Manufacturers' Shipments, Inventories, and Orders (M3), 1992-2019, and if available 2020-2026. These data will be merged with the CFS, CMF, and ASM to identify suppliers experiencing a drop in their shipments.
- Quarterly Survey of Capacity Utilization (QPC), 1976-2020, and if available 2021-2026. These data will be merged with the CFS, CMF, and ASM to identify suppliers experiencing a drop in their their capacity utilization.
- Longitudinal Firm Trade Transactions Database (LFTTD), 1992-2019, and if available 2020-2026. These data will be used to identify which firms experienced disruptions in their global supply chains.
- Research and Development Surveys (RADS), 1976-2018, and if available 2019-2026. This data will be merged with CFS, CMF, and ASM to identify suppliers using R&D extensively in their production.
- Census of Auxiliary Establishments (AUX), 1977, 1982, 1987, 1992, 1997, 2002, 2007, 2012, and 2017, 2022 if and when available. We will use information on non-production establishments, including headquarters.
- County Business Patterns Business Register (CBPBR), 1976-2019, and if available 2020-2026. We will extract the location of headquarters and other plants from the CBPBR, which will supplement the information contained in the LBD. Data from the CBPBR will be used to link the other CMF/ASM/LBD/iLBD/QFRCEN to Compustat.
- Compustat-Business Register bridge file, 1976-2016, and if available 2017-2026. This will be used to link the CMF/ASM/LBD/iLBD/QFRCEN to Compustat. We will link the Factset Revere data on supply chain relationships to CMF/ASM/LBD/iLBD/QFRCEN via the Compustat GVKEY identifier.

The years of data being requested, 1976-2026, are justified in each subsection. To summarize, in part #1, data from 1976 to 2026 is requested. The longer time frame will enable us to see more clearly the effects supply chain disruptions on the resource allocation and performance of customer firms. It will also allow us to make sure that our results are not driven by business cycles including the early 1990s or the 2001 recession. In part #2, data from 1992 to 2026 is being requested since our supply chain risk measure based on earnings call transcripts and company filings go back to 1992.

Non-Census Bureau Data (Provided by Researcher)

In addition to the Census Bureau data, we will use several external data sources. Most of the external sources concern U.S. publicly traded companies and have been linked together via the GVKEY identifier. Thus, these sources can be linked to the CMF/ASM/LBD/QFRCEN via the CBPBR-Compustat Bridge. More precisely,

- Standard & Poor's Compustat. Firm-level balance sheet and related variables for the years 1976 to 2021 This data will be used to construct firm-level financial variables – such as leverage and debt-to-cash flow ratios – used as controls in the estimating equations. It will also be used to identify suppliers of firms.
- Factset Revere Supply Chain Relationship Database. Factset Revere collects relationship information from primary public sources such as SEC 10-K annual filings, investor presentations, and press releases, and classifies them through relationship types (e.g., customer, supplier, competitor, different types of partnerships). Factset Revere spans the period 2002 – 2020. We will merge the customer and supplier information with financial information from Compustat.
- SHELDUS (Spatial Hazard and Loss Database for the United States) database compiled by the Center for Emergency Management and Homeland Security at Arizona State University for the years 1976 to 2020. These data will be used to identify the date and estimated dollar amount of damages for each natural disaster as well as the FIPS codes of affected counties.
- Product market fluidity measure, introduced by Hoberg, Phillips, and Prabhala (2014), from the Hoberg-Phillips Data Library. This proxy assesses the degree of competitive threat and product market change surrounding a firm. We use it to evaluate whether the reaction to natural disasters of firms facing different competitive threats varies in line with our hypotheses. The url for the data is: https://hobergphillips.tuck.dartmouth.edu/industryconcen.htm
- 8-K filings through the Security and Exchange Commission's (SEC) EDGAR website to construct our firm-level supply chain risk measure. The SEC requires firms to disclose any material information such as earnings projections, bankruptcy, officer departures, material definitive agreements, or shareholder vote results within four business days, which makes 8-K filings a critical source of information for investors and analysts.
  The url for the data is: https://www.sec.gov/edgar.shtml
- Earnings calls transcripts from Thomson Reuters' Street Events from 2002 to 2021. Generally, firms host an earnings call every fiscal quarter to share information they wish to disclose. Presentation by company executives is followed by a question-and-answer (Q&A) session with market participants. We will use the transcripts of these calls to construct our firm-level supply chain risk measure.
- Bureau of Economic Analysis's (BEA) Input-Output tables. The data will be used to identify whether plants in output industries are more likely to experience investment and employment cutbacks following an increase in supply chain risk.

To summarize, we will use plant-level data from the CMF in conjunction with the ASM, LBD, iLBD, QFRCEN, M3, QPC, CFS, LFTTD, SIRD, and BRDIS for the period 1976 to 2026 to estimate the effect of realized supply

chain disruptions and increasing supply chain risks on trade credit, investment, employment and payroll, profitability, productivity, plant sale and closure in equations (1) through (6). We will supplement these estimating equations with variables constructed from data in Compustat and Factset Revere. These variables will be linked to Census data via the Compustat-CBPBR Bridge (maintained by the Census Bureau) using GVKEY identifiers.

Note: Please specify datasets to be linked, whether linkages are at the record level (e.g., person, household, business), the purpose of the linkage (e.g., geographic/industry context), and provide basic information on how the linkages are to be performed.

Mixed Proposal – Data Linkages

We are requesting the following files:

- Employer Characteristics File (ECF, ECF T26) – 2014 snapshot and future snapshots through 2024 as available
- Employment History File (EHF) - 2014 snapshot and future snapshots through 2024 as available
- Individual Characteristics File (ICF, ICF T26) - 2014 snapshot and future snapshots through 2024 as available
- Unit to Worker File (U2W) - 2014 snapshot and future snapshots through 2024 as available
- American Community Survey (ACS) – 2005-2018 and 2019-2024 as available
- BOC PIK Crosswalk American Community Survey (CENSUS_CROSSWALK_ACS) – 2005-2018 and 2019-2024 as available

The ECF will allow us to identify firms which were subject to shocks, as well as firms which serve as controls. The EHF contains many of our desired outcome variables, including earnings and transition rates, and can enable us to examine whether charity workers' employment transitions occur within the charity sector or whether they tend to leave the charity sector to work at for-profit firms. The Individual Characteristics File is of interest to us because it provides a set of demographic controls, which we intend to include in our individual-level regressions. Demographics are of particular interest in the charity sector because earnings growth will be correlated with age and gender. The latter may occur because of statistical discrimination or possibly because of unobservable differences in hours which may follow from gendered differences in time required to care for family members. In either case, failure to control for employees' demographic characteristics may bias our results. We are therefore requesting access to the restricted American Community Survey file, so as to incorporate a comprehensive set of demographic characteristics into our analysis. Finally, the U2W file will facilitate matching between the ICF and EHF or ECF for multi-unit employers, which constitute between 30-40% of state-level employment.

Note: Please specify datasets to be linked, whether linkages are at the record level (e.g., person, household, business), the purpose of the linkage (e.g., geographic/industry context), and provide basic information on how the linkages are to be performed.

Addendum 6: User Provided Data

User Provided Data from Approved Census Proposals

Demographic Proposal – User Provided Data

User-Provided Dataset 1

Dataset name: Transitional Kindergarten (TK) – Indicators denoting timing of district-level adoption of TK in California, Michigan, and Washington. (15 MB)
Is dataset publicly available or proprietary? Source: URL provided for CA, proprietary for MI and WA • URL: https://www.cde.ca.gov/ds/ad/filestkdata.asp
Will PIKs be applied? No

User-Provided Dataset 2

Dataset name: Childcare Supply – The capacity of licensed or registered childcare providers at the census tract level in every state and Washington, D.C. in 2018. (30 MB) Is dataset publicly available or proprietary? Source: Center for American Progress • URL: https://www.americanprogress.org/article/americas-child-care-deserts-2018/
Will PIKs be applied? No

User-Provided Dataset 3

Dataset name: National Database of Childcare Prices – Dataset on childcare prices at the county level, currently available from 2008 to 2018. (30 MB)
Is dataset publicly available or proprietary? Source: U.S. Department of Labor • URL: https://www.dol.gov/agencies/wb/topics/childcare
Will PIKs be applied? No

User-Provided Dataset 4

Dataset name: Preschool Expansion in North Carolina – Indicators denoting timing of and counties assigned treatment with regard to the expansion of Smart Start in the 1990s. (10 MB)
Is dataset publicly available or proprietary? Source: Anders et al. (2023) • URL: https://www.aeaweb.org/articles?id=10.1257/pol.20200660&from=f
Will PIKs be applied? No

User-Provided Dataset 5

Dataset name: 1980s Medicaid Expansion – Indicators denoting timing of and states assigned treatment with regard to the expansion of Medicaid in the 1980s. (10 MB) Is dataset publicly available or proprietary? Source: East et al. (2023) • URL: https://www.aeaweb.org/articles?id=10.1257/aer.20210937
Will PIKs be applied? No

User-Provided Dataset 6

Dataset name: ACA Medicaid Expansion – Indicators denoting timing of and states assigned treatment with regard to the ACA expansion of Medicaid in 2014. (10 MB) Is dataset publicly available or proprietary? Source: Miller et al. (2020) • URL: https://www.aeaweb.org/articles?id=10.1257/app.20190176
Will PIKs be applied? No

User-Provided Dataset 7

Dataset name: School Finance Reforms – Indicators denoting timing of and states assigned treatment with regard to school finance reforms. (1 MB)
Is dataset publicly available or proprietary? Sources: Jackson, Johnson, and Persico (2016) and Lafortune, Rothstein, and Schanzenbach (2018) • URL: https://www.aeaweb.org/articles?id=10.1257/app.20160567
Will PIKs be applied? No

User-Provided Dataset 8

Dataset name: AFDC Participation Rates – AFDC participation rates for white and nonwhite children from 1958 to 1997. (10 MB)
Is dataset publicly available or proprietary? Source: Goodman-Bacon (2021) • URL: https://www.journals.uchicago.edu/doi/abs/10.1086/695528
Will PIKs be applied? No

User-Provided Dataset 9

Dataset name: Head Start Rollout – County-level rollout of community health centers and Head Start measured by grant funds. (20 MB)
Is dataset publicly available or proprietary? Source: Bailey and Goodman-Bacon (2015) • URL: https://www.aeaweb.org/articles?id=10.1257/aer.20120070
Will PIKs be applied? No

User-Provided Dataset 10

Dataset name: Food Stamps Rollout – Year of rollout of Food Stamps program at the county-level from 1961 to 1979. (20 MB)
Is dataset publicly available or proprietary? Source: Almond, Hoynes, and Schanzenbach (2016) • URL: https://www.aeaweb.org/articles?id=10.1257/aer.20130375
Will PIKs be applied? No

User-Provided Dataset 11

Dataset name: Payroll and Employment – Annual county-level estimates on total payroll and employment. (30 MB)
Is dataset publicly available or proprietary? Source: Bureau of Economic Analysis Region (BEAR) Economic Accounts • URL: https://www.bea.gov/data/economic- accounts/regional
Will PIKs be applied? No

User-Provided Dataset 12

Dataset name: County-Level Population – Annual county-level US population estimates at the single year of age level, from 1969 – 2020. (30 MB)
Is dataset publicly available or proprietary? Source: Surveillance, Epidemiology, and End Results (SEER) • URL: https://seer.cancer.gov/popdata/
Will PIKs be applied? No

User-Provided Dataset 13

Dataset name: Employment and Wages – Quarterly counts of employment and wages at various geographical units and industry. (30 MB)
Is dataset publicly available or proprietary? Source: Quarterly Census of Employment and Wages (QCEW) • URL: https://www.bls.gov/cew/
Will PIKs be applied? No

User-Provided Dataset 14

Dataset name: Payroll and Hours – Estimates of payrolls and hours at various geographical units. (30 MB)
Is dataset publicly available or proprietary? Source: Local Area Unemployment Statistics (LAUS) • URL: https://www.bls.gov/lau/
Will PIKs be applied? No

User-Provided Dataset 15

Dataset name: Historical Government Expenditures – Annual measures of government expenditures at the state and local level, from 1967 to 1992. (30 MB)
Is dataset publicly available or proprietary? Source: Annual Survey of State and Local Government Finances • URL: https://www.census.gov/programs-surveys/gov- finances/data.html
Will PIKs be applied? No

User-Provided Dataset 16

Dataset name: Current Government Expenditures – Annual measures of government expenditures at the state and local level, from 1993 to 2019. (30 MB)
Is dataset publicly available or proprietary? Source: Annual Survey of State and Local Government Finances • URL: https://www.census.gov/programs-surveys/gov- finances/data.html
Will PIKs be applied? No

User-Provided Dataset 17

Dataset name: 1980 Census Summary Tape Files – State- and county-level population size and demographic/economic characteristics. (40 MB)
Is dataset publicly available or proprietary? URL:
https://www.icpsr.umich.edu/web/ICPSR/series/15 Will
PIKs be applied? No

User-Provided Dataset 18

Dataset name: 1990 Census Summary Tape Files – State- and county-level population size and demographic/economic characteristics. (40 MB)
Is dataset publicly available or proprietary? URL:
https://www.icpsr.umich.edu/web/ICPSR/series/165 Will
PIKs be applied? No

User-Provided Dataset 19

Dataset name: 2000 Census Summary Tape Files – State- and county-level population size and demographic/economic characteristics. (40 MB)
Is dataset publicly available or proprietary? URL:
https://www.icpsr.umich.edu/web/ICPSR/series/166 Will
PIKs be applied? No

User-Provided Dataset 20

Dataset name: School District Finances – Enrollment, revenues, expenditures, debt, and assets of elementary and secondary public school systems from 1986 to 2020. (30 MB) Is dataset publicly available or proprietary? Source: Annual Survey of School System
Finances (F-33) • URL: https://www.census.gov/programs-surveys/school-finances.html • URL: https://nces.ed.gov/ccd/files.asp

Will PIKs be applied? No

User-Provided Dataset 21

Dataset name: Head Start Enrollment – State-level enrollment in Head Start from 1988- 2021, by age group. (30 MB)
Is dataset publicly available or proprietary? Source: Kids Count Data Center • URL: https://datacenter.kidscount.org/data#USA/2/0/char/0
Will PIKs be applied? No

User-Provided Dataset 22

Dataset name: Crime Data – County-level detailed arrest (1974-2016) and offense (1974- 2017) data. (40 MB)
Is dataset publicly available or proprietary? Source: Uniform Crime Reporting (UCR) Program Data from Kaplan (2019) • URL: https://www.openicpsr.org/openicpsr/project/108164/version/V3/view
Will PIKs be applied? No

User-Provided Dataset 23

Dataset name: Georgia Pre-K Enrollment – County-level enrollment in Georgia Pre-K in public and private provider sites, 1997, 1999, and 2003-present. (20 MB)
Is dataset publicly available or proprietary? 2003-present: Georgia Department of Early Care and Learning 1997 URL: https://www.issuelab.org/resources/4108/4108.pdf • 1999 URL: http://www.gunviolence.issuelab.org/resources/4134/4134.pdfpresent. (20 MB) Will PIKs be applied? No

Economic Proposal – User Provided Data

Non-Census Bureau Data (Provided by Researcher)

In addition to the Census Bureau data, we will use several external data sources. Most of the external sources concern U.S. publicly traded companies and have been linked together via the GVKEY identifier. Thus, these sources can be linked to the CMF/ASM/LBD/QFRCEN via the CBPBR-Compustat Bridge. More precisely,

- Standard & Poor's Compustat. Firm-level balance sheet and related variables for the years 1976 to 2021 This data will be used to construct firm-level financial variables – such as leverage and debt-to-cash flow ratios – used as controls in the estimating equations. It will also be used to identify suppliers of firms.
- Factset Revere Supply Chain Relationship Database. Factset Revere collects relationship information from primary public sources such as SEC 10-K annual filings, investor presentations, and press releases, and classifies them through relationship types (e.g., customer, supplier, competitor, different types of partnerships). Factset Revere spans the period 2002 – 2020. We will merge the customer and supplier information with financial information from Compustat.
- SHELDUS (Spatial Hazard and Loss Database for the United States) database compiled by the Center for Emergency Management and Homeland Security at Arizona State University for the years 1976 to 2020. These data will be used to identify the date and estimated dollar amount of damages for each natural disaster as well as the FIPS codes of affected counties.
- Product market fluidity measure, introduced by Hoberg, Phillips, and Prabhala (2014), from the Hoberg-Phillips Data Library. This proxy assesses the degree of competitive threat and product market change surrounding a firm. We use it to evaluate whether the reaction to natural disasters of firms facing different competitive threats varies in line with our hypotheses. The url for the data is: https://hobergphillips.tuck.dartmouth.edu/industryconcen.htm
- 8-K filings through the Security and Exchange Commission's (SEC) EDGAR website to construct our firm-level supply chain risk measure. The SEC requires firms to disclose any material information such as earnings projections, bankruptcy, officer departures, material definitive agreements, or shareholder vote results within four business days, which makes 8-K filings a critical source of information for investors and analysts.
  The url for the data is: https://www.sec.gov/edgar.shtml
- Earnings calls transcripts from Thomson Reuters' Street Events from 2002 to 2021. Generally, firms host an earnings call every fiscal quarter to share information they wish to disclose. Presentation by company executives is followed by a question-and-answer (Q&A) session with market participants. We will use the transcripts of these calls to construct our firm-level supply chain risk measure.
- Bureau of Economic Analysis's (BEA) Input-Output tables. The data will be used to identify whether plants in output industries are more likely to experience investment and employment cutbacks following an increase in supply chain risk.

Note: The SAP requests that researchers enter information individually for each dataset with a character limit of 200.

Mixed Proposal – User Provided Data

Dataset provided	URL	Description	Years	Census linking variable
IRS Form 990s	https://nccs-data.urban.org/index.php https://registry.opendata.aws/irs990/	Information returns for all 501(c)(3) charitable organizations required to file a Form 990	1989 - present	EIN, year
Chronicle of Philanthropy (Top US Donors)	https://www.philanthropy.com/interactives/philanthropy-50#id=browse_2018	Annual list of the top 50-60 US donors	1998 - 2016	EIN, year
Madoff victims	https://www.nytimes.com/interactive/projects/madoff/page/1	Names and addresses of victims of the Madoff scandal	2008	EIN
Foundation Directory Online	https://fconline.foundationcenter.org	The FDO maps foundations to their recipient organizations	2004 - present	EIN, year
DonorSearch	https://www.donorsearch.net	DonorSearch maps individual donors to their recipient organizations	2007 - present	EIN, year
NLSY97	https://www.bls.gov/nls/nlsy97.htm	NLSY97 follows a sample of 8,984 individuals born between 1980 and 1984	1997 - present	Probabilistic match

PSID	https://psidonline.isr.umich.edu	Nationally representative sample of over 18,000 individuals living in 5,000 families in the U.S.	2004 - present	Probabilistic match
Baccalaureate and Beyond Longitudinal Study	https://nces.ed.gov/surveys/b&b/	Nationally representative sample of postsecondary students	1993 - present	Probabilistic match
Beginning Postsecondary Students	https://nces.ed.gov/surveys/bps/	Nationally representative sample of students enrolled in their first year of postsecondary education	2003 - 2017	Probabilistic match

In order to facilitate the analysis outlined in section 2(b), we will provide a variety of forms of external data. First, we will match the employers in the ECF to the set of nonprofits which file IRS Form 990. These data will come from the National Center for Charitable Statistics Core Files and the database of IRS Form 990 Filings by e-filers. The former dataset is available from 1989 through 2015 and is hosted through the Urban Institute, while the latter covers the period from 2009 through the present and comes directly from the IRS via Amazon AWS. Each Form 990 record includes the organization's employer identification number (EIN). We therefore plan to match these organizations to their LEHD data records using their EINs. By matching these organizations to their tax forms, we vastly enrich the set of firm-level financial characteristics available for analysis.

We identify two types of shocks to nonprofits' finances: a positive shock from receipt of big gifts, and a negative shock from exposure to one of the largest Ponzi schemes in recent memory. The data on big gifts comes from the list of America's top 50 to 60 donors, produced by The Chronicle of Philanthropy and Slate Magazine on an annual basis. This dataset spans the years 1998 through 2016. This list was compiled by Mayo (2019) into a dataset of 218 big gifts, matched to 501(c)(3) nonprofits observed in the National Center for Charitable Statistics (NCCS) data.

Data on negative shocks will come from a combination of the publicly-available list of Bernie Madoff's victims, the Foundation Center's Foundation Directory Online, the aforementioned IRS Form 990 data, and DonorSearch data. The Madoff victims list comes from the filing made in the U.S. Bankruptcy Court in Manhattan by the firm AlixPartners (Chew (2009)), which lists names and addresses of victims. These victims include both individuals and institutions, such as foundations, trust funds, or pension funds. In the cases where victims are foundations, we plan to extract the name and address of the foundation and use

this information to search for the corresponding entry in a database of private foundations, such as the Foundation Center's Foundation Directory Online (FDO). The FDO maps each foundation to its recipient organizations; this mapping is available for most foundations beginning in 2004. We plan to leverage this crosswalk and then map the recipient organizations to their EINs based on their names. The NCCS data contain both EINs and names. In the cases when victims are individuals, we plan to extract the name and address of the individual and use this information both to link to any foundations that may bear the victim's name, and to search for the corresponding entry in a database of individual donations, such as the one provided by the company DonorSearch. The DonorSearch data derive mappings between individual charitable donors and recipient organizations from nonprofits' annual reports. These data are available from 2007 to the present. Having used this crosswalk between individual donors and their beneficiaries, we will once again use the names of the beneficiaries in conjunction with the NCCS data to identify the EINs of organizations which have been exposed to the Madoff scandal. The magnitude of the exposure can also be calculated, as the FDO data include the amount of foundations' grants to beneficiaries and the DonorSearch data include the amount of individual donors' gifts. The NCCS data include the amount of donations each nonprofit has received in a given year, as well as their total program service expenditures, total revenues, and total salaries. By calculating victims' pre-Madoff contributions to the beneficiary organization as a share of total donations (or program service expenditures, or revenues, or salaries), we are able to construct a measure of each nonprofit's exposure to the Madoff scandal.

Data on student loans will come from the NLSY97, which follows a sample of 8,984 individuals born between 1980 and 1984 through 18 rounds of annual follow-up surveys, beginning in 1997. Respondents, who primarily graduated college between 2004 – 2007, were asked about the amount of federal and non-federal student aid taken in order to finance their education; in subsequent years, they were also asked about the amount of these loans which they still owe. We plan to conduct a probabilistic match between the public-use NLSY97 data and the LEHD data, using the NLSY97 respondents' age, gender, highest level of educational achievement, race/ethnicity, employer industry, quarter of transition into and out of a particular job, Census region, employer size, and income. We will also use this link to retrieve information on students' fields of study from the NLSY97, which we view as an important conditioning variable in our regression of sectoral choice on loan amounts. This information will be supplemented with the PSID, which began asking respondents about their non-collateralized student debt in 2009. We will use the same probabilistic matching technique to link the PSID and LEHD data. We will also consider supplementing the data from NLSY97 with data from other public-use educational surveys, such as the Baccalaureate and Beyond Longitudinal Study or the Beginning Postsecondary Students Longitudinal Study, using similar probabilistic match techniques to facilitate the match to the LEHD.

Note: The SAP requests that researchers enter information individually for each dataset with a character limit of 200.

Addendum 7: Methodology

Methodology from Approved Census Proposals - Expected length: 5-10 pages

Demographic Proposal - Methodology

I will examine some public ECE programs collectively and others individually—sample sizes permitting and ensuring geographic areas of small populations will not be an issue. A non- exhaustive list of states with programs I might examine include Arizona, Arkansas California, Connecticut, Florida, Delaware, Georgia, Hawaii, Illinois, Iowa, Kansas, Kentucky, Louisiana, Michigan, New Jersey, New York, North Carolina, Ohio, Oklahoma, Oregon, South Carolina, Texas, Vermont, Virginia, Washington, West Virginia, Wisconsin, and the District of Columbia.

For ease of exposition, I will use Georgia as a focal state for the remainder of this document. While the exact methods necessary for each state or collection of states might vary slightly, the broad strokes should be very similar, and Georgia offers an illustrative example.

1.1 Defining Exposure to ECE

One of the major limitations of all the datasets I will use in this project is that I cannot link adults to their childhood characteristics. The most important implication of this limitation is that I cannot observe which adults participated in ECE as children. Therefore, I will examine the long-run outcomes of those *exposed* to ECE. Importantly, a large fraction of those exposed to ECE in Georgia participated.

I will approximate ECE exposure by sorting individuals into (state of birth) × (year of birth + 4) pre-K cohorts. All adults born in Georgia beginning with the 1995 pre-K cohort will be considered exposed. As an example, consider an individual observed at age 30 in the 2014 ACS. This person would be in the 1988 pre-K cohort because they would have been four years old in that year. Regardless of their state of birth, they would not have been exposed to Georgia ECE. Now consider another individual, also observed in the 2014 ACS, but at age 20. This person would be in the 1998 pre-K cohort. If they were born in Georgia, they would have been exposed to Georgia ECE.

My definition of exposure might not perfectly capture the population of interest since it's based on state of birth rather than state of residence at age four. In practice, however, the two definitions would likely produce similar results since 86 percent of children born in Georgia reside there at age four.[1]

1.2 Primary Analysis: Difference-in-Differences

In my primary analysis, I will estimate the long-run effects of ECE using a difference-in- differences (DiD) approach. The first "difference" is a comparison of people born in Georgia who turned four before and after the introduction of ECE. The second "difference" is a comparison of people born in Georgia and people born in states without ECE. Since the treatment is ECE exposure, this framework estimates intent-to-treat (ITT) effects.

Letting Y_{ict} denote an adult outcome for individual i from county of birth c in pre-K cohort t, my primary model specification will be of the form:

$$Y_{ict} = \beta_0 + \beta_1 ECE_i + \theta \boldsymbol{X_{ict}} + \alpha_c + \gamma_t + \epsilon_{ict} \quad (1)$$

[1] I calculate out-migration from Georgia using a 1-in-20 national sample from the 2000 Census.

where ECE_i is a binary indicator for being exposed to ECE in Georgia, α_c is a vector of county of birth fixed effects, and γ_t is a vector of pre-K cohort fixed effects. [2] $\boldsymbol{X_{ict}}$ is a vector that will control for individual-level characteristics including age, race/ethnicity, sex, and month of birth, as well as county of birth characteristics. The ITT effect of interest is β_1. I will estimate Equation 1 using data on several exposed and unexposed cohorts.

Following Zerpa (2021), I will limit my primary control group to the 18 states that had no more than a very small statewide pre-K program during the time of this analysis.[3] This restriction yields a cleaner comparison group and interpretation of β_1. I will also estimate models that use various other subsets of states as the control group to assess robustness.

Inference in this setting is tricky because my sample has too few states for reliable use of most asymptotic methods. With only 19 clusters (i.e., states), the standard cluster-robust variance estimator would not perform well. Moreover, some inference methods that perform well with a small number of clusters, like the Wild cluster bootstrap, would not perform well here because there is only one treated state (i.e., Georgia). Fortunately, there is a developing literature on methods that are more appropriate for my setting (e.g., Conley and Taber, 2011; Ferman and Pinto, 2019). I will conduct inference using one (or more) of these methods.

1.3 Secondary Analysis: Synthetic Control Method

I will also conduct analyses that uses the synthetic control method (SCM). SCM constructs a counterfactual for Georgia by taking a weighted average of "donor pool" states (Abadie et al., 2010). By construction, "synthetic Georgia" should closely approximate Georgia in the pre- treatment period. Differences in outcomes between Georgia and synthetic Georgia in the
treatment period can be interpreted as ECE's ITT effects. This research design is useful for investigating situations where there are concerns about parallel trends. It's also a natural setting for conducting permutation inference.

SCM is more credible with a longer pre-treatment period, so I will include more pre- treatment cohorts in this analysis than in the DiD analysis. I will again use the 18 states identified by Zerpa (2021) as the primary donor pool, but I may use other subsets of states in robustness checks. I will use the same data as in the DiD analysis, but aggregated to the (state of birth) × (year of birth + 4) level.

I will use a variation of the traditional SCM estimator to calculate ECE's effects. First, I will demean each state's outcomes using pre-treatment means. Then, I will apply the traditional SCM estimator to the demeaned data. Demeaning forces SCM to match on trends rather than levels, as with traditional DiD (Ferman and Pinto, 2021; Doudchenko and Imbens, 2016). This is important in my setting because Georgia's educational outcomes tend to be at the bottom of the donor pool distribution. I will evaluate robustness by estimating models with a variety of specifications.

I will conduct inference in the SCM analysis using a permutation method, as proposed by Abadie et al. (2010). This is essentially a placebo falsification test. This type of inference would complement the regression-based inference in the primary analysis. In the DiD analysis, inference will quantify uncertainty in the parameter estimate due to sampling variability. In the SCM analysis, inference will quantify uncertainty in the parameter estimate due to the possibility

[2] Note that the issues sometimes associated with two-way fixed effects models are not present in my setting because treatment adoption is not staggered, the control units are never treated, and treatment never turns off.

[3] These states are Alabama, Alaska, Arizona, Hawaii, Idaho, Indiana, Iowa, Minnesota, Mississippi, Montana, Nevada, New Hampshire, North Dakota, Rhode Island, South Dakota, Utah, Washington, and Wyoming.

of chance effects. These are different thought experiments. Conducting inference both ways will strengthen the causal interpretation of any estimated effects.

1.4 Identifying Assumptions

Some assumptions are required to identify ITT effects. In the DiD analysis, the key assumption is that Georgia and the control group would have followed parallel trends in the absence of ECE. I will evaluate this assumption using dynamic event study models (again using an appropriate method for inference). In the SCM analysis, the analogous assumption is that synthetic Georgia correctly shows how outcomes would have evolved in real Georgia in the absence of ECE. I will evaluate this assumption by examining how well synthetic Georgia matches real Georgia in the pre-treatment period.

Another assumption is that no other shocks occurred in Georgia (or control states) in between ECE exposure and the measurement of adult outcomes. This would conflate the effects of ECE with the effects of the shocks. To my knowledge, there are no shocks large enough to threaten a causal interpretation of the results. Note that Cascio and Schanzenbach (2013) make the same assumption to estimate the impact of ECE in Georgia on eighth-grade test scores (although the risk of conflation is smaller in their case because they examine medium-run outcomes).

1.5 Outcomes of Interest

The Census data will allow me to estimate the impact of ECE in Georgia on a rich set of adult outcomes. I plan to estimate impacts on all, or some subset, of the following: educational outcomes, such as high school graduation, college enrollment, associate degree attainment, and bachelor's degree attainment; labor market outcomes, such as employment, earnings, and having a professional job; crime outcomes, ranging from misdemeanors to incarceration; health outcomes, such as disability and death; receipt of government benefits, such as SNAP, TANF, SSI, and tenant rental assistance; geographic mobility; and life cycle outcomes, such as marriage and fertility.

1.6 Treatment Effect Heterogeneity

A key difference between my study and many in the literature is that I will examine ECE programs that serve a diverse population across large states. Therefore, an important contribution of mine will be examining treatment effect heterogeneity. Sample sizes permitting, I plan to estimate effects separately by sex, race/ethnicity, and county characteristics. Estimating differential effects for those born in rural vs. urban counties, or rich vs. poor counties, will be particularly informative since estimating such effects is not possible in studies that examine a single city (e.g., Gray-Lobe et al., 2023).

1.7 Estimating Effects on the Treated

The analyses outlined above can only estimate ECE's ITT effects. I will provide suggestive evidence of the effects on the treated (ATETs) by estimating how many children were induced to participate in ECE. There are a few research designs I might use to do this. The first would be a DiD/SCM approach analogous to my primary and secondary analyses. The second would be an

RD approach that utilizes plausibly exogenous variation in ECE eligibility among children born around a birthday eligibility cutoff. Other papers have used these strategies for this purpose, but the larger sample size afforded to me by combining the CPS, SIPP, and ACS (linked to the NUMIDENT) would give me the statistical power necessary for more precise estimates (Cascio and Schanzenbach, 2013; Fitzpatrick, 2010). A third approach would be to combine the first two in a "difference-in-discontinuities" analysis.

1.8 Mechanism Analysis

With ECE, the short- and medium-run effects on family members are some of the most important to understand for contextualizing long-run effects. To shed light on these mechanisms, I will examine the effects of ECE on parental outcomes, including employment, earnings, fertility, and educational enrollment and attainment. I will also analyze how these effects vary by the number/age of children in the household and by the availability and price of alternative childcare options. Outside of parental outcomes, I will examine how the care arrangements of a focal child and of their siblings change when the focal child participates in public ECE. I will perform these analyses using CPS, SIPP, ACS, and SSA data (linked to the NUMIDENT) and with one of the three research designs mentioned in Section 1.7 (i.e., DiD/SCM, RD, or differences-in-discontinuities). For mechanism analyses, I may examine ECE programs that are too recent for participants to have long-run outcomes. Doing so would increase statistical precision and allow me to examine mechanisms as they operate in modern-day ECE landscapes.

1.9 Identifying Shortcomings of Current Census Data Collection

As previously mentioned, one of the ways my project will benefit the Census Bureau is by identifying "shortcomings of current data collection" (Criterion 9). The first shortcoming my project will highlight is the inadequacy of collecting only point-in-time characteristics for evaluating long-run outcomes. For example, the ACS, CPS, and SIPP generally only ask respondents about their earnings and government benefit receipt in the year of the survey (and sometimes the prior year). Responses from one or two years can miss important information on long-run economic well-being if people's circumstances fluctuate over time. The extent of this problem is an open empirical question. To improve our understanding of the problem, I will conduct my analysis (as described above) using two datasets. First, I will use point-in-time outcome data from the ACS, CPS, and SIPP. Then, I will use longitudinal administrative data
from the SER/DER files, Tenant Rental Assistance file, etc. I will assess the importance of using longitudinal data by comparing the results.

The second shortcoming my project will highlight is the inadequacy of existing survey instruments for connecting childhood and family characteristics to long-run outcomes. Currently, cross-sectional Census products ask adults very few questions about their parents or their early lives, making it difficult to prepare estimates of long-run outcomes for subpopulations defined by these characteristics. With the advent of PIKing, it is sometimes possible to obtain childhood and family characteristics by linking children to parents via decennial censuses. This procedure makes it more possible than ever before to connect children to parents, but there are certain conditions that must be met for it to succeed. For one, the procedure requires that children and parents live together in a PIKed decennial census year. Second, it requires that we have
information on parents' characteristics in another Census product, or that an administrative

dataset has enough temporal and geographic coverage to include them. Unfortunately, for many subpopulations and outcomes of interest, these conditions are not met.

In my context, if adults were asked about their ECE participation as children, I would be able to estimate treatment effects on the treated, rather than the exposed. As another example, asking adults about their parents' educational attainment or occupations would allow researchers to estimate long-run effects separately for individuals who grew up in different social classes. To demonstrate the inadequacy of current data, I will document the following:

1) analyses I conduct that would not be possible without restricted data,
2) analyses I could conduct if Census products asked more questions about childhood and family background, and
3) survey questions on childhood and family background that would have been useful for my project.

Economic Proposal - Methodology

Firm-Level Analysis

Our goal is to estimate the effects of supply chain disruptions on firms' financing, investment, and employment decisions. We consider data from 1976 until 2026. We start from 1976 because our supplier data from Compustat starts in 1978 and we need 2 years before 1978 to conduct a pre-trends analysis for our difference-in-difference analysis. Furthermore, the longer time frame will enable us to ensure that our results are not driven by business cycle conditions of a short time period.

To analyze resource allocation decisions and resulting performance, we will first use a firm-level panel dataset. To create this panel, we will aggregate plant-level data in the Census of Manufacturers (CMF), Annual Survey of Manufacturers (ASM), the Longitudinal Business Database (LBD), the Integrated Longitudinal Business Database (iLBD), Quarterly Financial Report (QFRCEN), Commodity Flow Survey (CFS), Research and Development Surveys (RADS), Manufacturers' Shipments, Inventories, and Orders (M3), Quarterly Survey of Plant Capacity Utilization (QPC), Census of Auxiliary Establishments (AUX), Longitudinal Firm Trade Transactions Databse (LFTTD), and the Standard Statistical Establishment List (CBPBR). Each of these data sets are necessary to construct the key variables of interest.

The CMF covers all U.S. manufacturing plants with at least one paid employee and is conducted in "Census years" (years ending with either 2 or 7). The ASM covers a subset of CMF plants (those with greater than 250 employees and a randomly selected subset of smaller plants), and is conducted in non-Census years. The CMF and ASM will be used to provide information about key plant variables, such as industry sector, location, capital expenditures, employment, total assets, and the value of shipments and material inputs.

The LBD is an annual register of all U.S. business establishments (with at least one paid employee) and contains longitudinal establishment identifiers as well as data on employment, payroll, industry sector, location, and corporate affiliation. The longitudinal establishment identifiers in the LBD will be used to construct longitudinal linkages between the CMF and ASM at the plant level. Once the data are linked at the plant level they will then be aggregated to the firm level to construct our firm-level panel data set.

LBD includes only establishments with at least one paid employee. However, non-employer firms constitute an important part of the economy, mainly job creation. For this reason, we will also use iLBD which is an extension of LBD and includes also non-employer firms. The iLBD will be used to identify and analyze non-employer firms on the suppy chain.

The QFRCEN is a quarterly survey of firms that covers several sectors of the U.S. economy: mining, manufacturing, and wholesale and retail trade firms. Since 1982, firms with more than $250 million in book assets are sampled with certainty, whereas firms with between $250K and $250 million in assets are sampled randomly. The QFRCEN is critical to our study for two main reasons. Most importantly, unlike the CMF/ASM, the LBD, and the iLBD, it provides firm-level financial information including detailed information about debt and equity for both public and private firms. The QFRCEN also provides information about accounts receivable and accounts payable, which will help analyze how disruptions to supply chains affect the provision of trade credit.

Using this data, we will construct dependent variables that measure resource utilization: investment, employment and payroll, trade credit, establishment closure, establishment sale, and establishment purchase. We will construct our measure of investment as total capital expenditures divided by total capital stock, where total capital expenditures (stock) is the sum of capital expenditures (stock) across all of the firms' plants. Employment is constructed as the total number of employees (or logarithm of total number of employees) across all plants, and payroll is the average wage per employee.

We will construct firm-level measures of productivity and profitability as follows. Firm-level total factor productivity (TFP) is the capital-weighted average of individual plant-level TFPs across all of the firm's

plants. Plant-level TFP is the residual from estimating a log-linear Cobb-Douglas production function by ordinary least squares (OLS) separately for each 3-digit Standard Industrial Classification (SIC) industry and year (Foster et al., 2008). We will measure firm-level profitability first as the return on capital (ROC) computed as the sum of shipments minus labor and materials costs across all of the firm's plants (i.e., total profits) to total capital stock. We will also use the operating margin (OM) to measure firm-level profitability, computed as total profits to total shipments.

Other variables such as firm size and age will be constructed. Firm size will be computed as the logarithm of the total value of shipments across all plants, while firm age will be computed as the logarithm of one plus the number of years since the plant first appeared in the LBD or iLBD.

To identify suppliers, we will combine three data sets. First, we will use the Commodity Flow Survey (CFS). The CFS collects data on shipments originating from mining, manufacturing, wholesale, and catalog and mail-order retail establishments. Establishments provide information about their outbound shipments during a one-week period, four times per year. CFS data will be used to identify suppliers and their customers' locations. The second dataset that will be used to identify suppliers and customers is the Factset Revere Supply Chain Relationship database. Factset Revere collects relationship information from primary public sources such as SEC 10-K annual filings, investor presentations, and press releases, and classifies them by relationship type (e.g., customer, supplier, competitor, different types of partnerships). We identify supply chain relationships using companies' reported customers and suppliers. The third dataset is Compustat Customer Segment Files from 1978 to 2020. According to the Statement of Financial Accounting Standard (SFAS) rule No.131, firms have to report customers that account for at least 10% of the firm's sales.

To identify disruptions to suppliers' operations, we will use a combination of several data sets. First, we will use the SHELDUS (Spatial Hazard and Loss Database for the United States) database compiled by the Center for Emergency Management and Homeland Security at Arizona State University to identify the date and estimated dollar amount of damages of each natural disaster as well as the FIPS codes of affected counties. Second, we will use the Longitudinal Firm Trade Transactions Database (LFTTD). LFTTD provides firm-level exports and imports data, which will help us identify shocks to global supply chains. Use of the LFTTD will be crucial to our analysis given that most firms rely on global supply chains to reduce costs, which makes shocks to global supply chains, such as the 2011 great East Japan earthquake and the Thailand floods, and more recently the Covid-19 outbreak as disruptive as domestic supply chain shocks. Third, we will use Manufacturers' Shipments, Inventories, and Orders (M3). M3 provides monthly data on manufacturers' value of shipments, new orders, and inventories. We will use M3 to identify firms experiencing a sudden drop in their shipments. We will use this measure as a proxy for disruptions to suppliers' operations. Finally, we will use the Quarterly Survey of Plant Capacity Utilization (QPC) providing capacity utilization rates for U.S. manufacturing plants. We hypothesize that firms experiencing a disruption in their operations will decrease their capacity utilization, which will help us identify disruptions to suppliers' operations in a cleaner way.

Additional firm-level control variables for this firm-level analysis include leverage and profits. Leverage is defined as the sum of long-term and short-term debt, scaled by total assets. Profits are defined as income before extraordinary expenses, divided by total assets. These additional control variables are defined at the firm level using Compustat data and and will be linked to the LBD, iLBD, and CMF/ASM via the Compustat GVKEY identifier.

The effect of supply chain disruptions on firm-level dependent variables will be analyzed by estimating the following regression model:

$$\Delta \boldsymbol{y_{j,t,t+1}} = \boldsymbol{\alpha_j} + \boldsymbol{\alpha_t} + \boldsymbol{\beta_1 DISRUPT_{j,t}} + \boldsymbol{\gamma' X_{j,t}} + \boldsymbol{\varepsilon_{j,t}} \quad (1)$$

Where *j* indexes firm, *t* indexes year, α is a fixed effect (firm and year), *DISRUPT* is a dummy for whether at least one of the suppliers of firm *j* experiences a disruption in operations due to a disaster in year *t*, and *X* is a vector of firm-specific control variables. The dependent variable, the change in *y*, will be the change in one of the outcome variables discussed in Section II.A.1 (Trade Credit, Investment, Employment etc.). Throughout the analysis, we control for firm characteristics including size, leverage, age, and profits. We also

absorb unobserved heterogeneity by fixed effects. The inclusion of these fixed effects allows us to explore whether the behavior of firms in counties affected by natural disasters changes in the year following the disaster in comparison to other firms in the same year. We cluster standard errors at the firm level, which is particularly important because our dataset includes overlapping years.

The estimated coefficient β_1 indicates whether disruption of suppliers' operations correlates with firm performance and resource allocation after controlling for other relevant characteristics. The estimated economic magnitude of the coefficient as well its statistical significance (as indicated by the estimated t-statistic on the coefficient) will be of particular interest.

The data will cover the period between 1976 and 2018. We start from 1976 because our supplier data from Compustat starts in 1978 and we need 2 years before 1978 to conduct a pre-trends analysis for our difference-in-difference analysis.

After finding the estimated change in trade credit, investment, and employment, we will study whether change in these variables is related to cross-sectional industry and firm characteristics. The first characteristic we will examine is product market fluidity, introduced by Hoberg, Phillips, and Prabhala (2014), from the Hoberg-Phillips Data Library. This proxy assesses the degree of competitive threat and product market change surrounding a firm. We use it to evaluate whether facing different competitive threats results in different reactions to supply chain disruptions. The second characteristic we will analyze is input specificity. Barrot and Sauvagnat (2016) argue that effects of supply chain disruptions will be greater for specific inputs since it will be harder to find and switch to new suppliers when inputs are specific. To test this hypothesis, we will use R&D expenditures as a measure of input specificity. For R&D expenditures, we will rely on the Research and Development Surveys (RADS).[1] These surveys collected by the Census Bureau provide annual firm-level data on R&D expenditures, which will be critical to our cross-sectional analysis.

We will then estimate effects in cross-sectional firm data as follows:

$$\Delta \boldsymbol{y_{j,t,t+1}} = \boldsymbol{\alpha_j} + \boldsymbol{\alpha_t} + \boldsymbol{\beta_1 DISRUPT_{j,t}} + \boldsymbol{\beta_2 DISRUPT_{j,t} * CS} + \boldsymbol{\gamma' X_{j,t}} + \boldsymbol{\varepsilon_{jt}} \quad (2)$$

Where each of the variables and subscripts are the same as in equation (1), except for the inclusion of *CS,* which represents the cross-sectional source of variation including the product market fluidity measure or R&D expenditures of the supplier whose operations are disrupted. The estimated coefficient of interest, β_2, indicates whether product market competition or supplier's input specificity affects the customer firm's trade credit, investment, and employment policy.

Plant-Level Analysis

The plant-level analysis will use the disaggregated version of the firm-level panel described above and will proceed in two steps. First, we will generate estimates of the effect of supply chain disruptions on plant-level productivity, investment, employment, and payroll. In addition, we will consider plant closures, sales, and purchases. Second, we will examine the cross-section of plants (for a given firm) and ask *which* plants experience changes in performance and resource allocation.

The outcome variables of interest are productivity (TFP), profitability (ROC and OM), investment, employment, and payroll, and are described above in the firm-level analysis. In addition, we will consider plant closures, sales, and purchases. The sample of plant closures includes all sample plants that are coded as "death" in the LBD or iLBD. In addition, we will consider plant ownership changes (asset sales and purchases). This will allow us to see how firms reorganize their operations following disruptions to their suppliers' operations.

We will estimate the following regression model:

$$\Delta \boldsymbol{y_{i,j,t,t+1}} = \boldsymbol{\alpha_i} + \boldsymbol{\alpha_t} + \boldsymbol{\beta_1 DISRUPT_{j,t}} + \boldsymbol{\gamma' X_{j,t}} + \boldsymbol{\varepsilon_{jt}} \quad (3)$$

Where *i* indexes plant, *j* indexes firm, *t* indexes year, α is a fixed effect (plant and year), *DISRUPT* is a dummy for whether at least one of the suppliers of firm *j* experiences a disruption in operations due to a disaster in

[1] Foster et al. (2016) provide an excellent summary and analysis of BRDIS and SIRD.

year *t*, and *X* is a vector of plant-specific control variables (plant age, size, and so on). The dependent variable, change in *y*, will be the change in one of the outcome variables above (investment, employment, etc.). For plant closures, sales, and purchases, we use dummy variables as dependent variables. We cluster standard errors at the firm level, which is particularly important because our dataset includes overlapping years.

The estimated coefficient β_1 indicates whether disruption of suppliers' operations correlates with plant performance, investment, employment, closure, sale, purchase, firm performance, or resource allocation after controlling for other relevant characteristics. The estimated economic magnitude of the coefficient as well their statistical significance (as indicated by the estimated t-statistic on the coefficient) will be of particular interest.

Next, we seek to understand *which* types of assets or plants are sold, bought, closed, and allocated more resources following disruptions of suppliers' operations. To this end, we will define a plant-level indicator variable called *YES (NO)* that will categorize each plant for a given firm. *YES (NO)* is set equal to one (zero) if the attribute in consideration is satisfied by a plant at the beginning of a year t. We will consider two types of attributes—plant industry (input or output) and plant distance to the supplier—measured as follows. First, a plant is considered to be an output industry if it is in the output linkage for the supplier plant whose operations are disrupted. Second, a plant is considered to be close (far) if it is less (more) than 100 miles away from the supplier plant whose operations are disrupted.

We will then estimate effects in cross-sectional plant data as follows:

$$\Delta \boldsymbol{y_{i,j,t,t+1}} = \boldsymbol{\alpha_i} + \boldsymbol{\alpha_t} + \boldsymbol{\beta_1 DISRUPT_{j,t}.YES} + \boldsymbol{\beta_2 DISRUPT_{j,t}.NO} + \boldsymbol{\gamma' X_{j,t}} + \boldsymbol{\varepsilon_{jt}} \quad (4)$$

where each of the variables and subscripts are the same as in equation (3), except for the inclusion of the *YES(NO)* interaction variable, which permits identification of a differential effect of disruptions by plant industry and distance. The estimated difference $\beta_1 - \beta_2$ is of particular interest as it indicates whether there is a differential impact of the supply chain distruptions across the plant types.

Another advantage of the plant-level Census data, CMF and ASM, LBD, and iLBD is that they will allow us to see whether supply chain disruptions incentivize vertical ownership. We can observe the industry of each plant, which will allow us to examine whether firms acquire or start plants in the same industries as the industries of the disrupted suppliers.

Overall, the plant-level analysis will provide a novel exploration of the effect of supply chain disruptions on internal resource allocation and the boundaries of the firm.

II.B. Part #2: Supply Chain Risk, Relationships, and Asset Redeployment

II.B.1 Conceptual Framework and Existing Literature

In this part, we adopt an "ex ante" approach and focus on how firms react to an increase in the *perceived* probability of disruption. Not only a realized disruption of supply chains but also an increasing risk of disruption of supply chains can make firms change their policies (Giglio et al., 2021). Given the large costs that supply chain disruptions impose, an increase in the perceived operating risk that a firm is subject to may determine changes in the way a supply chain operates and can ultimately increase its fragility (Elliott and Golub, 2021). On the one hand, even small shocks hampering the reliability of a supplier may lead firms to sever their relationships. If firms indeed react by severing their existing relationships and looking for new customers and suppliers, changes in the perceived risks of disruption may affect firms' and regions' comparative advantage and long-term growth. On the other hand, firms have incentives to invest in order to increase the reliability of the supply chain and limit any long-term damages (Elliott, Golub, and Leduc, 2020). Yet, we know little of the mechanisms limiting the fragility of supply chains.

We use textual analysis of quarterly earnings conference-call transcripts and SEC current report filings (i.e., 8-Ks) to construct firm-level measures of supply chain risk faced by public firms in the U.S. between 1992 and 2026. The majority of firms listed in the U.S. hold regular earnings conference calls to inform investors and analysts about the firm's performance. They also respond to questions from call participants.

To quantify supply chain risk faced by firms, we combine earnings conference calls transcripts with SEC current report filings and calculate the share of conversations and discussions that center around risks related to supply chains (Hassan et al. 2019). To operationalize this, we train a library of supply chain related words and count the number of times these supply chain related words are used in conjunction with words representing "risk."

Using this measure of supply chain risk, we will first analyze how supply chain risks faced by firms affect their relationships with other suppliers. Firms may increase their multisourcing efforts for key inputs to make sure that disruptions to their suppliers' operations do not cause reductions in their output (Elliott and Golub, 2021). Using the Commodity Flow Survey, we plan to examine whether firms look for new suppliers to mitigate increasing supply chain risks.

Next, we plan to analyze how increasing supply chain risk affects internal resource allocation across establishments. The disaggregated nature of the Census data, the Census of Manufacturers (CMF), Annual Survey of Manufacturers (ASM), the Longitudinal Business Database (LBD), and the Integrated Longitudinal Business Database (iLBD) will allow us to observe all of a firm's plants that differ across several dimensions (location, focus, input-output), which will allow us to see which plants experience investment and employment cutbacks.

Finally, we will analyze how increasing supply chain risk affects merger and acquisition activity, i.e., firm's boundaries. Firms might buy or open new plants operating in input industries to decrease their dependence on suppliers and manage the risk of disruption to their production. The disaggregated plant-level data will allow us to track the boundaries of the firm as well as to see clearly how firms re-organize their operations when faced with incrasing supply chain risks.

II.B.2 Empirical Methodology

In Part #1, we analyze how firms respond to realized supply chain disruptions. However, we do not know much about how firms prepare themselves for such risks, which constitutes the main objective of Part #2. We will analyze what firms do when supply chain-related risks increase.

On the theoretical front, the effect of increasing supply-chain related risks on investment, employment, and asset redeployment is not obvious. On the one hand, increasing risks might lead firms to decrease their investment and production (Hassan et al. 2019). On the other hand, to manage supply chain risks, firms might invest in, buy, or start plants in input industries. In other words, they might increase their presence in upstream production. The Census data has two main advantages. First, at the intensive margin, it enables us to study how firms reallocate resources when faced with supply chain risks because we can see plants and their industry (input or output) characteristics. Second, at the extensive margin, we can analyze whether firms open, close, or sell plants when they think there is an increasing probability that suppliers' production might be disrupted. These two layers of analysis are almost impossible to conduct with conventional firm-level datasets.

To quantify supply chain risks faced by each firm, we do a textual analysis of earnings conference-call transcripts and SEC current report filings between 1992 and 2020. We use earnings call transcripts because the majority of U.S. firms organized earnings conference calls where management gives investors and analysts its view on the firm's past and future performance and responds to questions from call participants. We use the same approach as Hassan et al. (2019) and count the number of times these supply chain related words are used in conjunction with words representing "risk."

We will conduct our analysis using the CMF and the ASM between 1992 and 2026. The CMF and ASM will be used to provide information about key plant variables, such as industry sector, location, capital expenditures, employment, total assets, and the value of shipments and material inputs. We will use the LBD and the iLBD to identify plant openings, closures, and sales.

To analyze the investment and risk-management behavior of firms, we will estimate the following regression:

$$\boldsymbol{y_{i,j,t} = \alpha_i + \alpha_t + \beta_1 SCRISK_{j,t} + \gamma' X_{j,t} + \varepsilon_{jt}} \quad (5)$$

Where *i* indexes plant, *j* indexes firm, *t* indexes year, *α* is a fixed effect (plant and year), *CSRISK* is the year supply chain risk constructed out of earnings conference-call transcripts and SEC current report filings between 1992 and 2020. *X* is a vector of plant-specific control variables (plant age, size, and so on). The dependent variable, *y*, denotes the outcome variables: investment, employment, payroll, and plant opening, closure, and sale. We cluster standard errors at the firm level.

We will then analyze *which* types of plants are sold, bought, closed, and allocated more resources when supply-chain related risks increase. As in Section II.B.1, we will define a plant-level indicator variable called *YES (NO)* that will categorize each plant for a given firm. *YES (NO)* is set equal to one (zero) if the attribute in consideration is satisfied by a plant at the beginning of a year t. We will consider two industry-related types of attributes— (input or output) and (core or peripheral)—measured as follows. First, a plant is considered to be an output industry if its industry is an output industry for the suppliers of the firm. For core vs. peripheral, we follow Maksimovic and Phillips (2002) and, for each firm, classify a three-digit SIC industry as core (peripheral) if its shipments summed across plants is more (less) than 25 percent of the firm's total shipments.

We will then estimate effects in the cross-section of plants as follows:

$$\boldsymbol{y_{i,j,t} = \alpha_i + \alpha_t + \beta_1 SCRISK_{j,t}.YES + \beta_2 SCRISK_{j,t}.NO + \gamma' X_{j,t} + \varepsilon_{jt}} \quad (6)$$

where each of the variables and subscripts are the same as in equation (4), except for the inclusion of the *YES(NO)* interaction variable, which permits a differential effect of supply chain risk by plant industry. The estimated difference $\beta_1 - \beta_2$ is of particular interest as it indicates whether there is a differential impact of the supply chain disruptions across the plant types.

Mixed Proposal

a. Assessing the quality of the data on non-profit organizations and characterizing non-profit employees

The first step in this project is to assess the accuracy of the non-profit data. In order to do this, we will create our own measure of charity versus non-charity employment. This distinction will be based on whether or not the employer files an IRS Form 990. In other words, we would merge Census data with IRS data on tax-exempt organizations, using Employer Identification Numbers (EINs). As such, we will be able to assess the quality of the NAICS/SIC classifications by testing how many times the LEHD and IRS classifications agree. In addition, the IRS Form 990 data include information on the number of employees (and volunteers) at each nonprofit organization, enabling us to assess the consistency between the two data sources.

We will also assess whether missing or seemingly erroneous values in either data source can be imputed or corrected using the other source. To do this, we will examine the use of imputation methods such as those suggested by Peytchev (2012).

Next, we will use the LEHD to document some basic facts about charity employees in the United States. In particular, we are interested in questions such as, what is the age and tenure distribution of charity employees? How does the education level of non-profit workers differ from for-profit workers? What do transition rates look like both within non-profit sectors and across sectors? Do transition rates vary by charity size and age? These questions will mainly be answered using cross tabulations. There is a great deal that can be learned with some basic tabulations and assessing how these statistics change over time. Most of the variables of interest will come directly from the data (e.g., age, education and wages), but others must be constructed. For example, tenure will be calculated using employment start and end dates, and transition rates will be inferred from information on employment spells.

b. Do shocks to a nonprofit's finances or financial prospects have an impact on wage growth or the probability of separation?

While there exists a healthy literature exploring the existence and magnitude of the nonprofit wage differential, this paper would be among the first to examine how the motivation for this wage differential affects the dynamic relationship between nonprofit workers and their firms. Researchers have found evidence of a positive correlation between wages and firm performance on the firm side (Bell and Van Reenen (2012)), but no such evidence exists on the charity side. Indeed, we would be the first to specify a wage-setting model for the nonprofit sector. We plan to develop a principal-agent model of the relationship between the nonprofit manager and the nonprofit employee. The employee would have warm-glow preferences; their outside options would include employment at other nonprofit firms as well as employment at a for-profit firm, from which they receive no warm glow. The manager would set wages subject to the nonprofit's revenues and a non-distribution constraint[2].

Since nonprofit employees may receive a warm glow from their work, the theory of compensating differentials leads us to believe that non-profit managers may respond to negative financial shocks at the

[2] In the for-profit sector, wages are positively correlated with firm performance due to rent sharing. In the non-profit sector, rent sharing is still consistent with the non-distribution constraint, but workers now receive a monotonic transformation of “profits”.

organizational level by lowering wage growth among employees. This intuition is corroborated with recent empirical evidence that nonprofit work is an amenity (Bell (2019)). We intend to test these predictions by gathering information on large exogenous shocks to non-profit finances and tracing out their impact on employees. We are interested in both positive and negative shocks. The positive shocks we will examine include events such as big gifts from private donors (we have information on both lifetime gifts and bequests), or natural disasters that result in a spike in donations. The negative shocks we consider are scandals and loss of major donors. For example, we have information on Bernie Madoff's victims, many of whom were philanthropists or heads of foundations. We anticipate that there may be a significant degree of asymmetry in a charity's response to positive or negative shocks.

Information on organization-level shocks will be matched to the LEHD using EINs. We will examine two main outcomes of interest: earnings growth and the probability that a worker separates from their employer. We propose to use a generalized event-study framework where the identifying assumption is that worker outcomes would evolve similarly in the absence of any shocks, and that there are no contemporaneous changes that affect "treated" and "non-treated" nonprofits differentially.

We would estimate equations such as:

$$Y_{it} = \sum ß_{is} T_{it} \{t = s\} + X'_{it}\gamma + \alpha_i + \lambda_t + \varepsilon_{it}$$

where Y_{it} represents the outcome of interest, T_{it} represents an organization-level shock, X'_{it} represent organization-specific covariates, α_i and λ_t represent charity and time fixed effects respectively, and ε_{it} is the error term.

c. How do government policies affect the decision to work in the non-profit sector and the wages earned?

The first government policy we plan to evaluate is the Public Service Loan Forgiveness (PSLF) program. This program, initiated in 2007, enabled students with William D. Ford Federal Direct Loans to have the remaining balance forgiven if they made "120 qualifying monthly payments under a qualifying repayment plan while working full-time for a qualifying employer," where "qualifying employers" include government organizations, 501(c)(3) nonprofits, and certain non-tax-exempt nonprofit organizations (Federal Student Aid (2019)). We hypothesize that PSLF made it more attractive to work in the non-profit sector, increasing the number of applicants. This would have created downward pressure on wages. However, the fact that PSLF was introduced at the height of the Great Recession, means that there were likely other forces at work. On the one hand, the Great Recession may have resulted in fewer donations (as people had less disposable income), meaning charities had less money to spend on wages. On the other hand, the recession also likely increased demand for charitable services, possibly increasing wages. Which effects dominate is an empirical question.

We plan to use difference-in-differences-type regression models to examine this question. This class of models is used to identify the effect of a law change for a treatment group as compared to a control group[3]. We would run regressions such as:

[3] We would construct two different control groups: (i) Nonprofit employers that do not qualify for the PSLF program, and (ii) employees with no student debt

$$PSLF_{gi} = \alpha_i + \beta\ DirectLoan_i * Post2007_t + \delta_t + X'_{igt}\gamma + \varepsilon_{igt}$$

$$Ln(Y_{igt}) = \alpha_i + \beta\ PSLF_{gi} * Post2007_t + \delta_t + X'_{igt}\gamma + \nu_{gt} + \varepsilon_{igt}$$

where i indexes the individual, g indexes whether the employer qualifies for the PSLF program or whether the employee has student debt, and t indexes time. Post2007 is a binary variable equal to 1 for years after 2007 and 0 before, and $PSLF_{gi}$ is a binary variable equal to 1 if the employer qualifies for the PSLF program. The first specification explores the impact of the PSLF program on the individual's choice of employer. We anticipate that individuals who have larger balances of direct loans will face a greater incentive to enter the charity sector under the PSLF program relative to individuals with smaller direct loan balances, or no direct loans at all. We intend to specify $DirectLoan_i$ initially as a dummy variable for the presence of any direct loans in 2007, and follow up with a heterogeneity analysis in which $DirectLoan_i$ encodes quantiles of the amount of such loans that an individual may have had in 2007.[4] We anticipate that this effect will be non-monotonic in the amount of student debt, as larger educational loans have been shown to induce individuals to choose higher-paying jobs (Rothstein and Rouse (2011)). We are therefore interested in the vector of coefficients represented by β, which yields the difference in the probability that an individual with greater loan balances will choose to work in the charity sector relative to an individual with smaller balances, or no loans at all, after the introduction of the PSLF program relative to before. The first specification is an individual-level regression, where α_i captures an individual-level fixed effect, X'_{igt} include individual-level covariates, such as college major choice, and ε_{igt} is an individual-level error term. We plan to include a cohort-level fixed effect as well, where cohorts are defined relative to the date of labor market entry.

In the second specification, we are interested in three different outcomes (Y_{it}): number of employees, number of volunteers, and wages, where wages would be the average real value of wages at organization i in time t. δ_t represents a year effect, α_i is an organization fixed effect, ν_{gt} captures unobserved sector-time effects, X'_{igt} are organization-specific covariates, and ε_{igt} is the organization-specific error term. Standard errors would be clustered at the organization level. β is the coefficient of interest, capturing the difference in wages/employment between sectors following the introduction of PSLF. In this example of the type of difference-in-difference model we would employ, the identifying assumption would be that employers that are and are not eligible for PSLF have similar trends in the absence of treatment.

[4] Data on direct loans comes from the NLSY97.

Addendum 8: List of References

List of References from Approved Census Proposals

Demographic Proposal

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 2010, *105* (490), 493–505.

Anders, John, Andrew C. Barr, and Alexander Smith, "The Effect of Early Childhood Education on Adult Criminality: Evidence from the 1960s through 1990s," *American Economic Journal: Economic Policy*, 2023, *15* (1), 37-69.

Cascio, Elizabeth and Diane Schanzenbach, "The Impacts of Expanding Access to High-Quality Preschool Education," *Brookings Papers on Economic Activity*, 2013, *44* (2), 127–178.

Conley, Timothy and Christopher Taber, "Inference with 'Difference in Differences' with a Small Number of Policy Changes," *The Review of Economics and Statistics*, 2011, *93* (1), 113–125.

Doudchenko, Nikolay and Guido Imbens, "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Technical Report Working Paper 22791, National Bureau of Economic Research, Cambridge, MA October 2016.

Ferman, Bruno and Cristine Pinto, "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity," *The Review of Economics and Statistics*, 2019, *101* (3), 452–467.

Fitzpatrick, Maria, "Preschoolers Enrolled and Mothers at Work? The Effects of Universal Prekindergarten," *Journal of Labor Economics*, 2010, *28* (1), 51–85.

Gray-Lobe, Guthrie, Parag Pathak, and Christopher Walters, "The Long-Term Effects of Universal Preschool in Boston," *The Quarterly Journal of Economics*, 2023, *138* (1), 363–411.

Henry, Gary, Craig Gordon, and Dana Rickman, "Early Education Policy Alternatives: Comparing Quality and Outcomes of Head Start and State Prekindergarten," *Educational Evaluation and Policy Analysis*, 2006, *28* (1), 77–99.

Weiland, Christina and Hirokazu Yoshikawa, "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills," *Child Development*, 2013, *84* (6), 2112–2130.

Zerpa, Mariana, "Short and Medium Run Impacts of Preschool Education: Evidence from State Pre-K Programs," Technical Report, Mimeo February 2021.

References

Acemoglu, Daron, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, 2012. "The Network Origins of Aggregate Fluctuations," *Econometrica*, 80, 1977–2016.

Amberg, Niklas, Tor Jacobson, Erik von Schedvin, and Richard Townsend, 2021. "Curbing Shocks to Corporate Liquidity: The Role of Trade Credit," *Journal of Political Economy*, 129, 182–242.

Auboin, Marc, 2009. "Boosting the Availability of Trade Finance in the Current Crisis: Background Analysis for a Substantial G20 Package," *CEPR Policy Insight* 35, 1–7.

Barrot, Jean-N., and Julien Sauvagnat, 2016. "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks," *Quarterly Journal of Economics*, 131, 1543–1592.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Ahmed Tahoun, 2021. "Supply chain Disruptions: Evidence from the Great East Japan Earthquake," *Quarterly Journal of Economics*, 136, 1255–1321.

Census Bureau, 2009. "Research Opportunities at the U.S. Census Bureau."

Elliott, Matt, Ben Golub, and Matthew V. Leduc, 2020. "Supply Network Formation and Fragility," Working Paper.

Elliott, Matt, and Ben Golub, 2021. "Networks and Economic Fragility," Working Paper.

Foster, Lucia, John Haltiwanger, and Chad Syverson, 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*, 98, 394-425.

Foster, Lucia, Cheryl Grim, and Nikolas Zolas, 2016. "A Portrait of Firms that Invest in R&D," Working Papers 16-41, Center for Economic Studies, U.S. Census Bureau.

Giannetti, Mariassunta, 2003. "Do Better Institutions Mitigate Agency Problems? Evidence from Corporate Finance Choices," *Journal of Financial and Quantitative Analysis*,38, 185–212.

Giannetti, Mariassunta, Mike Burkart, and Tore Ellingsen, 2011. "What You Sell is What You Lend? Explaining Trade Credit Contracts," *Review of Financial Studies*, 24, 1261–1298.

Giglio, Stefano, Matteo Maggiori, Krishna Rao, Johannes Stroebel, and Andreas Weber, 2021. "Climate Change and Long-Run Discount Rates: Evidence from Real Estate," *Review of Financial Studies*, 34, 3527–3571.

Giroud, Xavier, and Holger M. Mueller, 2019. "Firms' Internal Networks and Local Economic Shocks," *American Economic Review*, 109, 3617–49.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Julien Sauvagnat, 2019. "Firm-Level Political Risk: Measurement and Effects," *Quarterly Journal of Economics*, 134, 2135–2202.

Hines, Paul, Jay Apt, and Sarosh Talukdar, 2008. "Trends in the History of Large Blackouts in the United States," IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 1–8.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala, 2014. "Product Market Threats, Payouts, and Financial Flexibility," *Journal of Finance*, 69, 293–324.

Maksimovic, Vojislav and Gordon Phillips, 2001. “The Market for Corporate Assets: Who Engages in Mergers and Asset Sales and Are There Efficiency Gains?,” *Journal of Finance*, 56, 2019–2065.

Maksimovic, Vojislav and Gordon Phillips, 2002. “Do Conglomerate Firms Allocate Resources Inefficiently Across Industries? Theory and Evidence,” *Journal of Finance*, 57, 721-767.

Petersen, Mitchell and Raghuram G. Rajan, 1997. “Trade Credit: Theories and Evidence,” *Review of Financial Studies*, 10, 661–691.

Rajan, Raghuram and Luigi Zingales, 1995. “What Do we Know about Capital Structure? Some Evidence from International Data,” *Journal of Finance*, 50, 1421–1460.

Wilner, Benjamin S., 2000. “The Exploitation of Relationships in Financial Distress: The Case of Trade Credit,” *Journal of Finance*, 55, 153–178.

References

Bell, A. (2019). "Job Amenities & Earnings Inequality." Working Paper.

Buurman, M., Delfgaauw, J., Dur, R. and Van der Bossche, S. 2012. "Public sector employees: Risk averse and altruistic?" *Journal of Economic Behavior & Organization*, 83: 279-291

Chew, R. 2009. "The Bernie Madoff Client List is Made Public." *TIME Magazine.* Retrieved from http://content.time.com/time/business/article/0,8599,1877414,00.html

Duquette, N. 2016. "Do tax incentives affect charitable contributions? Evidence from public charities' reported revenues." *Journal of Public Economics*, 137: 51-69

Duquette, N. 2019. "Do share-of-income limits on tax-deductibility of charitable contributions affect giving?" *Economics Letters*, 174: 1-4

Federal Student Aid, 2019. "Public service loan forgiveness." Retrieved from https://studentaid.ed.gov/sa/repay-loans/forgiveness-cancellation/public-service

Gregg, P., Grout, P., Ratcliffe, A., Smith, S. and Windmeijer, F. 2011. "How important is pro-social behaviour in the delivery of public services?" *Journal of Public Economics*, 95: 758-776

Houston, D. 2000. "Public-service motivation: a multivariate test." *Journal of Public Administration Research and Theory*, 10(4): 713-727

Houston, D. 2006. "'Walking the walk' of public service motivation: public employees and charitable gifts of time, blood, and money." *Journal of Public Administration Research and Theory*, 16(1): 67-86

The Independent Sector, 2020. "The Charitable Sector." Retrieved from https://independentsector.org/about/the-charitable-sector/

Mayo, J. 2019. "How do big gifts affect rival charities and their donors?" Working Paper

Millard, B. and Machin, A. 2007. "Characteristics of public sector workers." *Economic and Labour Market Review*, 1(5): 45-55

Peytchev, Andy. 2012. "Multiple Imputation for Unit Nonresponse and Measurement Error." *Public Opinion Quarterly*, June, 76(2): pp. 214–237.

Rothstein, J., & Rouse, C. E. (2011). "Constrained after college: Student loans and early-career occupational choices." *Journal of Public Economics*, 95(1–2), 149–163. https://doi.org/10.1016/j.jpubeco.2010.09.015

Addendum 9: Requested Output

Requested Output from Approved Census Proposals - Expected length: 2-15 pages

Demographic Proposal – Requested Output

My primary results in this project will be 1) summary statistics, 2) estimates of ECE's long-run impacts, and 3) estimates of ECE's short- and medium-run impacts. I will obtain these results using individual-level microdata. I am aware of the risks associated with using confidential data, and I will take appropriate steps to ensure that the risk of disclosure is minimal.

The methodology and output described in this proposal pose very little disclosure risk for several reasons. First, I will only report results at levels of aggregation that keep individual-level data safe. Most of my estimates will be at the state- or multi-state-level. When I conduct subgroup analyses, the most granular impact estimates I will report will be disaggregated along one or two of the following dimensions: sex, race/ethnicity, or a county characteristic (e.g., rural/urban). I will never disaggregate impact estimates by more than two of these dimensions at a time. I am aware of the Census policy concerning estimates of Geographic Areas of Small Populations, and I will not report fixed effect estimates for birth counties or other sub-state geographies.

Second, I will work closely with Census Bureau staff to ensure that I follow best practices and that all output passes disclosure review. I will further aggregate or not report any results that pose a disclosure risk.

Economic Proposal – Requested Output

The output for all of the research described in Section II will be model-based, meaning that the bulk of the research output on the real economic consequences of supply chain risks and disruptions would be regression output, i.e., coefficient estimates from the models described above. All estimation will include all observations for which there is data. While the models described above include firm, plant, and year fixed effects, and sometimes MSA and industry fixed effects, their estimated coefficients will not be reported. Presentations would be mostly limited to figures and displays of estimated coefficients and their standard errors in order to minimize disclosure issues. In more detail, the expected output will contain:

- Summary statistics. For each of the parts, the means and standard deviations of the key dependent and independent variables will be shown at the firm- and plant-levels. We will limit the production of summary statistics and only use these to support modeling/regression output described below.
- Regression results from equation (1), a model of firm behavior (investment, employment and trade credit) and performance (productivity, and profitability) in which dummy variables indicating natural disasters and supply chain disruptions as well as firm size, leverage, age, and profits are key independent variables.
- Regression results from equation (2), where independent variables from model (1) will be interacted with firm-level cross sectional measures of product market fluditiy and R&D expenditures.
- Regression results from equations (3) and (4), a model of plant behavior and performance in which dummy variables indicating natural disasters and supply chain disruptions as well as firm size, leverage, age, and profits are key independent variables. These independent variables are interacted with plant classification.
- Regression results from equation (5), a model of plant employment, investment, and redeployment (sale or closure) following increasing supply chain risk. The supply chain risk measure constructed through a textual analysis of conference-call transcripts and SEC current report filings (i.e., 8-Ks) will

be the main independent variable. In addition, plant size and plant age will be the additional independent variables.

- Regression results from equation (6), a model of plant employment, investment, and redeployment (sale or closure) following increasing supply chain risk, which will be interacted with an indicator variable (*YES*) that is turned on if the plant operates in a core industry for the seller or if the plant operates in an output industry for the firm.

In addition, all output must pass disclosure review by the RDC Administrator and/or the Disclosure Review Officer. To facilitate this process, we are willing to exclude results in cases where disclosure concerns may arise.

Mixed Proposal – Requested Output

The main output of the project will be estimated coefficients of regression models and summary statistics. No individual microdata will be disclosed. Furthermore, any geographic or time fixed effects will only be used as control variables and will not be reported. This minimizes the risk of inadvertent disclosure in either the summary statistics or the regression results. Summary statistics will be presented in the form of tabulations and graphs. We understand that any tables we produce must have a large enough sample size and not contain any information that may risk revealing an employee or employer's identity. The types of statistics we are interested in include means, medians (and other quantiles) and density functions. We hope to examine these at various levels of disaggregation e.g. by age, geography, sector (NTEE classification) and education level.

In addition, all output must pass disclosure review by the RDC Administrator. In order to facilitate this process, our results may be further aggregated or not reported in cases where disclosure concerns may arise.

Addendum 10: Census Benefits

Census Benefits from Approved Census Proposals – Character limit 10,000 per benefit

Demographic Proposal – Census Benefits

Identifying further data needs

Criterion 9: Identifying shortcomings of current data collection programs and / or documenting new data collection needs.

The first shortcoming of current data collection my project will highlight is the inadequacy of collecting only point-in-time characteristics for evaluating long-run outcomes. For example, the American Community Survey (ACS), Current Population Survey (CPS), and Survey of Income and Program Participation (SIPP) generally only ask respondents about their earnings and government benefit receipt in the year of the survey (and sometimes the prior year). Responses from one or two years can miss important information on long-run economic well-being if people's circumstances fluctuate over time. The extent of this problem is an open empirical question. To improve our understanding of the problem, I will conduct my analysis using two datasets. First, I will use point-in-time outcome data from the ACS, CPS, and SIPP. Then, I will use longitudinal administrative data from the SSA Detailed Earnings Record (DER), the SSA Summary Earnings Record (SER), the Tenant Rental Assistance files, and more. I will assess the importance of using longitudinal data by comparing the results.

The second shortcoming my project will highlight is the inadequacy of existing survey instruments for connecting childhood and family characteristics to long-run outcomes. Currently, cross-sectional Census products ask adults very few questions about their parents or their early lives, making it difficult to prepare estimates of long-run outcomes for subpopulations defined by these characteristics. Although PIKing sometimes allows one to link children to parents, it remains a challenge in many cases. In my context, if adults were asked about their ECE participation as children, I would be able to estimate treatment effects on the treated, rather than the exposed. As another example, asking adults about their parents' educational attainment or occupations would allow researchers to estimate long-run effects separately for individuals who grew up in different social classes. I will detail my findings on both shortcomings in a Technical Memo to the Census Bureau. Regarding the first shortcoming, I will show the extent to which the long-run outcomes of children exposed to public ECE differ when I use point-in-time and longitudinal data. This information will benefit the Census Bureau as it decides which questions to include in surveys and which data products to make publicly available.

Regarding the second shortcoming, I will document 1) the analyses I conduct that would not be possible without restricted data, 2) the analyses I could conduct if Census products asked more questions about childhood and family background, and 3) the survey questions on childhood and family background that would have been useful for my project. Again, this information will inform the Census Bureau as it plans for future surveys and data collection efforts.

Describing population

Criterion 11: Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5.

My project will benefit the Census Bureau under Criterion 11 by analyzing the long-run outcomes of individuals exposed to public ECE. The share of children in the United States that participate in ECE has become quite large, but despite this growth, estimates of long-run outcomes are rare since ECE participants have only recently become old enough for their long- run outcomes to have materialized. My project will address this gap in the literature. My project is highly related to several papers in the Center for Economic Studies (CES) Working Paper Series. In the past two years alone, three papers have been released that examine the adult outcomes of subpopulations defined by educational experiences. Working Paper Number CES- 21-19 (Foote, 2021) compares the earnings of Title IV graduates and all graduates; Working Paper Number CES-22-11 (Anstreicher, Fletcher, and Thompson, 2022) examines educational and employment outcomes for those who were and weren't exposed to court-ordered desegregation as children; and Working Paper Number CES-21-14 (Davison et al., 2021) examines a holistic set of adult outcomes for students who were and weren't suspended in high school. The Census Bureau has also supported research on public ECE in the past. Working Paper Number CES-08-04 (Fitzpatrick, 2008) examined differences in employment between mothers whose children were and weren't eligible for ECE in Georgia, Florida, and Oklahoma. My project would extend Fitzpatrick's by examining the outcomes of children after a long time horizon.

To isolate differences in outcomes between subpopulations exposed to ECE and subpopulations not exposed, I will use a difference-in-differences framework and the synthetic control method. This analysis will require several Census datasets that are not publicly available. My core sample will come from the ACS, CPS, and SIPP. There are two primary reasons I need access to the restricted versions of these datasets. The first is that no survey on its own is large enough for me to conduct subgroup analyses. The restricted version of the ACS contains a larger sample size, and as I will discuss below, I cannot use the CPS or the SIPP at all without linking them to the NUMIDENT. The second reason is that the restricted versions can be linked to individuals' counties of birth (via the NUMIDENT) and to other datasets that contain longitudinal accounts of the outcomes I plan to examine.

I will assign adults to counties of birth and pre-K cohorts by linking the core datasets to the NUMIDENT file. Having county of birth is crucial for three reasons. First, with only public data, I would have extremely limited knowledge of individuals' pre-treatment characteristics. With county of birth, I could merge on county-level characteristics (rural/urban, median income, school district spending per pupil, etc.) and control for them in my analyses. Second, without county of birth characteristics, I could not sufficiently assess balance between states with and without public ECE. If I find that the two groups are unbalanced on important dimensions, I might employ a re-weighting strategy or construct a control group sample matched on
individual- and county-level characteristics. Third, without county of birth from the NUMIDENT I would be unable to use the CPS and SIPP since they do not contain the information on birthplace required to define ECE exposure. Finally, I also need the NUMIDENT file for exact birthdate. Having exact birthdate will allow me to accurately categorize people into

pre-K cohorts, and it will serve as the running variable in regression discontinuity analyses that exploit birthday eligibility cutoffs.

The CPS School Enrollment Supplement is necessary for understanding ECE enrollment over time. The monthly CPS files contain educational attainment, but not enrollment. Information on enrollment by cohort and state is essential for converting intent-to-treat effects into treatment effects on the treated. More generally, understanding ECE enrollment patterns is essential for giving context to long-run outcomes. The Master Address Files, the Household Composition Key, and the Decennial Censuses are necessary for linking the adults in my sample to their parents. These datasets contain information on shared residency between parents and their children that can facilitate these links. This linking is important to my project for three reasons:

1) it would provide an expanded range of childhood controls, 2) it would allow me to estimate impacts on educational and economic mobility across generations, and 3) it would allow me to estimate treatment effect heterogeneity based on characteristics like parental education or family income. I will also use the Master Address File for assigning individuals to county of residence so that I can merge on local demographic and economic conditions. I will use these measures to examine how the short- and medium-run effects of ECE vary with local context, mediating any long-run outcomes.

The remaining datasets I request are for obtaining long-run outcome data. I request many datasets so that I may examine a large set of adult outcomes. I will use the ACS, CPS, and SIPP for educational attainment, marriage, and fertility; the SSA SER and DER for employment and earnings; the Criminal Justice Administrative Records System (CJARS) for criminal behavior; the CPS March Supplement and Food Security Supplement for receipt of government benefit programs; the Tenant Rental Assistance file for housing assistance; the CPS Fertility Supplement for teen pregnancy and other fertility outcomes; and the CPS and SIPP Supplemental Security (SSR) Extracts for participation in the Supplemental Security Income program. Importantly, the administrative datasets provide longitudinal accounts of the outcomes of interests, which is critical for tracking long-run outcomes. Publicly available Census data would require that I exclude anyone observed at too young an age (which would reduce precision), and it would force me to estimate point-in-time benefit receipt.

Economic Proposal – Census Benefits

Criterion 5: Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census, or estimate.

This project will contribute to the ongoing analysis of the quality of Census data on two dimensions: careful comparisons with internal and external data. The first dimension will consist of comparing internal Census datasets with each other. First, the Quarterly Financial Report Census Years (QFRCEN) will be compared with other internal Census data. Internal Census personnel have reported that the QFRCEN is a little used data set, and thus comparing it with other Census datasets and providing information on its quality would constitute a benefit to the Census. An anonymous reviewer from the Center for Economic Studies confirmed that "there is a potential for a strong benefit to the Census Bureau from utilizing the QFR." To achieve part #1, this project will first merge the QFRCEN with the Economic Census to provide a check on data values. The QFRCEN contains data items such as the total value of shipments appearing in the Census of Manufactures (CMF) and Annual Survey of Manufactures (ASM), which will help this project make a direct comparison between the QFRCEN and the Economic Census and write a technical memorandum to the Center for Economic Studies staff documenting the match and subsequent quality assessment of data values. Furthermore, the project will improve the quality of firm-level identifiers in the QFRCEN by linking firms across different quarters and surveys. To do so, researchers will use the employer identification number (EIN) of firms along with matches based on firm name and location of firm headquarters.

Second, to achieve both parts, this project will carefully compare the Census data with external sources. Given the nature of the Compustat data, this comparison will be at the firm level only, but will nevertheless be informative and may potentially be used to improve LBD estimates. Particular attention will be paid to variables in both the LBD and Compustat: employment, investment, payroll, and industry classification. Statistics of the distributions of these variables will be explored to determine how similar they are. Another important variable in Compustat that will be used for comparison is the location of headquarter. For this comparison, we will do the following two things: First, LBD will be supplemented with Census of Auxiliary Establishment (AUX). The AUX contains information on non-production (''auxiliary'') establishments, including information on headquarters. Combining LBD and AUX will help the authors identify the headquarter establishments and compare it with the location of headquarters in Compustat. Second, the authors will use the Management and Organizational Practices Survey (MOPS). One related question in MOPS is whether the establishment is co-located with the firm's headquarters. The researchers will use the answer to this question to identify the location of headquarters as well provide information on the quality of the data.

Furthermore, to achieve part #1, the authors will use the QFRCEN for comparison purposes with external data. Many of the balance sheet and income statement items in the QFRCEN, including accounts receivable and accounts payable, are collected by Compustat too. The project will provide a direct comparison between these two data sets, which will help improve the quality of the QFRCEN estimates. Given that QFRCEN is a little used data set, this comparison would constitute a benefit to the Census Bureau in terms of providing information about the quality of the dataset.

Finally, to achieve parts #1 and #2, the project will compare supply chain information in the Commodity Flow Survey (CFS) with the supply chain information in Compustat and Factset Revere Databases. CFS constitutes the main source of supply chain information produced by the Census Bureau. The project will compare CFS with two other external sources of supply chain information: Compustat and Factset Revere Database. First, Compustat provides Customer Segment Files from 1978 to 2020. According to the Statement of Financial Accounting Standard (SFAS) rule No.131, firms must report customers that account

for at least 10% of the firm's sales. Shipment data in the CFS will be compared with the sales data in Compustat, which will benefit the Census Bureau in terms of providing information about the quality of the CFS dataset. Second, Factset Revere collects relationship information from primary public sources such as SEC 10-K annual filings, investor presentations, and press releases, and classifies them through relationship types (e.g., customer, supplier, competitor, different types of partnerships). The authors will identify supply chain relationships using companies' reported customers and suppliers between 2002 and 2020. The project will provide a direct comparison between CFS and these two data sets, which will help improve the quality of the CFS estimates.

Another benefit of matching Census data with Factset Revere is related to the quality of firm identifiers. Factset details all organizational changes as well as legal subsidiaries created by firms that have EINs but no employees. This will permit identification of the timing and nature of changes in organization. Merging Factset Revere may first provide some additional benefits to the Business Register through additional information which may inform EIN/status changes. The results of these comparisons will be reported to CES staff through a technical memorandum. Second, in addition to EIN matching, we will match based on business names, which we believe has the potential to increase the match rate of EIN based matching algorithms. The authors will document whether this approach increases the match rate though a technical memorandum. Two anonymous reviewers from the Center for Economic Studies confirmed that improving the firm identifier would be a strong benefit.

Any discrepancies between databases will be investigated and reported to the relevant Census Bureau staff. The resulting datasets from these exercises will be kept on the Census RDC secure server for Census Bureau use. This project will summarize its findings in an internal technical memorandum documenting the quality of firm identifiers, adding to the efforts of the Census Bureau and of previous researchers that addressed these issues in the past.

Criterion 11: Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5.

This project will provide estimates of both firm- and establishment-level investment, employment, financing, productivity, and birth and death. The Census Bureau has an ongoing interest in understanding economy-wide establishment dynamics (formation, closure, growth, contraction, and performance) and has dedicated significant efforts towards understanding how these dynamics vary across sectors and which firm--level factors are influential. This project will provide estimates of the effects of an actual disruption as well as an increasing probability of disruption to suppliers' operations on customer firms' investment, employment, financing, birth and death, and performance across metropolitan statistical areas. Both public and private supplier and customer firms will be examined. In doing so, these estimates of investment, employment, and productivity by firm, metropolitan area, and industry, will contribute to the Census Bureau's understanding of: 1) the role of geography in economic modeling, 2) the role of entrepreneurial and small size enterprises in the economy.

This project consists of two parts. Part #1 estimates how disruptions of suppliers' production affect customers' investment, employment, and asset redeployment decisions as well as their productivity. By using natural disasters as shocks to suppliers' operations, the first model of part #1 will focus on investment, employment, payroll, and firm exit as well as firm profitability and productivity. The researchers will use 4 datasets to identify suppliers whose operations were disrupted. First, The Commodity Flow Survey (CFS) will be used to identify the locations of suppliers. Second, the Longitudinal Firm Trade Transactions Database (LFTTD) providing firm-level exports and imports data will be used to identify firms that experienced disruptions in their global supply chains. LFTTD will be crucial to the

analysis since most firms rely on global supply chains, which makes shocks to global supply chains, such as the 2011 East Japan earthquake and the Covid-19 outbreak as disruptive as domestic supply chain shocks. Third, the Manufacturers' Shipments, Inventories, and Orders (M3) will be used identify firms experiencing a sudden drop in their shipments. Finally, the Quarterly Survey of Plant Capacity Utilization (QPC) providing capacity utilization rates for U.S. manufacturing plants will be used to identify firms experiencing disruptions in their capacity utilization due to disasters. These three datasets, CFS, M3, and QPC, will make feasible the identification of smaller-sized and younger suppliers, in addition to publicly traded U.S. firms. Therefore, this project will be consistent with the Census Bureau's interest, as stated in the *Research Opportunities at the U.S. Census Bureau* document, in studying the role of small and young businesses in the U.S. economy.

Once suppliers whose operations are disrupted are identified, the relation between firm economic performance and actual disruption of suppliers' operations will be documented. Next, the analysis will use plant-level data and will focus on internal restructuring and asset redeployment among firms whose suppliers' operations were disrupted. First, part #1 will examine at which plants investment and employment cutbacks take place. Second, part #1 will examine which plants are more likely to be sold following disruptions to suppliers' operations. To understand which firm-level characteristics are important to understand the effect of supply chain disruptions, the authors will use the Research and Development Surveys (RADS) providing firms' R&D expenditures. The authors will try to understand whether firms who conduct more R&D show a greater reaction when their supply chains are disrupted. This analysis will benefit the Census Bureau's efforts towards understanding which sector- and firm-level characteristics are influential in driving economy-wide firm and establishment dynamics.

Part #2 complements the "ex post" analysis in the first part with an "ex ante" analysis and estimates the effects of increasing supply chain risks on firms' investment, employment and asset redeployment decisions. Using textual analysis of quarterly earnings, conference-call transcripts, and SEC current report filings (i.e., 8-Ks) the authors will construct firm-level measures of the extent and type of supply chain risk faced by public firms in the U.S. between 1992 and 2026. The first model of part #2 looks at the intensive margin and estimates whether firms increase or decrease investment and employment when faced with supply chain risks. The second model, on the other hand, looks at the extensive margin and estimates what plants are sold or closed when the firm faces supply chain risks. Asset redeployment decisions (establishment closures and sales) constitute an important area where size and composition of firms constantly change. The authors will prepare estimates about the sample of firms whose supply chain risks increase as well as provide estimates about the characteristics and performance of sold establishments as a result of increasing supply chain risks.

The project will also use Integrated Longitudinal Business Database (iLBD) to examine private and non-employer customer and supplier firms. Non-employer firms constitute an important part of the economy and the project will provide estimates of the effects of shocks on the non-employer firms. Therefore, this project will be consistent with the Census Bureau's interest, as stated in the *Research Opportunities at the U.S. Census Bureau* document, in studying the role of small and young businesses in the U.S. economy.

In summary, this project will analyze patterns of economic expansion and contraction at both the firm and plant levels and attempt to uncover how actual disruption and an increasing probability of disruption to suppliers' operations affect the real economic activity. In doing so, the project will prepare estimates that will inform the Census Bureau about the patterns of firm growth, firm entry and exit, firm investment and employment behavior, as well as firm productivity. The benefits in this criterion will be documented in the form of CES-Working Papers and publications in peer-reviewed journals.

Mixed Proposal – Census Benefits

Criterion 5: Understanding and / or improving the quality of data produced through a Title 13, Chapter 5 survey, census or estimate.

To our knowledge, nonprofit organizations have rarely been studied within the LEHD data. Therefore, it will be of great benefit to the Census Bureau to have researchers examine these data carefully. Given that the nonprofit sector is the third largest US employer behind retail and manufacturing, it is important to ensure that the Census Bureau has access to accurate information. Our examination of the LEHD will be far more extensive than can be carried out in the routine internal consistency checks during survey processing. The examination should lead both to an assessment of this aspect of data quality, and to recommendations for directions for improvement.

Good empirical analysis often begins with tasks such as examining where records or items are missing[5], where responses are extreme, or take on inconsistent values. One example of where the LEHD might be providing an incomplete picture is charity size. Both the LEHD and the Form 990 data include a measure of the number of employees working at each nonprofit. In the Form 990, nonprofit organizations are requested to list both the "Total number of individuals employed in the calendar year", and the "Total number of volunteers". Comparison of these numbers to the analogue in the LEHD data will allow the researchers to assess the quality of the data produced through the LEHD. The fact that many charities rely heavily on volunteers means that focusing solely on employees (as counted in the LEHD) may result in labelling some charities as small when in fact they are not. Thus, adding IRS Form 990 data on volunteers will help provide a more accurate measure of charity size. This is important not only for Census records but also for making better-informed government policy. In particular, mischaracterizing charity size, and the implied provision of charitable services, may result in misallocation of government resources.

Linking the LEHD data to IRS Form 990 data will offer another way to assess the quality of the data produced through the LEHD. By adding information from the publicly available IRS Form 990 data, the researchers will be able to compare the LEHD to another more closely examined and value-added data.

In particular, the researchers will be able to examine the quality of NAICS and SIC classifications. Any organization that files a Form 990 should be classified as a nonprofit organization in the LEHD data. Thus, by recording the discrepancies between the two data sources, the researchers will be enhancing the Census Bureau's knowledge of its data collection programs. In particular, IRS Form 990 data are classified according to an alternative system: the National Taxonomy of Exempt Entities (NTEE). The NTEE classifications have been hand-checked for accuracy and quality by the NCCS. The researchers will construct a mapping between the NTEE codes and the NAICS and SIC classifications, and confirm whether the NAICS codes assigned to establishments in the LEHD data correspond to the appropriate NTEE code as assigned by the NCCS, where the appropriateness of the match is defined by the mapping. Both the NAICS and SIC classification systems were developed to describe for-profit organizations, but may not align with the NTEE classification used by the IRS for a number of reasons. For example, a nonprofit legal aid organization may be characterized by NAICS under its catch-all nonprofit category, Sector 813, ("Religious, Grantmaking, Civic, Professional, and Similar Organizations"), or it may be considered by NAICS to be part of the legal services industry, Sector 5411 ("Legal Services"). While such a nonprofit may simultaneously be considered a civic organization and a legal services organization, it is

[5] Any missing or seemingly-implausible values may be able to be imputed using methods such as those suggested by Peytchev (2012).

more likely to compete with other legal services providers than it is to compete with providers of non-legal civic services. Subsequent research which relies on NAICS codes to identify legal services firms may be incompletely capturing the legal services industry if nonprofit legal services providers are assigned NAICS code 813 rather than NAICS code 5411. Comparison with the NTEE would allow us to identify these misclassifications, and to provide a remedy for future researchers.

In this way, comparing measures from independent sources that should be similar, or that should differ in predictable ways, increases the Census Bureau's knowledge of its data collection programs.

The researchers will prepare and submit to the Census Bureau technical memoranda summarizing the ways in which the data sources differ, and how the quality might be improved. They will also submit interim reports on the benefits to the Census Bureau every 12 months, and a Post Project Certification document at the end of the project.

Criterion 7: Enhancing the data collected in a Title 13, Chapter 5 survey or census.

By linking the LEHD data to information from the IRS Form 990 Filings, the researchers will be adding several new variables to the existing data. Specifically, the researchers will produce crosswalks between the set of nonprofit organizations included in the Employer Characteristics File of the LEHD Data and the IRS Form 990 Filings. These filings come from two sources: (i) the National Center for Charitable Statistics (NCCS) provides data from Form 990 on a sample of 501(c)(3) nonprofit organizations from 1989 through 2015; and (ii) the IRS provides access to the universe of nonprofit organizations which file the Form 990 electronically, for all years between 2009 and 2018. This crosswalk will greatly expand the set of characteristics available for these employers. For example, the Form 990 includes information on their revenues and expenditures by source, their use of tax preparers and lobbyists, their total assets and liabilities by the end of each fiscal year, and more.

This is a crucial part of the project. Linking external data to Census Bureau data enhances the data for the Census Bureau and its partners. In particular, the Census Bureau will be able to make more informed inferences about economic and social relationships using these linked data. Those inferences may improve imputations for non-response or provide information about the quality of sampling frames and data collection techniques. Furthermore, insights drawn from analyzing these enhanced data may also provide information on potential improvements to future data collections.

The researchers will prepare and submit to the Census Bureau technical memoranda summarizing the ways in which the data has been enhanced. They will also submit interim reports on the benefits to the Census Bureau every 12 months, and a Post Project Certification document at the end of the project.

Criterion 11: Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5.

To our knowledge, nonprofit organizations have rarely been studied within the LEHD data. Therefore, this project will be yields statistics beyond those that have already been published by the Census Bureau.

The researchers will characterize the population of nonprofit employees. Research questions will include: (1) What is the age and tenure distribution of charity employees? (2) How does the education level of non-profit workers differ from for-profit workers? (3) What do transition rates look like both within non-profit sectors and across sectors (where sectors are defined using the National Taxonomy of Exempt Entities (NTEE) codes)? (4) Do transition rates vary by charity size and age?

These questions are important to answer in helping to characterize the nonprofit sector, and it is therefore crucial to have access to a comprehensive set of demographic characteristics, such as those available in the American Community Survey (ACS). Very little is known about the characteristics of nonprofit employees, and how they differ from for-profit employees, or align with the demographics of the populations served by their organizations. Knowing more about this sector not only enhances our knowledge of one of the largest employers in the US, but also helps inform policy on how best to support and retain nonprofit employees. Comparison of nonprofit employees' demographics with those of their organizations' clients will serve to document the differences between givers and receivers of charitable services.

This project will also study the beneficiaries of the Public Service Loan Forgiveness (PSLF) plan. Policies such as PSLF are aimed at attracting and retaining workers in the nonprofit (and public) sectors. As well as producing regression estimates regarding the effectiveness of this policy, the researchers will also be able to describe the characteristics of the workers at whom this policy is targeted. This may help contribute to the larger literature on student loans and inequality.

These questions will mainly be answered using cross tabulations. There is a great deal that can be learned with some basic tabulations and assessing how these statistics change over time. Most of the variables of interest will come directly from the data (for example, age, education and wages), but others must be constructed by the researchers. For example, tenure will be calculated using employment start and end dates, and transition rates will be inferred from information on employment spells.

Such statistics include summary statistics about specific variables (means, medians, moments), and coefficient estimates that summarize behavior of subgroups of the population. These statistics increase the information available about these populations, subpopulations, and their characteristics.

The researchers will prepare and submit to the Census Bureau technical memoranda summarizing these statistics regarding nonprofit employers, as well as papers for journal publication. They will also submit interim reports on the benefits to the Census Bureau every 12 months, and a Post Project Certification document at the end of the project.