The IRS Databank: Developing a Population Panel Dataset for Tax Policy Research

Raj Chetty, Harvard and NBER
John N. Friedman, Harvard and NBER
Nathaniel Hilger, Harvard
Emmanuel Saez, UC Berkeley and NBER
Danny Yagan, Harvard

November 2011

Prepared as part of IRS contract TIRNO-09-R-00007
1. Overview of Databank

2. Example: constructing a panel of EITC filers

3. Application: Uncovering Impacts of EITC on Wage Earnings Distribution
Individual level taxpayer data may only be accessed by those with statutory authority

Must be used for purposes related to tax administration as defined in Internal Revenue Code 6013

Two sources of data: population files and SOI samples
SOI prepares “perfected” random samples

- Distributes to internal customers and makes a few public-use files available
- These samples have fewer errors and contain more variables than population data

Population data has two advantages relative to SOI samples

- Longitudinal: can follow people over time without attrition
- Spatial: can “zoom in” on policy experiments that affect a narrow slice of the U.S. population yet retain substantial power
Using Population Data For Tax Policy Research

- Existing structure of population files is not optimized for statistical research on tax policy
  
  - Most datasets organized by household filing units, which change over time
  
  - Individuals often switch between primary earners, secondary earner, and not filing taxes at all
  
  - Many datasets have more than one billion rows, which makes merges very time consuming and often infeasible

- Our team has tackled these problems by creating a simple, unified dataset that we call the IRS Databank
The databank reorganizes an important subset of these data

Key elements:

- **Complete Individual-level Panel**: Contains one row per person per year for every person listed on a tax form during 1996-2009

- **Pre-Merged Household Links**: Contains spouse and dependent masked TIN, as well as key variables

- **Commonly Used Variables for Sample Selection**: income, location, major program eligibility (e.g. EITC)

Constructing the databank took six months

But now takes about one week to add new variables to the databank
The IRS Databank

- The databank is organized into “modules” that contain commonly used groups of key variables. Current modules include:
  - Linkage of individual to household filing unit (and spouse if present)
  - 1040 information
  - Dependent identities (masked TINs and DOBs)
  - Information Returns: W-2 wage earnings for both husband and spouse and other information returns, such as 1098-T and 1099

- Modules exist (and can be updated) independently but have identical number of rows in identical order
  - Very fast to combine modules into complete datasets
  - Very fast to select samples from the full population
Question: How does EITC eligibility affect low-income individuals?

Study Population: All years of data for all individuals who, at some point, were EITC-eligible
Question: How does EITC eligibility affect low-income individuals?

Study Population: All years of data for all individuals who, at some point, were EITC-eligible

- Household EITC Eligibility (F1040)
- Individuals within each Eligible Household (Taxpayer Info)
- Households for Eligible Individuals in Other Years (Taxpayer Info)
- Household Income (F1040)
- Household Dependents (Depend)
Question: How does EITC eligibility affect low-income individuals?

Study Population: All years of data for all individuals who, at some point, were EITC-eligible

Equivalent databank code (runs in 24-48 hours):

```r
data local.base;
  set databank.morp_merge_all_spine_mskd;
  databank.morp_merge_all_1040;
  databank.morp_merge_all_depend;

data sample_select;
  set local.base;
  if (CONDITIONS) then eic_elig = 1;
  if tax_yr = 2009 & eic_elig = 1 then output;
  keep tinx;

data local.analysis;
  merge local.base sample_select (in=a);
  by tinx;
  if eic_elig = 1;
```

Assembles required Databank modules
Selects individuals for sample
Constructs analysis dataset
Databank does not alleviate key limitations of population data files

- Users must always carefully consider if population files are appropriate for the research purpose

Databank remains a work in progress

- Update the databank as tax returns are filed
- Support additional users
  - Add new variables for other projects
  - Databank is large (4 TB, more than 5 bil. rows), so writing efficient SAS code remains critical

Construction of an analogous corporate databank currently in progress
Application: Effects of EITC on Labor Supply

- We estimate effects of EITC on wage earnings by exploiting differences across neighborhoods in knowledge about EITC

- Lack of counterfactuals has made it difficult to identify intensive-margin impacts of EITC in prior work

- Our idea: use cities with low levels of information about tax policies as counterfactuals for behavior in the absence of tax policy

- Proxy for knowledge: fraction of self-employed recipients reporting income exactly at EITC refund-maximizing kink

- We compare wage earnings distribution across high vs. low self-emp. bunching areas to measure “real” labor supply impacts of EITC

- Audit evidence reveals that W-2 income rarely manipulated
Earned Income Tax Credit Schedule for Single Earners with One Child

- **Percent of EITC Eligible Wage-Earners**
  - Taxable Income
  - EITC Credit Amount

- **Taxable Income**
  - $0
  - $10K
  - $20K
  - $30K

- **EITC Credit Amount**
  - $0
  - $1K
  - $2K
  - $3K
  - $4K
Income Distribution for Single Wage Earners with One Child

Percent of EITC Eligible Wage-Earners

EITC Credit Amount

Taxable Income

- $0
- $10K
- $20K
- $30K

- $0
- $1K
- $2K
- $3K
- $4K
Income Distribution for Single Wage Earners with One Child

Is the EITC having an effect on this distribution?

Percent of EITC-Eligible Wage-Earners

Taxable Income

EITC Credit Amount

Income Distribution for Single Wage Earners with One Child

Is the EITC having an effect on this distribution?
IRS Databank population file yields the sample size needed to identify impacts of EITC on local earnings distribution

Sample restriction: individuals who at least once between 1996-2009: (1) file a tax return, (2) have income < $40,000, (3) claim a dependent

Sample size after restrictions:
- 77.6 million individuals
- 1.09 billion person-year observations on income
Outline of Empirical Analysis

- Step 1: Develop a proxy for knowledge about the EITC in each neighborhood using sharp bunching among self-employed
2008 Federal EITC Schedule for a Single Filer with Children

Taxable Income (Real 2010 $)

- $0
- $10K
- $20K
- $30K
- $40K

EITC Credit

- $0
- $1K
- $2K
- $3K
- $4K
- $5K

- One child
- Two children
Income Distribution for EITC-Eligible Households with Children in 2008

Percent of EITC-Eligible Households

Taxable Income (Real 2010 $)

- One child
- Two children
Empirical Implementation: Proxy for Knowledge

- Proxy for knowledge using the fraction of EITC filers who report income at first (refund maximizing) kink and have self-employment income

- Our proxy is a noisy measure of true knowledge
  - Differences across cities in sharp bunching may be due to other determinants of tax compliance rather than knowledge
  - This measurement error attenuates estimate of the impact of taxes on wage earnings

→ Lower bound on estimated impact of EITC on wage earnings
Self-Employed Sharp Bunching by State in 2008

The map shows the self-employed sharp bunching by state in 2008, with shades indicating the range of values from 0.0084 to 0.0341. States with darker shades have higher values, while lighter shades indicate lower values. The map includes states from all regions of the United States.
Self-Employed Sharp Bunching in 2008 by 3-Digit Zip Code in Kansas, Louisiana, Oklahoma, and Texas
Outline of Empirical Analysis

- Step 1: Develop a proxy for knowledge about the EITC in each neighborhood using sharp bunching among self-employed

- Step 2: Establish learning as a mechanism for differences in sharp bunching across neighborhoods
Event Study of Bunching for Movers, by Destination Area

<table>
<thead>
<tr>
<th>Event Year</th>
<th>Self-Emp. Sharp Bunching for Movers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Movers to Lowest Information Areas</td>
</tr>
<tr>
<td></td>
<td>Movers to Medium Information Areas</td>
</tr>
<tr>
<td></td>
<td>Movers to Highest Information Areas</td>
</tr>
</tbody>
</table>

$\Delta \varepsilon = 0.41\% (0.05\%)$
Outline of Empirical Analysis

- Step 1: Develop a proxy for knowledge about the EITC in each neighborhood using sharp bunching among self-employed

- Step 2: Establish learning as a mechanism for differences in sharp bunching across neighborhoods

- Step 3: Compare wage earnings distributions across low- and high-knowledge neighborhoods to uncover impacts of EITC on earnings
Income Distributions for Single Wage Earners with One Child

Is the EITC having an effect on this distribution?
Wage Earnings Distributions in High vs. Low Information Areas
Single Individuals with One Child

Percent of EITC-Eligible Wage-Earners

Income

High Information Neighborhoods

Low Information Neighborhoods
EITC Credit Amount for Single Wage Earners with Two Children vs. Neighborhood Bunching

- $3200
- $3250
- $3300
- $3350

- 0.0%
- 0.8%
- 1.6%
- 2.4%
- 3.2%
- 4.0%

Neighborhood Self-Emp. Sharp Bunching
Outline of Empirical Analysis

- Step 1: Develop a proxy for knowledge about the EITC in each neighborhood using sharp bunching among self-employed

- Step 2: Establish learning as a mechanism for differences in sharp bunching across neighborhoods

- Step 3: Compare wage earnings distributions across low- and high-knowledge neighborhoods to uncover impacts of EITC on earnings

- Step 4: Compare impacts changes in EITC subsidies on earnings across low vs. high knowledge nbhds. to account for omitted variables
Child Birth as a Source of Tax Variation

- To identify causal impacts of EITC, need variation in tax incentives
  - Birth of first child $\rightarrow$ substantial change in EITC incentives
  - Although birth affects labor supply directly, cross-neighborhood comparisons provide good counterfactuals
Earnings Distributions in the Year Before First Child Birth for Wage Earners

Percent of Households

Income

Lowest Information Neighborhoods

Medium Information Neighborhoods

Highest Information Neighborhoods

$0

$10K

$20K

$30K

$40K

Percent of Households
Earnings Distributions in the Year of First Child Birth for Wage Earners

Percent of Households

Income

$0  $10K  $20K  $30K  $40K

Lowest Information Neighborhoods
Medium Information Neighborhoods
Highest Information Neighborhoods
Earnings Distributions in the Year of First Child Birth for Wage Earners
Individuals Working at Firms with More than 100 Employees

Percent of Households

Income

Lowest Information Neighborhoods

Medium Information Neighborhoods

Highest Information Neighborhoods
Simulated EITC Credit Amount for Wage Earners Around First Child Birth
Individuals Working at Firms with More than 100 Employees

Highest Information Neighborhoods
Medium Information Neighborhoods
Lowest Information Neighborhoods

Age of Child

Simulated EITC Credit

- $1400
- $1300
- $1200
- $1100
- $1000
- $900

-4
-2
0
2
4

Age of Child

Lowest Information Neighborhoods
Medium Information Neighborhoods
Highest Information Neighborhoods
Our estimates can be used to characterize impact of EITC on income distribution taking into account behavioral responses.

Use neighborhoods with little self-employment bunching as counterfactual for earnings distribution without EITC.
Impact of EITC on Income Distribution for Single Earners with 2+ Children

Percent of EITC-Eligible Wage-Earners vs. Total Income

- No EITC
- Counterfactual

Total Income:
- $0
- $10K
- $20K
- $30K
- $40K
Impact of EITC on Income Distribution for Single Earners with 2+ Children

Percent of EITC-Eligible Wage-Earners

Total Income

No EITC
Counterfactual
EITC, No Behavioral
Response
Impact of EITC on Income Distribution for Single Earners with 2+ Children

The graph illustrates the impact of the Earned Income Tax Credit (EITC) on the income distribution of single earners with 2+ children. The x-axis represents total income, ranging from $0 to $40K, while the y-axis shows the percent of EITC-eligible wage-earners.

Three scenarios are depicted:
- **No EITC Counterfactual** (blue line)
- **EITC, No Behavioral Response** (green line)
- **EITC with Behavioral Response** (red line)

The curves indicate that the EITC has a significant effect on income distribution, with a peak impact observed at around $20K total income. The presence of behavioral response further enhances the effect, especially noticeable at lower income levels.

The graph highlights the effectiveness of the EITC in mitigating income inequality, particularly for lower-income earners.