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

Outline

1. Overview of Databank

2. Example: constructing a panel of EITC filers

3. Application: Uncovering Impacts of EITC on Wage Earnings Distribution

United States Tax Data

- 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

United States Tax Data

- 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 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

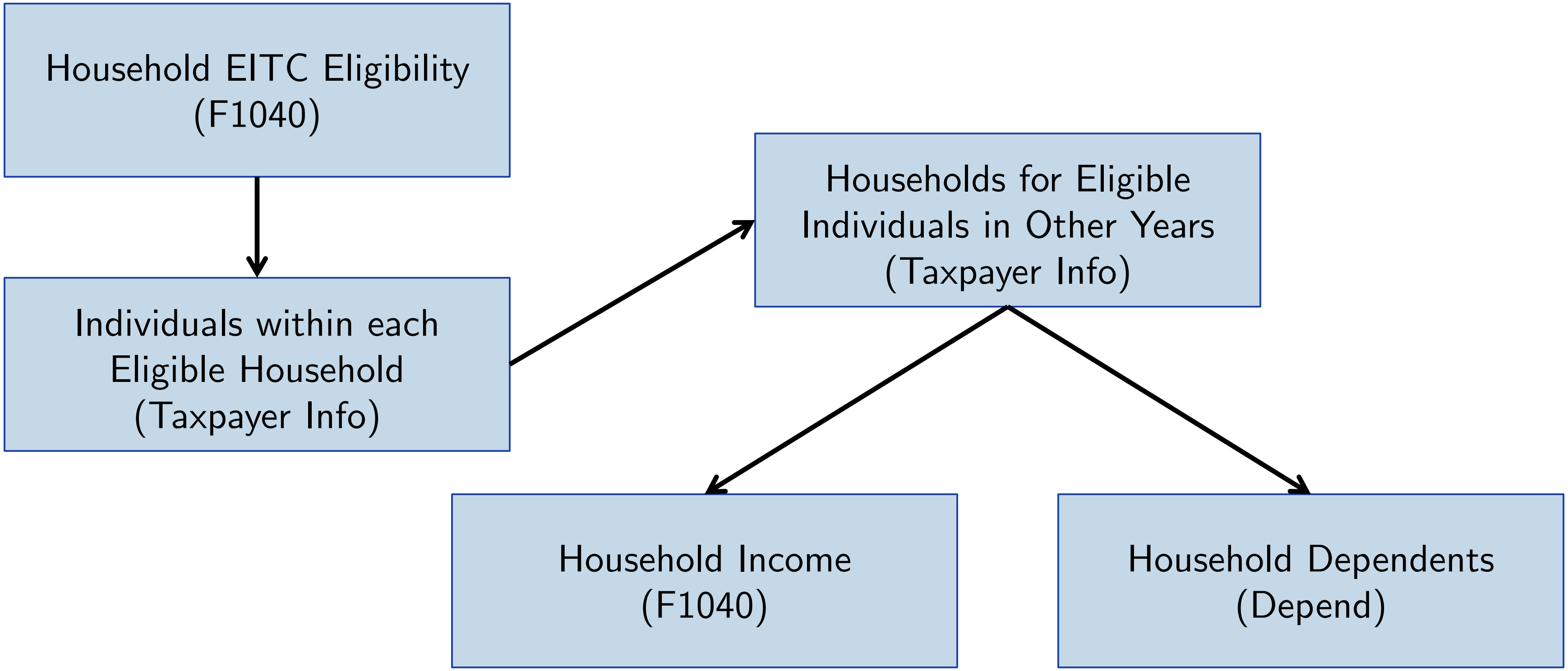
- The databank is organized into “modules” that contain commonly used groups of key variables. Current modules include:
  - Linkage of individual to household filing unit (and spouse if present)
  - 1040 information
  - Dependent identities (masked TINs and DOBs)
  - Information Returns: W-2 wage earnings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

Example: Constructing an EITC analysis sample

- 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

Constructing an EITC sample from CDW files

- 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)

Constructing an EITC sample from the Databank

- 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):

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

Assembles required Databank modules

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

Selects individuals for sample

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

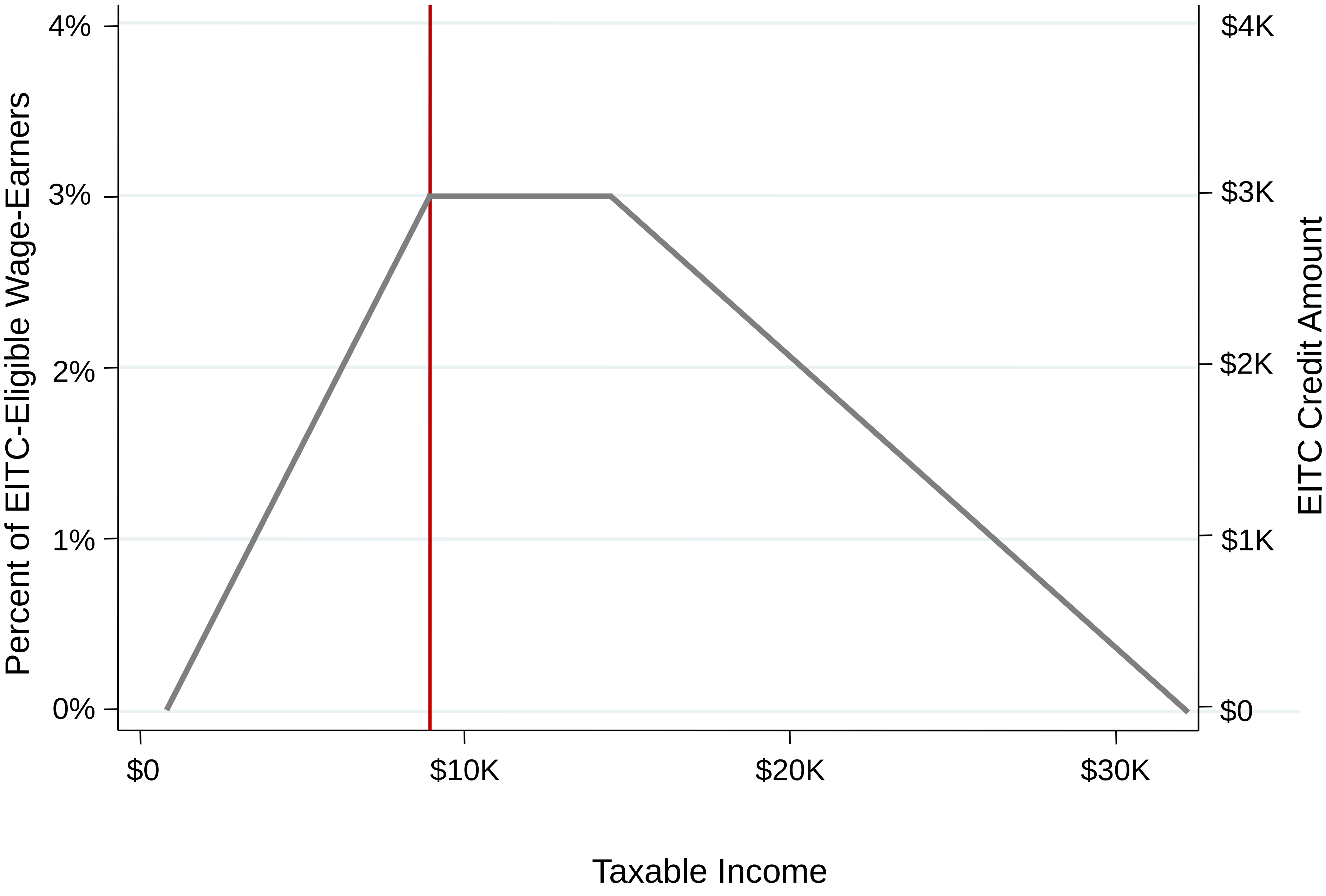
Constructs analysis dataset

IRS Databank in the Long Run

- 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

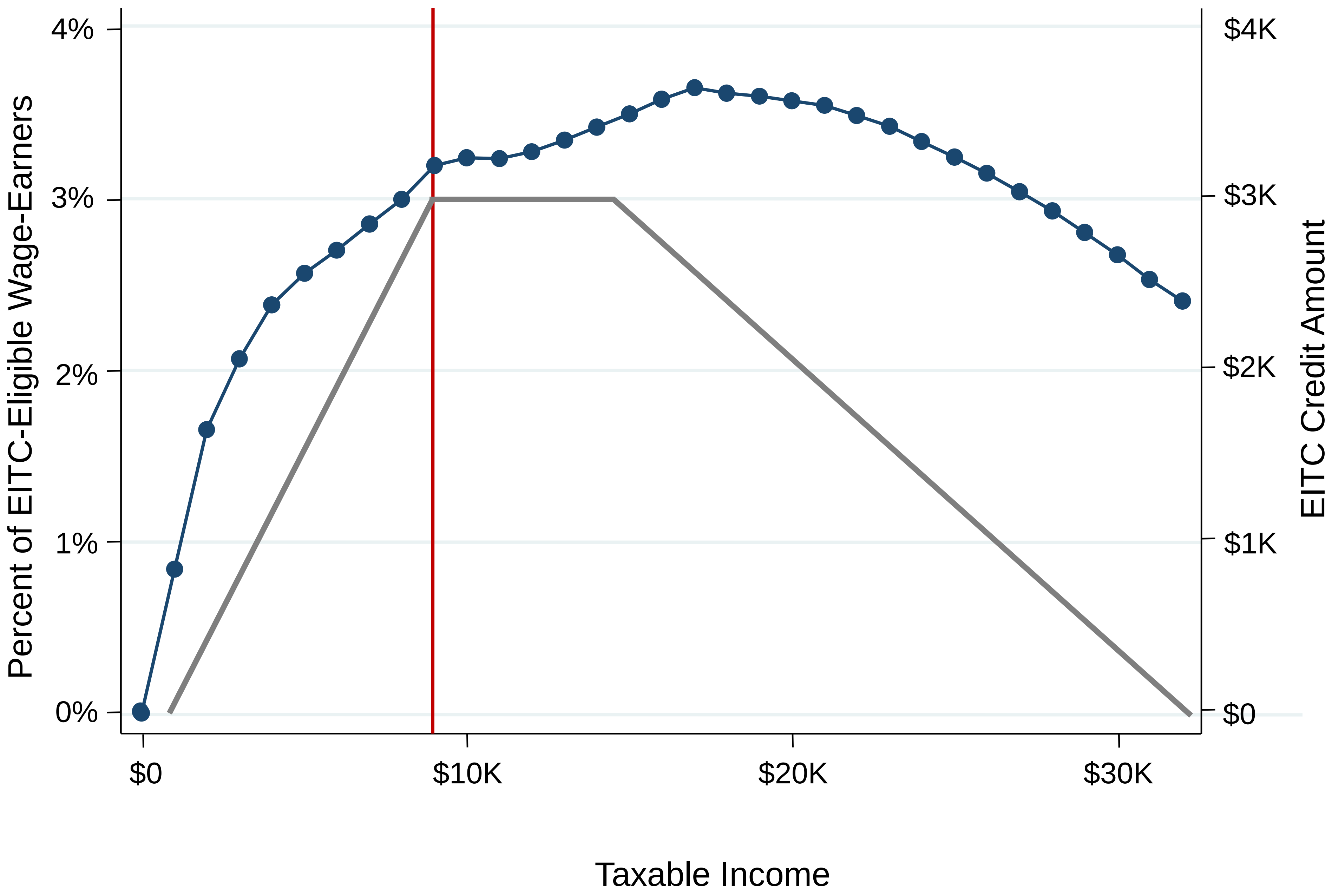
Application: Effects of EITC on Labor Supply

- 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

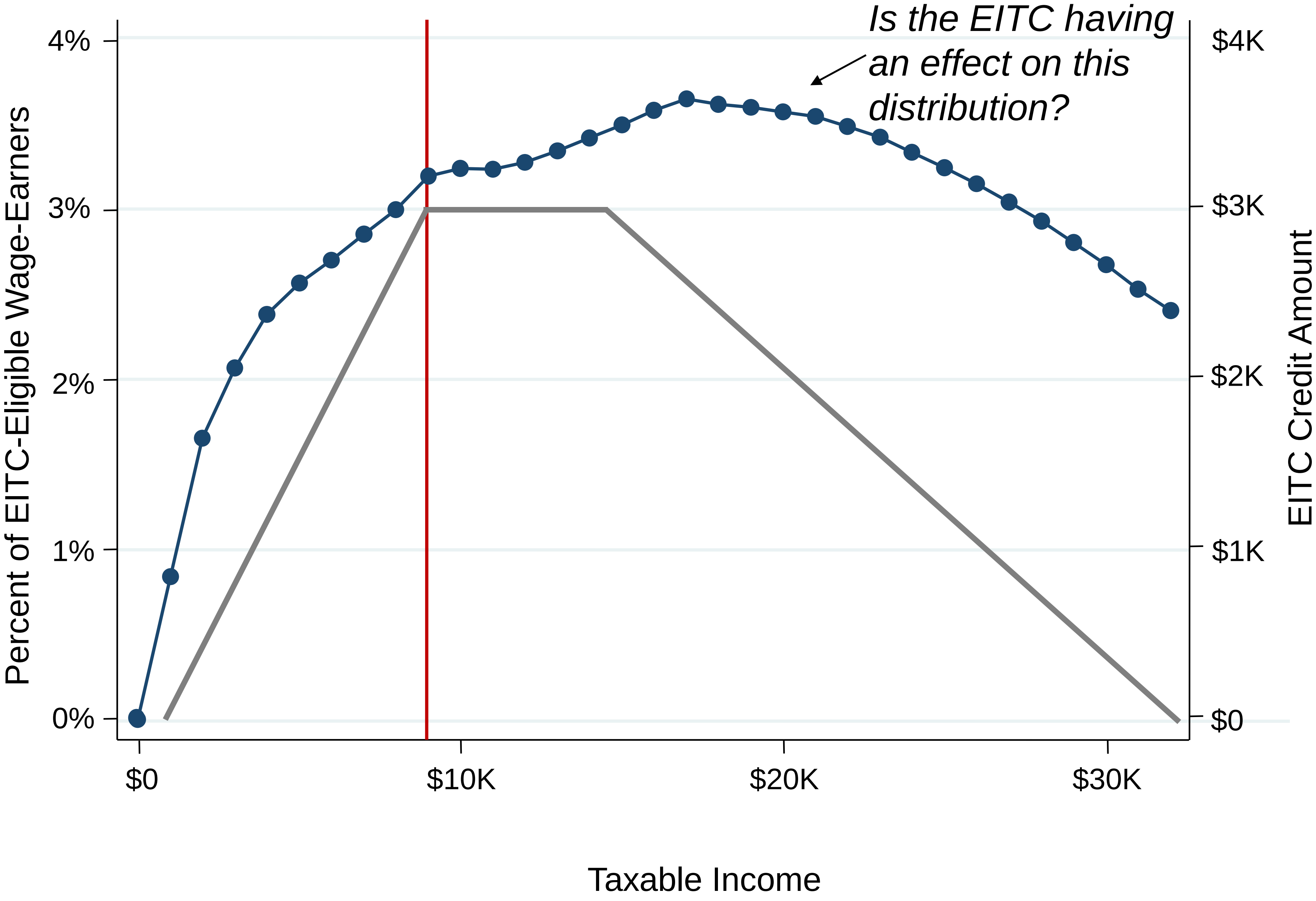


Earned Income Tax Credit Schedule for Single Earners with One Child

Income Distribution for Single Wage Earners with One Child



Income Distribution for Single Wage Earners with One Child



Is the EITC having an effect on this distribution?

Data and Sample Definition

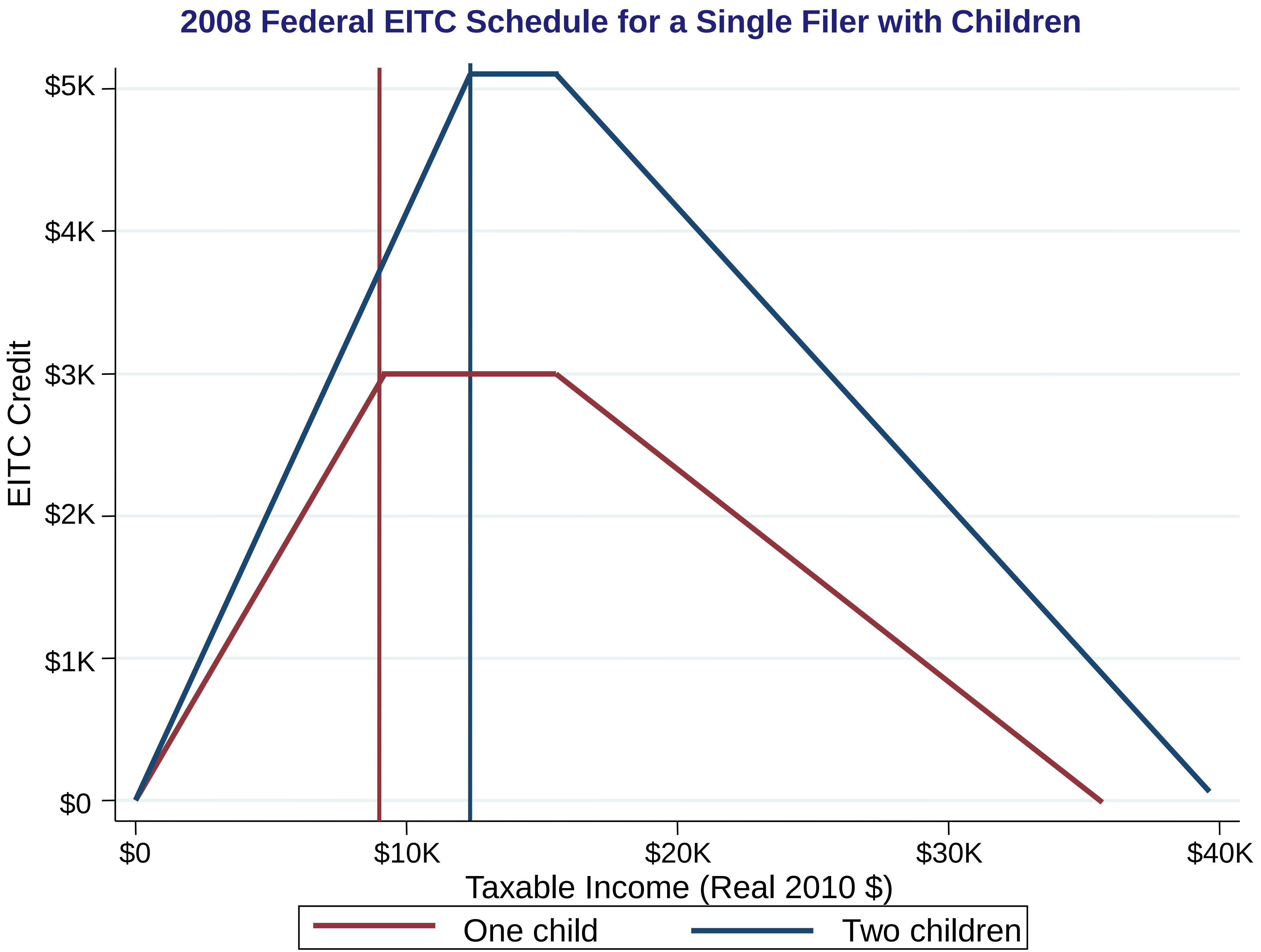
- IRS Databank population file yields the sample size needed to identify impacts of EITC on local earnings distribution

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

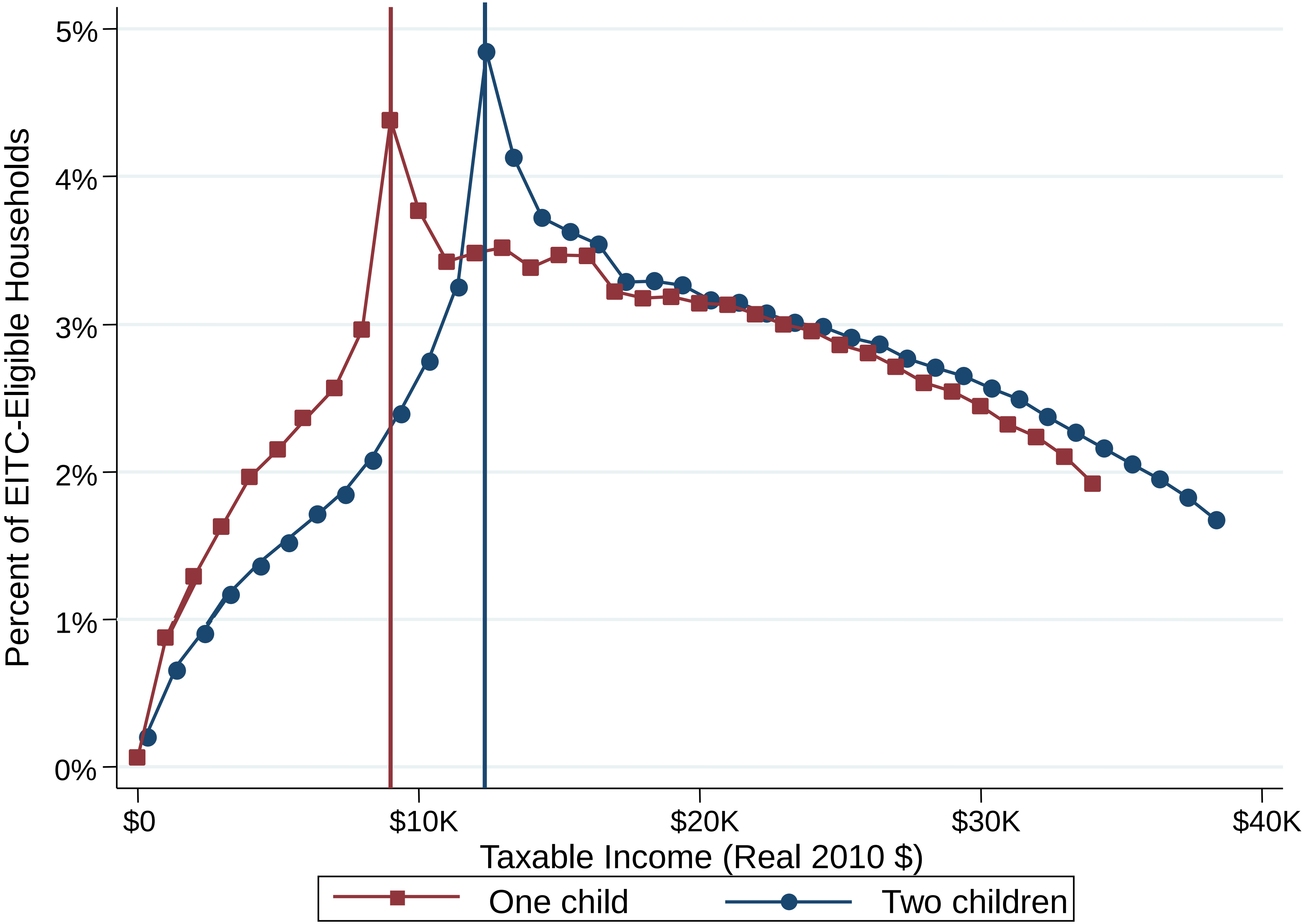
- Sample size after restrictions:
    - 77.6 million individuals
    - 1.09 billion person-year observations on income

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
EITC Credit
$5K
$4K
$3K
$2K
$1K
$0
$0
$10K
$20K
$30K
$40K
Taxable Income (Real 2010 $)
One child
Two children

Income Distribution for EITC-Eligible Households with Children in 2008



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

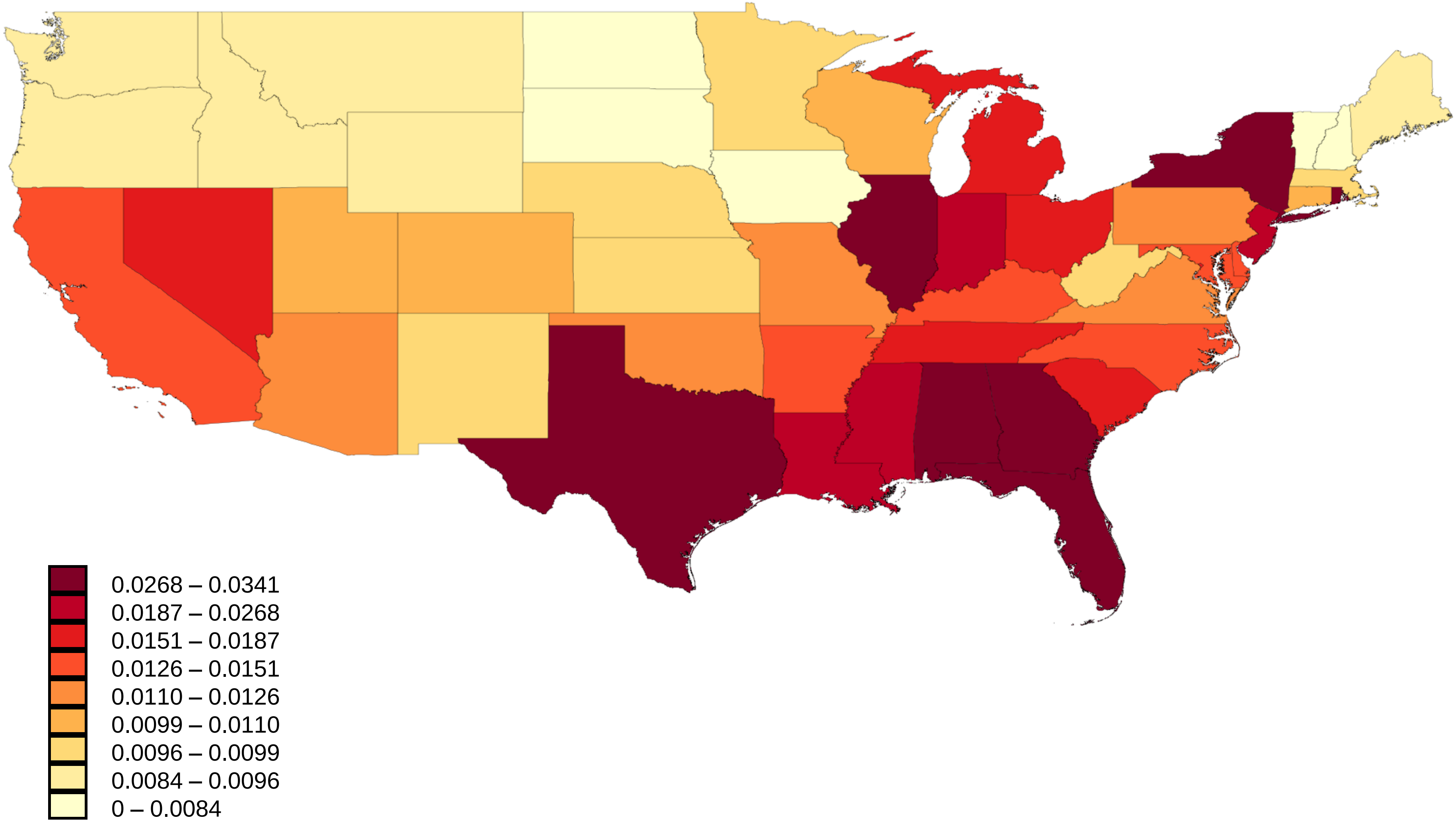
- Our proxy is a noisy measure of true knowledge

    - Differences across cities in sharp bunching may be due to other determinants of tax compliance rather than knowledge

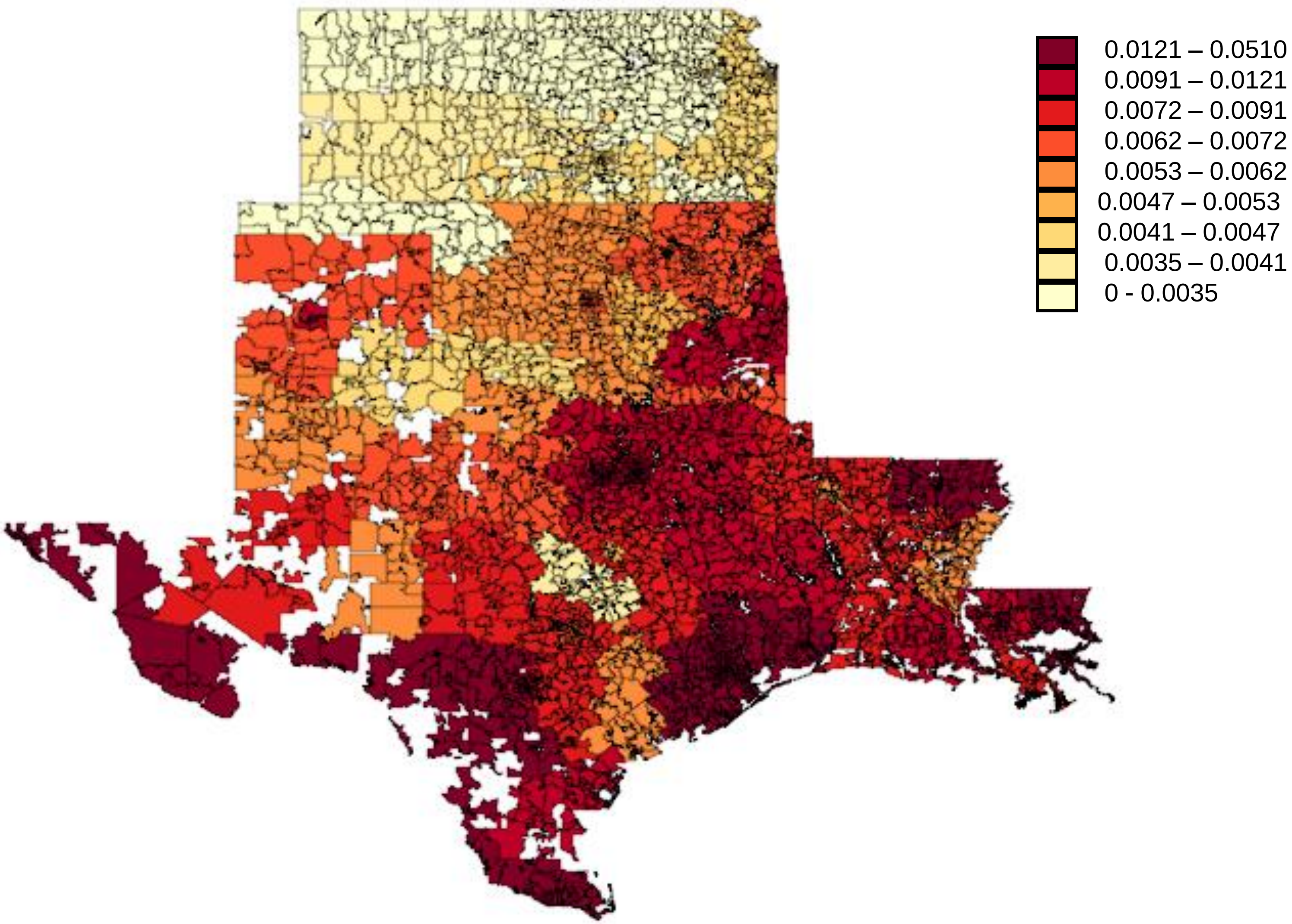
    - This measurement error attenuates estimate of the impact of taxes on wage earnings

  → Lower bound on estimated impact of EITC on wage earnings

Self-Employed Sharp Bunching by State in 2008



Self-Employed Sharp Bunching in 2008
by 3-Digit Zip Code in Kansas, Louisiana, Oklahoma, and Texas

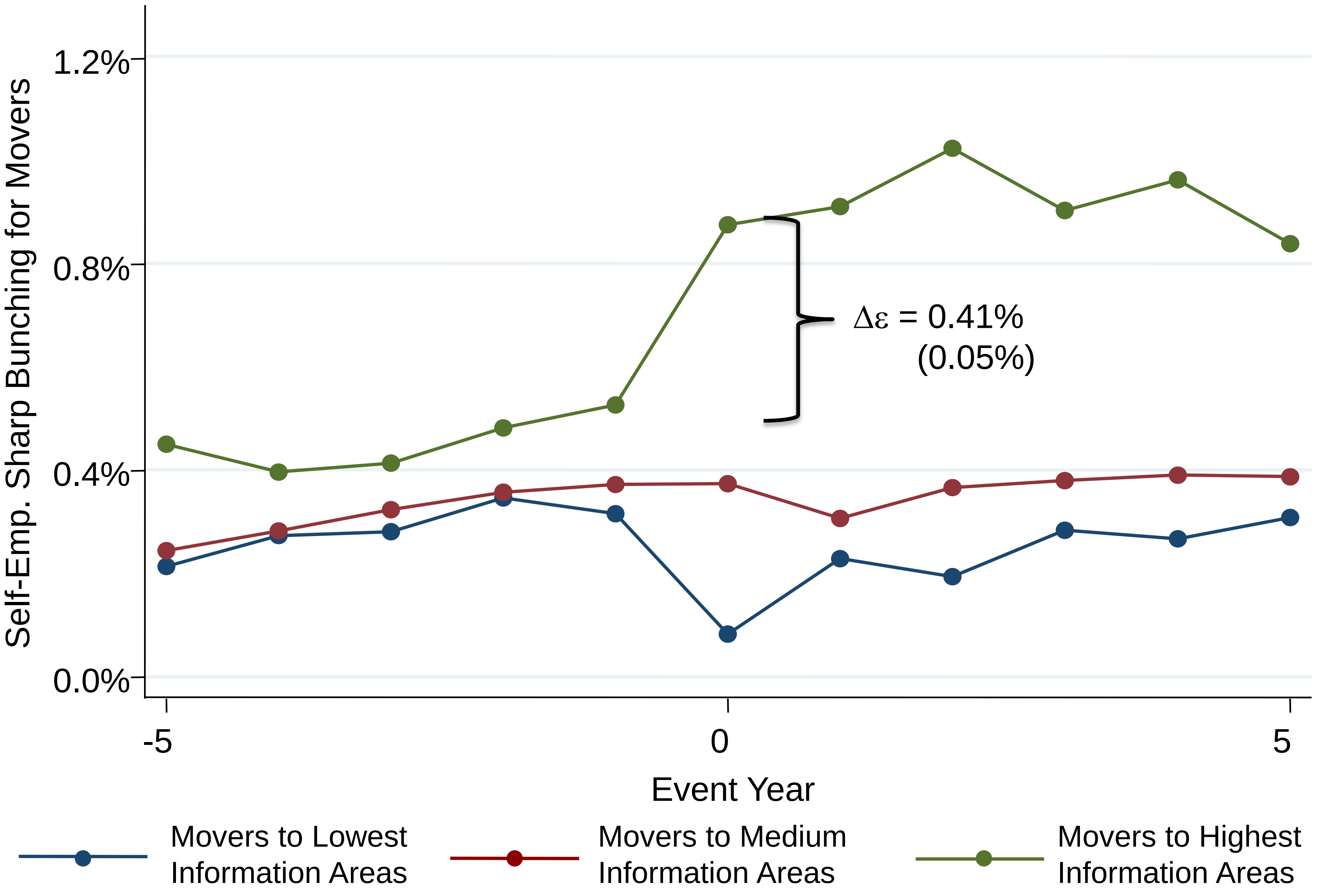


Outline of Empirical Analysis

- 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



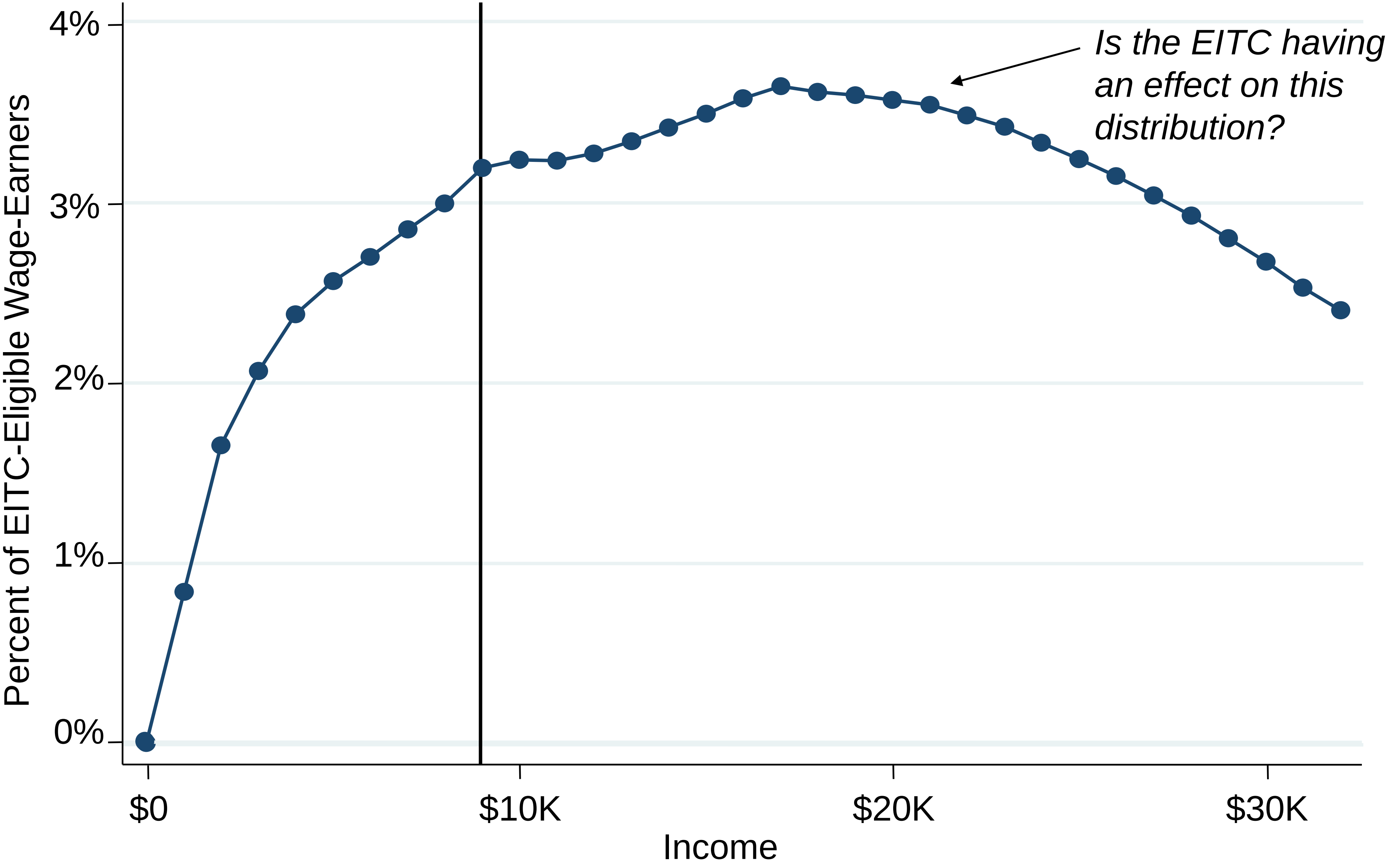
Outline of Empirical Analysis

- 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

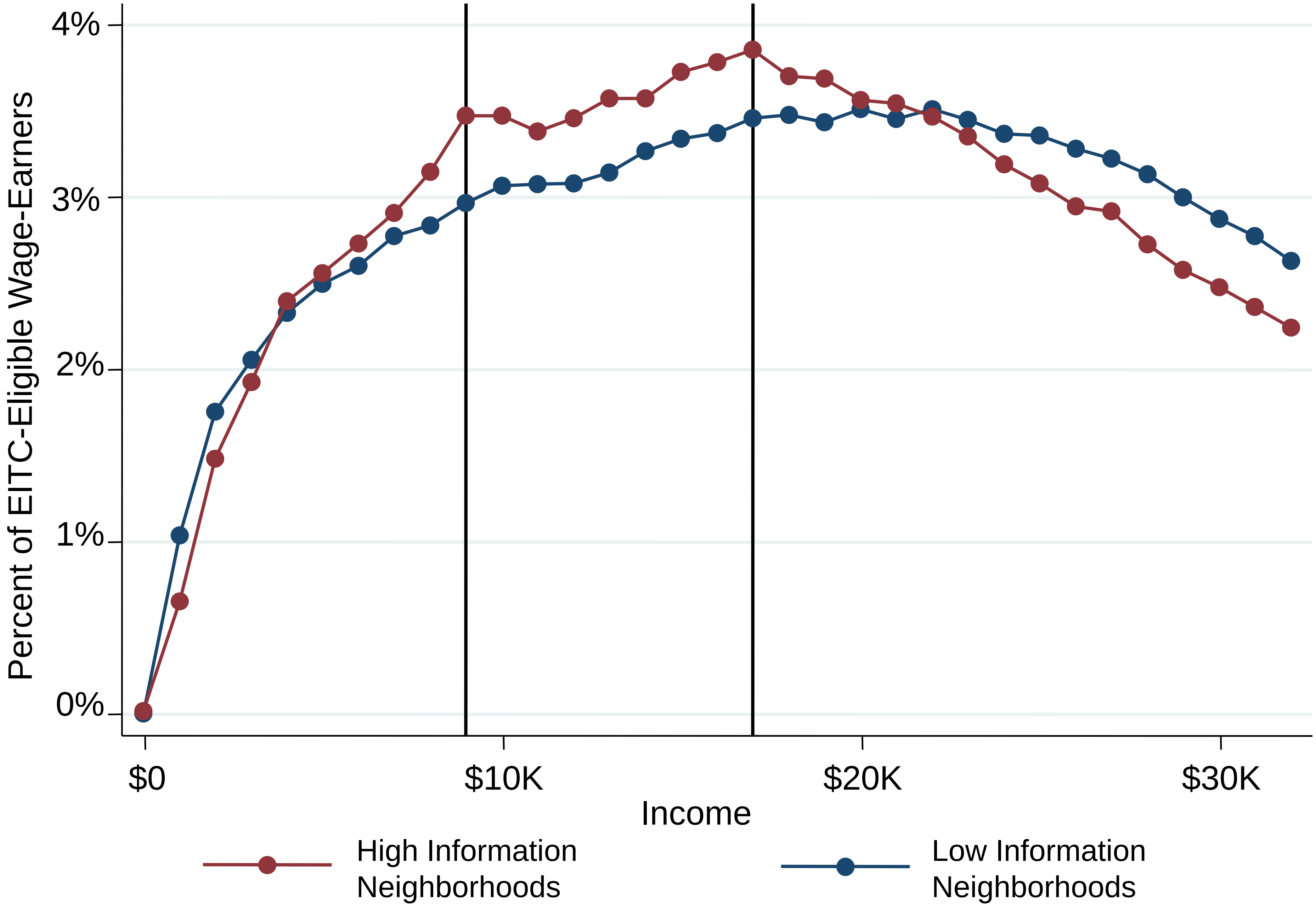
- Step 3: Compare wage earnings distributions across low- and high-knowledge neighborhoods to uncover impacts of EITC on earnings

Income Distributions for Single Wage Earners with One Child

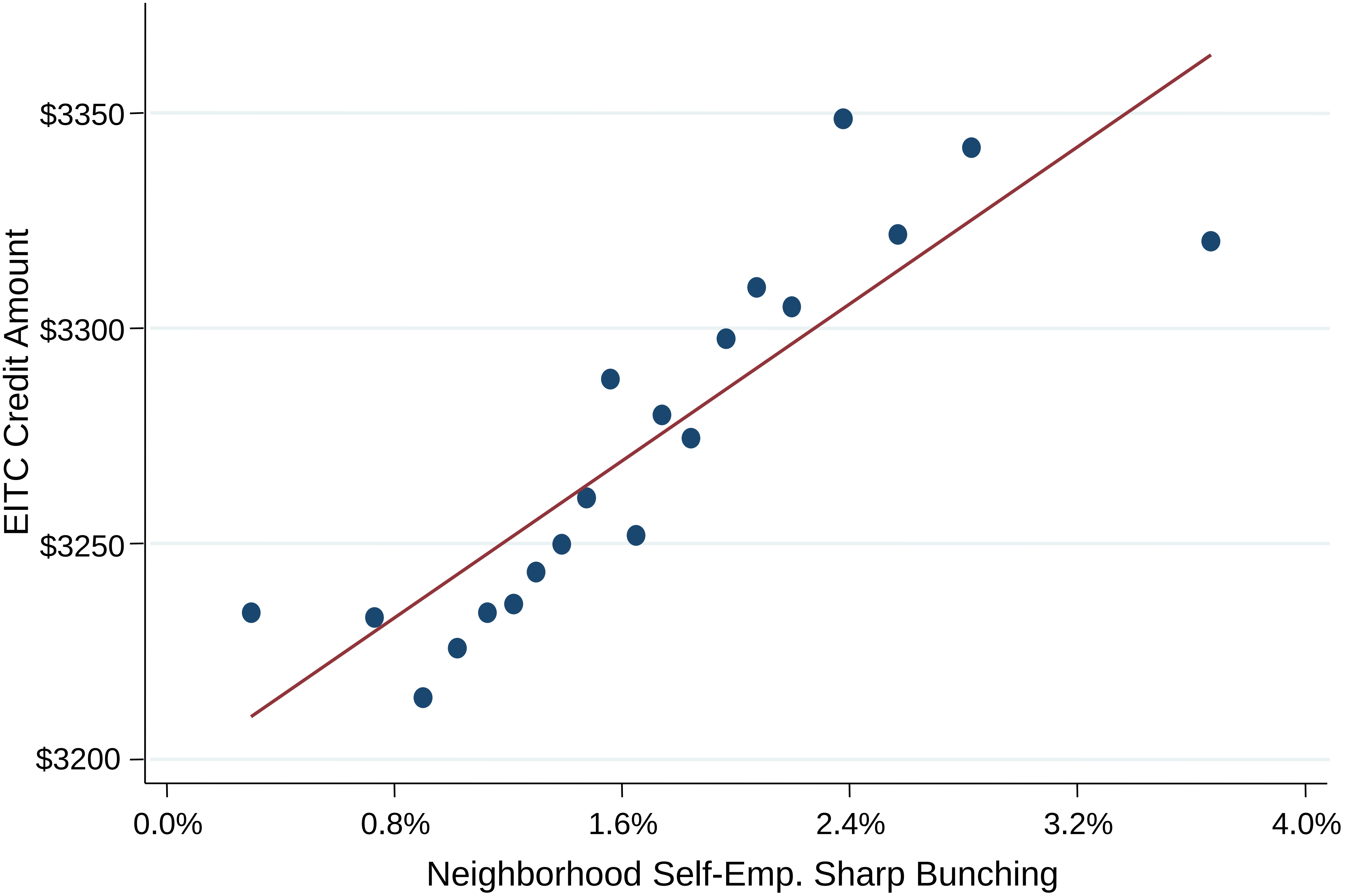


Is the EITC having an effect on this distribution?

Wage Earnings Distributions in High vs. Low Information Areas
Single Individuals with One Child



EITC Credit Amount for Single Wage Earners with Two Children vs. Neighborhood 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

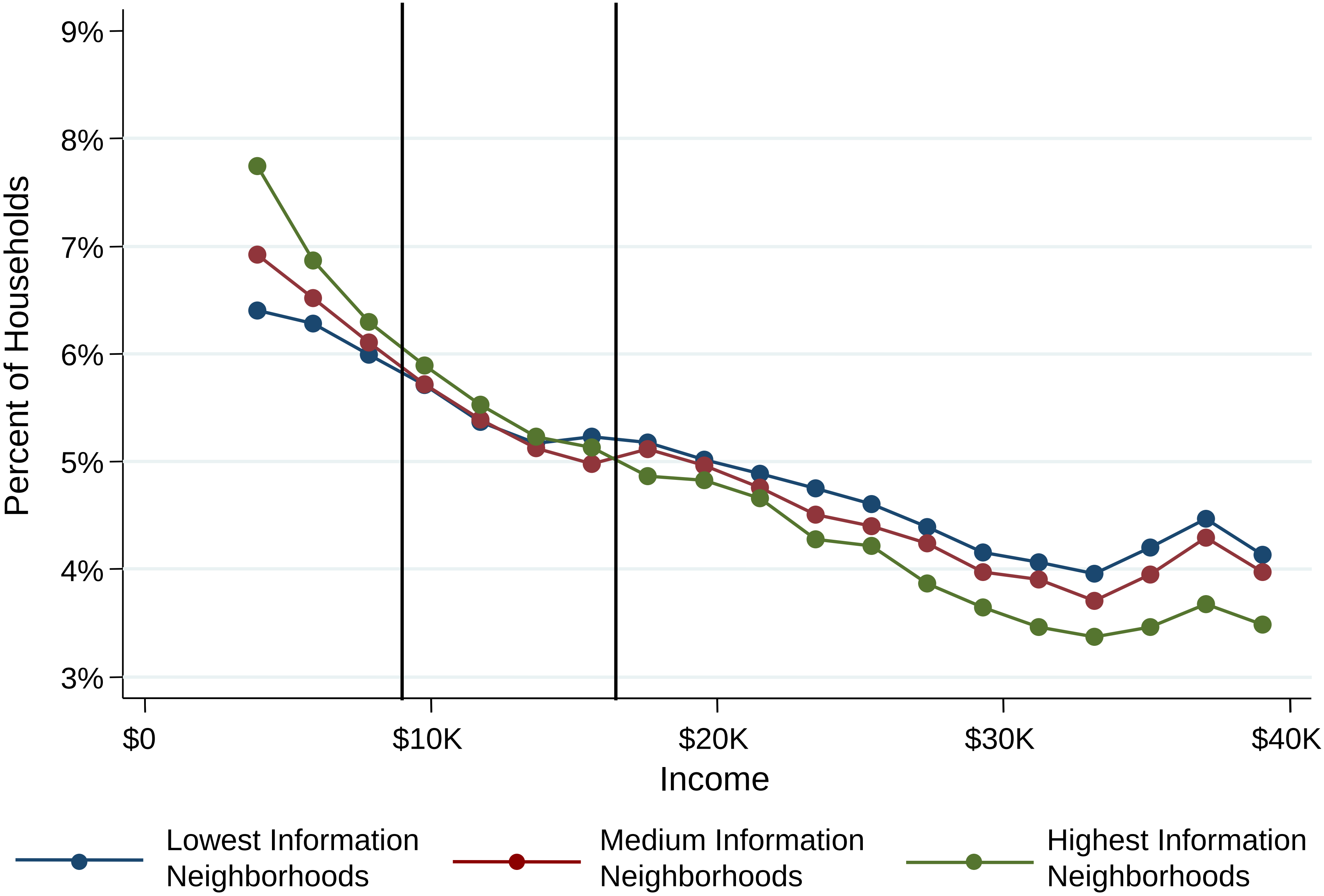
- Step 3: Compare wage earnings distributions across low- and high-knowledge neighborhoods to uncover impacts of EITC on earnings

- Step 4: Compare impacts changes in EITC subsidies on earnings across low vs. high knowledge nbhds. to account for omitted variables

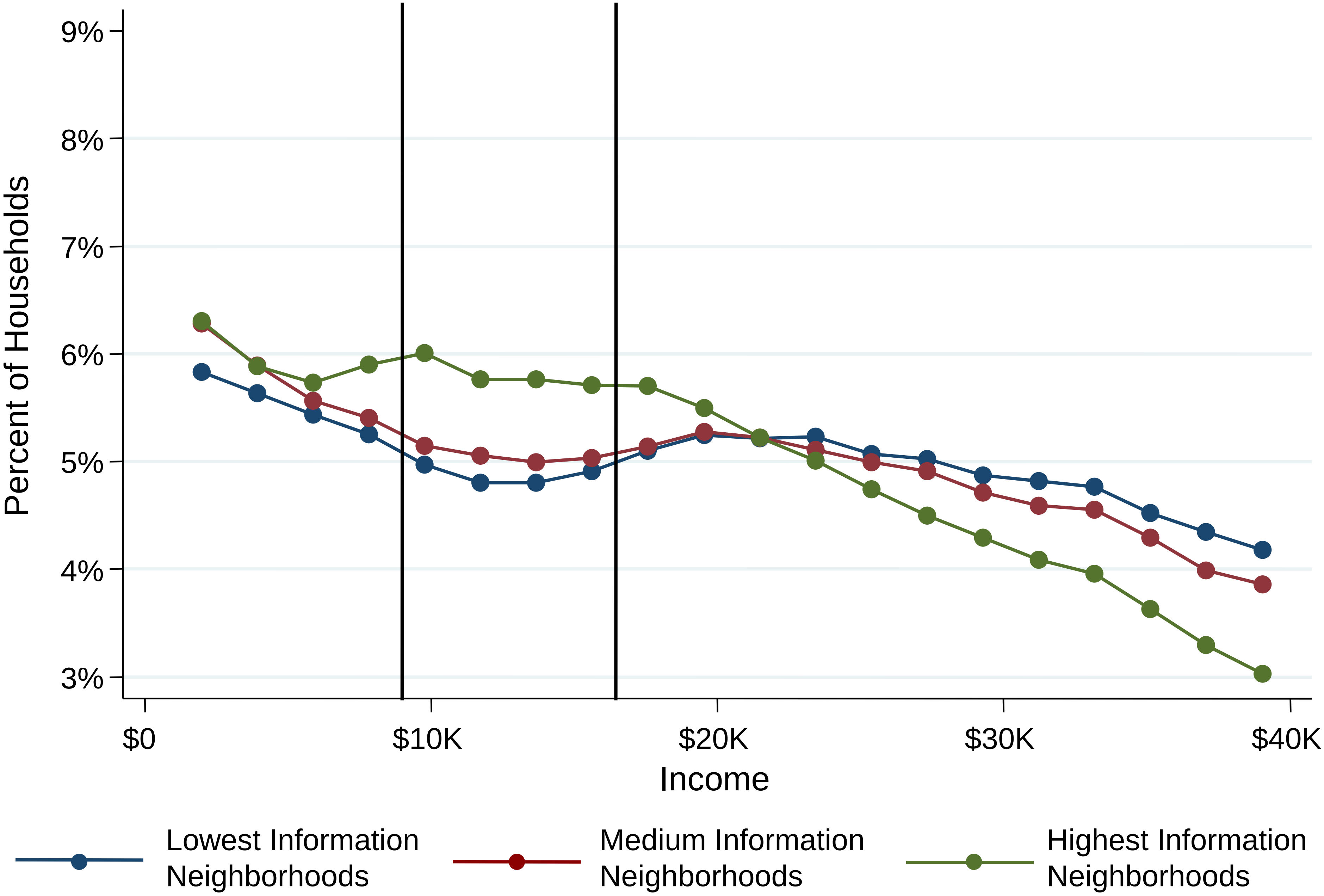
Child Birth as a Source of Tax Variation

- To identify causal impacts of EITC, need variation in tax incentives
  - Birth of first child → substantial change in EITC incentives
  - Although birth affects labor supply directly, cross-neighborhood comparisons provide good counterfactuals

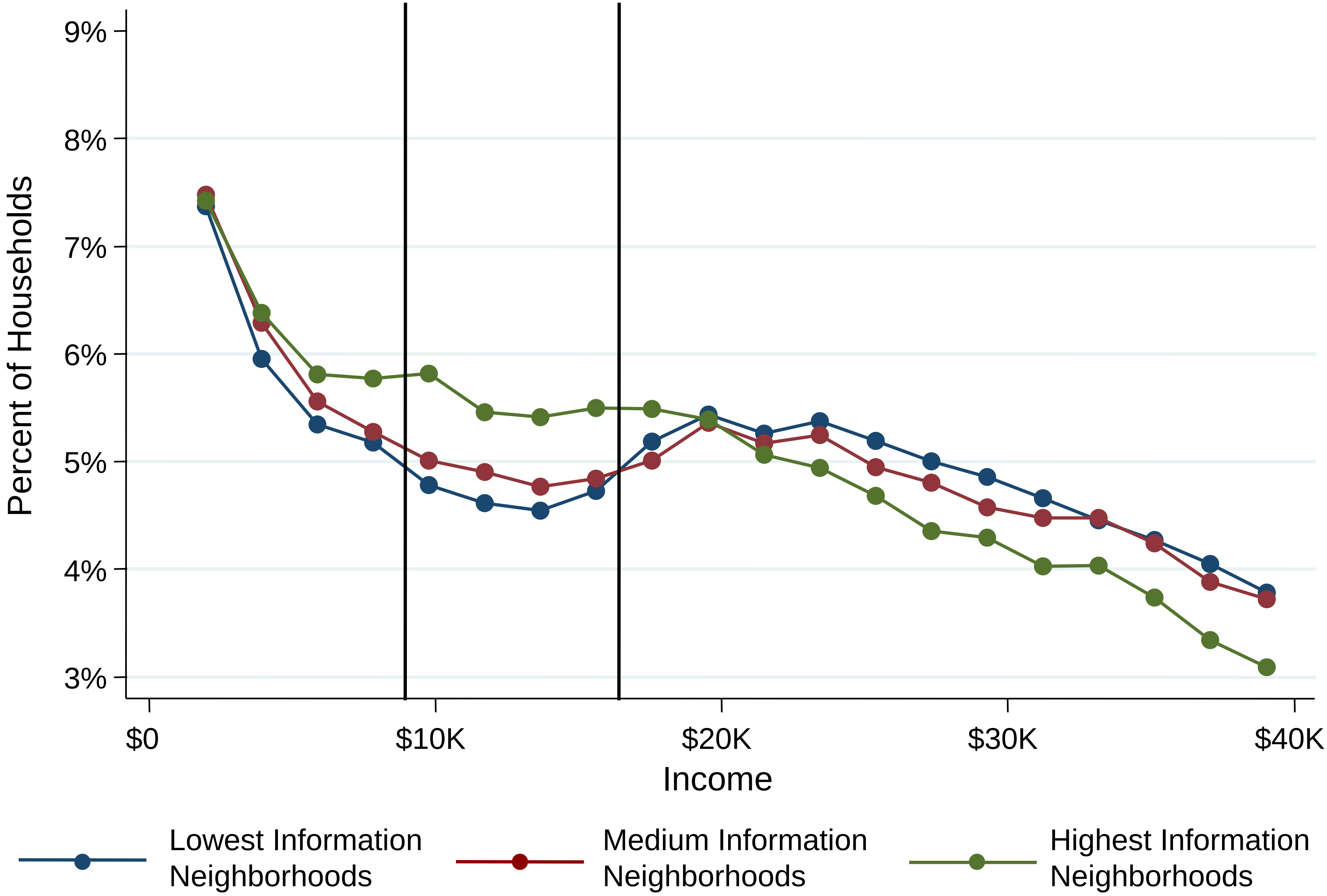
Earnings Distributions in the Year Before First Child Birth for Wage Earners



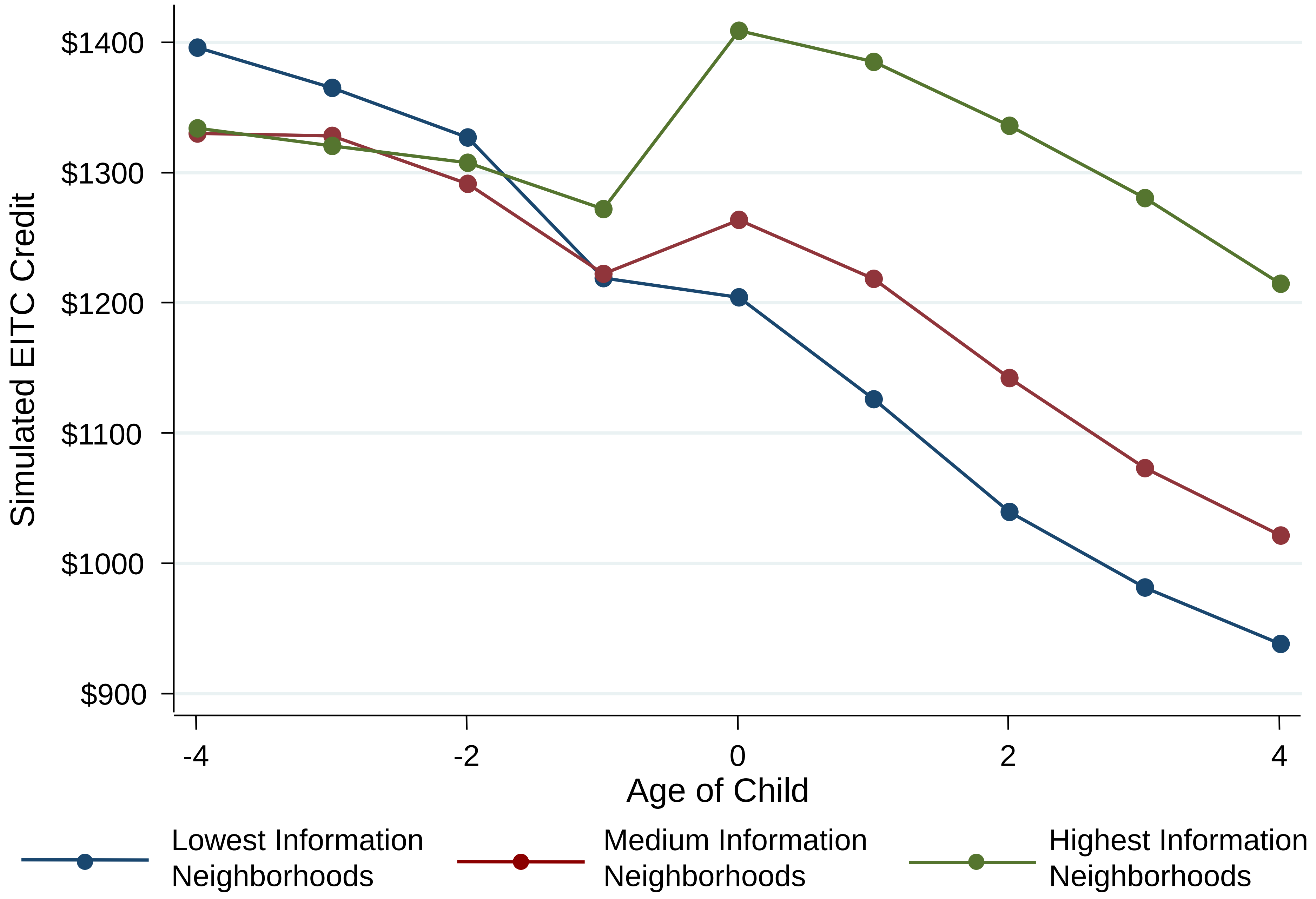
Earnings Distributions in the Year of First Child Birth for Wage Earners



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



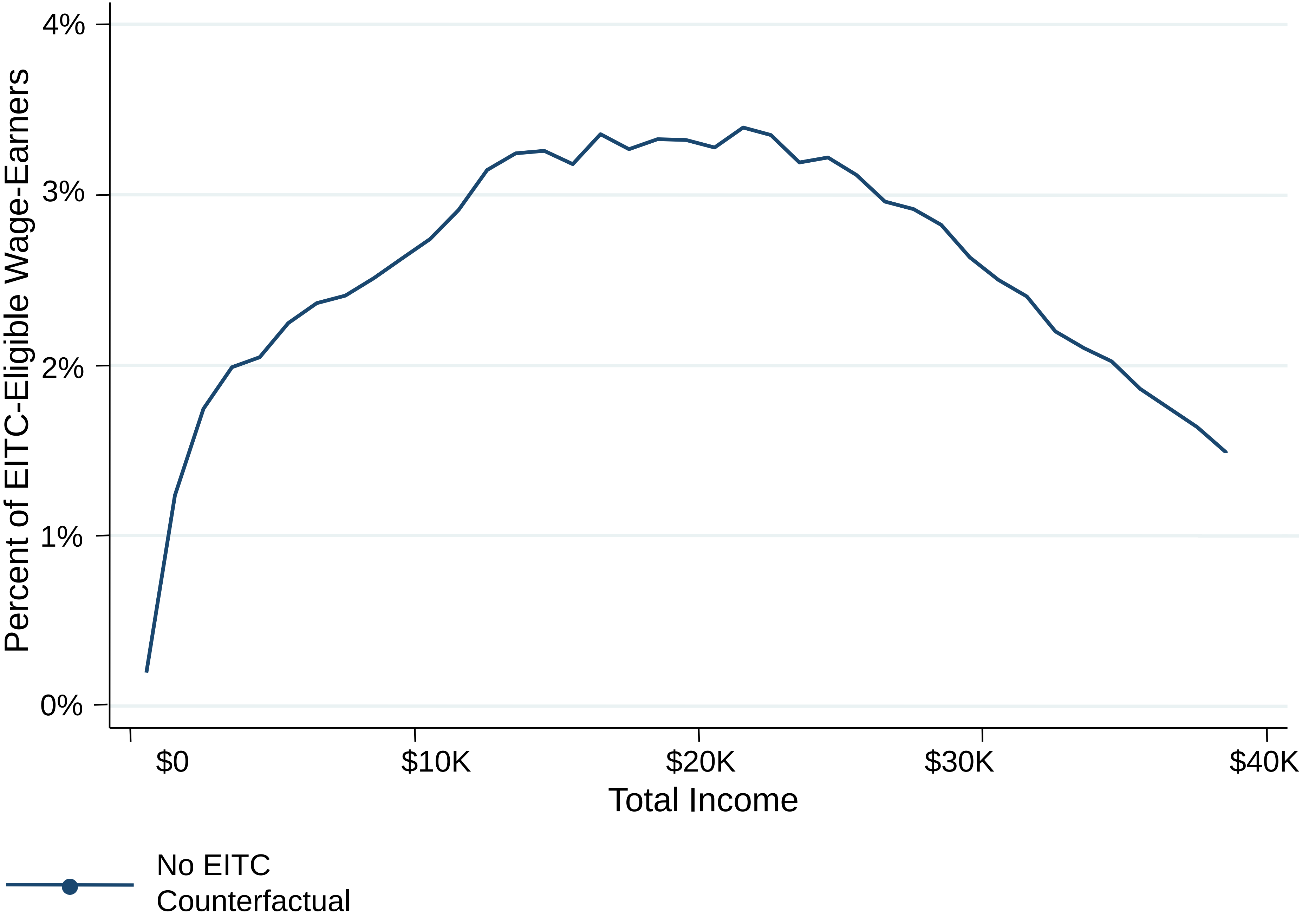
Simulated EITC Credit Amount for Wage Earners Around First Child Birth
Individuals Working at Firms with More than 100 Employees



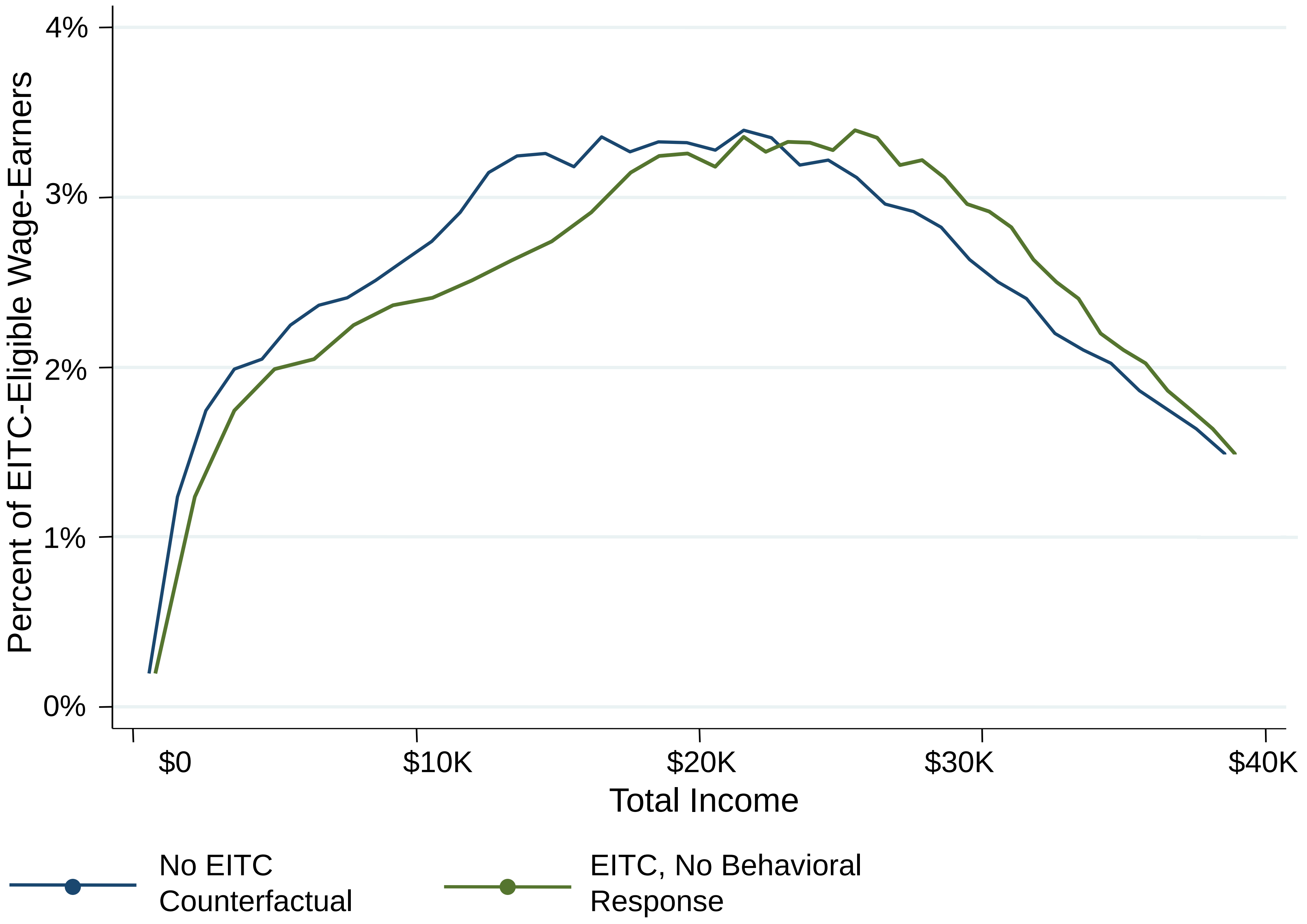
Tax Policy Implications

- 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

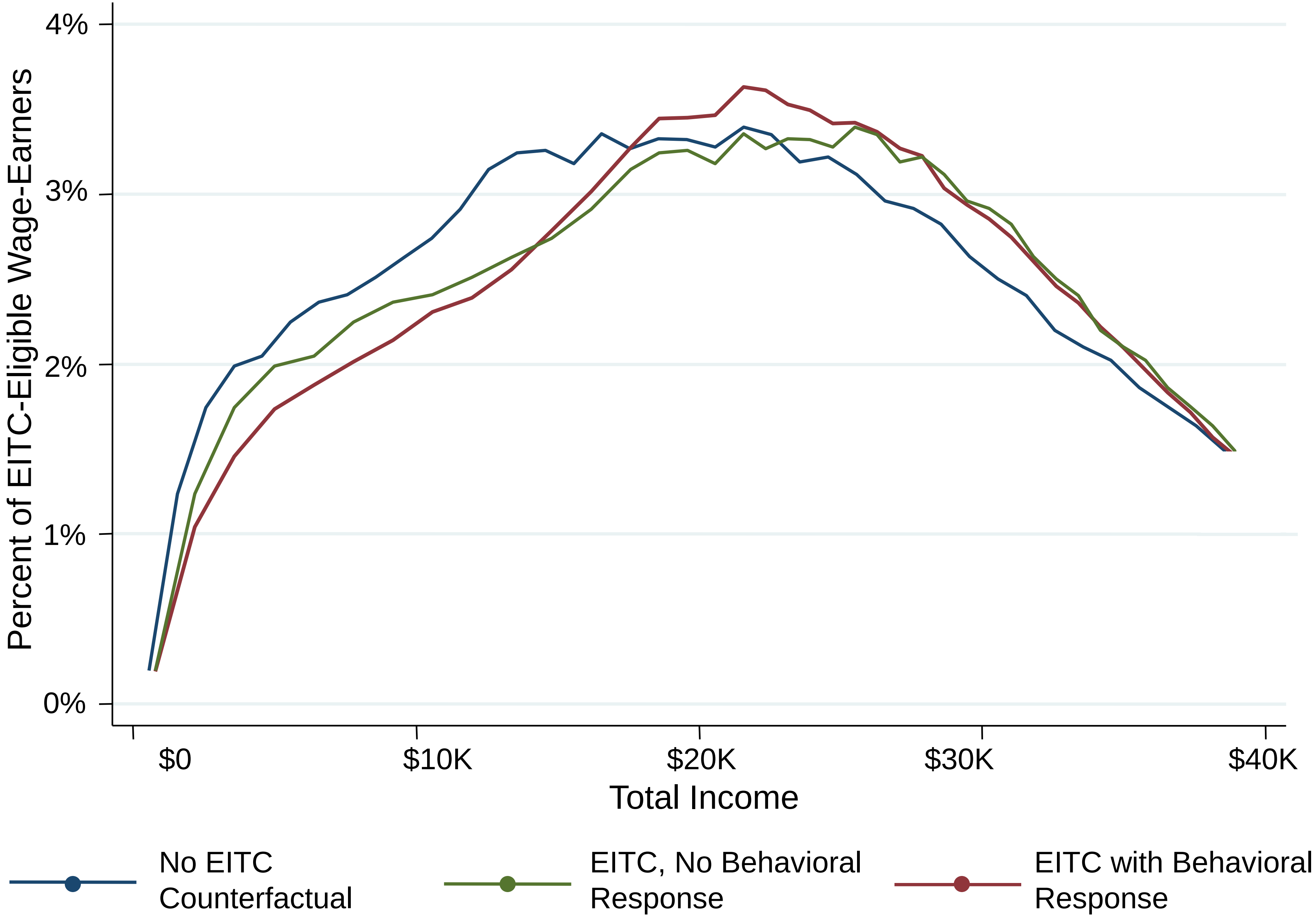
Impact of EITC on Income Distribution for Single Earners with 2+ Children



Impact of EITC on Income Distribution for Single Earners with 2+ Children



Impact of EITC on Income Distribution for Single Earners with 2+ Children


Percent of EITC-Eligible Wage-Earners
$0
$10K
$20K
$30K
$40K
Total Income
No EITC Counterfactual
EITC, No Behavioral Response
EITC with Behavioral Response