



## Session 2. Nonfiling: IRS-Census Data Comparisons

**Moderator:**

***Mike Weber***

*IRS, RAAS, Statistics of Income*

**What Drives Filing Compliance?**

***Brian Erard***

*B. Erard and Associates*

**Searching for Ghosts Redux: Improved  
Methodologies for Estimating the Nonfiling  
Tax Gap**

***Mark Payne***

*IRS, RAAS, Office of Research*

**Handling Respondent Rounding of Wages  
Using the IRS and CPS Matched Dataset**

***Minsun Riddles***

*Westat, Inc.*

**Discussant:**

***Amy O'Hara***

*U.S. Census Bureau*

# What Drives Filing Compliance?

Brian Erard (*B. Erard & Associates*)  
and John Guyton, Pat Langetieg, Mark  
Payne, and Alan Plumley (IRS RAAS)\*

\*The views expressed in this presentation are those of the authors  
and do not necessarily represent the opinion of the IRS

# Filing Requirements

---

- Not everyone is required to file a tax return
- One is required to file if ...
  - Income is above a threshold that varies according to filing and dependency status
  - Owe special taxes (AMT, social security tax on unreported tips, recapture taxes, etc.)
  - Have net self-employment earnings of more than \$400
  - Received advance premium tax credits payments
  - Certain other special cases

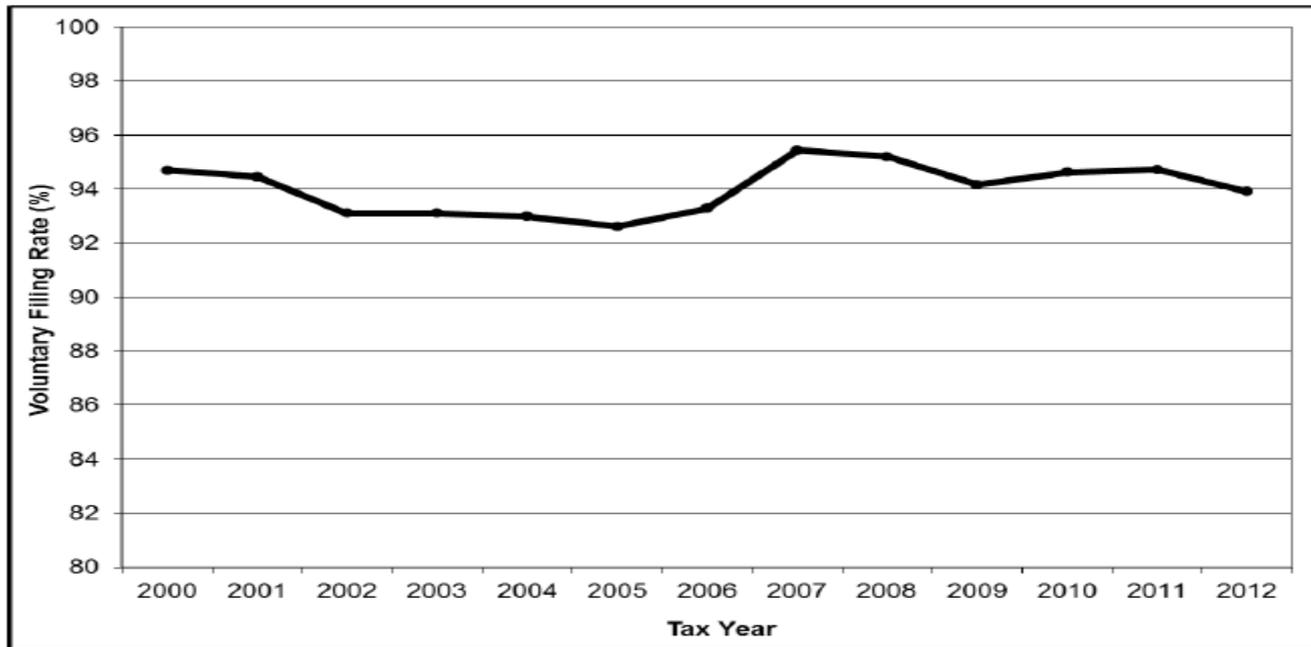
# IRS Concepts:

## Nonfilers and Voluntary Filing Rate

---

- Nonfiler (Ghost)
  - Return is required but not filed in a timely manner
  - Concept includes late filers who eventually file on their own or in response to enforcement action
- Voluntary Filing Rate (VFR)
  - Ratio of timely filed required returns to total required returns in the population

# Individual Income Tax Voluntary Filing Rate (VFR), Tax Years 2000-2012



# Nonfiling Tax Gap

---

- Individual income tax gap
  - Individual income taxes owed by taxpayers but not paid in a timely manner
- Individual nonfiling tax gap
  - Portion of the individual income tax gap attributable to net taxes owed by nonfilers
  - \$26 billion per year (spread across 6-7 million nonfilers)
  - 8.15% of overall individual income tax gap
- Nonfiling also accounts for an estimated \$4 billion per year in unpaid self-employment taxes

# What Drives Filing Compliance?

---

- Some potential factors:
  - Filing burden
  - Balance due/refund
  - Potential for refundable credits
  - Income level
  - State income tax filing requirement
  - Enforcement risk/Income visibility
  - Demographics, attitudes, tax knowledge

# Traditional Approach to Examining the Drivers of Behavior

---

- Qualitative Choice Model of Participation

$$P^* = \beta'X + \varepsilon \quad P = \begin{cases} 1 & \text{Participant} \\ 0 & \text{Non-participant} \end{cases}$$

- Estimate model using method of maximum likelihood (logit, probit, etc.)
- Then interpret estimate coefficients  $\beta$  to understand the how the various “drivers”  $X$  impact participation behavior.

# Problem with Traditional Approach

---

- For a traditional logit or probit analysis, we would require:
  1. A representative sample of both filers and nonfilers
  2. An indicator for filing status
- We can draw a representative sample of filers from tax return data, but this sample would only include filers
  - *Requirement 1 not satisfied*
- We can draw a representative sample of both filers and nonfilers from a Census survey, but this sample would not identify which respondents filed and which did not file
  - *Requirement 2 not satisfied*

Maybe there is a way to combine these two data sources...



# Calibrated Qualitative Choice Framework

---

As with a standard probit or logit model, let the conditional probability of filing be:

$$P(\text{File} = 1 | X_F) = P(\beta' X_F)$$

The relationship between the VFR and the conditional probability of filing is:

$$VFR = \sum_{x_F \in X_F} h(x_F) P(\beta' x_F),$$

where  $h(X_F)$  is the joint probability distribution of  $X_F$ .

# Probability of Inclusion in a Filer-Only Sample

---

Suppose we were to draw only a representative sample of filers. The probability that a taxpayer with characteristics  $x_{Fi}$  would be included in such a sample is:

$$\frac{\text{Pr(in population)} * \text{Pr(file|in population)}}{\text{Pr(file)}} = \frac{h(x_{Fi})P(\beta'x_{Fi})}{VFR},$$

The numerator represents the joint probability that a taxpayer would both file and have these characteristics, while the denominator represents the unconditional probability of filing in the population (i.e., the VFR).

The denominator of the above expression accounts for the fact that the sample is restricted only to filers.

# Estimation if $h(X_F)$ Were Known

Rather remarkably, if we knew  $h(X_F)$ , the joint distribution of  $X_F$ , we could actually estimate our model based on a filer-only subsample. The likelihood function would be:

$$\mathcal{L} = \prod_{i=1}^{N_1} \frac{P(\beta' x_{Fi})}{VFR}.$$

So we would solve:

$$\max_{\beta} \sum_{i=1}^{N_1} \ln[P(\beta' x_{Fi})]$$

subject to the constraint:

$$VFR = \sum_{x_F \in X_F} h(x_F) P(\beta' x_F)$$

# Estimation When $h(X_F)$ is Unknown

---

To impose the constraint in the preceding optimization problem:

$$VFR = \sum_{x_F \in X_F} h(x_F) P(\beta' x_F),$$

one would need to know  $h(x_F)$ , which is implausible.

To address this problem, we replace the exact constraint with its analog constructed from a supplementary random sample of filers and nonfilers based on the CPS:

$$VFR = \frac{1}{N_2} \sum_{j=1}^{N_2} w_j P(\beta' x_{Fj})$$

# Calibrated Qualitative Choice Model

---

So we solve:

$$\max_{\beta} \sum_{i=1}^{N_1} \ln[P(\beta' x_{Fi})]$$

subject to the constraint:

$$VFR = \frac{1}{N_2} \sum_{j=1}^{N_2} w_j P(\beta' x_{Fj}),$$

where  $N_1$  is the size of our filer-only sample from tax returns and  $N_2$  is the size of our supplementary sample of filers and nonfilers based on the CPS.

Can you see why we call this a “Calibrated Qualitative Choice Model”?

# Pooling Cross-Sections

---

We estimate our model using a time series of cross-sections covering tax year 2000 through tax year 2012.

So we actually solve:

$$\max_{\beta} \sum_{t=1}^T \sum_{i=1}^{N_{1t}} \ln[P(\beta' x_{Fit})]$$

subject to the  $T$  constraints:

$$VFR_t = \frac{1}{N_{2t}} \sum_{j=1}^{N_{2t}} w_{jt} P(\beta' x_{Fjt}) \quad t = 1, \dots, T$$

# Some Issues

---

- Need to restrict the CPS sample to required returns
  - Done based on our team's prior work in measuring the VFR
- Explanatory variables need to be present in both data sources
  - Rules out certain variables, such as presence of a refund/balance due
- Explanatory variables also need to be consistently measured across the two data sources
  - Rules out some additional variables, such as filing status and earned income credit eligibility

# Key Findings

---

- Demographics
  - Filing is more relatively likely if elderly and less likely if married
- Income
  - Filing is positively associated with gross income and negatively associated with being self-employed
  - Filing is relatively less likely if gross income is near the filing threshold
- Burden
  - Filing is negatively associated with the burden of preparing and filing a return, but less so near filing threshold

# Key Findings, Continued

---

## ■ Incentives

### – Economic Stimulus

- Controlling for other factors, filing increased by about 1 percentage point in tax year 2007 in response to the Economic Stimulus, but the impact was temporary.

### – Expanded EIC for 3 or more children

- Controlling for other factors, filing increased among households with 3 or more children from tax year 2009 on (period with extra earned income credit for 3<sup>rd</sup> child).



# **Searching for Ghosts Redux: Improved Methodologies for Estimating the Nonfiling Tax Gap**

***Pat Langetieg, Mark Payne, and Alan Plumley***  
*(IRS Research, Applied Analytics, and Statistics: Office of Research)*

***IRS-TPC Research Conference***  
***June 23, 2016***

# Individual Income Tax Nonfiling Gap Estimation

**Nonfiling Gap:** True tax liability not paid on time by those who do not file on time

- Includes both Late Filers and Not-filers

	Administrative Data Method	Census Method
Dataset	IRS population (1040s, information documents)	CPS data matched to IRS data using improved unique identifiers
Income	Use information documents; Impute SE income	Impute income to CPS based on IRS data
Tax units	Impute based on CPS profiles	Use CPS demographics
Tax benefits	Impute	Impute

# IRS Administrative Method: Not-Filers

- **Impute self-employment income** (based on \$ reported on filed returns)
- **Randomly assign individuals to families / tax units**
  - Based on demographic profiles (gender, age group, marital status, no. of dependents) in CPS-ASEC.
- **Impute** adjustments, deductions and credits (based on NRP data)
- **Calculate tax and balance due**
- **Derive stochastic averages**
  - Five replicates for family unit assignment with unique imputations
  - Extreme outliers removed (bad data)
  - Estimate is the average of the middle seven of 25 one-percent samples

# Administrative Method Estimates of the Not-Filer Gap, Tax Years 2008-2010<sup>†</sup>

Key Items	Amount (\$B)
Number of required returns among not-filers (millions)	7.4
Total income	\$216.3
Total adjustments, exemptions, and deductions that offset income*	\$93.0
Total taxable income	\$123.2
Tentative income tax after nonrefundable credits*	\$19.7
Self-employment tax	\$6.9
Net tax due	\$26.6
Total prepayments* refundable credits	\$9.0
<b>Total nonfiling gap of not-filers</b>	<b>\$17.6</b>

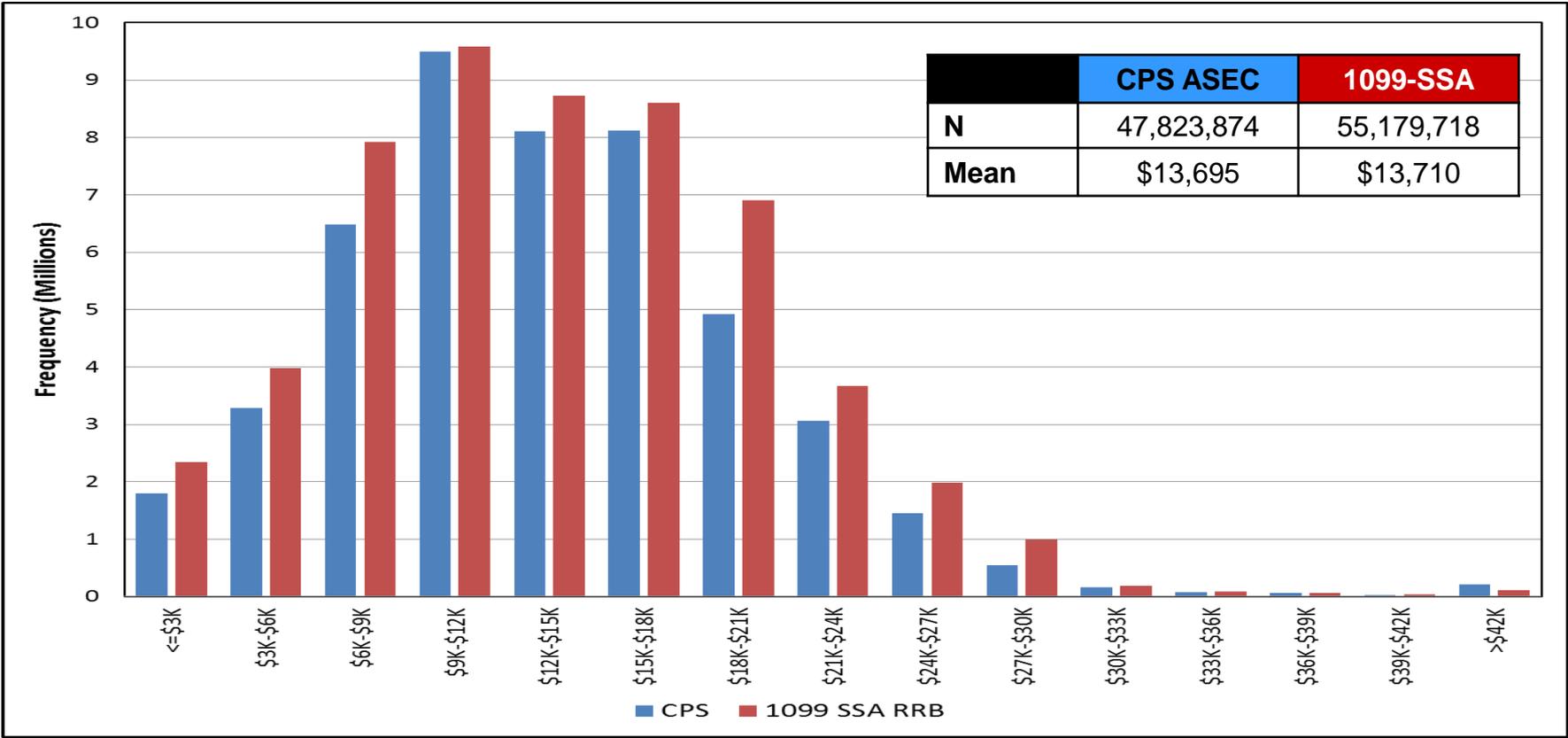
<sup>†</sup> Estimates averaged over Tax Years 2008 through 2010.

\* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

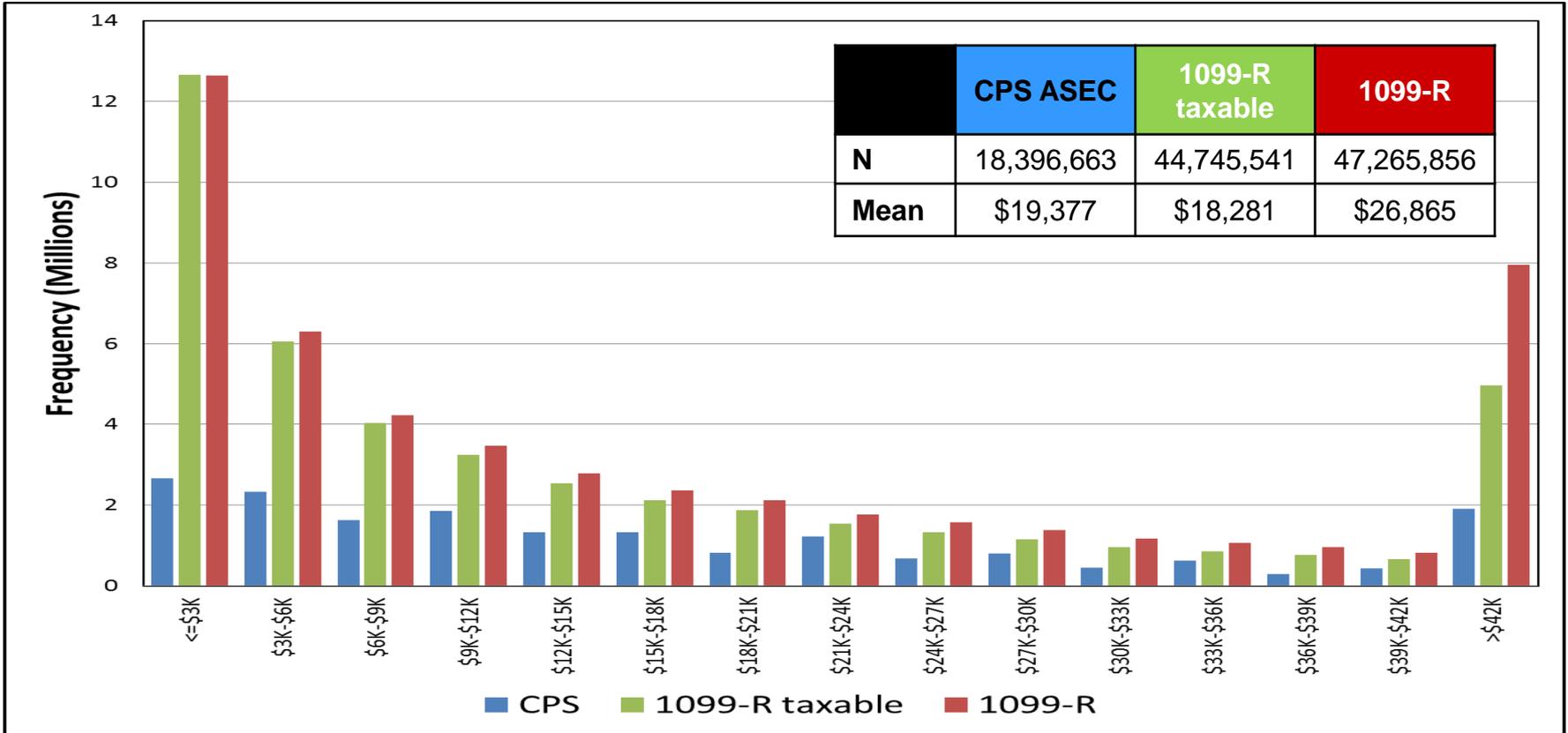
# Census Method: Not-Filers

- **Restrict to matched CPS-IRS data** (good matches with unique income) and **re-weight the sample**
- **Impute income** (based on IRS data and matched third-party information)  
Why??

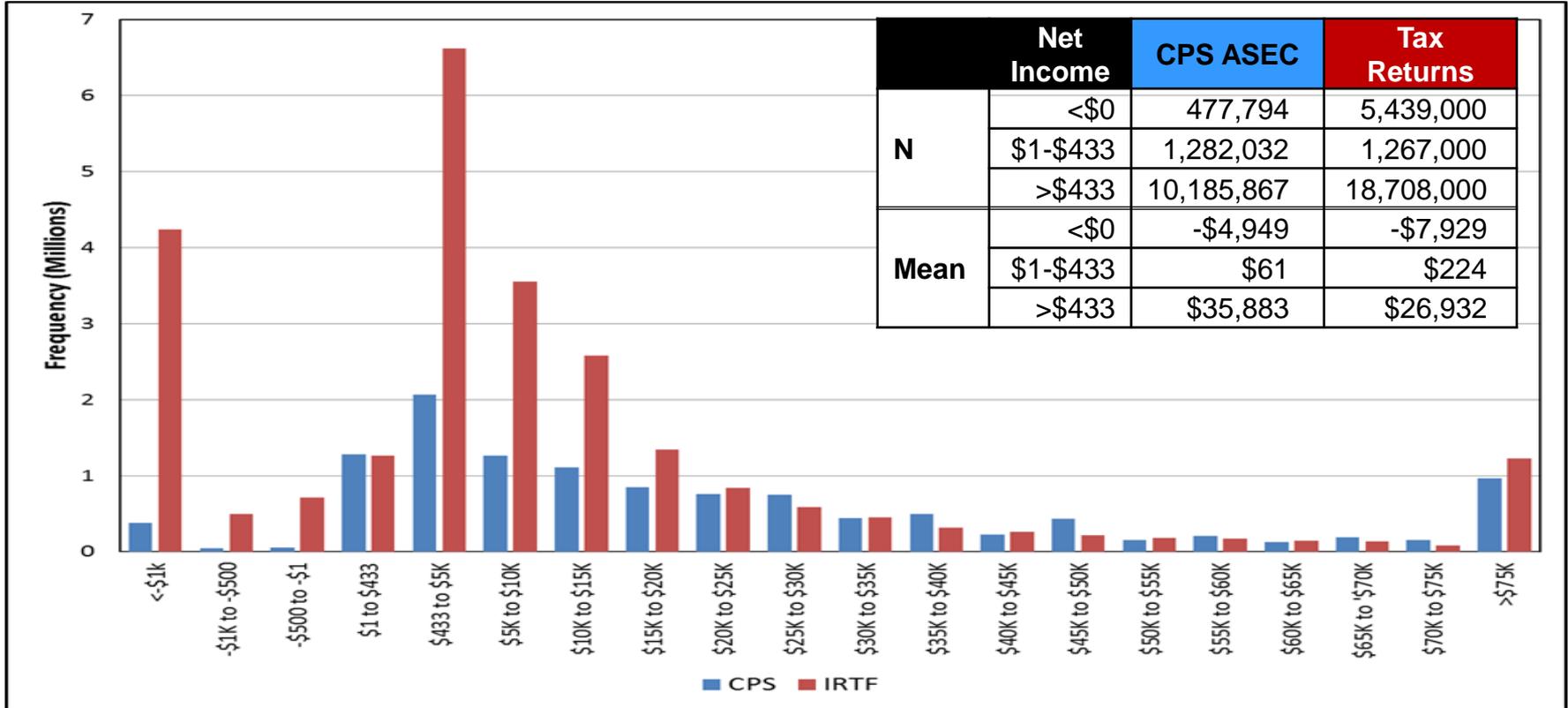
# Social security income not heavily underreported in the CPS



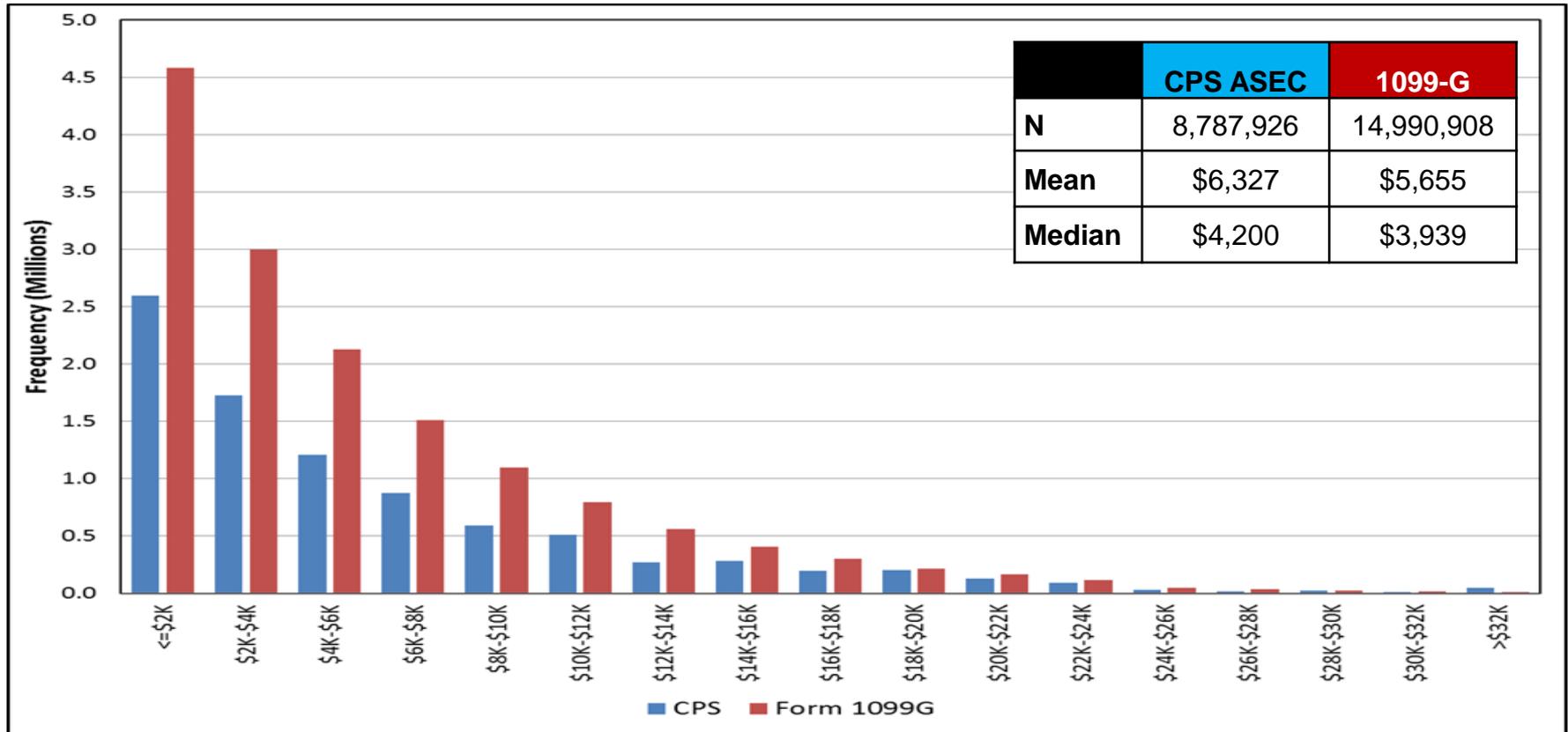
# But incidence of pension income much lower in the CPS



# Self-employment income underreported even compared to tax returns



# Unemployment compensation also underreported in the CPS



# Census Method: Not-Filers (continued)

- Restrict to matched CPS-IRS data (good matches with unique income) and re-weight the sample
- Impute income (based on IRS data and matched third-party information)
- **Group individuals into tax units** (based on CPS household records)
- **Impute tax benefits** (using models developed from NRP data)
- **Estimate prepayments and refundable credits** (using aggregate ratios of these to total tax from the Administrative Data Method)

# Census Method Estimates of the Not-Filer Gap, Tax Years 2008-2010<sup>†\*</sup>

Key Items	Amount (\$B)
Number of required returns among not-filers (millions)	10.9
Total income	\$421.4
Total adjustments, exemptions, and deductions that offset income*	\$143.4
Total taxable income	\$278.0
Tentative income tax after nonrefundable credits*	\$46.6
Self-employment tax	\$8.6
Net tax due	\$55.2
Total prepayments* refundable credits	\$28.8
<b>Total contribution to the nonfiling gap</b>	<b>\$26.4</b>

<sup>†</sup> Estimates averaged over Tax Years 2008 through 2010.

\* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

# Estimating the Tax Gap of Late Filers

- **Adjust income** by matching to information returns
- **Recalculate tax**
- **Recalculate balance due** after prepayments
- **Derive estimates** from the average of multiple large samples from the population to lessen the effects of data errors

# Estimates of tax gap for late filers, tax years 2008-2010<sup>†\*</sup> (\$ Billions)

Key Items	Census Method (Early Late Filers)	Admin Method (All Late Filers)
Number of required returns among not-filers (millions)	3.1	6.8
Total taxable income	\$119.9	\$242.2
Net tax due	\$25.6	\$50.7
Total prepayments* and refundable credits	\$21.3	\$39.4
<b>Total contribution to the nonfiling gap</b>	<b>\$4.4</b>	<b>\$11.3</b>

<sup>†</sup> Estimates averaged over Tax Years 2008 through 2010.

\* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

# Individual Income Tax and Self-Employment Tax Nonfiling Gap Estimates (\$ Billions)

	Average 2008-2010
<b>Final Nonfiling Gap Estimate*</b>	<b>29.8</b>
<b>Census Method</b>	<b>30.8</b>
Not-Filers	26.4
Late Filers	4.4
<b>Administrative Data Method</b>	<b>28.9</b>
Not-Filers	17.6
Late Filers	11.3

\*The portion of this attributable to self-employment is \$3.8 billion, assuming that payments are allocated to income tax and self-employment tax proportional to the magnitude of tax liability

## Number of required returns and total income are similar between the two methods <sup>§ †</sup>

Key Items	Administrative Data Method	Census Method	Amount Difference	Percentage Difference
Number of required returns (millions)	14.2	14.0	0.2	1.7%
Wages	\$356.2	\$368.8	-\$12.6	-3.5%
Interest	\$11.5	\$13.2	-\$1.7	-14.9%
Dividends	\$10.9	\$29.8	-\$18.8	-171.9%
Schedule C net income	\$70.2	\$65.6	\$9.5	13.5%
Schedule D net income	\$18.0	\$9.4	\$8.6	47.9%
Taxable IRA and pension income	\$61.3	\$63.6	-\$2.3	-3.7%
Schedule E net income	\$21.3	\$18.8	\$2.4	11.4%
Unemployment compensation	\$14.1	\$11.9	\$2.2	15.4%
Taxable SSI income	\$12.7	\$12.2	\$0.5	4.2%
Other income	\$3.9	-\$1.1	\$5.0	128.0%
<b>Total income<sup>†</sup></b>	<b>\$591.8</b>	<b>\$604.7</b>	<b>-\$12.9</b>	<b>-2.2%</b>

<sup>§</sup> Estimates averaged over Tax Years 2008 through 2010.

<sup>†</sup> The Total Income amount is slightly larger than the sum of the components because Total Income cannot be less than zero on any given return.

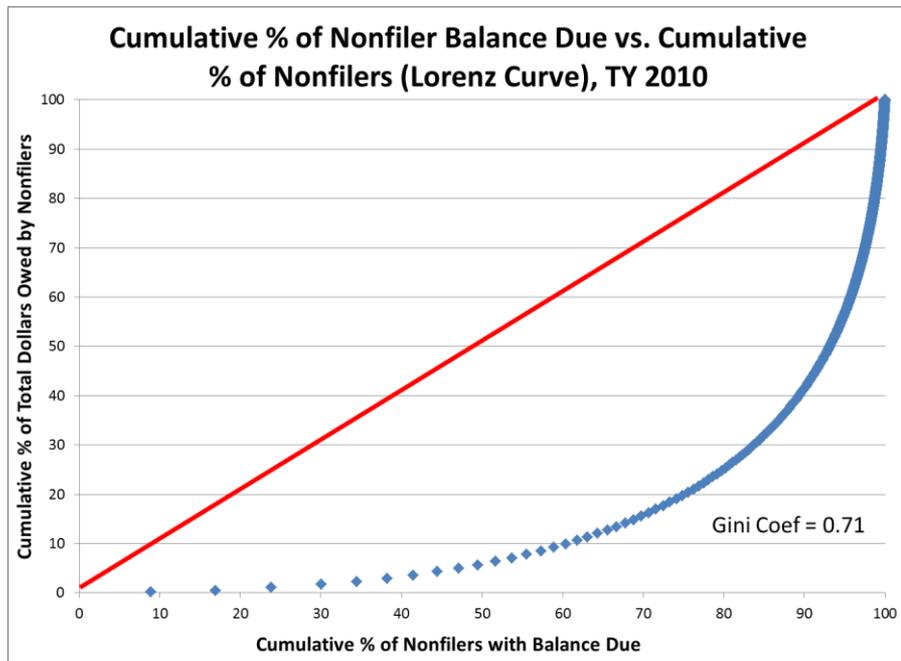
## The nonfiling tax gap estimates for the two methods are close <sup>§</sup>

Key Items	Administrative Data Method	Census Method	Amount Difference	Percentage Difference
Taxable income	\$365.4	\$397.9	-\$32.5	-8.9%
Net tax due	\$77.3	\$80.8	-\$3.6	-4.6%
Total payments of tax*	\$48.4	\$50.1	-\$1.7	-3.5%
<b>Total nonfiling gap</b>	<b>\$28.9</b>	<b>\$30.8</b>	<b>-\$1.9</b>	<b>-6.5%</b>

<sup>§</sup> Estimates averaged over Tax Years 2008 through 2010.

\* Income (and tax) offsets were limited to the amount needed to reduce income (or tax) to zero.

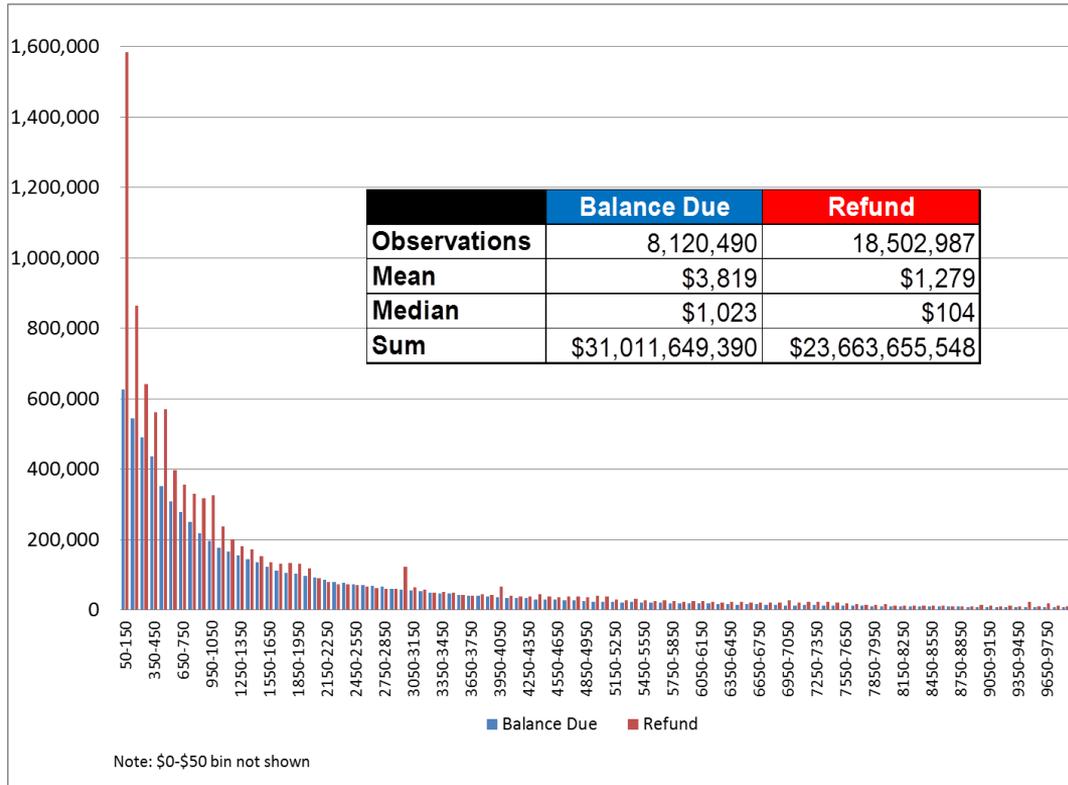
## A small share of nonfilers is responsible for a large share of tax owed



Percentile	Balance Due by Decile All Nonfilers (\$ Billions)	Balance Due by Decile Not-filers Only (\$ Billions)	% Share of Balance Due by Decile All Nonfilers
10	0.0	0.0	0.1%
20	0.2	0.2	0.5%
30	0.2	0.2	0.8%
40	0.4	0.3	1.4%
50	0.7	0.5	2.2%
60	1.0	0.6	3.4%
70	1.5	0.9	5.3%
80	2.4	1.5	8.2%
90	4.3	2.7	14.4%
100	18.7	11.4	63.7%

Top 10% responsible for 64% of tax gap; top 20% responsible for 78%.

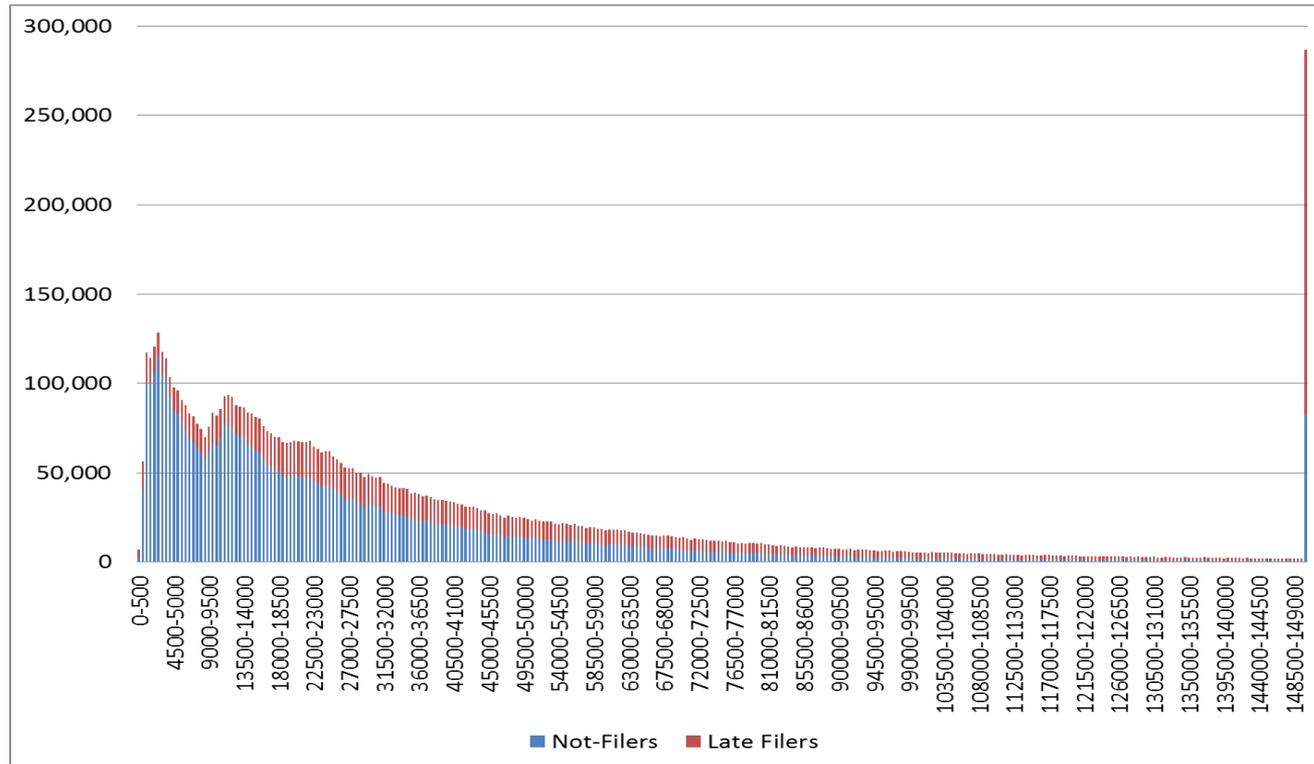
# If refunds considered, nonfiler returns would increase net revenues by \$7.4 Billion in TY 2010;



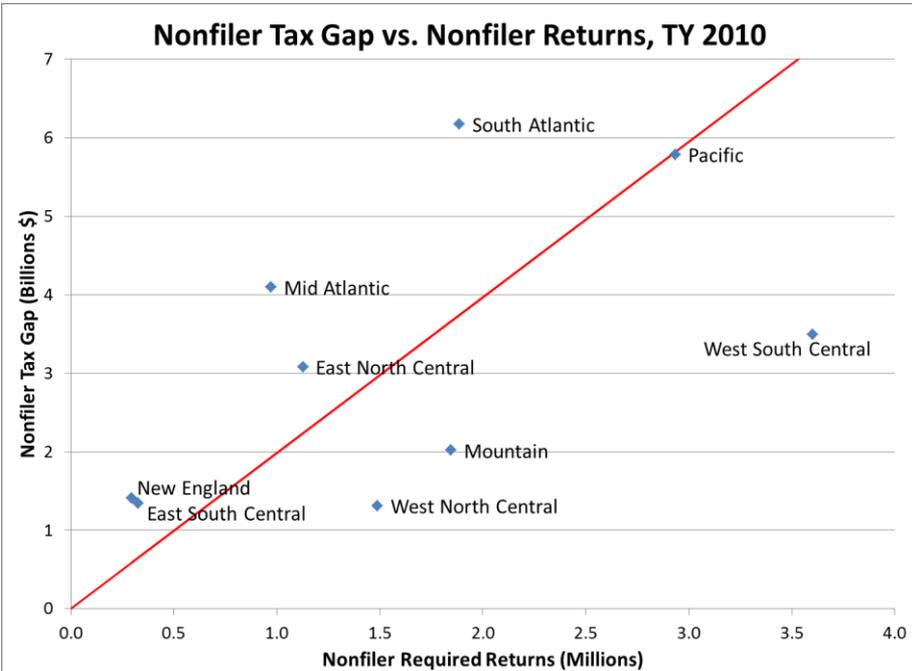
## Net Balance Due After Refunds

**Not-filers**    \$13.8 B  
**Late filers**   -\$6.4 B  
**Net**                \$7.4 B

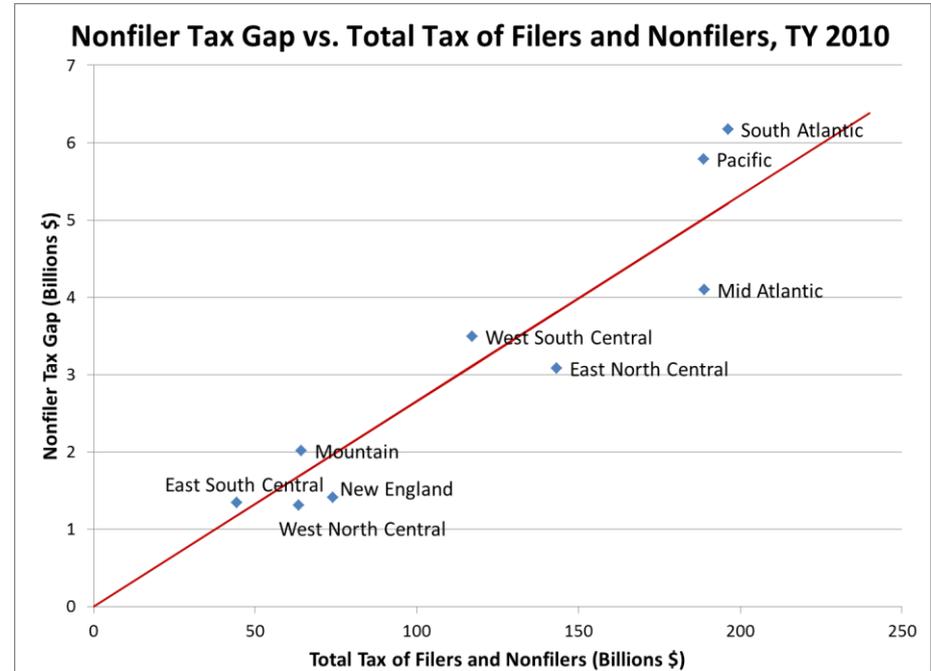
# Distribution of total income among nonfilers has long tail (Over 600,000 with more than \$100,000 in income) (TY2010)



# Nonfiler Tax Gap vs. Required Returns and vs. Total Tax of Filers and Nonfilers, by Region

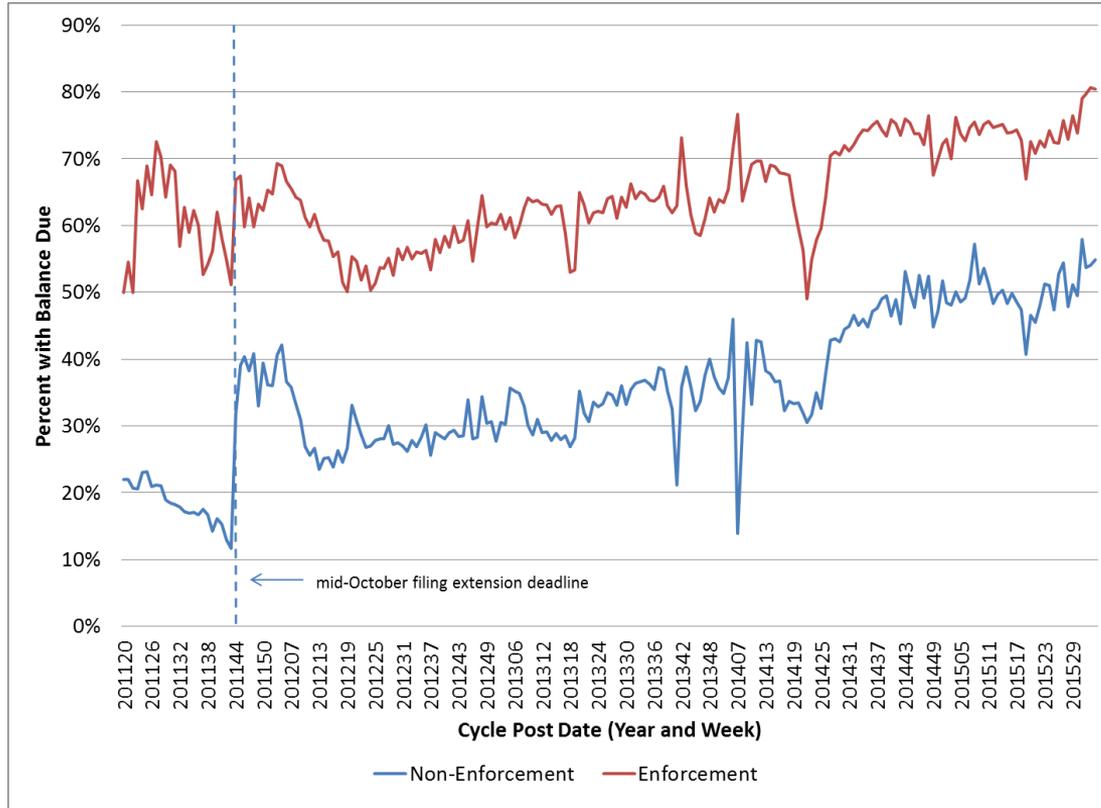


Avg tax gap per required return = \$1,984

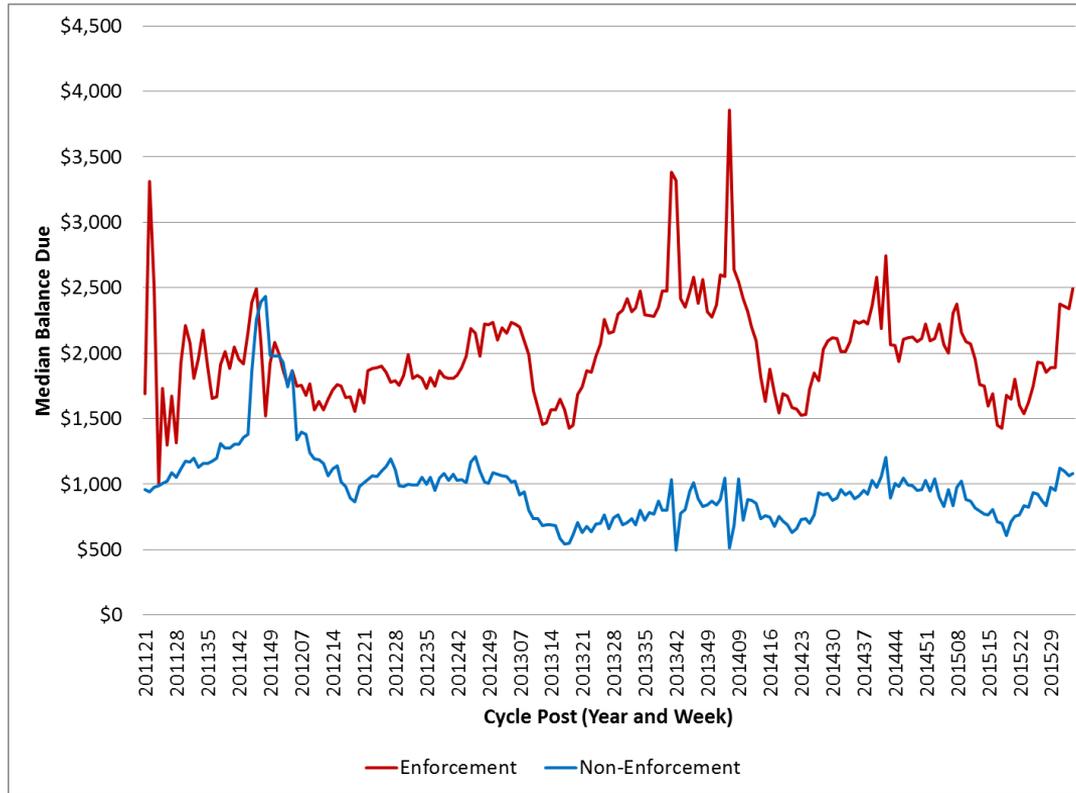


Avg nonfiler tax gap percent of total tax = 2.7%

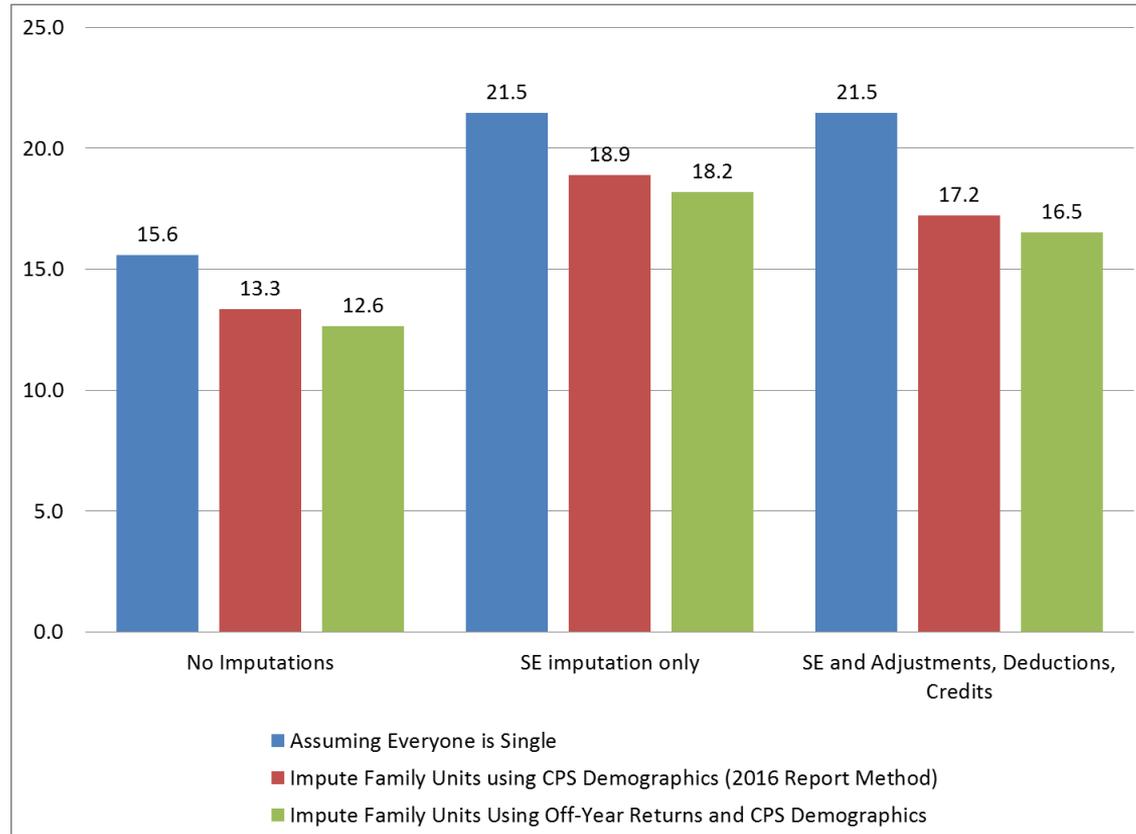
# The % of late returns with a balance due is higher for enforcement cases and is greater for later cycle post dates



# The median balance due for late filers targeted with a nonfiler notice is higher



# Not-Filer Portion of TY2010 Gap under Different Assumptions, Admin Method



# Future Directions

- Improve imputations of tax units in IRS Administrative Data Method using tax return and SSA data
- Explore ways to improve nonfiler workload selection and to identify nonfilers who may be eligible for benefits
- Explore methods for correcting income rounding in CPS.



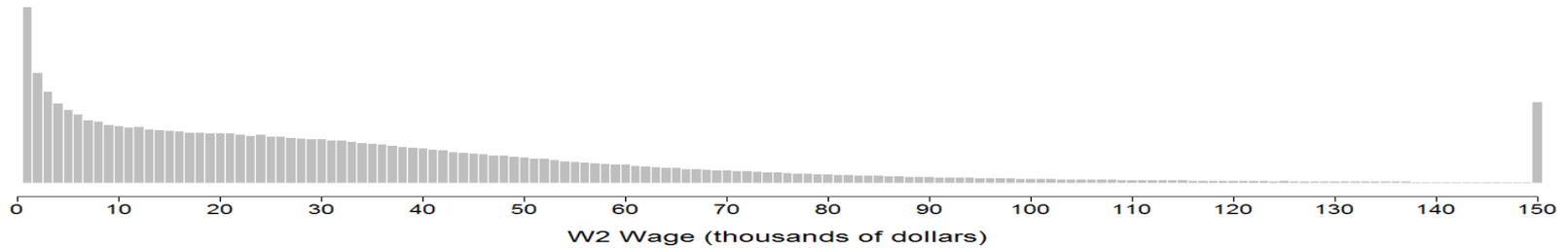
# Handling Respondent Rounding of Wages Using the IRS and CPS Matched Dataset

Minsun K. Riddles, Sharon L. Lohr, J. Michael Brick, Westat  
Patrick T. Langetieg, John M. Payne, Alan H. Plumley, IRS

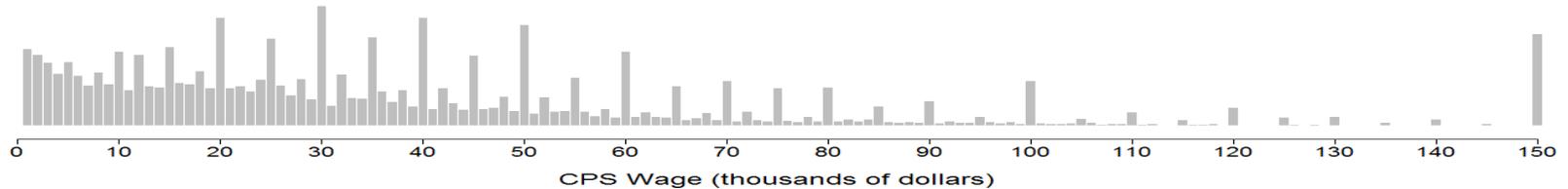
6th Annual IRS-TPC Joint Research Conference on Tax Administration  
June 23, 2016

# Distribution of Wage Income: Two Sources

## W-2 Forms



## Current Population Survey



# Sources of Wage Data: **W-2** Forms

- Employer files W-2 for each employee who
  - Had wages from employer, including noncash payments
  - Had withheld income, Social Security, or Medicare tax
  - Would be subject to income tax withholding if he or she had not claimed additional allowances or exemptions
- Not rounded wages

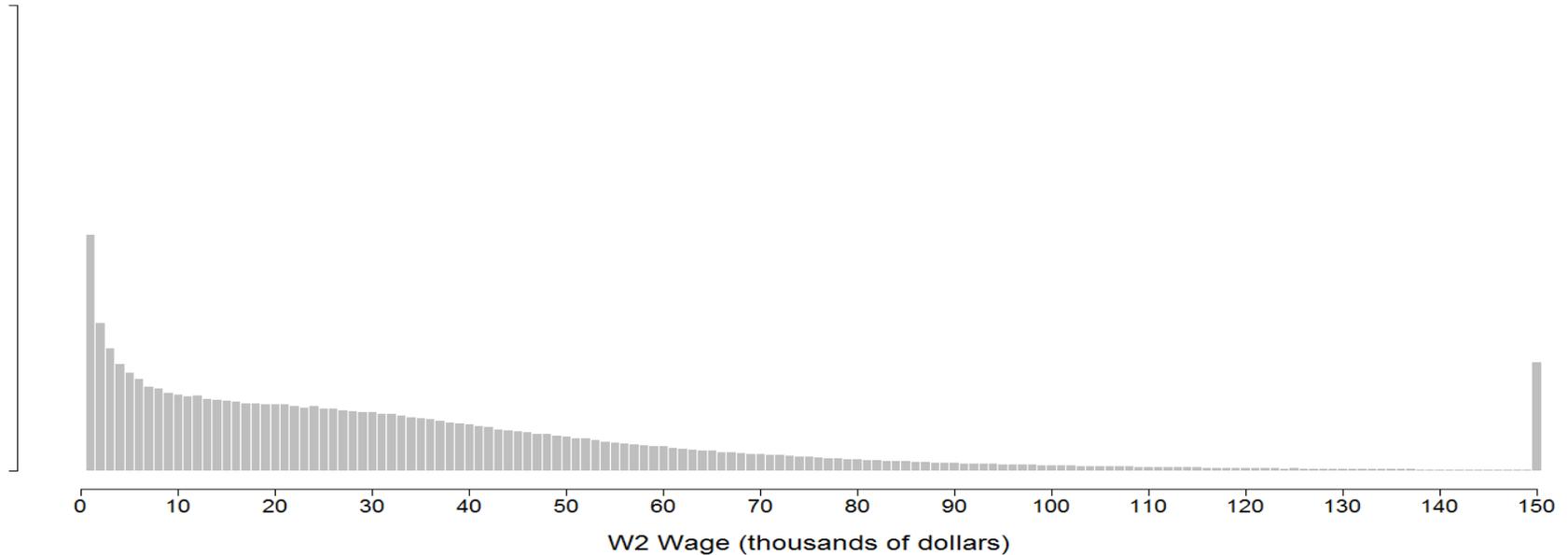
# Current Population Survey

- Household survey: primary source of labor force statistics
- Civilian non-institutionalized population age 16+
- Earnings before tax and deductions
  - Wage earnings from longest job
  - Other wage earnings from any other work
- Wage distribution has spikes

# Motivating Problem

- How close are estimated distributions of wages from
  - **W-2** forms (unrounded)
  - **CPS** data (rounded, with measurement error)
- Approach
  - Model density of **W-2** wages
    - Full population
    - Records that are in **CPS**, using **CPS** weights
  - Model density of **CPS** wages
    - Component 1: density of “true” values
    - Component 2: rounding mechanism for respondents

# W-2 Wage Distribution: W-2 Population (N=151M)

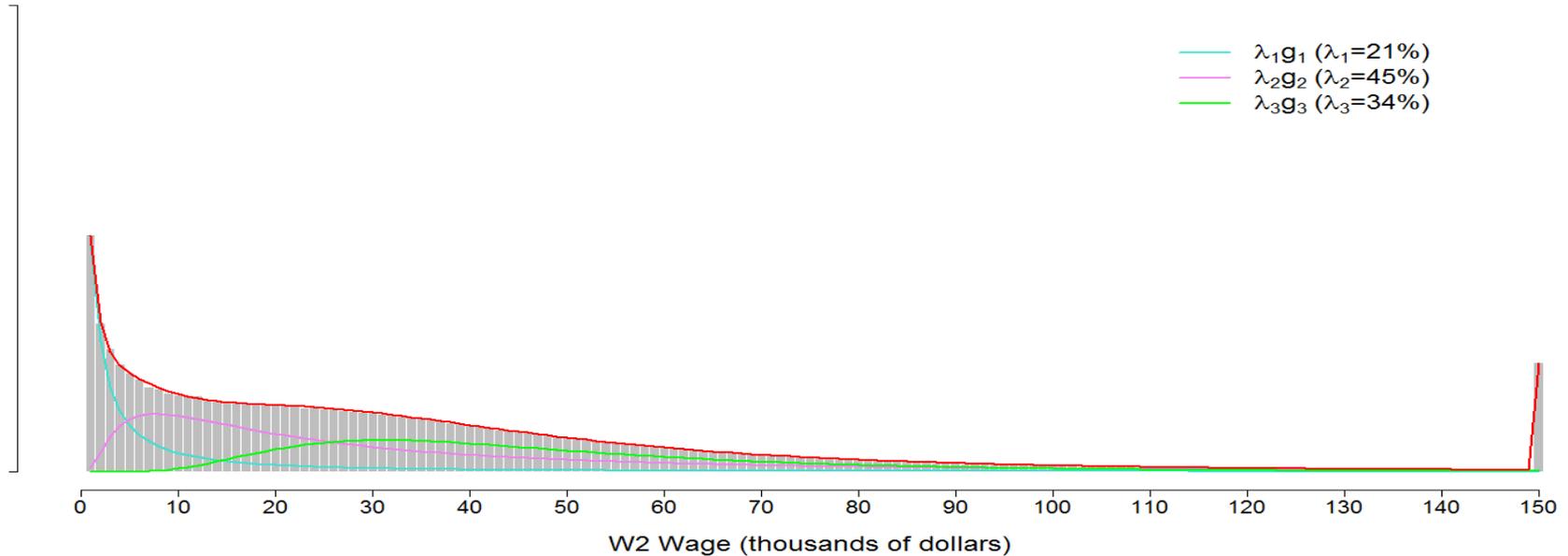


# Maximum Likelihood: Density for W-2 Wages

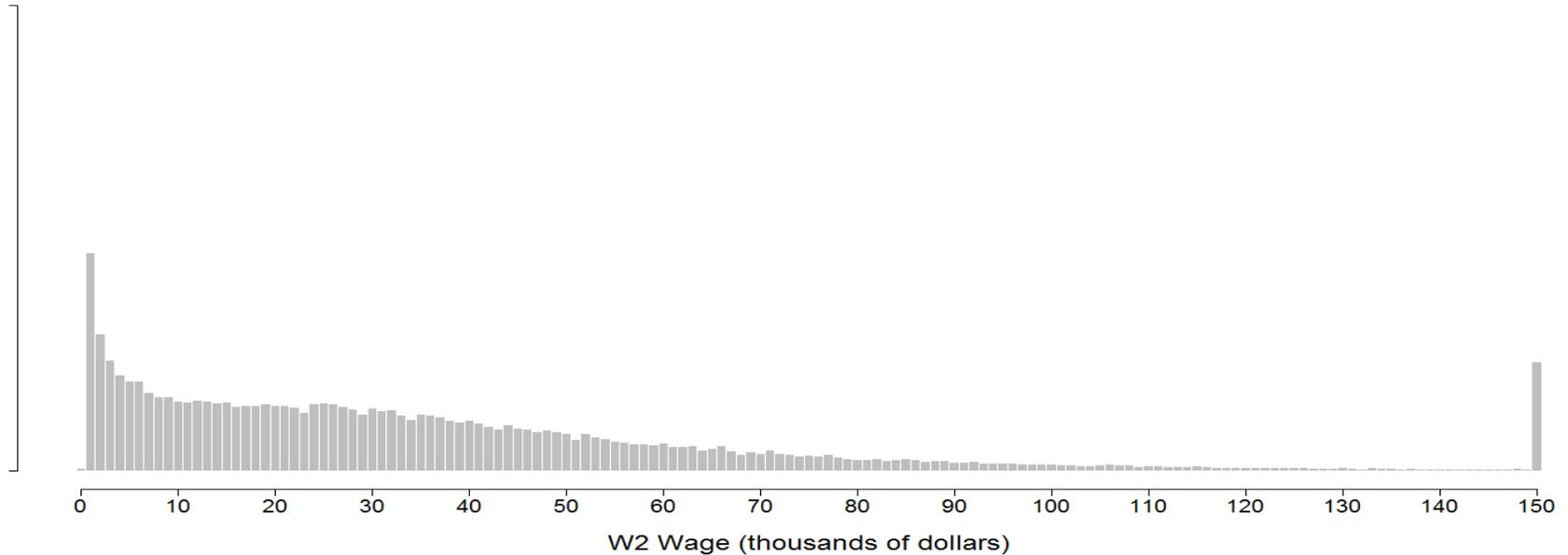
- X: True value of wages
- Parametric density  $f(x|\theta)$
- Mixture of three lognormal distributions
  - Smooth
  - Highly skewed
  - Multimodal

# Underlying Parametric Density

## Fitted W-2 Wage Distribution: W-2 Population

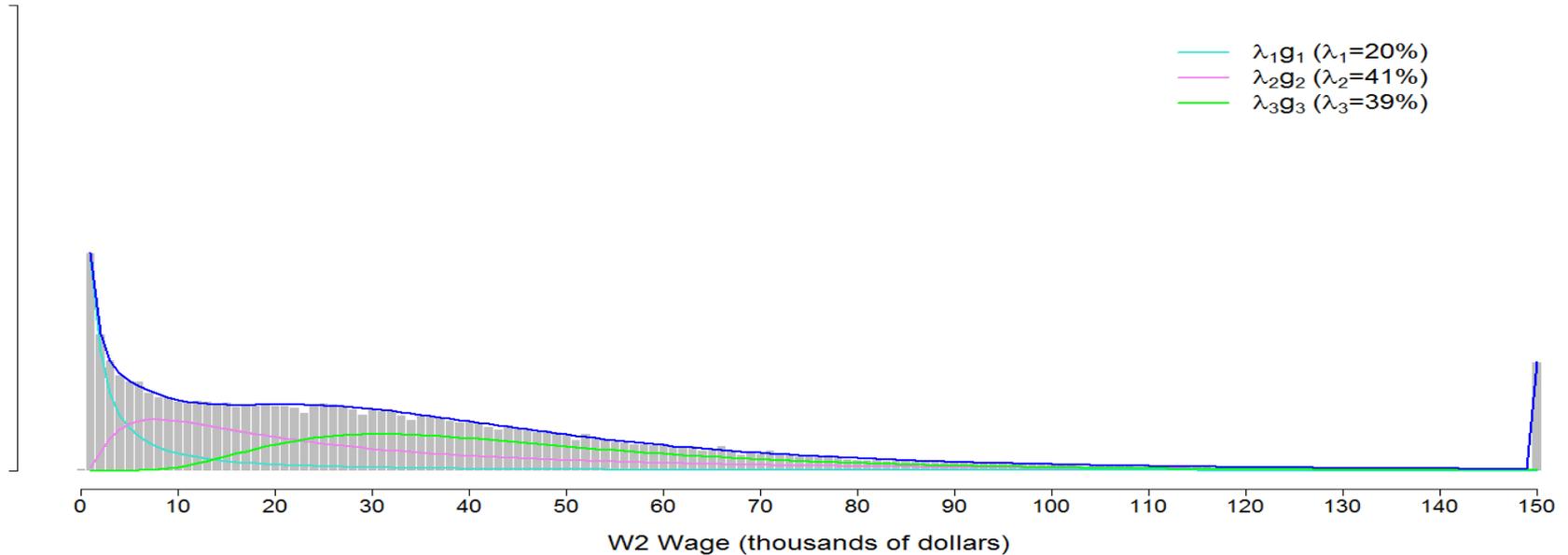


# W-2 Wage Distribution: CPS-IRS Matched (n=78K)



# Underlying Parametric Density

## Fitted W-2 Wage Distribution: CPS-IRS Matched



# Smoothing the CPS Distribution

## 1. Underlying parametric density $f(x|\theta)$

- $x$  = true value of wages
- Mixture of three lognormal distributions

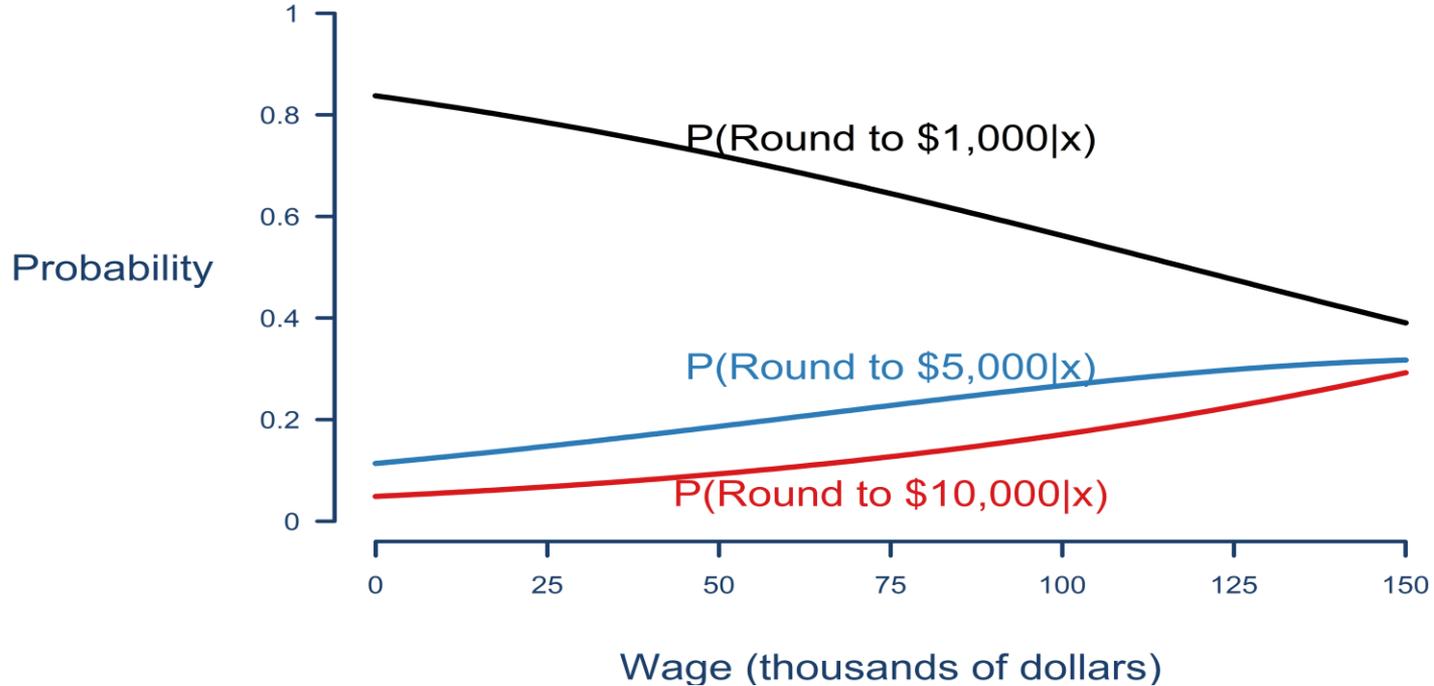
## 2. Model for rounding mechanism

$$\left\{ \begin{array}{l} P(\text{round to nearest } \$1,000 | x, \gamma) \\ P(\text{round to nearest } \$5,000 | x, \gamma) \\ P(\text{round to nearest } \$6,000 | x, \gamma) \\ P(\text{round to nearest } \$10,000 | x, \gamma) \\ P(\text{round to nearest } \$50,000 | x, \gamma) \end{array} \right.$$

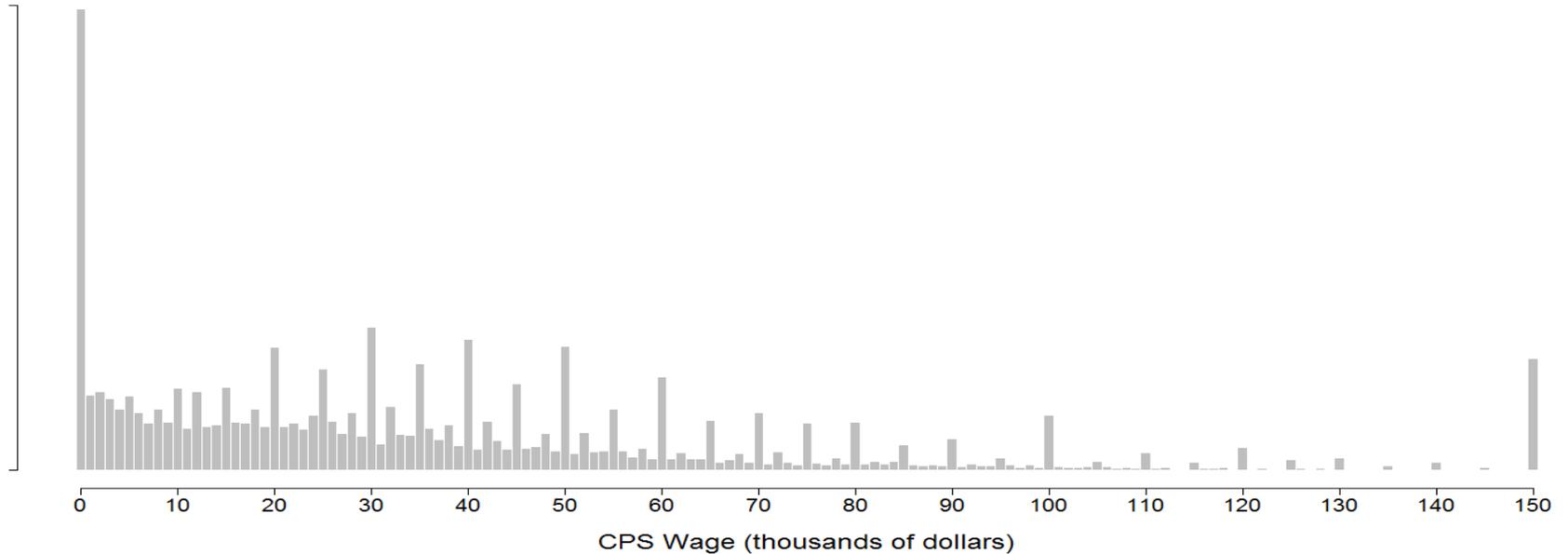
- Depends on “true” value of  $x$  for the person

# Model for Rounding Mechanism

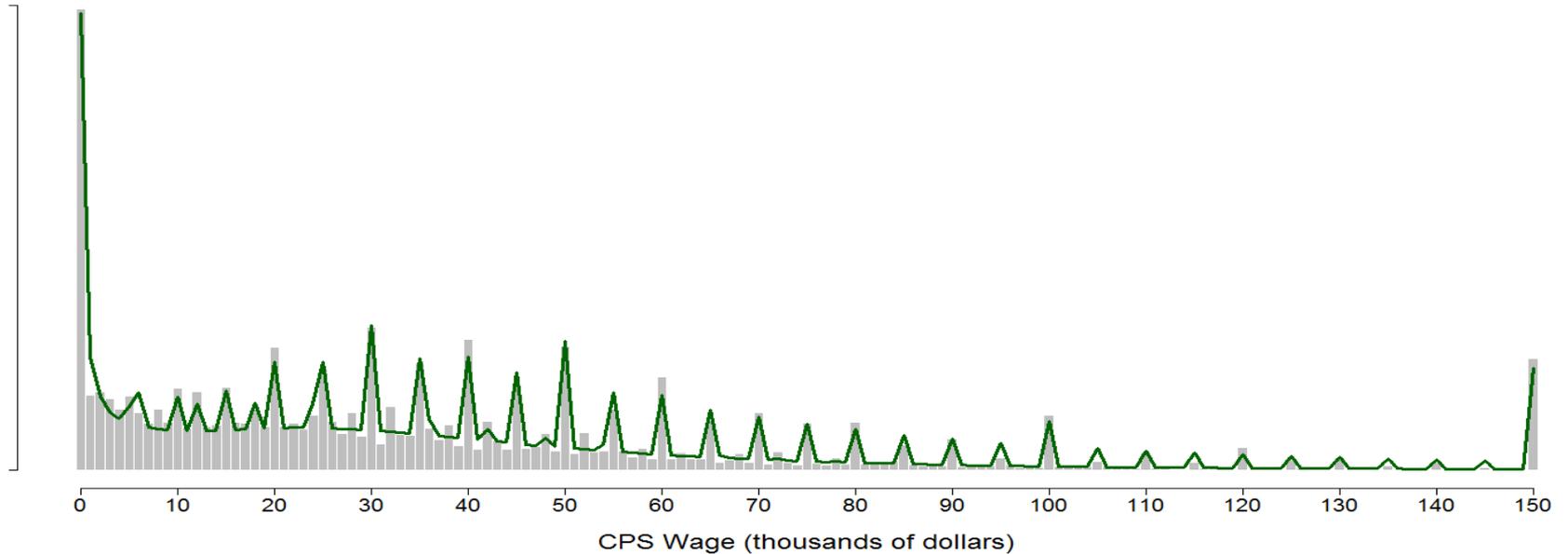
assume higher wage, more rounding



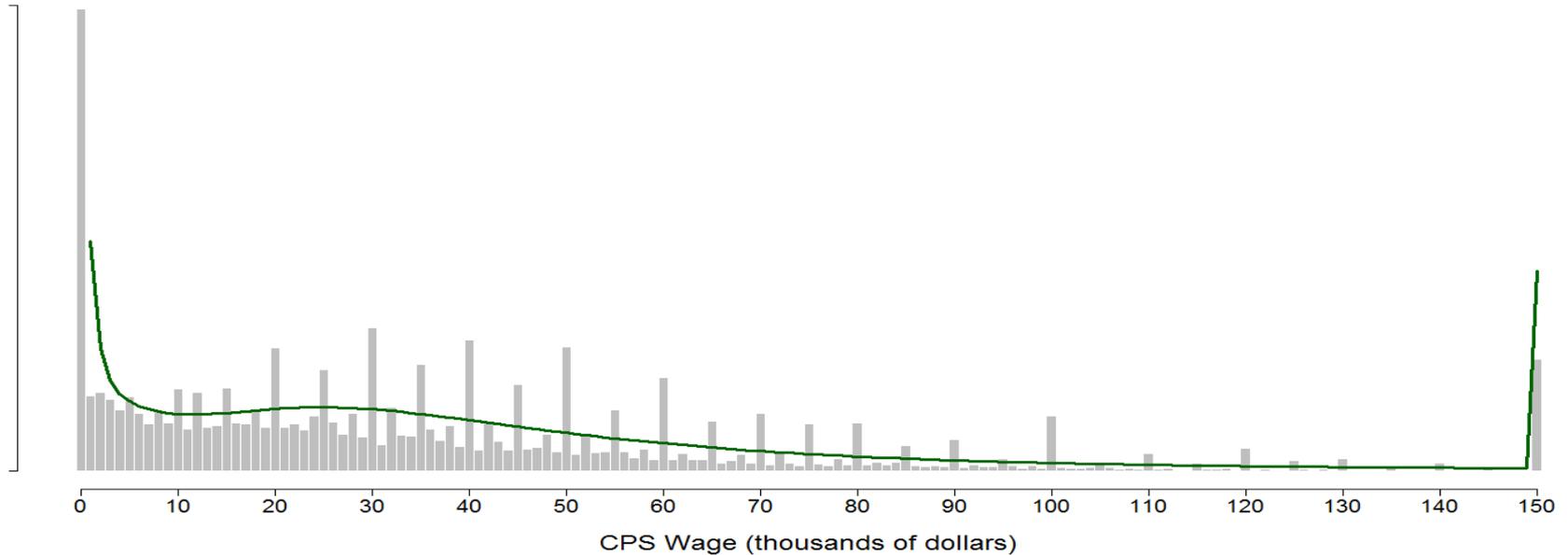
# CPS Wage Distribution: CPS-IRS Matched



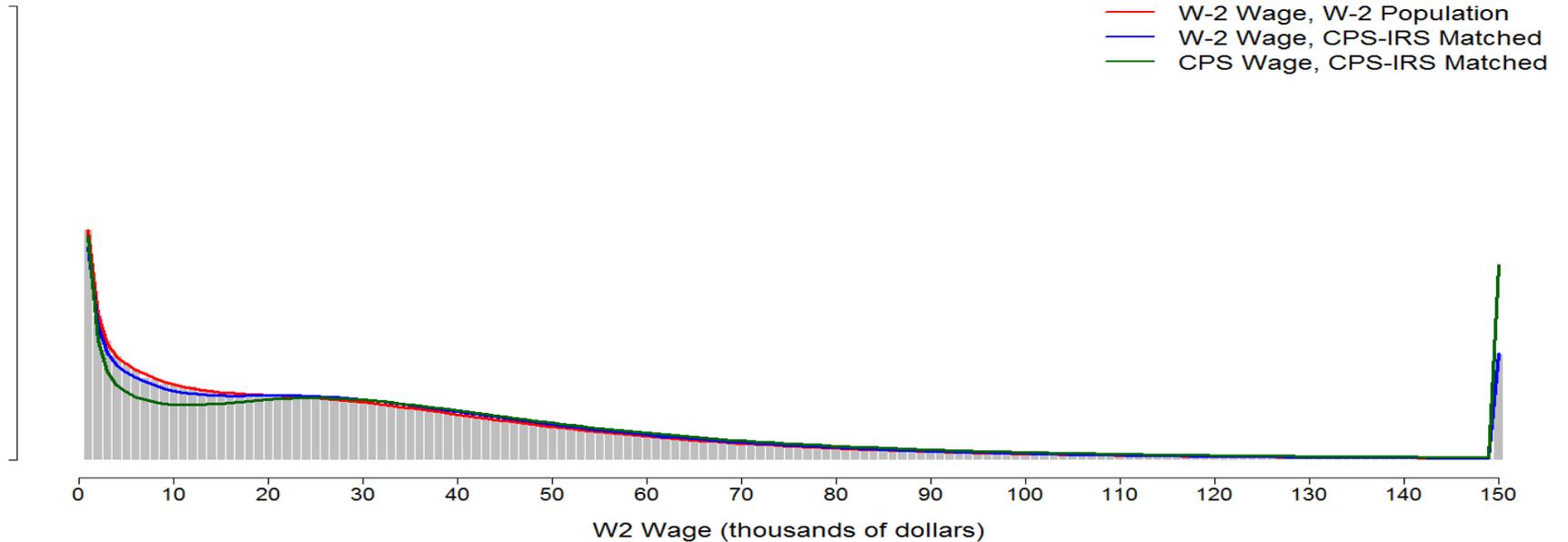
# Fitted CPS Wage Distribution: CPS-IRS Matched



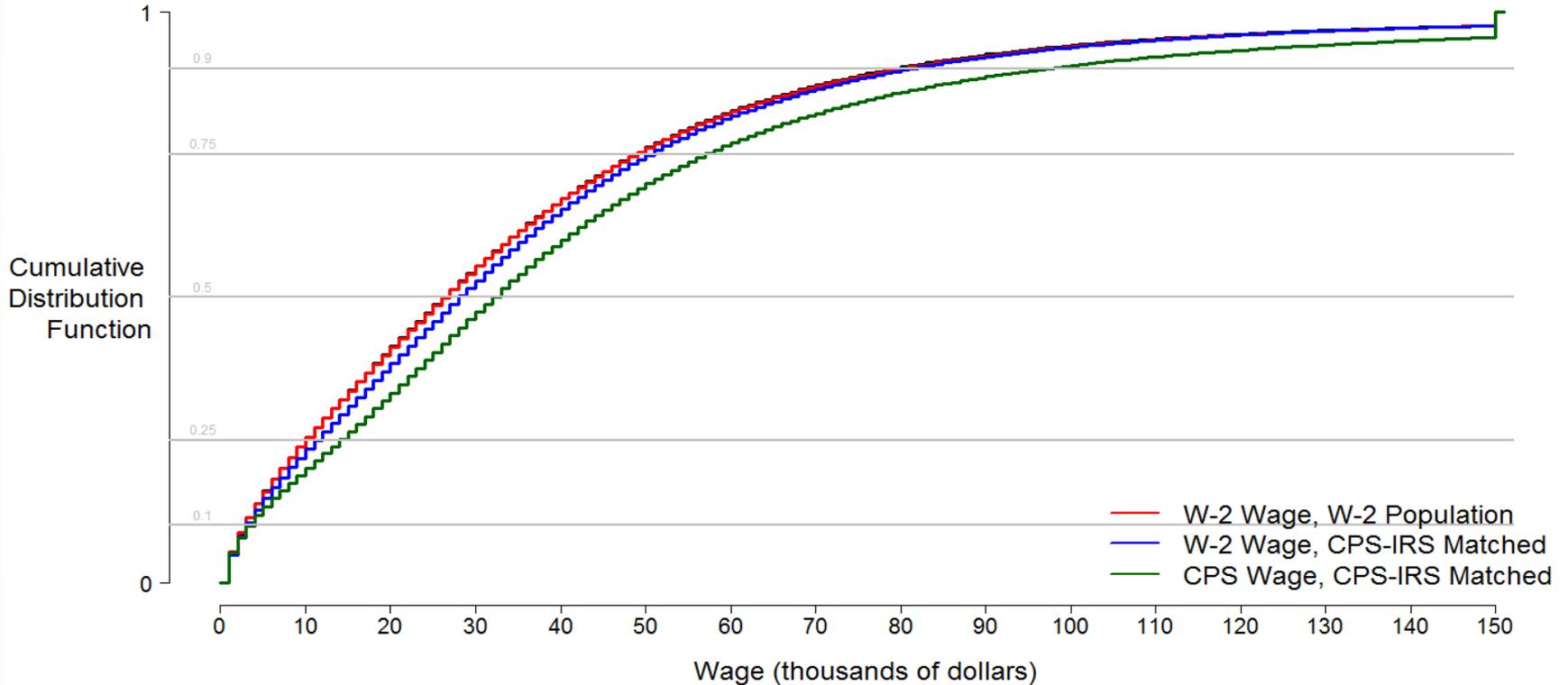
# CPS Underlying Parametric Density $f(x|\hat{\theta})$



# Fitted W-2 Wage Distribution: W-2 Population



# Fitted Wage Distribution



# Conclusions and Future research

- Estimated density of W-2 data: mixture of 3 lognormals
- Estimated density of CPS data
  - Mixture of 3 lognormals (different parameters)
  - Model the rounding mechanism
- All models fit the data well
- Models are flexible, can include other covariates if desired (e.g. occupation)
- Allows exploration of differences between W-2 and CPS distributions

# Contact Information

Minsun K. Riddles [MinsunRiddles@westat.com](mailto:MinsunRiddles@westat.com)