Welcome
Statistics of Income
2015 Consultants Panel Meeting

Barry Johnson
Director Statistics of Income Division
Thanks to the Committee on National Statistics for hosting

Special thanks to Dr. Connie Citro, Director of the Committee on National Statistics, National Academy of Science
A few logistics

• Facilities
• Lunch
• Travel Vouchers (please see Wanda Robinson)
• Thanks for help organizing meeting:
  • SOI Staff: Brian Balkovic, Paul Bastuscheck, Melissa Belvedere, Rose Defalco, Martha Harris, Wayne Kei, Melissa Ludlum, Clay Moulton, David Paris, Wanda Robinson, Georgette Walsh
  • NSF Staff: Eileen LeFurgy, Jesse Willis
• If you need anything during the day, please see Brian Balkovic, Rose Defalco, or Wanda Robinson
SOI Panel Members

Michael Allen
Rosanne Altshuler
Jenny Bourne
Peter Brady
Len Burman
Martin David
Daniel Feenberg
John Graham

Julia Lane
Jim Nunns
George Plesko
Fritz Scheuren
Lin Smith
Robert Strauss
Michael Udell
Patricia Whitridge
(guest panelist)
Updates

RAS
• Rosemary Marcuss retired, Alain Dubois is now acting Director
• Research, Analysis and Statistics/Office of Compliance Analytics redesign

SOI
New or improved products detailed on preread
• Resources
• Communications strategy
  • Web redesign efforts
  • Infographics
  • Updated SOI Bulletin articles
• .Net conversion
• 5-Year Business Plan
  • Interviews with SOI staff and internal stakeholders
  • Interviews with external stakeholders
• Acting Director July 20-October 2
# 2015 Consultants Panel Agenda

## What's New at SOI?

<table>
<thead>
<tr>
<th>Session</th>
<th>Presenter</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welcome</td>
<td>Barry Johnson</td>
<td>9:00 am</td>
</tr>
<tr>
<td>Joint Statistical Research Program Update</td>
<td>Mike Weber</td>
<td></td>
</tr>
<tr>
<td>A New Approach to Producing Migration Data</td>
<td>Kevin Pierce</td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>Panel</td>
<td></td>
</tr>
</tbody>
</table>

## Break

<table>
<thead>
<tr>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:55 am</td>
</tr>
</tbody>
</table>

## Thinking Big About SOI Data

<table>
<thead>
<tr>
<th>Session</th>
<th>Presenter</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big SOI</td>
<td>Fritz Scheuren</td>
<td>10:05 am</td>
</tr>
<tr>
<td>Partnership Data Research</td>
<td>Danny Yagan, Eric Zwick, Owen Zidar</td>
<td></td>
</tr>
<tr>
<td>What is SOI's value added</td>
<td>Jim Nunns</td>
<td></td>
</tr>
<tr>
<td>AAPOR Big Data Report</td>
<td>Julia Lane</td>
<td></td>
</tr>
<tr>
<td>Canada Revenue Agency's Business Intelligence Strategy and Agency Data Program</td>
<td>Patricia Whitridge</td>
<td></td>
</tr>
<tr>
<td>LEI update and session discussion</td>
<td>Arthur Kennickell</td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>Larry May, Ralph Rector. Panel</td>
<td></td>
</tr>
</tbody>
</table>

## Lunch

<table>
<thead>
<tr>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:05 pm</td>
</tr>
</tbody>
</table>
## 2015 Consultants Panel Agenda (continued)

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:20 pm</td>
<td>Are Piketty and Zucman Getting it Right? Evaluating Distributional Statistics Based on Aggregate Data</td>
<td>Jenny Bourne</td>
</tr>
<tr>
<td></td>
<td>More Than They Realize: The Income of the Wealthy and the Piketty Thesis</td>
<td>Jenny Bourne</td>
</tr>
<tr>
<td></td>
<td>Measuring Income at the Top</td>
<td>John Sabelhaus</td>
</tr>
<tr>
<td></td>
<td>Mortality Differentials - How Much Longevity Can Money Really Buy?</td>
<td>Brian Raub</td>
</tr>
<tr>
<td></td>
<td>Discussant</td>
<td>Len Burman</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>Panel</td>
</tr>
<tr>
<td>2:30 pm</td>
<td>A Productive Partnership, Joint Work with Stanford</td>
<td>David Grusky</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>Panel</td>
</tr>
<tr>
<td>3:10 pm</td>
<td>An Overview of the SOI Consultants Panel</td>
<td>George Plesko</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>Panel</td>
</tr>
</tbody>
</table>

**Adjorn**
2014 SOI JOINT STATISTICAL RESEARCH PROGRAM
2014 Joint Statistical Research Program

- 87 Proposals were submitted (43 from grad students)
- 13 focused on compliance
- 13 focused on corporate tax issues
- 57 focused on individual tax issue
- 10 focused on a combination of the two
- Over 20 proposals required direct matching to outside data sets that raised legal and privacy concerns
- Several papers required data that are not available
- At least 12 did not credibly connect research to tax administration
2014 Joint Statistical Research Program

- 12 proposals were approved
  - 7 Tax Policy focused proposals (to be conducted by SOI)
  - 5 Compliance focused proposals (to be conducted by the Research Division)
Approved Projects (SOI)

Expanding SOI Data Products on Flow-Through Entities
Joseph Rosenberg - Tax Policy Center
James Nunns - Tax Policy Center

Tax incentives and changes in labor and capital income inequality
Wojciech Kopczuk – Columbia University
Approved Projects (SOI)

Effect of Estate Tax on Wealth Accumulation, Labor Supply, and Cross-State Migration
Jon Bakija - Williams College

James J. Heckman – University of Chicago
Magne Mogstad, - University of Chicago
Bradley Setzler - University of Chicago
Approved Projects (SOI)

A Protocol for Classifying the Taxpayer's Occupation

David B. Grusky - Stanford University
Michael Hout - New York University
David Johnson - Bureau of Economic Analysis
Michelle Jackson - Stanford University
Jonathan Fisher - Stanford University
Pablo Mitnik - Stanford University
Approved Projects (SOI)

The Impact of Income Volatility on Measured Cross-Section Income Inequality
Jeffrey P. Thompson - Federal Reserve Board
John Sabelhaus - Federal Reserve Board

Distribution of Tax Expenditures from a Permanent Income Perspective
Katharine Abraham - University of Maryland
Approved Projects (Research)

Estimating the Causal Effect of Third-Party Reporting on Small-business Tax Compliance
James Alm – Tulane University
Bibek Adhikari – Tulane University

Nonprofit Taxable Activities - How and which nonprofit organizations use taxable revenues to supplement other revenue streams within their operations
Steven Balsam – Temple University
Eric Harris – Rutgers University
Approved Projects (Research)

(Non)disclosure of subsidiary locations and corporate tax behavior
Scott Dyreng - Duke University
Jeff Hoopes - Ohio State
Jaron Wilde - University of Iowa

The Effect and Effectiveness of Tax Auditors
Joel Slemrod - University of Michigan
Ugo Troiano - University of Michigan
Shlomo Yitzhaki - Hebrew University of Jerusalem
Approved Projects (Research)

The Impact of the Offshore Voluntary Disclosure
Joel Slemrod - University of Michigan
Jeffrey Hoopes - Ohio State University
Daniel Reck - University of Michigan
SOI MIGRATION DATA: A NEW APPROACH
What are the Migration Data?

Migration data show the movement of individuals, via the address listed on Form 1040, over a two-year period.
What are the Migration Data?

Migration data show the movement of individuals, via the address listed on Form 1040, over a two-year period.

As an example…

Calendar Years

Migration Data 2011 - 2012

Tax Years: 2010 2011
What are the Migration Data?

Migration data show the movement of individuals, via the address listed on Form 1040, over a two-year period.

As an example...

Calendar Years

Migration Data 2011 - 2012

Tax Years: 2010 2011

The data are available at the State or county level:

a. Number of inflows – residents moving in
b. Number of outflows – residents moving out
What are the Migration Data?

<table>
<thead>
<tr>
<th>State level</th>
<th>County level</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) State-to-State Inflow</td>
<td>(1) County-to-County Inflow</td>
</tr>
<tr>
<td>(2) State-to-State Outflow</td>
<td>(2) County-to-County Outflow</td>
</tr>
</tbody>
</table>
What are the Migration Data?

State level

(1) State-to-State Inflow
(2) State-to-State Outflow

County level

(1) County-to-County Inflow
(2) County-to-County Outflow

<table>
<thead>
<tr>
<th>Origin from Alabama (State Code)</th>
<th>Destination into</th>
<th>Number of returns</th>
<th>Number of exemptions</th>
<th>Adjusted gross income (AGI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State Code</td>
<td>State</td>
<td>State Name</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>01</td>
<td>96</td>
<td>AL</td>
<td>AL Total Migration US and Foreign</td>
<td>51,971</td>
</tr>
<tr>
<td>01</td>
<td>97</td>
<td>AL</td>
<td>AL Total Migration US</td>
<td>50,940</td>
</tr>
<tr>
<td>01</td>
<td>98</td>
<td>AL</td>
<td>AL Total Migration Foreign</td>
<td>1,031</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>AL</td>
<td>AL Non-migrants</td>
<td>1,584,665</td>
</tr>
</tbody>
</table>
Previous Migration Data and New Migration Data

Previous (Census) Methodology

1980s 1990s 2000s 2010-2011 2011-2012 2012-2013...
Previous Migration Data and New Migration Data

Previous (Census) Methodology

New (SOI) Methodology

<table>
<thead>
<tr>
<th>1980s</th>
<th>1990s</th>
<th>2000s</th>
<th>2010-2011</th>
<th>2011-2012</th>
<th>2012-2013...</th>
</tr>
</thead>
</table>

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
Previous Migration Data and New Migration Data

- Based on partial-year data
Previous Migration Data and New Migration Data

<table>
<thead>
<tr>
<th>Previous (Census) Methodology</th>
<th>New (SOI) Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on partial-year data</td>
<td>Based on full-year data</td>
</tr>
</tbody>
</table>

- 1980s - 1990s - 2000s - 2010-2011 - 2011-2012 - 2012-2013...
Previous Migration Data and New Migration Data

Previous Methodology
(Census)

| 1980s | 1990s | 2000s | 2010-2011 | 2011-2012 | 2012-2013... |

New Methodology
(SOI)

- Based on partial-year data
- Matched on primary TIN

- Based on full-year data
Previous Migration Data and New Migration Data

<table>
<thead>
<tr>
<th>Previous (Census) Methodology</th>
<th>New (SOI) Methodology</th>
</tr>
</thead>
</table>

- Based on partial-year data
- Matched on primary TIN

- Based on full-year data
- Matched on primary, secondary, and dependent filer TINs
Previous Migration Data and New Migration Data

Previous (Census) Methodology

- Based on partial-year data
- Matched on primary TIN

New (SOI) Methodology

- Based on full-year data
- Matched on primary, secondary, and dependent filer TINs
- Inclusion of summary flows by AGI and age of the primary taxpayer
FROM PARTIAL-YEAR DATA TO FULL-YEAR DATA
Percentage of Returns and AGI received Before and After late September, Calendar Year 2012

<table>
<thead>
<tr>
<th>Percentage of returns</th>
<th>Percentage of adjusted gross income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before late September</td>
<td>After late September</td>
</tr>
<tr>
<td>95.9</td>
<td>88.8</td>
</tr>
<tr>
<td>4.1</td>
<td>11.2</td>
</tr>
</tbody>
</table>
EXPANDED YEAR-TO-YEAR RETURN MATCHING
## Expanded Matching between filers

<table>
<thead>
<tr>
<th>YEAR 1</th>
<th>YEAR 2</th>
<th>Percent of the total matched returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Primary filer</td>
<td>→ Primary filer</td>
<td></td>
</tr>
</tbody>
</table>

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
## Expanded Matching between filers

<table>
<thead>
<tr>
<th>YEAR 1</th>
<th>YEAR 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Primary filer</td>
<td>→ Primary filer</td>
</tr>
<tr>
<td>2. Primary filer</td>
<td>→ Secondary filer</td>
</tr>
<tr>
<td>3. Secondary filer</td>
<td>→ Primary filer</td>
</tr>
<tr>
<td>4. Secondary filer</td>
<td>→ Secondary filer</td>
</tr>
<tr>
<td>5. Dependent filers</td>
<td>→ Primary filer</td>
</tr>
<tr>
<td>6. Dependent filers</td>
<td>→ Secondary filer</td>
</tr>
</tbody>
</table>

Percent of the total matched returns
### Expanded Matching betweenfilers

<table>
<thead>
<tr>
<th>YEAR 1</th>
<th>YEAR 2</th>
<th>Percent of the total matched returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Primary filer → Primary filer</td>
<td></td>
<td>94.6%</td>
</tr>
<tr>
<td>2. Primary filer → Secondary filer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Secondary filer → Primary filer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Secondary filer → Secondary filer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Dependent filers → Primary filer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Dependent filers → Secondary filer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Expanded Matching between filers

<table>
<thead>
<tr>
<th>YEAR 1</th>
<th>YEAR 2</th>
<th>Percent of the total matched returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Primary filer → Primary filer</td>
<td>2. Primary filer → Secondary filer</td>
<td>94.6%</td>
</tr>
<tr>
<td>3. Secondary filer → Primary filer</td>
<td>4. Secondary filer → Secondary filer</td>
<td>0.8%</td>
</tr>
<tr>
<td>5. Dependent filers → Primary filer</td>
<td>6. Dependent filers → Secondary filer</td>
<td>1.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>less than 0.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>less than 0.1%</td>
</tr>
</tbody>
</table>
## Expanded Matching between filers

<table>
<thead>
<tr>
<th>YEAR 1</th>
<th>YEAR 2</th>
<th>Percent of the total matched returns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Primary filer → Primary filer</td>
<td>94.6%</td>
<td></td>
</tr>
<tr>
<td>2. Primary filer → Secondary filer</td>
<td>0.8%</td>
<td></td>
</tr>
<tr>
<td>3. Secondary filer → Primary filer</td>
<td>1.7%</td>
<td></td>
</tr>
<tr>
<td>4. Secondary filer → Secondary filer</td>
<td>less than 0.1%</td>
<td></td>
</tr>
<tr>
<td>5. Dependent filers → Primary filer</td>
<td>2.8%</td>
<td></td>
</tr>
<tr>
<td>6. Dependent filers → Secondary filer</td>
<td>less than 0.1%</td>
<td></td>
</tr>
</tbody>
</table>
THE GROSS MIGRATION FILE
## Gross Migration File

### Migration Flows

<table>
<thead>
<tr>
<th></th>
<th>Total Matched Returns (1)</th>
<th>Non-migrant Returns (2)</th>
<th>Outflow Returns (3)</th>
<th>Inflow Returns (4)</th>
<th>Same State Returns (5)</th>
</tr>
</thead>
</table>

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
# Gross Migration File

## Migration Flows

<table>
<thead>
<tr>
<th>Total Matched Returns (1)</th>
<th>Non-migrant Returns (2)</th>
<th>Outflow Returns (3)</th>
<th>Inflow Returns (4)</th>
<th>Same State Returns (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 26 (1)</td>
<td>26 under 35 (2)</td>
<td>35 under 45 (3)</td>
<td>45 under 55 (4)</td>
<td>55 under 65 (5)</td>
</tr>
<tr>
<td>65 and over (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Age Categories

<table>
<thead>
<tr>
<th>Age Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 26</td>
</tr>
<tr>
<td>26 under 35</td>
</tr>
<tr>
<td>35 under 45</td>
</tr>
<tr>
<td>45 under 55</td>
</tr>
<tr>
<td>55 under 65</td>
</tr>
<tr>
<td>65 and over</td>
</tr>
</tbody>
</table>
Gross Migration File

Migration Flows

<table>
<thead>
<tr>
<th>Total Matched Returns (1)</th>
<th>Non-migrant Returns (2)</th>
<th>Outflow Returns (3)</th>
<th>Inflow Returns (4)</th>
<th>Same State Returns (5)</th>
</tr>
</thead>
</table>

Age Categories

<table>
<thead>
<tr>
<th>Under 26 (1)</th>
<th>26 under 35 (2)</th>
<th>35 under 45 (3)</th>
<th>45 under 55 (4)</th>
<th>55 under 65 (5)</th>
<th>65 and over (6)</th>
</tr>
</thead>
</table>

State

- Alabama
- Alaska
- Arizona
- Arkansas
- California
- ....
- Wyoming

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
Gross Migration File

Migration Flows

<table>
<thead>
<tr>
<th>Migration Flows</th>
<th>Total Matched Returns (1)</th>
<th>Non-migrant Returns (2)</th>
<th>Outflow Returns (3)</th>
<th>Inflow Returns (4)</th>
<th>Same State Returns (5)</th>
</tr>
</thead>
</table>

Age Categories

<table>
<thead>
<tr>
<th>Age Categories</th>
<th>Under 26 (1)</th>
<th>26 under 35 (2)</th>
<th>35 under 45 (3)</th>
<th>45 under 55 (4)</th>
<th>55 under 65 (5)</th>
<th>65 and over (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>Alabama</td>
<td>Alaska</td>
<td>Arizona</td>
<td>Arkansas</td>
<td>California</td>
<td>. . .</td>
</tr>
</tbody>
</table>

AGI Categories

<table>
<thead>
<tr>
<th>AGI Categories</th>
<th>$1 under $10,000</th>
<th>$10,000 under $25,000</th>
<th>$25,000 under $50,000</th>
<th>$50,000 under $75,000</th>
<th>$75,000 under $100,000</th>
<th>$100,000 under $200,000</th>
<th>$200,000 or more</th>
</tr>
</thead>
</table>

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
COMPARING OLD VS. NEW MIGRATION DATA
Percentage Change in Number of Returns for SOI Migration Data, by AGI, Calendar Years 2011-2012

Size of adjusted gross income

- $200,000 or more: 24.6%
- $100,000 under $200,000: 7.6%
- $75,000 under $100,000: 4.9%
- $50,000 under $75,000: 4.2%
- $25,000 under $50,000: 3.2%
- $10,000 under $25,000: 2.7%
- $1 under $10,000: 4.5%

Percentage change

0 5 10 15 20 25 30
Distribution of Matched Returns, by AGI, Calendar Years 2011-2012

SOI Migration Data: A New Approach

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
Net-Migration Rate

\[
\text{Net-Migration Rate} = \frac{\text{In-Migrants} - \text{Out-Migrants}}{\text{(Non-migrants} + \text{Out-Migrants})}
\]
Net-Migration Rate

Net-Migration Rate \(=\) \(\frac{\text{In-Migrants} - \text{Out-Migrants}}{(\text{Non-migrants} + \text{Out-Migrants})}\)

- Computed for all 50 States, plus DC
- Previous and new migration data
Number of States, by Percentage Difference of the Net-Migration Rates of Old and New Data

Number of states

<table>
<thead>
<tr>
<th>Percentage difference</th>
<th>Number of states</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater than 0 percent</td>
<td>50</td>
</tr>
<tr>
<td>under 5 percent</td>
<td></td>
</tr>
<tr>
<td>5 percent under 10 percent</td>
<td>5</td>
</tr>
<tr>
<td>10 percent or more</td>
<td>1</td>
</tr>
</tbody>
</table>
Top 5 States with Largest Net-Migration Rate Differential, Calendar Years 2011-2012

<table>
<thead>
<tr>
<th>State</th>
<th>Net migration rate previous migration data</th>
<th>Net migration rate new migration data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wyoming</td>
<td>0.77</td>
<td>1.05</td>
</tr>
<tr>
<td>South Dakota</td>
<td>0.56</td>
<td>0.66</td>
</tr>
<tr>
<td>Louisiana</td>
<td>-0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>Mississippi</td>
<td>-0.29</td>
<td>-0.21</td>
</tr>
<tr>
<td>Florida</td>
<td>0.21</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Panel Discussion

What’s New At SOI?

Discussion Question

• Are we doing enough to ease the transition to the new migration data series, or is there something else that would be useful?
Break:  
10 minutes

Next:  
Thinking Big About SOI Data
THINKING BIG ABOUT SOI DATA
BIG SOI
Fritz Scheuren
NORC at the University of Chicago
SOI-O-O-O-O BIG!!!

• Time to transform SOI again
• Into the new Big Data World
• Now visible to all!!!
• Title taken from a 1924 Edna Ferber Novel
• Of course, the “I” was put in
• To give it an SOI connection!!!
Big Data as Challenge?

• SOI must change drastically to survive
• To doubt that is to be asleep!!!
• But the change is one SOI has been getting ready for
• All SOI has to do is want to change more/bigger
Big Data as Opportunity?

- Opportunity or Adventure?
- Planning is the Difference?
- Commit NOW then do!
- One small step at a time!
- Pilot and then pilot again?
- But keep moving, fast!!!
- The field is full of other players
- Keep a Wise head and young heart!
The five Eras of SOI so far?

• The Pre SOI Age at IRS
• The Pre Electronic SOI Age
• Golden Age of SOI Statistics
• Golden Age of SOI “Data”
• Big SOI Metadata Age is now
• Two Analytic Program Examples
The Pre SOI Age

• SOI was part of a General Reform of Government
• Greater transparency was part of that movement
• Statistical Summaries of the tax system were needed
The Pre Electronic SOI Age

• SOI was the only major source of Tax Return Information until the 1950/60s
• SOI Statistics for individuals and corporations were annual then
• But corporations were not even sampled until the 1950s
Golden Age of SOI Statistics

• At the beginning SOI customers only could use and, hence, only wanted SOI Statistics
• Of course, more timely statistics and always in more detail
• Special SOI studies were an answer for a while
But Change was Occurring

• The 1960s were changing America, mostly for the better IRS/OTA too
• The IRS Master File was improving and SOI’s clients were upgrading their computer systems, like now, sadly faster that SOI
• This led to a demand for SOI Data over SOI Statistics/Analytics
Golden Age of SOI “Data”

• SOI expands its program reach beyond IRS enumerative samples
• Linkages were strengthened with SSA and FRB/Census survey data
• But still very difficult and expensive and ad hoc
Why Not More?

- As successful as SOI was it did not adapt/adopt fast enough to the growing data dense world
- Is SOI committed to really major changes now. It must be?
- This is the time to break free of the parts of the past where SOI was kept back and “Run to Glory”
What is Big Data?

• “Big Data” buzz has everyone looking at new options, SOI too
• But what is “Big Data” really?
• No fixed definition yet
• I’ll use Rob Kitchin’s definition
• And examples from June Journal
Defining SOI in Big Data Era

• Big Data is Big N and Big P?
• Big N (number of units) is way more than typically was used just a few years ago
• Making P (number of items per unit) large is still hard to come by and a place for record linkage
SOI Going BIG Little by Little

• Barriers abound bureaucratic, statistical, and financial
• But SOI has already started and needs to expand relentlessly
• SOI can walk the way ahead!!!
• An Inter/intra agency respect-respect approach is needed
Elements on SOI “Going Big”

• Building SOI Data Relationship across Agencies
• Handling Confidentiality Issues which grow with linkage
• Attempting to Measure increases in Linkage uncertainty
Confidentiality in Linkage Issues

• Very Hard at every stage
• Merging, Matching, Sharing Results
• Several small-scale precedents
• No routine practice exists
• BLS/Census Synthetic Data Project mentioned last year offers hope!
Measuring Linkage Uncertainty

- Again, very hard at every stage
- Linkage costs unaffordable, even undoable without errors
- Trading Sampling Error for Big Data linkage errors?
- May be a good trade but it depends? Sometimes not!
Try anyway! Experience will tell?

• The sub-systems to be “bigged” were not designed to be linked
• These are statistical problems and can be satisficed/if not solved!
• Beyond our pay grade are the turf problems but once people work together more, who knows?
Two Partial Big Data Examples

• Individual and organizational examples chosen -- one each
• The “Shock of Recognition” – We have been “bigging” for a while
• Still more than just scaling up?
• Two “bigging” to Start SOI with!
Individual’s Little “Bigging”

- Medicare Public Use (PUF) case
- SCF/CPS expanding the larger sample over years/record types
- Expanding the Main SOI public sample as in Medicare Example
- Synthetic Expansion of SCF/CPS data too larger SOI data sets
Partnership’s Little “Bigging”

• Partnering with OTA to further the BLS linkages with the unemployment Tax System
• Already broadly covered by same IR Code as SOI
• Census/BLS Synthetic precedent exists and could fit nicely
When to Select What Sample?

• SOI Partnership Program goals:
  • Editing/interpreting dollar totals?
  • Studying business demography?
  • Partnership income recipients?
  • Flexible cross-section and longitudinal tools, new and old!!!
  • Before/after linkages up and down?
Bless You’s and Thanks to All

• First to Barry who has allowed me to speak
• Then to the Panel that has stood by SOI all these years
• Then to my sisters and brothers at SOI my IRS family
Joint Business Tax Data Project: Lessons from Estimating Average Partnership Tax Rates

Danny Yagan, UC Berkeley and NBER
Owen Zidar, Chicago Booth and NBER
Eric Zwick, Chicago Booth and NBER

June 2015

This work is preliminary and does not necessarily reflect the views of the Treasury Department.
General need to link firms to ultimate owners

- OTA need: Estimate tax rate and revenue consequences of tax reform
- Growing reality: Business taxation occurs at owner level, after flowing through ownership chains
- Pilot: Link partnerships to ultimate owners to estimate current average tax rate on partnership income

This presentation:

1. Our partnership project and findings
2. Two actionable suggestions for SOI work going forward
Partnerships in the United States: Who Owns Them and How Much Tax They Pay

Michael Cooper, U.S. Treasury Department
John McClelland, U.S. Treasury Department
James Pearce, U.S. Treasury Department
Richard Prisinzano, U.S. Treasury Department
Joseph Sullivan, U.S. Treasury Department
Danny Yagan, UC Berkeley and NBER
Owen Zidar, Chicago Booth and NBER
Eric Zwick, Chicago Booth and NBER

May 2015

This work is preliminary and does not necessarily reflect the views of the Treasury Department.
Motivation: What does this transformation mean?

Percentage of Total Net Business Income by Entity Type, 1980-2010

Business activity lies increasingly outside of the C-corporate sector
Tax reform requires facts on tax rates for each sector

We provide:

1. Linked partnership-partner data
2. Facts on the distribution of business income
3. Underlying tax rates for the partnership sector
Part 1: Partnerships file a business inc. tax return...

**U.S. Return of Partnership Income**

For calendar year 2011, or tax year beginning ____________, 2011, ending ____________, 20__

- **A** Principal business activity
- **B** Principal product or service
- **C** Business code number

**G** Check applicable boxes:

1. Initial return
2. Final return
3. Name change
4. Address change
5. Amended return
6. Technical termination - also check (1) or (2)

**H** Check accounting method:

1. Cash
2. Accrual
3. Other (specify)

**I** Number of Schedules K-1. Attach one for each person who was a partner at any time during the tax year

**J** Check if Schedules C and M-3 are attached

---

**Caution. Include only trade or business income and expenses on lines 1a through 22 below. See the instructions for more information.**

<table>
<thead>
<tr>
<th>Income</th>
<th>1a</th>
<th>1b</th>
<th>1c</th>
<th>1d</th>
<th>1e</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

6
...which lists allocations only by partner type...

<table>
<thead>
<tr>
<th>Analysis of Net Income (Loss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Net income (loss). Combine Schedule K, lines 1 through 11. From the result, subtract the sum of Schedule K, lines 12 through 13d, and 16l.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis by partner type:</th>
<th>(i) Corporate</th>
<th>(ii) Individual (active)</th>
<th>(iii) Individual (passive)</th>
<th>(iv) Partnership</th>
<th>(v) Exempt organization</th>
<th>(vi) Nominee/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>General partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited partners</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
...but are reflected in K-1s (issued *per partner*).
Data challenges: Owners can be one of many entity types, and different kinds of returns are processed by different systems.

Our procedure: link K-1’s to partnership returns by merging on the Document Locator Number (linking 25m K-1’s to 3m partnerships).

High coverage: >96% of business income.
Part 2: Who owns partnerships
Partnership participation is very concentrated

Participation Rates by Business Income Type

Individual partners

- Sole proprietorship income
- C-corp/foreign dividend income
- Partnership income
Partnership income is exceptionally concentrated
High-income partners own finance / prof. services

Partnership Income Shares by Selected Partnership Industries
Individual partners

- Hospitality (NAICS 72)
- Professional services (NAICS 54)
- Finance and real estate (NAICS 52, 55, and 531)
Part 3: Estimating partnership tax rates

- Assign tax rates to each partner
  - For tiered partnerships this requires tracking flows through the tiers to a final owner
  - For S-corporation partners this requires assigning a tax rate to each shareholder of the S-corporation

- Aggregate circular tiered partnerships into single partnerships
- Aggregate these partner tax rates to the partnership level
Case 1: Linear tiers

Triangle A connects to B and C with the values 1040/1120. B and C also connect to triangle A with the same values, forming a linear hierarchy.
Case 2: Circular tiers (treat as single partnership)

1040/1120 ← D ← E → 1040/1120

1040/1120 ← F →
Assign tax rates to partners

- Each partner’s income tax liability is recalculated w/o the allocated amount of each income type appearing on the K-1 from a given partnership

- OTA has CDW-based tax calculators for 1040, 1041, 1120-C and 1120-S

- Those without a calculator:
  - 1120-F: 35%
  - 1120-L: 35%
  - 1120-PC: 35%
  - 1120-REIT: 35%
  - 1120-RIC: 22%
  - 990: 35% (UBIT) or 0%
  - Unidentified: 35%
Aggregate partner tax rates to a partnership

- Define: Tax rate on income type $i$ to partner $k$ from partnership $p$ is:
  \[ T_{i,k,p} = \frac{TAX_{i,k,p}}{D_{i,k,p}} \]

- These rates can be aggregated up for a partnership tax rate:
  \[ T_p = \frac{1}{D_p} \left[ \sum_{i=1}^{I} \sum_{k=1}^{K} D_{i,k,p} \times T_{i,k,p} \right] \]

- Note: Definitions matter

  - Tax rates technically unbounded but vast majority in $[0\%,35\%]$.
  - Weighting by actual amount vs. absolute value matters at top.
Results: Entity income tax rates by entity form

**Tax Rates by Business Organizational form**

<table>
<thead>
<tr>
<th>Form</th>
<th>Tax Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sole Proprietorships</td>
<td>13.6</td>
</tr>
<tr>
<td>Partnerships</td>
<td>19.7</td>
</tr>
<tr>
<td>S-corps</td>
<td>24.7</td>
</tr>
<tr>
<td>C-corps</td>
<td>22.7</td>
</tr>
</tbody>
</table>

*Source: Based on 18.5% average U.S. dividend rate estimated in Poterba (2004)*
1. **Who owns them?:** Partnership income accrues to high-earners even more disproportionately than C-corporate income (60% to the top 1%)

2. **How much tax do they pay?:** We estimate a partnership tax rate of 19.7%, lower than the C-corporate and lower for the largest partnerships
Future OTA partnership work will rely on links to ultimate owners

The two key uses of our linked data could hopefully be incorporated into SOI’s partnership study file (and potentially other business study files):

1. Adding an ownership file, similar to S-corporation study
2. Including all K-1 fields in ownership file
1: Adding an ownership file, similar to S-corp. study

- S-corporation study file: Comes with information on S-corporation owners
- Partnerships: Much harder because of partnership tiers

- Suggestion: Provide one “entity” file and one “ownership” file
  - Entity file: Current partnership study file
  - Ownership file: Direct and indirect owners of entity file partnerships

- Sample: Owners of the partnerships in the entity file and owners at least three tiers down
- Rows: One row per owner
- Columns: K-1 fields and indicator of partnership tier
2: Including all K-1 fields in ownership file
What is SOI’s Value Added?

Lessons from the Non-Filer Project

June 5, 2015
Jim Nunns
SOI Consultants Panel Meeting
SOI’s Value Added: Context

- CDW creates vast menu for new value added
  - New products, improved population data
- SOI eager to enhance products and services
- But, SOI faces severe resource constraints
  - Tight budgets into foreseeable future
  - Days of “easy” productivity gains over
  - Human capital hard to develop & maintain
  - Ongoing demand for current SOI products
- Tradeoffs and careful study design required
SOI’s Value Added: Origins of Non-Filer Project

- Non-Filer Project shows how CDW adds value
- Origins
- Most of population represented on tax returns
  - “Filers” may be taxpayers or dependents
- Information returns cover most “Non-Filers”
- Early SOI studies on non-filer population
- Recent micro-data studies
  - JCT and OTA for microsimulation models
  - RAS for National Research Program (NRP)
SOI’s Value Added: Elements of Non-Filer Project

- Collaborative effort between SOI & customers
- Based on sample of information returns
- Clean data; add SSA data; flag nonresidents
- Match to CDW to identify filers
- Map information into “return” for each person
- Final products:
  - Micro data files for INSOLE and PUF
  - Tabulations covering filers and non-filers
SOI’s Value Added: Lessons from Non-Filer Project

- Creates new low-cost, high-value products
- Draws on comparative advantages
  - “Experimental” work by JCT, OTA, RAS
  - “Gold standard” production by SOI
- Builds on existing SOI products
- CDW essential to success
- Methodology well documented
- Straightforward extensions to population files
  - Improve data on CDW and Data Bank
  - Adds geographic and longitudinal depth
THANK YOU

For more information please contact:

Jim Nunns
jnunns@urban.org

Visit us at:
www.taxpolicycenter.org
AAPOR Report on Big Data

AAPOR Big Data Task Force

*February 12, 2015*
Task Force Members:

Lilli Japec, Co-Chair, Statistics Sweden
Frauke Kreuter, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB
Marcus Berg, Stockholm University
Paul Biemer, RTI International
Paul Decker, Mathematica Policy Research
Cliff Lampe, School of Information at the University of Michigan
Julia Lane, American Institutes for Research
Cathy O’Neil, Johnson Research Labs
Abe Usher, HumanGeo Group

Acknowledgement: We are grateful for comments, feedback and editorial help from Eran Ben-Porath, Jason McMillan, and the AAPOR council members.
The report has four objectives:

1. to educate the AAPOR membership about Big Data (Section 3)
2. to describe the Big Data potential (Section 4 and Section 7)
3. to describe the Big Data challenges (Section 5 and 6)
4. to discuss possible solutions and research needs (Section 8)
Big Data
AAPOR Task Force

Source: Frauke Kreuter
until recently

three main data sources
Survey Data

Administrative Data

Experiments

Source: Frauke Kreuter
now
Social media sentiment (daily, weekly and monthly) in the Netherlands, June 2010 - November 2013. The development of consumer confidence for the same period is shown in the insert (Daas and Puts 2014).
Big Data

Volume

Data at Rest
Terabytes to exabytes of existing data to process

Velocity

Data in Motion
Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms
Structured, unstructured, text, multimedia

http://www.rosebt.com/blog/data-veracity
Hope that found/organic data

Can replace or augment expensive data collections
More (= better) data for decision making
Information available in (nearly) real time

Source: Frauke Kreuter
But (at least) one more V

http://www.rosebt.com/blog/data-veracity
fkreuter@umd.edu

Thank You!
CHANGE IN PARADIGM AND RISKS INVOLVED

Julia Lane
New York University
American Institutes for Research
University of Strasbourg
Big Data definition

• “Big Data” is an imprecise description of a rich and complicated set of characteristics, practices, techniques, ethics, and outcomes all associated with data. (AAPOR)

• No canonical definition

• By characteristics: Volume Velocity Variety (and Variability and Veracity)

• By source: found vs. made

• By use: professionals vs. citizen science

• By reach: datafication

• By paradigm: Fourth paradigm

Source: Julia Lane
Motivation

• New business model
  • Federal agencies no longer major players
• New analytical model
  • Outliers
  • Finegrained analysis
  • New units of analysis
• New sets of skills
  • Computer scientists
  • Citizen scientists
• Different cost structure

Source: Julia Lane
New Frameworks

Source: Ian Foster, University of Chicago
New kinds of analysis

Academic Science is A Network Form of Organization

Source: Jason Owen Smith and UMETRICS data
Access for Research

How much should a nation spend on science? What kind of science? How much from private versus public sectors? Does demand for funding by potential science performers imply a shortage of funding or a surfeit of performers?.... A new “science of science policy” is emerging, and it may offer more compelling guidance for policy decisions and for more credible advocacy.

Source: Julia Lane
Value in other fields

Source: Julia Lane
Privacy, Big Data, and the Public Good
Frameworks for Engagement

Edited by Julia Lane
Victoria Stodden
Stefan Bender
Helen Nissenbaum

Source: Julia Lane
Core Questions

• What is the legal framework?
• What is the practical framework?
• What is the statistical framework?

Source: Julia Lane
Legal Framework

• Current legal structure inadequate

"The recording, aggregation, and organization of information into a form that can be used for data mining, here dubbed 'datafication', has distinct privacy implications that often go unrecognized by current law (Strandburg)

• Assessment of harm from privacy inadequate

• Privacy and big data are incompatible
• Anonymity not possible
• Informed consent not possible

Source: Julia Lane
Public lives of others

- “With friends like these...”
  - Social networks: what friends reveal implicitly & explicitly
- Tyranny of the Minority
  - Inference from representative sample
  - “multiple attributes can be inferred globally when as few as 20% of the users reveal their attribute information.”
  - Mislove et al., “You Are Who You Know: Inferring User Profiles in Online Social Networks.”

Source: Julia Lane
Statistical Framework

- Importance of valid inference
  - The role of statisticians/access

- Inadequate statistical disclosure limitation
  - Diminished role of federal statistical agencies
  - Limitations of surveys

- New analytical framework:
  - Mathematically rigorous theory of privacy
  - Measurement of privacy loss
  - Differential privacy

Source: Julia Lane
Some suggestions

Recommendations: Data Protection

Do not adopt HIPAA as the standard for data protection
Use an array of data protection approaches (Rec 5.1), such as:
- Plan with the concept of a portfolio approach considering safe people, safe projects, safe data, safe settings, and safe outputs
- Use a range of statistical methods to reduce disclosure risk
- Consult resources and data protection models, such as: university research data management service groups, individual IT/protection experts, and specialized institutions
- Use existing standards for data protection promulgated by the National Institute of Standards and Technology
- Develop a national center to define and certify information risk of different types of studies and corresponding data protection plans to minimize risks (Rec 5.2)

Recommendations for Research on Minimal Risk and Expedited Review

- Build evidence of risks in daily life and age-indexed routine medical, psychological, or educational examinations, tests, or procedures of the general population
- Develop appropriate algorithms for calculating risk from both the probability and magnitude of harm
- Encourage evidence for effective procedures for minimizing potential harms to no-more-than-minimal risk
- Study effects of social and behavioral research on research participants for evidence-based assessments of “known and foreseeable” risk

Source: Julia Lane
And a reminder of why

Source: Julia Lane
Comments and questions

• Julia.lane@nyu.edu
Skills Required to Integrate Big Data into Public Opinion Research

Abe Usher
Chief Technology Officer, HumanGeo
• Big data demystified
• Four layers of big data
• Skills required
• Easter eggs

Source: Abe Usher
Big Data Today

Courtesy of Google Trends: [http://goo.gl/4H8Ttd](http://goo.gl/4H8Ttd)
Big Data & Public Opinion vs Pop Culture

Courtesy of Google Trends: http://goo.gl/QHIQcN
What is big data?

What is Hadoop File System? (HDFS)

What is Hadoop MapReduce? (MR)

Source: Abe Usher
How can you identify a legitimate hip-hop artist (versus someone who just gets up and rhymes)?

http://www.npr.org/2005/08/30/4824690/original-gangster-rapper-and-actor-ice-t

Source: Abe Usher
How can you identify a legitimate hip-hop artist (versus someone who just gets up and rhymes)?

“Game knows game, baby.”

Source: Abe Usher
How can you identify a legitimate hip-hop artist (versus someone who just gets up and rhymes)?

“If you have expert knowledge, then you are capable of answering complex questions by interpreting domain specific information.” [paraphrased]

Source: Abe Usher
Peter Gibbons hatches a plot to write a computer virus that grab fractions of a penny from a corporate retirement account.

http://goo.gl/rDg1U

**Takeaway point:** Little bits of value (information) provide deep insights in the aggregate

Source: Abe Usher
Takeaway point: Hadoop simplifies the creation of massive counting machines

Source: Abe Usher
"Ogres are like onions. Ogres have layers, onions have layers..."

You get it? We both have layers.

Source: Abe Usher
Big Data: Layers

Data Source(s)

- Web Scraping
- Text Analysis
- Networks

Data Storage

- Example: Hadoop Distributed File System
- Example: Hadoop MapReduce

Data Analysis

- Example: map visualization

Data Output

- Structured Data (different ways of collecting data)
- Text data (making use of new types of data)
- Relationship data (creating new units of economic activity)
- Output data (creating innovation measures)
- Sense making (from the population)
- Analysis (capturing social and economic dynamics)
- Conveying information (in statistically valid context)

Source: Abe Usher
Four roles related to big data: each provide different skills

- **DOMAIN EXPERT**
  User, analyst, or leader with deep subject matter expertise related to the data, its appropriate use, and its limitations.

- **SYS ADMIN**
  Team member responsible for defining and maintaining a computation infrastructure that enables large scale computation.

- **RESEARCHER**
  Team member with experience applying formal research methods, including survey methodology and statistics.

- **COMPUTER SCIENTIST**
  Technically skilled team member with education in computer programming and data processing technology.

Source: Abe Usher
Computer scientist
• Data preparation
• MapReduce algorithms
• Python/R programming
• Hadoop ecosystem

System Administrator
• Storage systems (MySQL, Hbase, Spark)
• Cloud computing:
  • Amazon Web Services (AWS)
  • Google Compute Engine
• Hadoop ecosystem

Source: Abe Usher
Big data enables new insights into human behavior

Geolocated social media activity in Washington DC during a 15 minute time period generated by MR. TweetMap

Source: Abe Usher
Abe Usher
abe@thehumangeo.com
http://www.thehumangeo.com/
Google “do a barrel roll”

“Google gravity”

Google search in Klingon [www.google.com/?hl=kn](http://www.google.com/?hl=kn)

Source: Abe Usher
Big Data Veracity: Error Sources and Inferential Risks

Paul Biemer
RTI International
and
University of North Carolina

Source: Paul Biemer
Errors in Big Data: An Illustration

Suppose 1 in 1,000,000 people are terrorists
The Big Data Terrorist Detector is 99.9% accurate
The detector says your friend, Jack is a terrorist.
What are the odds that Jack is really a terrorist?

Source: Paul Biemer
Suppose 1 in 1,000,000 people are terrorists
The Big Data Terrorist Detector is 99.9 accurate
The detector says your friend, Jack is a terrorist.
What are the odds that Jack is really a terrorist?

Answer: 1 in 1000 i.e., 99.9% of the terrorist detections will be false!

Source: Paul Biemer
Some questions regarding Big Data veracity

- What constitutes a Big Data error?

Systematic error in Google Flu Trends data

Source: Paul Biemer
Some questions regarding Big Data veracity

- What constitutes a Big Data error?
- What are the sources and causes of the errors?
- Do the error distributions vary by source?
- Are the errors systematic or variable or both?
- How do the errors affect data analysis such as
  - Classifications
  - Correlations
  - Regressions
- How can analysts mitigate these effects?

Source: Paul Biemer
Total Error Framework for Traditional Data Sets

**Typical File Structure**

<table>
<thead>
<tr>
<th>Record #</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$\ldots$</th>
<th>$V_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Paul Biemer

variables or features
Total Error Framework for Traditional Data Sets

**Typical File Structure**

<table>
<thead>
<tr>
<th>Record #</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>...</th>
<th>$V_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**variables or features**

**Source:** Paul Biemer

$\text{total error} = \text{row error} + \text{column error} + \text{cell error}$
**Possible Column and Cell Errors**

### Typical File Structure

<table>
<thead>
<tr>
<th>Record #</th>
<th>$V_1$</th>
<th>$V_2$</th>
<th>$\ldots$</th>
<th>$V_K$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Variables or features**

- Misspecified variables = specification error
- Variable values in error = content error
- Variable values missing = missing data

Source: Paul Biemer
Possible Row Errors

Typical File Structure

<table>
<thead>
<tr>
<th>Record #</th>
<th>( V_1 )</th>
<th>( V_2 )</th>
<th>...</th>
<th>( V_K )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

variables or features

- Missing records = **undercoverage error**
- Non-population records = **overcoverage**
- Duplicated records = **duplication error**

Source: Paul Biemer
Shortcomings of the Traditional Framework for Big Data

- Big Data files are often not rectangular
  - hierarchically structure or unstructured
- Data may be distributed across many data bases
  - Sometimes federated, but often not
- Data sources may be quite heterogeneous
  - Includes texts, sensors, transactions, and images
- Errors generated by Map/Reduce process may not lend themselves to column-row representations.

Source: Paul Biemer
Big Data Process Map

Generate
- Source 1
- Source 2
- Source K

ETL
- Extract
- Transform (Cleanse)
- Load (Store)

Analyze
- Filter/Reduction (Sampling)
- Computation/Analysis (Visualization)

Source: Paul Biemer
Big Data Process Map

**Generation**
- Source 1
- Source 2
- Source K

**ETL**
- Errors include: low signal/noise ratio; lost signals; failure to capture; non-random (or non-representative) sources; metadata that are lacking, absent, or erroneous.

**Analyze**
- Filter/Reduction (Sampling)
- Computation/Analysis (Visualization)

**Load (Store)**

Source: Paul Biemer
Errors include: specification error (including, errors in meta-data), matching error, coding error, editing error, data munging errors, and data integration errors.
Big Data Process Map

Generation

Source 1

Source 2

Source K

Errors include: sampling errors, selectivity errors (or lack of representativity), modeling errors

Transform (Cleanse)

Load (Store)

Filter/Reduction (Sampling)

Computation/Analysis (Visualization)

Data are filtered, sampled or otherwise.

Source: Paul Biemer
Errors include: modeling errors, inadequate or erroneous adjustments for representativity, computation and algorithmic errors.
Implications for Data Analysis

Stork Die-off Linked to Human Birth Decline

Source: Paul Biemer
Implications for Data Analysis

- Study of rare groups is problematic
- Biased correlational analysis
- Biased regression analysis
- Coincidental correlations
- Noise accumulation – inability to identify correlates
- Incidental endogeneity – $\text{Cov}$(error, covariates)

Source: Paul Biemer
Implications for Data Analysis

- Study of rare groups is problematic
- Biased correlational analysis
- Biased regression analysis
- Coincidental correlations
- Noise accumulation – inability to identify correlates
- Incidental endogeneity – $\text{Cov(error, covariates)}$
  - These latter three issues are a concern even if the data could be regarded as error-free.
  - Data errors can considerably exacerbate these problems.

**Current research is aimed at investigating these errors.**

Source: Paul Biemer
Recommendations

1. **Surveys and Big Data are complementary data sources not competing data sources.** There are differences between the approaches, but this should be seen as an advantage rather than a disadvantage.

2. **AAPOR should develop standards for the use of Big Data in survey research when more knowledge has been accumulated.**
3. AAPOR should start working with the private sector and other professional organizations to educate its members on Big Data

4. AAPOR should inform the public of the risks and benefits of Big Data.
5. AAPOR should help remove the barrier associated with different uses of terminology.

6. AAPOR should take a leading role in working with federal agencies in developing a necessary infrastructure for the use of Big Data in survey research.
CRA Data Program

Patricia Whitridge, Canada Revenue Agency
SOI Panel Meeting - June 5, 2015
Why the need for a CRA Data Program?

- Better deal with needs for data that cross program and organizational boundaries, and external data
- Currently, no single entity responsible for the data – ITB managed in past
- Provide leadership in new types of data usage (e.g. analytics) and new data directions (e.g. Open Data)
- Provide a coordinated whole-of-Agency approach to acquiring, using, sharing, managing, and publishing data
Program Scope

- Builds on existing functions:
  - Data provisioning services (per data sharing agreements, and ad hoc requests)
  - Business Intelligence Centre of Expertise
  - Maintain data catalogue, supported by Managed Metadata Environment
  - Data Stewardship

- New functions
  - Identify new data sources and opportunities
  - Data governance
  - Data policies and standards (business-oriented)
  - Talent management
  - Oversee – and report on – state of Agency’s data
Proposed Program Model

- Centralized leadership and coordination, with cross-Agency involvement
- Existing functional and technical responsibilities for managing data are maintained
- Data Program works in partnership with branches/regions and ITB
- Accountability for Data Program is assigned to a new role: Chief Data Officer
Opinions expressed are those of the presenter and do not necessarily reflect the views of the Federal Reserve or its staff.
LEI: Review and Update

• LEI: identification system for entities
  • “Entities” broadly construed
  • ISO 17442: 2012 standard serves as basis

• Fixed identification number
  • Unique code: 20 characters, including 2 check digits
    • Example: U.S. Sugar Corp.: 549300VV3SF28T3NK585
  • Exclusive assignment
  • Persistent
  • No persistent embedded intelligence

• Reference data
  • Variable data
  • Data to support precise identification of individual entities
LEI Scope

• ISO 17442: “The term legal entities includes, **but is not limited to** unique parties that are legally or financially responsible for the performance of financial transactions or have the legal right in their jurisdiction to enter independently into contracts…”
  • Excludes natural persons acting as natural persons
  • Otherwise very broad, in principle—even government entities
  • Many edge cases: quasi-entities, contract-based structures, large individual actors in markets
  • In practice, proceeding cautiously to avoid confusion

• On-going work with ISO to define a standard for legal form
  • Could be used to define eligibility for LEI
Global LEI System (GLEIS)

- Regulatory Oversight Committee (ROC)
  - Committee on Evaluation and Standards (CES)
- Global LEI Foundation (GLEIF)
- Local Operating Units (LOUs)
- Registrants
- Local regulators
- Users
Funding Model

• Use is free
• Nonprofit, cost-recovery principles
• Charge for registration and annual maintenance
  • Initial charge currently about $200
  • Maintenance fee about $100
  • Scale economies expected to drive down cost
• A fixed fee charged to LOUs for each registered LEI supports the GLEIF
Data Quality

• Highest priority: **Information must be reliable and timely**

• Validation
  • Entity exists, person applying authorized, entity within scope
  • Reference data confirmed
  • Responsibility for accuracy rests primarily with the registrant

• Public challenge facility
Data on Organizational Relationships

• Relationship data needed for aggregating exposures or for tracing flows of money or information
  • Essential both for regulators and private risk managers to aggregate disparate sources of data to understand risk
  • Will also support broader transparency
    • E.g., Anti money laundering
  • First relationships considered: “Direct parent”/”Ultimate parent”
    • Need for extensible structure to accommodate other relationships in the future
• Other elements of relationship need to be considered
  • Numerous technical complications
  • Privacy issues and jurisdictions with opaque ownership structures
First Phase of Organizational Relationship Data

• Based on accounting consolidation definitions
• Open questions
  • Collect from “parents” or “children”—or both
  • Roles of the LOUs and the GLEIF in collecting/consolidating data
  • Timeliness
  • How encourage compliance in absence of universal regulatory mandate
  • How to address “holes” in organizational structures in the short run
    • Both “missing LEIs” for one side of relationship and entire missing branches of organizations
  • Appropriate sources of data for validation
• Iterative proof of concept planned
• On-going consultation with regulators and private sector
• First phase of implementation expected around end of 2015
Standardization

• LEI is a classic example of the role of standardization in making big data useable
  • ISO TC 68 deals with all standards for the financial industry
• Other standards bodies
• Broad role for standardization
• Importance of metadata
  • “local standardization”
  • Data point modeling
“Big data hubris”

• Big Data: Are we making a big mistake? (Tim Harford in the *Financial Times*)
  • Often clash with accepted statistical procedures
  • Informative sampling, **nonstationarities**, ambiguous definitions/frame of reference, fuzzy provenance, other little-data problems writ large, etc
  • Are we looking for scientific results or are we surfing on short-horizon buzz?

• Almost certainly not a mistake in general, but:

**There is no substitute for thinking about what you are doing**
Are we all Bayesians now?

• Modeling, matching, simulating
• “Priors” over unknown population definitions or data content?

• SOI data may provide a key universe anchor for other parts of the “data collage”
Big Data Skills Mix

• **Domain expert**
  • Traditional strength of that SOI embodies as an organization
  • Data curation and matching
  • Also able to do other things: role in analytics

• **Researcher** (methodology, statistics, mathematics)

• **Computer scientist**

• **System administrator**

• **Significant interdependencies**
Thanks!
Thinking Big About SOI Data

Panel Discussion

Discussion Questions

• Given current resources, what changes in current products or production methods should SOI consider in order to free resources for new work?

• What changes to public data releases should be developed using linked flow-through data?
Lunch: 1 hour 15 minutes

Next: Are Piketty and Zucman Getting it Right?
# 2015 Consultants Panel Agenda (afternoon sessions)

<table>
<thead>
<tr>
<th>Session</th>
<th>Speaker</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are Piketty and Zucman Getting it Right? Evaluating Distributional Statistics Based on Aggregate Data</td>
<td>Jenny Bourne</td>
<td>1:20 pm</td>
</tr>
<tr>
<td>More Than They Realize: The Income of the Wealthy and the Piketty Thesis</td>
<td>John Sabelhaus</td>
<td></td>
</tr>
<tr>
<td>Mortality Differentials - How Much Longevity Can Money Really Buy?</td>
<td>Brian Raub</td>
<td></td>
</tr>
<tr>
<td>Discussant</td>
<td>Len Burman</td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>Panel</td>
<td></td>
</tr>
<tr>
<td>A Productive Partnership, Joint Work with Stanford</td>
<td>David Grusky</td>
<td>2:30 pm</td>
</tr>
<tr>
<td>Discussion</td>
<td>Panel</td>
<td></td>
</tr>
<tr>
<td>An Overview of the SOI Consultants Panel</td>
<td>George Plesko</td>
<td>3:10 pm</td>
</tr>
<tr>
<td>Discussion</td>
<td>Panel</td>
<td></td>
</tr>
<tr>
<td>Adjorn</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ARE PIKETTY AND ZUCMAN GETTING IT RIGHT?
More Than They Realize: The Income of the Wealthy and the Piketty Thesis

Jenny Bourne, Carleton College, Economics Department
Eugene Steuerle and Ellen Steele, Urban Institute
Brian Raub and Joseph Newcomb, Statistics of Income, IRS
Measuring Inequality

Realized Capital Income ➔ Wealth

The Wealthy Differ from You and Me

• Higher economic returns (portfolio holdings, selection bias)
• Lower realized returns (ability to re-categorize and time income)
• Effective tax rate on economic income from capital <10%
Data

Decedents from 2007 who filed estate tax return Form 706 (total N=36,889; stratified sample N=12,296)

Federal tax returns for 2002-2007 Form 1040

Total gross estate $229 billion
Mean gross estate $6.2 million, mean net estate $6 million
Median gross estate $3.2 million, median net estate $3.15 million
# Descriptive Information

**Demographic trait**

- Age at death 70 years or older: 80.3
- Male: 57.0
- Married at death: 49.3

**Marital status 2002--2007**

- Always married: 45.6
- Always not married: 38.8
- Mixed: 5.6

**Net estate category ($million)**

- 0--2: 3.5
- 2--5: 70.9
- 5--10: 17.1
- 10--50: 7.7
- 50--100: 0.5
- Over 100: 0.3
# Three Measures of Capital Income

<table>
<thead>
<tr>
<th>CAPY1</th>
<th>CAPY2</th>
<th>TAXY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxable interest</td>
<td>Taxable interest</td>
<td>Taxable interest</td>
</tr>
<tr>
<td>+Tax-exempt interest</td>
<td>+Tax-exempt interest</td>
<td></td>
</tr>
<tr>
<td>+Capital gains</td>
<td>+Capital gains</td>
<td>+Taxable capital gains</td>
</tr>
<tr>
<td>+Dividends</td>
<td>+Dividends</td>
<td>+Dividends</td>
</tr>
<tr>
<td>+Gains from sale of</td>
<td>+Gains from sale of</td>
<td>+Gains from sale of</td>
</tr>
<tr>
<td>business property</td>
<td>business property</td>
<td>business property</td>
</tr>
<tr>
<td>+½ Schedule C</td>
<td>+½ Schedule C</td>
<td>+½ Schedule C</td>
</tr>
<tr>
<td>+¾ Schedule E</td>
<td>+¾ Schedule E</td>
<td>+¾ Schedule E</td>
</tr>
<tr>
<td>+½ Schedule F</td>
<td>+½ Schedule F</td>
<td>+½ Schedule F</td>
</tr>
<tr>
<td>+½ IRA distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+½ Pensions&amp;annuities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-Interest deduction</td>
<td>-Interest deduction</td>
<td>-Interest deduction</td>
</tr>
</tbody>
</table>
CAPY1 as a Proportion of Total Income, by Status (2002-2006)
CAPY1 as a Proportion of Total Income, by Wealth Category (2002-2006)
Capital Gains as a Proportion of CAPY1, by Status (2002-2006)

old

young
Capital Gains as a Proportion of CAPY1, by Wealth Category (2002-2006)
TAXY as a Proportion of CAPY1, by Wealth Category (2002-2006)
Net Capital Income as Proportion of Net Estate (2002-2006)

**CASY1 (old)**
- Married female
- Married male
- Single female
- Single male

**TAXY (old)**

**CASY1 (young)**

**TAXY (young)**
Measures of Net Return to Capital by Net Estate Category, Separately by Year

2002
(S&P -21.97)

2003
(S&P 28.36)

2004
(S&P 10.74)

2005
(S&P 4.83)

2006
(S&P 15.61)
Taxable Capital Income as a Percentage of Net Estate by Wealth Category

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
Taxable Capital Income as a Percentage of Net Estate by Wealth Category (non-homeowners)
Proportion of Realized Capital Gains Relative to 5-Year Average, by Wealth Class, 2002--2006
Percentages of Estates with Taxable Capital Income in a Particular Range, by Wealth Category

TAXY (2002)

TAXY (2003)

TAXY (2004)

TAXY (2005)

TAXY (2006)

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
Regression Analysis

\[ \ln( \text{TAXY}) = \alpha + \beta_1 \ln \text{net estate} + \beta_2 \text{age} + \beta_3 \text{age squared} + \beta_4 (D \text{ always married}) + \beta_4(D \text{ always single}) + \beta_5(D \text{ male}) + \beta_6(D \text{ male*always married}) + \beta_7(D \text{ male*always single}) + \beta_8 \ln \text{charitable deduction} + \beta_9 \text{homepct} + \beta_{10}(D \text{ estate}>$50 \text{ million}) + \beta_{11}(D \text{ estate}>$50 \text{ million} * \ln \text{net estate}) \]

Robust standard errors

Adjusted R-squared 0.304-0.340
Elasticity of Taxable Capital Income with respect to Net Estate, by Year and Wealth Category
Tax Policy Implications

Stock market gains 2003 = 28.36%
Long-term real return on stock ~ 7%
Typical realized return by wealthy individual ~ 2%
Top marginal income tax rate = 35%

Effective marginal income tax rate = 10%
Caveat: does not account for other taxes paid
The Wisdom of Warren Buffett

“I still pay a lower tax rate than my secretary”

Compares capital gains tax rate to rate on labor income

Doesn’t account for discretion in realization

Wealthy have higher economic returns and lower realized returns than non-wealthy

Inferring wealth distribution from realized capital income understates wealth inequality
Goals for this Paper

• U.S. top income and wealth shares are high and rising, but how high, and how fast?

• Widely-cited top shares estimates based on administrative income tax data diverge from Survey of Consumer Finances (SCF)
  o Piketty and Saez (2003, updated)
  o Saez and Zucman (2014)

• Primary goal is to understand why the two approaches diverge, and solve for biases
Sources: Survey of Consumer Finances (SCF) and Piketty and Saez (2003 + updates). SCF incomes are collected for the calendar year prior to each triennial survey. See Appendix B for details on Administrative, SCF Bulletin, and SCF Market income concepts. Income thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Sources: Survey of Consumer Finances (SCF) and Piketty and Saez (2003 + updates). SCF incomes are collected for the calendar year prior to each triennial survey. See Appendix B for details on Administrative, SCF Bulletin, and SCF Market income concepts. Income thresholds for identifying the top 0.1% of households and tax units are reported in Appendix C.
Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Top Wealth Shares Reconciliation

• Why do SCF and tax-based “Gross Capitalization” top wealth share estimates diverge?
  o Capitalized approach uses taxable SOI incomes for income-generating assets, imputations for rest
  o Capitalized calibrated to Financial Accounts (FA)
  o SCF and FA balance sheet concepts diverge
  o SCF and FA estimated aggregates diverge
  o Some implied capitalization factors problematic

• 160 million tax units versus 120 million families

• SCF (by rule) does not survey Forbes 400
Gross Capitalization (GC) Approach

• Given taxable capital income type \( k = 1, \ldots, 9 \) along with estimates of wealth that do not generate taxable income, for family \( i \)

\[
\text{wealth}^G_{i} = \sum_{\forall k} \frac{SOI \ income^k_{i}}{ror^k} + \text{nonfinancial}_{i}
\]

• In practice, Saez and Zucman (2014) compute \( ror \) for each asset \( k \) to calibrate to FA aggregates

\[
ror^k = \frac{\sum_{\forall i} \left( SOI \ income^k_{i} \right)}{FA \ asset^k}
\]
Reconciling Survey of Consumer Finances (SCF) and Administrative Data
Top 1% Wealth Shares

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Reconciling Survey of Consumer Finances (SCF) and Administrative Data
Top 1% Wealth Shares

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Reconciling Survey of Consumer Finances (SCF) and Administrative Data
Top 1% Wealth Shares

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Reconciling Survey of Consumer Finances (SCF) and Administrative Data
Top 1% Wealth Shares

Percent Shares

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
Reconciling Survey of Consumer Finances (SCF) and Administrative Data

Top 1% Wealth Shares

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C.
95% Confidence Interval on Reconciled Top 1% Wealth Shares

- Administrative Data
- SCF Bulletin Wealth, Households
- SCF Benchmarked to FA Values, Tax Units, Plus Forbes 400

Wealth Reconciliation at the Very Top

• Recent top 1% wealth shares largely reconciled, some remaining trend divergence

• Still, recent top 0.1% wealth share is greater in capitalized administrative tax data
Top 0.1% Wealth Shares

- **Administrative Data**
- **SCF Bulletin**
- **SCF Benchmarked to FA Values, Tax Units, Plus Forbes 400**
Administrative Data -- Top 0.1% Asset Composition
Wealth Reconciliation at the Very Top

• Fixed income explains why top 0.1% wealth share greater in capitalized administrative tax data

• Look closer at asset composition and RoR
  o Fixed-income assets were 25%, now 45% of assets
  o Bonds ≈ 1/3\textsuperscript{rd}, deposit accounts are the other 2/3\textsuperscript{rds}.
  o Do the top 0.1 really hold savings deposit accounts?

• Rate of return on fixed-income = 1 pct. (for all)
  • → capitalization factor of 100x for interest income
  • Compare to market rates of return
Capitalize Top 0.1% Interest Income with SZ methodology (i.e. 1 pct. RoR in 2012)

Administrative Data

SCF Benchmarked to FA Values, Tax Units, Plus Forbes 400

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C. Shaded area represents 95% confidence interval based on sampling and imputation variance.
Capitalize Top 0.1% Interest Income with 10-year Treasury Yield (i.e. 2 pct. RoR ‘12)

Sources: Survey of Consumer Finances (SCF) and Saez and Zucman (2014). See Appendix B for details on SCF and FA wealth concepts. Wealth thresholds for identifying the top 1% of households and tax units are reported in Appendix C. Shaded area represents 95% confidence interval based on sampling and imputation variance.
Conclusions

• Estimates of top income and wealth shares from SCF can be reconciled with estimates derived directly from administrative tax data

• SCF suggests that administrative-based top share estimates too high and rising too fast

• Reconciliations offer direction for future work, as broader income and wealth measures are likely to further reduce estimated top shares
Thanks!

jesse.bricker@frb.gov
alice.henriques@frb.gov
jake.krimmel@frb.gov
john.sabelhaus@frb.gov
If Time: Expanded Income Concept

• Cannot distribute all of NIPA personal income, but can at least bracket top income shares

• Assume that missing income in every year, starting in 1970, is allocated per tax unit
  o Top 1 percent gets only 1% of the missing income

• Top 1% income levels and growth much more muted, and tax unit adjustment would add

• Extreme assumption, but brackets truth: missing incomes are transfers, non-wage compensation, retirement saving
Figure 10. Effect of Allocating Missing Personal Income on Top 1% Income Shares

Sources: Bureau of Economic Analysis and Piketty and Saez (2003 + updates). Adjustment assumes all missing NIPA income (government transfers, unreported income, retirement saving, employer-provided health) are allocated to top share group in proportion to numbers of units, not in relation to other incomes. See Appendix B for a discussion of the mismatch between NIPA and administrative data concepts.
Mortality Differentials – How Much Longevity Can Money Really Buy?

Barry Johnson

Brian Raub

Statistics of Income, IRS
SOI Personal Wealth Study - Background

- Uses Federal estate tax data to estimate wealth of the living population with wealth at or greater than filing threshold
- Based on well-established “Estate Multiplier Technique”

\[
\text{MULT} = \frac{1}{(p \cdot r)} \text{ where:}
\]

\[
p = \text{probability of selection to the estate tax sample,}
\]

\[
r = \text{mortality rate appropriate to wealthy individuals,}
\]

- Assumes that estate tax decedents are random sample of the living wealthy population
Mortality Rates

• Proper mortality rates for the wealthy are key
  • Estimates are sensitive to rates
• The wealthy have lower mortality than population as a whole
• For past decade, SOI has used mortality rates for annuitants produced by Society of Actuaries
  • Annuitant 2000 mortality tables (A2000)
Mortality Rates

• Two new sources of mortality rates for the wealthy:
  • Annuitant 2012 mortality tables (A2012)
    • Successor to the A2000 tables
    • Based on study from 2000-2004
  • Saez-Zucman mortality rates (SZ)
    • Based on modelling wealth using capitalized income tax data
    • Linked to Social Security data to identify deaths
Mortality Rates

Figure 1: Mortality, Males in General Population

- NCHS Data 2008
- Saez Tax Return Estimates
### Mortality Rates

#### Mortality Differentials – How Much Longevity Can Money Really Buy?

<table>
<thead>
<tr>
<th>Period 2004-2008</th>
<th>Males</th>
<th></th>
<th></th>
<th>Females</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 10%</td>
<td>Top 5%</td>
<td>Top 1%</td>
<td>Top 10%</td>
<td>Top 5%</td>
<td>Top 1%</td>
</tr>
<tr>
<td>Age 30-49</td>
<td>0.52</td>
<td>0.44</td>
<td>0.53</td>
<td>0.51</td>
<td>0.57</td>
<td>0.40</td>
</tr>
<tr>
<td>Age 50-64</td>
<td>0.61</td>
<td>0.53</td>
<td>0.43</td>
<td>0.67</td>
<td>0.57</td>
<td>0.71</td>
</tr>
<tr>
<td>Age 65-79</td>
<td>0.77</td>
<td>0.71</td>
<td>0.60</td>
<td>0.76</td>
<td>0.73</td>
<td>0.69</td>
</tr>
<tr>
<td>Age 80+</td>
<td>0.97</td>
<td>0.91</td>
<td>0.92</td>
<td>0.97</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>SZ Diff from NCHS</strong></td>
<td>0.58</td>
<td>0.55</td>
<td>0.56</td>
<td>0.56</td>
<td>0.53</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Preliminary Results

Figure 2a: Number of Top Wealth Holders with $2 Million or More in Assets, 2007

<table>
<thead>
<tr>
<th>Methodology</th>
<th>2000</th>
<th>2012</th>
<th>SZ 95P</th>
<th>SZ 99P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actuary</td>
<td>2.3</td>
<td>2.7</td>
<td>2.8</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Figure 2b: Total Net Worth of Top Wealth Holders with $2 Million or More in Assets, 2007

<table>
<thead>
<tr>
<th>Methodology</th>
<th>2000</th>
<th>2012</th>
<th>SZ 95P</th>
<th>SZ 99P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actuary</td>
<td>12.1</td>
<td>14.2</td>
<td>14.7</td>
<td>15.6</td>
</tr>
</tbody>
</table>
Preliminary Results

Figure 3: Mortality Rate, Wealthy Males

- Actuary Data
- SZ/NCHS Data
- SZ Data

Mortality rate per 1000

Age

The content of this presentation is the opinion of the writer(s) and does not necessarily represent the opinion of the Internal Revenue Service.
Figure 4a: Number of Estate Tax Decedents, 2007

Preliminary Results
Preliminary Results – Sex Distribution

Figure 5a: Age Distribution of Females, Actuary Data

- 90 and older: 3%
- 80 under 90: 12%
- 70 under 80: 17%
- 60 under 70: 21%
- 50 under 60: 25%
- Under 50: 22%

Figure 5b: Age Distribution of Females, Saez-Zucman Data

- 90 and older: 6%
- 80 under 90: 8%
- 70 under 80: 20%
- 60 under 70: 21%
- 50 under 60: 15%
- Under 50: 30%
Preliminary Results – Age Distribution for Males

Figure 6b: Smoothed Male Mortality Compared to Actuary Data
Preliminary Results – Age Distribution for Females

**Figure 7a:** Number of Top Wealth Holders with $2 Million or More in Assets, 2007

- Actuary data: 2.7
- SZ Predicted: 2.7

**Figure 7b:** Total Net Worth of Top Wealth Holders with $2 Million or More in Assets, 2007

- Actuary data: 14.2
- SZ Predicted: 14.3
Conclusion

• After smoothing to account for broad age categories, preliminary findings that Saez-Zucman and actuarial mortality rates produce very similar wealth estimates

• Use of broad age categories has significant impact on estimates

• Use of actuarial rates may be preferable due to less age bias
Discussion Question

• Which approach to estimating mortality for the wealthy do you think is the most methodologically sound for SOI?
A PRODUCTIVE PARTNERSHIP, JOINT WORK WITH STANFORD
THREE SOI-STANFORD PROJECTS

NEW ESTIMATES OF INTERGENERATIONAL PERSISTENCE

EXPLOITING THE OCCUPATION FIELDS IN FORM 1040

BUILDING A NEW INTERGENERATIONAL PANEL
NEW ESTIMATES OF INTERGENERATIONAL PERSISTENCE

U.S. HAS LONG-STANDING INTEREST IN ECONOMIC MOBILITY ... AND SOME TAX PROVISIONS ARE PARTLY JUSTIFIED AS MOBILITY-INCREASING (E.G., LOW-INCOME TAX CREDITS, ESTATE TAX)

IT IS ACCORDINGLY IMPORTANT TO MONITOR INTERGENERATIONAL ECONOMIC MOBILITY AND ASSESS WHETHER RAMPED-UP TAX POLICY MIGHT BE NEEDED

THE PROBLEM: THE ADMINISTRATIVE DATA REVOLUTION HAS NOT RESOLVED HOW MUCH INTERGENERATIONAL PERSISTENCE THERE IS
THE EVIDENCE DEFICIT

**PROBLEM #1:** A WIDE RANGE OF ESTIMATES (FROM SURVEY AND ADMINISTRATIVE DATA)

- Solon’s (2008) meta-analysis of survey and administrative evidence: “inter-generational earnings elasticity in the U.S. may well be as large as 0.5 or 0.6”
- Chetty et al. (2014): Preferred estimate of 0.34 (for intergenerational income elasticity)
- A wide range: are one-third of percent income differences transmitted from one generation to next? Or is it two-thirds?

**PROBLEM #2:** Paucity of tax-return evidence ... because of methodological problems in estimating intergenerational elasticity (IGE)

We need to get it done: IGE is key workhorse measure (expected percent change in children’s income given a one percent increase in parental income)
HOW DO WE GET IT DONE?

A SIMPLE GOAL: ESTABLISH HOW MUCH INTERGENERATIONAL PERSISTENCE THERE IS IN THE U.S. TODAY

TWO PROBLEMS NEED TO BE SOLVED

• DATA PROBLEM
• METHODS PROBLEM
### TABLE 1: CONSTRUCTION OF SOI-M PANEL

<table>
<thead>
<tr>
<th>DATA</th>
<th>PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAX RETURNS FROM SOI FAMILY PANEL (1987-1996)</td>
<td>SOURCE OF PARENTAL INCOME DATA AND PARENT-CHILD SOCIAL SECURITY LINKS (WITH CLAIMED CHILDREN THEN TRACED FORWARD)</td>
</tr>
<tr>
<td>POPULATION OF TAX RETURNS (1997-1998)</td>
<td>INCOME DATA FOR 1987 PARENTS (INCOME SECURED UP TO YEAR WHEN CHILD BECOMES 23 YEARS OLD)</td>
</tr>
<tr>
<td>POPULATION OF TAX RETURNS (1998-2010)</td>
<td>INCOME DATA FOR CHILDREN AND THEIR SPOUSES</td>
</tr>
<tr>
<td>W2 FORMS (1999-2010)</td>
<td>EARNINGS OF CHILDREN, INCLUDING NONFILING CHILDREN</td>
</tr>
<tr>
<td>1040SE FORMS (1999-2010)</td>
<td>SELF-EMPLOYMENT INCOME</td>
</tr>
<tr>
<td>SSA DATA MASTER FILE</td>
<td>DEMOGRAPHIC INFORMATION (AGE AND GENDER OF PARENTS AND CHILDREN, YEAR OF DEATH OF CHILDREN)</td>
</tr>
<tr>
<td>1099G FORMS</td>
<td>UNEMPLOYMENT INCOME OF NONFILING CHILDREN</td>
</tr>
<tr>
<td>CURRENT POPULATION SURVEY (CPS)</td>
<td>IMPUTED INCOME FOR NONFILING CHILDREN WITHOUT W-2 OR UI DATA (USING CPS FILING STATUS VARIABLES)</td>
</tr>
</tbody>
</table>

CHETTY ET AL.’S (2014) ANALYSIS OF 1996-2012 TAX DATA YIELDS CHILDREN 29-32 YEARS OLD IN 2011-12 ... TOO EARLY IN CAREER TO YIELD GOOD IGE ESTIMATES?

**SOLUTION: CONSTRUCT SOI-M PANEL**

- **POPULATION OF INTEREST: CHILDREN BORN 1972-75 WHO WERE LIVING IN U.S. IN 1987**
- **START WITH SOI 1987-96 FAMILY PANEL**
- **ADD OTA REFRESHMENT SEGMENT TO CAPTURE 1987 NONFILERS AND ADD 1997-98 IRTF DATA FROM CDW TO COMPLETE DATA FOR PARENTS**
- **USE CDW TO OBTAIN CHILDREN’S IRTF DATA FOR 1998-2010**

**WHY SOI-M PANEL IS SO ATTRACTION**

- **REDUCES LIFECYCLE BIAS BY EXAMINING MOBILITY OF CHILDREN AGES 35-38 IN 2010**
- **ADDRESSES ATTENUATION BIAS BY USING 9 YEARS OF PARENTAL INFORMATION**
ADDRESSING THE METHODS PROBLEM

OLS LOG-LOG ESTIMATOR IS METHODOLOGICAL CONVENTION:

\[ E(\ln Y |x) = \beta_0 + \beta_1 \ln x \]

TWO CHOICES – BOTH BAD – IF ONE OPTS FOR OLS LOG-LOG ESTIMATOR

• DROP CHILDREN WITHOUT EARNINGS OR INCOME → SELECTION BIAS (AN IGE THAT PERTAINS TO “WHEN THINGS ARE GOING WELL”)
• KEEP CHILDREN WITHOUT EARNINGS OR INCOME AND ASSIGN ARBITRARY POSITIVE VALUE → ESTIMATES ARE EXTREMELY SENSITIVE TO CHOSEN VALUES

SOLUTION: DEFINE THE ESTIMAND CORRECTLY (IGE_e)

\[ \ln E(Y|x) = \alpha_0 + \alpha_1 \ln x \]

ELIMINATES SELECTION BIAS AND EXTREME SENSITIVITY OF ESTIMATES
DATA AND METHODOLOGICAL FIXES REDUCE BIASES

LATE THIRTIES SAMPLE (VIA SOI-M PANEL) → REDUCES LIFECYCLE BIAS

NINE YEARS OF PARENTAL INFORMATION (VIA SOI-M PANEL) → REDUCES ATTENUATION BIAS

CORRECT ESTIMATOR → REDUCES SELECTION BIAS

RELAX CONSTANT-ELASTICITY ASSUMPTION → REDUCES FUNCTIONAL-FORM BIAS
INCOME AND EARNINGS MEASURES

ANNUAL TOTAL INCOME OF PARENTS AND CHILDREN

ANNUAL AFTER-FEDERAL-TAX INCOME ("DISPOSABLE INCOME") OF PARENTS AND CHILDREN

INDIVIDUAL EARNINGS OF CHILDREN (INCLUDING EARNINGS FROM SELF-EMPLOYMENT)
GLOBAL IGE, 0.56, IS AT UPPER END OF ESTIMATES

LEFT TAIL IS FLAT BUT THEN SLOPE INCREASES: CONVEX CURVE

GLOBAL IGE: 0.56

Log of men’s expected earnings as a function of log parental disposable income
MEN’S TOTAL INCOME CURVE

Log of men's expected total income as a function of log parental total income

GLOBAL IGE: 0.52

MEN’S TOTAL INCOME IGE, 0.52, AT UPPER END OF RANGE OF ESTIMATES

ANOTHER CONVEX CURVE WITH ELASTICITY BETWEEN P50 AND P90 SUBSTANTIALLY LARGER THAN THAT BETWEEN P10 AND P50
WOMEN’S TOTAL INCOME CURVE

Log of women’s expected total income as a function of log parental total income

GLOBAL IGE: 0.47

GLOBAL IGE ONLY SLIGHTLY LOWER

AGAIN CONVEX: ELASTICITY BETWEEN P50 AND P90 SUBSTANTIALLY LARGER THAN BETWEEN P10 AND P50
CONCLUSION

IF ONE CORRECTS FOR SELECTION, ATTENUATION, LIFECYCLE, AND FUNCTIONAL-FORM BIASES

→ TOTAL INCOME IGES AND MEN’S EARNINGS IGE ARE VERY HIGH ... AND AT UPPER END OF EXISTING ESTIMATES
NEXT STEPS

BETTER DOCUMENTATION OF NEW SOI-M PANEL

UPDATE SOI-M PANEL WITH POST-2010 DATA

DEVELOP PROTOCOL FOR ANNUAL REPORTING OF KEY IGES BY SOI
THREE PROJECTS

NEW ESTIMATES OF INTERGENERATIONAL PERSISTENCE

EXPLOITING THE OCCUPATION FIELDS IN FORM 1040

BUILDING A NEW INTERGENERATIONAL PANEL
GOALS OF SOI-STANFORD OCCUPATION PROJECT

EXPLOIT OCCUPATION FIELDS ON FORM 1040 BY DEVELOPING CODING SCHEME AND APPLYING IT TO SOI SAMPLES

TAX POLICY USES: IMPROVED ESTIMATES OF INTERGENERATIONAL MOBILITY
• OCCUPATIONS CONVEY INFORMATION ON LIFETIME EARNINGS AND INCOME
• COMBINING ECONOMIC AND OCCUPATION REPORTS CORRECTS FOR UNDERESTIMATES OF INTERGENERATIONAL PERSISTENCE

CENSUS USES: OCCUPATION FIELDS FROM FORM 1040 MAY BE USEFUL FOR FILLING IN MISSING CENSUS, ACS, AND CPS REPORTS (ASSUMING REG CHANGES)

GENERAL SURVEY USES: ESTABLISH VIABILITY OF SHORT-RESPONSE OCCUPATION ITEMS
FIVE STEPS

**STEP 1 (LINKING):** LINK SELECTED FIELDS FROM 2011-12 TAX YEARS TO TRAINING SET DRAWN FROM 2011-12 CPS ASEC AND 2011-12 ACS (WHICH INCLUDE 2010 CENSUS OCCUPATION CODES)

**STEP 2 (MACHINE LEARNING ON TRAINING SET):** APPLY MACHINE LEARNING TO TRAINING SET TO DEVELOP CODING ALGORITHM USING VARIABLES ON FORM 1040, FORM W-2, SSA MASTER FILE, AND OTHER SOURCES

**STEP 3 (TEST AGAINST BALANCE OF DATA):** TEST RESULTING PROTOCOL AGAINST BALANCE OF ASEC AND ACS DATA

**STEP 4 (DEVELOP CODING SCHEME):** IF RESULTS ARE SATISFACTORY, DEVELOP AGGREGATED VERSION OF 2010 OCCUPATION SCHEME THAT YIELDS ACCEPTABLY LOW ERROR RATES (ALSO ALLOW FOR MULTIPLE IMPUTATION)

**STEP 5 (APPLY TO SOI AND POPULATION DATA):** APPLY SCHEME TO CURRENT AND HISTORICAL SOI FILES AND RECENT POPULATION FILES
VARIABLES FOR MACHINE LEARNING

**FORM 1040:** NAME; ADDRESS; FILING STATUS; WAGES, SALARIES, & TIPS; BUSINESS INCOME; CAPITAL GAIN OR LOSS; RENTAL INCOME; FARM INCOME OR LOSS; UNEMPLOYMENT COMP.; SS BENEFITS; EDUCATOR EXPENSES; BUSINESS EXPENSES; STUDENT LOAN INTEREST DEDUCTION; TUITION AND FEES; PAID PREPARER FLAG

**SCHEDULES A, C, D, E:** UNREIMBURSED EMPLOYEE EXPENSES; NAME OF PROPRIETOR; PRINCIPAL BUSINESS; BUSINESS NAME; EIN; GROSS RECEIPTS; EXPENSES FOR BUSINESS USE OF HOME; NET SHORT TERM GAIN OR LOSS FROM PARTNERSHIPS, S CORPORATIONS, ESTATES AND TRUSTS; RENTS AND ROYALTIES

**FORM 1099-MISC:** PAYER’S NAME, ADDRESS, FEDERAL ID NUMBER; RENTS; ROYALTIES

**FORM W-2:** EMPLOYER NAME AND ADDRESS, EIN; INDUSTRY

**SSA MASTER FILE:** GENDER; AGE

**VARIABLES FROM PRIOR FILING YEARS:** LAST YEAR’S OCCUPATION (AND MANY OTHERS)
THREE PROJECTS

NEW ESTIMATES OF INTERGENERATIONAL PERSISTENCE

EXPLOITING THE OCCUPATION FIELDS IN FORM 1040

BUILDING A NEW INTERGENERATIONAL PANEL
THE AMERICAN OPPORTUNITY STUDY (AOS)

THE U.S. HAS AN *UNASSEMBLED PANEL* … AND THE AMERICAN OPPORTUNITY STUDY (AOS) IS A NEW INITIATIVE TO ASSEMBLE IT

ALTHOUGH TAX DATA ARE KEY RESOURCES IN ADDRESSING LABOR MARKET ISSUES, THE AOS WOULD ALLOW US TO BETTER ADDRESS PROBLEMS ARISING FROM NONFILING AND MISSING DATA (E.G., RACE)

AND OF COURSE TAX DATA CAN ONLY BE USED FOR ANALYSES DIRECTLY RELEVANT TO TAX POLICY AND TAX ADMINISTRATION
STEP #1: LINKING RECORDS FROM ACS AND CENSUS ACROSS YEARS

ASSIGN PROTECTED IDENTIFICATION KEYS (PIKS) TO RECORDS

USE VARIABLES FROM ACS OR CENSUS LONG FORM (E.G., FIRST NAME, LAST NAME, YEAR OF BIRTH, ADDRESS, SEX) TO FIND SSN IN SSA NUMIDENT FILE

RESULT: INSTANT PANEL DATA THAT IS AUTOMATICALLY REFRESHING (I.E., NO TOP-DOWN, ARTIFICIAL DEFINITION OF POPULATION AT POINT OF CONSTRUCTION)
STEP #2: ADDING IN ADMINISTRATIVE DATA

1960-90 CENSUS (INC. EDUC., OCC., WRK. STATUS, FAMILY COMP.)

2000 CENSUS (INC., EDUC., OCC., WRK. STATUS, FAMILY COMP.)

2008-12 ACS (INCOME, EDUC., OCC., WRK. STATUS, FAMILY COMP.)

2013-18 ACS (INCOME, EDUC., OCC., WRK. STATUS, FAMILY COMP.)

SSA EARNINGS RECORDS (1978-2012)

IRS 1040 DATA (1995-2012)

PROGRAM DATA (E.G., UI, SNAP, TANF)

ONCE PIKS ARE ASSIGNED, LINKAGES TO ADMINISTRATIVE DATA CAN ALSO BE MADE (CONDITIONAL OF COURSE ON APPROVALS TO DO SO)
STEP #3: LINKING CHILDREN WITH PARENTS

1960-90 CENSUS (INC. EDUC., OCC., WRK. STATUS, FAMILY COMP.)

2000 CENSUS (INC., EDUC., OCC., WRK. STATUS, FAMILY COMP.)

2008-12 ACS (INCOME, EDUC., OCC., WRK. STATUS, FAMILY COMP.)

2013-18 ACS (INCOME, EDUC., OCC., WRK. STATUS, FAMILY COMP.)

SSA EARNINGS RECORDS (1978-2012)

IRS 1040 DATA (1995-2012)

PROGRAM DATA (E.G., UI, SNAP, TANF)

KIDLINK FILES (E.G., FORM SS-5 IF HISTORICAL FORMS AVAILABLE)

CORESIDENCY IN ACS, CENSUS LONG FORM

PARENTAL REPORTS OF CHILDREN’S SSN TO IRS

THE RESULT: AOS
STEP #4: SLIPPING IN THE SURVEY

SURVEYS WITH IDENTIFIERS CAN BE SLIPPED IN THE SURVEY AS A LEAN AND MEAN VALUE-ADDED INSTRUMENT
CRITICISMS AND CHALLENGES

“THE AOS IS LARGE AND UNGAINLY AND WILL SINK UNDER ITS OWN WEIGHT”

“THE AOS IS A REGISTER ON THE SLY”

“THE FORMIDABLE SECURITY ISSUES WILL BE THE UNDOING OF THE AOS”

“THE AOS POPULATION IS POORLY SPECIFIED”

“APPROVALS TO LINK TO ADMINISTRATIVE DATA WILL NEVER BE SECURED”
“THE AOS IS A REGISTER ON THE SLY”

TECHNICAL RESPONSE: ON-DEMAND DATA LINKAGE SERVICE ... NOT A REGISTER

HEAD-ON RESPONSE: THE PAPERWORK REDUCTION ACTS OF 1980 AND 1985 MANDATE THE AOS
“FORMIDABLE SECURITY ISSUES WILL BE UNDOING OF AOS”

FIRST-ORDER CONCERNS: LEGITIMATE THREATS TO SECURITY

SECOND-ORDER CONCERNS: FALLOUT FROM UNWARRANTED PUBLIC WORRIES

FIRST-ORDER CONCERNS

IDENTIFIERS ARE JUST PRODUCTION TOOLS … NO NEW CONCERNS

RELEASED TO CAREFULLY VETTED RESEARCH AND RESEARCHERS … NO NEW CONCERNS

ANALYZED IN RDCs OR, IN THE CASE OF ESPECIALLY SENSITIVE DATA, RDCs ON STEROIDS … NO NEW CONCERNS

CANNOT RULE OUT LEGITIMATE WORRIES (AND HENCE OPEN DISCUSSION IS NEEDED)

SECOND-ORDER CONCERNS

STANDARD PRESCRIPTION FOR MISINFORMATION: OPEN DISCUSSION
NEXT STEPS

DIGITIZING 1960-90 CENSUSES
IMPROVING PIKING METHODOLOGY
IMPROVING INTERGENERATIONAL LINKAGES
REDUCING SECURITY CONCERNS
FUNDING
PAYOFF TO AOS

IMPROVED EVIDENCE ON LABOR MARKET OUTCOMES AND INTERGENERATIONAL MOBILITY
LOW COST POLICY AND PROGRAM EVALUATION
REDUCED RELIANCE ON SURVEYS
CROSS-SOURCE MISSING DATA FILL-INS
SAMPLING FRAME
Panel Discussion
A Productive Partnership, Joint Work With Stanford
AN OVERVIEW OF THE SOI CONSULTANTS PANEL
Panel Discussion

An Overview of the SOI Consultants Panel
The SOI Panel
What does the SOI Panel look like?

Background
Expertise
Service

Survey responses of 13 panel members
What does the SOI Panel look like?

Background
Expertise
Service
## 2. Current position (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Academic</td>
<td>4</td>
<td>31%</td>
</tr>
<tr>
<td>2</td>
<td>US federal government</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>3</td>
<td>Other government</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td>4</td>
<td>Non-profit</td>
<td>5</td>
<td>38%</td>
</tr>
<tr>
<td>5</td>
<td>Private industry</td>
<td>2</td>
<td>15%</td>
</tr>
<tr>
<td>6</td>
<td>Retired</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>13</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

15 invitations to the survey
4. Have you ever worked in the US federal government? (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>12</td>
<td>92%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>
5. If yes, please specify (check all that apply):

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Executive Branch: Treasury (OTA)</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Executive Branch: IRS</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Executive Branch: Commerce (including Census)</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>Executive Branch: Other</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>Legislative Branch: JCT</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Legislative Branch: CBO</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Legislative Branch: Other</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>Federal Reserve</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Treasury “other”, and other agencies (e.g., NSF) omitted
9. Research area(s) / areas of expertise: (required - check all that apply)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Individual</td>
<td>10</td>
<td>77%</td>
</tr>
<tr>
<td>2</td>
<td>Corporate</td>
<td>6</td>
<td>46%</td>
</tr>
<tr>
<td>3</td>
<td>Partnership</td>
<td>3</td>
<td>23%</td>
</tr>
<tr>
<td>4</td>
<td>International</td>
<td>4</td>
<td>31%</td>
</tr>
<tr>
<td>5</td>
<td>Estate &amp; Gift</td>
<td>6</td>
<td>46%</td>
</tr>
<tr>
<td>6</td>
<td>State &amp; Local</td>
<td>4</td>
<td>31%</td>
</tr>
<tr>
<td>7</td>
<td>Tax-exempt</td>
<td>3</td>
<td>23%</td>
</tr>
<tr>
<td>8</td>
<td>Other</td>
<td>3</td>
<td>23%</td>
</tr>
</tbody>
</table>

Statistic | Value
---|---
Tax Administration
Statistics
record linkage and Mathematical Statistics
11. How many years have you been a member of the panel? (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-3</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>2</td>
<td>4-6</td>
<td>4</td>
<td>31%</td>
</tr>
<tr>
<td>3</td>
<td>7-9</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td>4</td>
<td>10-12</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td>5</td>
<td>13-15</td>
<td>1</td>
<td>8%</td>
</tr>
<tr>
<td>6</td>
<td>15-20</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>7</td>
<td>More than 20</td>
<td>6</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>

More than 185 total years of Panel participation
Average tenure greater than 14 years
13. Prior to joining the panel, did you have access to, and experience with, non-public SOI data? (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>11</td>
<td>85%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>2</td>
<td>15%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>
14. Since joining the panel, have you had access to, and experience with, non-public SOI data? (required)

<table>
<thead>
<tr>
<th></th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>10</td>
<td>77%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>3</td>
<td>23%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>
15. Have you ever published an article in the SOI Bulletin? (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>3</td>
<td>23%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>10</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>
17. Have you ever co-authored a paper with a SOI staff member? (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>7</td>
<td>54%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>6</td>
<td>46%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>
What do you think is the optimal size of the panel (please enter a whole number)?
What do you think is the optimal size of the panel (please enter a whole number)?

- Under 20, larger than 7 or 8.
- 15+
- 12
- 8
- 12-15
- Approximately 10
- A prime number 11 to 17, with members appointed for a fixed number of years
- 15
- 10 (or so).
- 10
- 12

Estimated mean = 12-ish
16. Do you think panel members should be appointed to serve a fixed-length term? (required)

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>6</td>
<td>46%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>7</td>
<td>54%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>13</td>
<td>100%</td>
</tr>
</tbody>
</table>
27. If terms were fixed, how many years should a term be?

(please enter a whole number)

Text Response

Not sure about fixed terms. There is the prior matter of overlapping terms of appointees that is not dealt with in the questionnaire.

5 years

5 to six years

3/5 years

6

Maybe 5 (with rotating terms).

10 years

4

5
27. If terms were fixed, how many years should a term be? 
(please enter a whole number)

Text Response

Not sure about fixed terms. There is the prior matter of overlapping terms of appointees that is not dealt with in the questionnaire.

5 years
5 to six years
3/5 years
6
Maybe 5 (with rotating terms).
10 years
4
5

Estimated mean = 6-ish
20. If panel members were appointed to fixed terms, should the appointment be renewable?

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>10</td>
<td>83%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>2</td>
<td>17%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>12</td>
<td>100%</td>
</tr>
</tbody>
</table>
26. Do you think that the panel should have a non-IRS chair or co-chair? 

<table>
<thead>
<tr>
<th>#</th>
<th>Answer</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>12</td>
<td>100%</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>12</td>
<td>100%</td>
</tr>
</tbody>
</table>
Open-ended Questions

Why do you serve on the panel
What characteristics/qualifications are necessary
What responsibilities should members have
Minimum commitments
Process to identify new members
Metrics to judge the panel’s effectiveness
Challenges in being effective
Other thoughts
2015 Consultants Panel Meeting

Closing Remarks