

---

# **An Assessment of the Need for a Redesign of the SOI Individual Tax Sample**

**June 13, 2014**

**Presentation to the SOI Advisory Panel**

**John L. Czajka**

**Coauthors: Amang Sukasih and Brendan Kirwan**

**MATHEMATICA**  
**Policy Research**

# Background

---

- **Last formal redesign in the late 1980s**
- **Collaborative effort among SOI, OTA, and Mathematica**
- **Sample designed with a target size of 95,000; implemented for TY 1991 with 126,000 returns, including 12,000 supplemental foreign returns; exceeded 330,000 returns for 2011**
- **Even without this growth a review is overdue**

## Background cont'd

---

- **Major factors contributing to sample growth**
  - Growth in the filing population
  - Increase in minimum sampling rate from 2 in 10,000 to 5 in 10,000 in 1998 and 10 in 10,000 in 2005
  - Upward shift in the income distribution only partially offset by indexing
  
- **Excluding the excess foreign returns, the increase between 1991 and 2011 was 203,000**
  - 30,000 due to population grown alone
  - 77,000 due to sampling rate changes
  - 31,000 due to growth in high-income nontaxables
  - 20,000 due to growth in other certainty strata
  - 45,000 due to additional income growth

# Meetings with JCT, OTA, BEA, CBO, TPC, NBER

---

- **Customers addressed the following issues:**
  - Income stratification
  - High-income nontaxable returns
  - Income growth and indexing
  - Interesting Return Indicator
  - Size of Continuous Work History Sample (CWHS)
  - Late filers
  - Timeliness of Complete Report
  - Usefulness of advance data
  - Usefulness of Compliance Data Warehouse (CDW)
  - Design of the public use file

# Meetings with JCT, OTA, BEA, CBO, TPC, NBER

---

- **Customers also noted the following:**
  - Person-level database of non-filers would be highly valuable to all customers
  - Sales of Capital Assets (SOCA) data have become less comprehensive with growth of pass-throughs
  - Better state estimates would be useful to some users, but other customers have no need for state estimates
  - Any post-audit information would be helpful
  - OTA, at least, is increasingly pulling together data across different SOI files

# Overview of Recommendations

---

- **Mathematica made 25 recommendations, reflecting discussions with customers and empirical analyses**
- **Unit costs of sample have declined due to combination of increased efficiency and growth in electronic filing**
- **Nevertheless, a reduction in sample size could free up resources to address other needs identified by customers**
- **Statistical efficiency of sample could be improved although optimizing the design across a wide range of items is a challenge**

## Overview of Recommendations cont'd

---

- **Several of the recommendations involve little or no change**
- **A revised design will require new stratum boundaries, a target size, and sampling rates and a suitable index**
- **At the same time, SOI needs to look at its longstanding approach to advance estimates and to SOCA**

## Recommendations requiring little or no change

---

- **Retention of key elements of the current sample design**
  - Minimum sampling rate of 1 in 1,000
  - Gross positive and negative income but replacement of some of their current components and addition of others
  - Form type as secondary stratifier but with uniform rates across types except in foreign study years
  - Selection of electronic and paper returns at the same rate (that is, no sub-stratification)
  - Certainty selection for returns with high business receipts
  - Selection on primary SSN
  - Prior year returns but with interpretation consistent with the data

## Recommendations requiring little or no change

---

- **Retention of other elements of file production**
  - Procedures for handling misclassification error, which is likely to be reduced by recommended changes in the income stratifier
  - Procedures for handling missing returns, which are rare and becoming more so
  - Current release schedule for the final file

## Recommendations requiring minimal new work

---

- **Revision of elements of the current sample design**
  - Elimination of sub-stratification by degree of interest
  - Certainty selection of high-income nontaxables only if legally required; otherwise, sample by stratum at rates sufficient to meet annual reporting requirement
  - Assessment of whether gross positive income should be replaced by AGI when the latter is larger
- **Development of comprehensive documentation of the sample design**
- **Make available to CDW users the recent comparison of SOI and CDW aggregates**

# Recommendations requiring additional research

---

- **Sample redesign**
  - Revision of the income stratum boundaries
  - Optimal allocation of the sample with these new boundaries and a new sample size
  - Replacement of the current index, based on GDP, with one based on personal income
- **Develop as an annual product a person-level database of non-filers**
- **Determine how any sample changes might be reflected in the public use file and communicate to the major users**
- **Explore evidence of declining quality in SSNs**

## Recommendations requiring customer input

---

- **Follow up on customer comments about the declining usefulness of advance estimates**
- **Follow up on customer comments about the declining values of the SOCA study**
- **Consider ways to assess the quality of CDW items that are too rare to estimate precisely with the Individual sample**
- **Ascertain what post-audit data are available and whether it could provide some sense of what the Individual sample data would look like if it were post-audit**