Jim Nunns *Urban-Brookings Tax Policy Center and* SOI *IPA mobility program*

Dan Feenberg *NBER and* SOI *IPA mobility program*

Victoria Bryant *Statistics of Income*

John Czajka *Mathematica Policy Research*

BALANCING UTILITY AND PRIVACY: PROPOSED CHANGES TO THE PUF

Background

SOI's Public Use File (PUF) has provided high-quality microdata for tax policy analysis since 1960

SOI and Mathematica perform rigorous nondisclosure checks on the PUF for each year

Periodic in-depth reviews and strengthening of PUF nondisclosure procedures

Annual, and periodically more in-depth reviews of PUF quality

In Fall 2010, Susan Boehmer formed Working Group to review PUF nondisclosure procedures and quality and utility of PUF data

Working Group headed by Dave Paris and includes Dan Feenberg, John Czajka, Victoria Bryant and other SOI staff

Background - Continued

The PUF is a subsample of the INSOLE (INdividual and SOLE proprietor) file

- Returns filed for years more than three years prior to the current year are included in the INSOLE but excluded from the PUF

INSOLE is a highly stratified sample of returns filed each year

SOI uses the INSOLE to create tables in the annual *Individual Complete Report* and for other statistical purposes

JCT and OTA use the INSOLE for their microsimulation models and other tax analysis

Most recently released PUF is for 2008

Changes to PUF to be implemented for 2009

Filing Population and INSOLE Sample for 2009

Strata	Strata Boundaries in 2009$			Population		INSOLE Sample	
				(000)	%	(000)	% Pop.
10-16	$1	under	$173,508	129,595	92.17	129.5	0.1
7-9, 17&18	-$1 $173,508	under/over	|$361,475|	8,394	5.97	26.2	0.3
5&6, 19&20	|$361,475|	under/over	|$1,445,900|	2,285	1.63	27.9	1.2
3&4, 21&22	|$1,445,900|	under/over	|$7,229,500|	263	0.19	48.9	18.6
101 (HINTS)		N/A		35	0.03	35.2	100.0
201, 1&2, 23&24	|$7,229,500|	under/over		27	0.02	27.3	100.0
Total				140,599	100.00	295.1	0.2

“Extreme” Records, Aggregation, Subsampling

In the 2008 design, roughly 100 records in the INSOLE with “extreme” values for certain variables were excluded from the PUF sample

In the proposed design, about 1,200 records in the INSOLE with the largest values for most variables are aggregated

- Returns included in the aggregation are selected by ranking all returns by each variable, and taking largest 10 to 400
- Aggregation greatly reduces disclosure risk, and preserves total values of variables

Subsampling reduces disclosure risk in both designs

- In proposed design, some additional subsampling
- But in lowest-income strata, subsampling rate would rise

PUF Sample Designs:2008 and Proposed

<table>
<tr><th rowspan="2">Strata</th><th colspan="2">PUF (2008 Design)</th><th colspan="2">PUF (Proposed)</th></tr>
<tr><th>CWHS Returns</th><th>Other Returns</th><th>CWHS Returns</th><th>Other Returns</th></tr>
<tr><td>10-16</td><td rowspan="3">Subsample to 3 of 10 endings</td><td>N/A</td><td rowspan="3">Subsample to 8 of 10 endings</td><td>N/A</td></tr>
<tr><td>7-9, 17&18</td><td rowspan="2">No Subsampling</td><td>Exclude</td></tr>
<tr><td>5&6, 19&20</td><td>No Subsampling</td></tr>
<tr><td>3&4, 21&22</td><td rowspan="3">Subsample to 3 of 10 endings, then subsample like other returns</td><td rowspan="2">Subsample to achieve a 10% rate</td><td rowspan="3">Subsample to 8 of 10 endings, then subsample like other returns</td><td>Subsample to 10% rate</td></tr>
<tr><td>101 (HINTS)</td><td>Restratify & subsample</td></tr>
<tr><td>201, 1&2, 23&24</td><td>Delete "extreme"; subsample rest at 10%</td><td>Aggregate "largest"; subsample rest at 10%</td></tr>
</table>

PUF Samples with 2008 and Proposed Designs

Strata	PUF (2008 Design)		PUF (Proposed)	
	(000)	% INSOLE	(000)	% INSOLE
10-16	38.9	30.0	103.7	80.0
7-9, 17&18	20.4	77.6	6.7	25.6
5&6, 19&20	26.3	94.3	27.5	98.4
3&4, 21&22	26.3	53.7	26.3	53.7
101 (HINTS)	3.5	10.0	2.5	7.1
201, 1&2, 23&24	2.7	10.0	2.6	9.6
Total	118.1	40.0	169.3	57.4

Deleting, Modifying and Blurring Variables

Current PUF disclosure avoidance procedures include deleting, modifying and blurring variables

Proposed design would retain all of these approaches, but with modifications

Under the 2008 design, stricter procedures apply to records with over $200,000 (in absolute value) of AGI or a selection probability over 10%

- Returns with quite high levels of positive income offset by losses may not be subject to the stricter procedures

Under the proposed design, the stricter procedures would apply to all returns selected above the CWHS rate

Deleted Variables

<table>
<tr><th rowspan="2">Strata</th><th colspan="2">PUF (2008 Design)</th><th rowspan="2">PUF (Proposed)</th></tr>
<tr><th>AGI <|$200K|</th><th>AGI >|$200K|</th></tr>
<tr><td>10-16</td><td rowspan="3">None</td><td>N/A</td><td rowspan="2">State code</td></tr>
<tr><td>7-9, 17&18</td><td rowspan="2">State code; sales tax deduction; alimony</td></tr>
<tr><td>5&6, 19&20</td><td rowspan="4">State code; sales tax deduction; alimony paid and received; marital status on aggregate record</td></tr>
<tr><td>3&4, 21&22</td><td colspan="2" rowspan="3">State code; sales tax deduction; alimony paid and received</td></tr>
<tr><td>101 (HINTS)</td></tr>
<tr><td>201, 1&2, 23&24</td></tr>
</table>

Deleted Variables - Continued

The key new deletion under the proposed design is of state code

State codes, in combination with other information on returns or available from other sources, increase disclosure risk

- The risk would increase with the proposed addition of new variables

In addition, state codes cannot be used to provide reliable state-by-state estimates from the PUF

- The sample is not designed to be representative by state

The Working Group is exploring alternatives for facilitating state-by-state analysis

Modified Variables

<table>
<tr><th rowspan="2">Strata</th><th colspan="2">PUF (2008 Design)</th><th rowspan="2">PUF (Proposed)</th></tr>
<tr><th>AGI <|$200K|</th><th>AGI >|$200K|</th></tr>
<tr><td>10-16</td><td rowspan="3">None</td><td>N/A</td><td rowspan="2">Cap total number of depends and separate caps on providing age of depends (in ranges)</td></tr>
<tr><td>7-9, 17&18</td><td rowspan="2">Marital status; # depends; exemptions</td></tr>
<tr><td>5&6, 19&20</td><td rowspan="4">Marital status; cap number of dependents by type; cap personal exemption amounts; aggregate record contains uncapped means</td></tr>
<tr><td>3&4, 21&22</td><td colspan="2" rowspan="3">Marital status; cap number of dependents by type; cap personal exemption amounts</td></tr>
<tr><td>101 (HINTS)</td></tr>
<tr><td>201, 1&2, 23&24</td></tr>
</table>

Blurring Variables

<table>
<tr><th rowspan="2">Strata</th><th colspan="2">PUF (2008 Design)</th><th rowspan="2">PUF (Proposed)</th></tr>
<tr><th>AGI
<|$200K|</th><th>AGI
>|$200K|</th></tr>
<tr><td>10-16</td><td rowspan="3">Univariate</td><td>N/A</td><td rowspan="2">Univariate</td></tr>
<tr><td>7-9, 17&18</td><td rowspan="2">Multivariate; see box below</td></tr>
<tr><td>5&6, 19&20</td><td rowspan="4">Multivariate; 10 categories of filing status and number of children at home; grouped by presence of variables; distance metric on normalized variables within categories</td></tr>
<tr><td>3&4, 21&22</td><td colspan="2" rowspan="3">Multivariate; 13 categories of filing status and number of children at home; grouped by presence of variables; distance metric on normalized variables</td></tr>
<tr><td>101 (HINTS)</td></tr>
<tr><td>201, 1&2, 23&24</td></tr>
</table>

Blurring Variables - Continued

Variables that are univariate blurred under both designs:

- Alimony paid and received
- Salaries and wages
- Medical and dental expenses
- Real estate taxes
- State and local income taxes (Wisconsin only)

Variables that are multivariate blurred under both designs:

- Salaries and wages
- Real estate taxes
- State and local income taxes

Rounding

Under the current design, amount fields are rounded to the four most significant digits

- For example, $228,867 would be rounded to $228,900

Amounts under $10,000 are not rounded using this procedure

Under the proposed design, amounts of $10,000 or more are rounded to the four most significant digits, and amounts under $10,000 are rounded to the nearest $10

Rebalancing Returns

Under the current design, the effects of deleting, modifying, blurring and rounding variables are included in (implied) residual variables

- The two key implied residuals contain certain items of income plus certain above-the-line deductions, and personal exemptions plus total deductions (standard or itemized)

Under the proposal, the effects of procedures would be removed by recomputing AGI, personal exemptions, itemized deductions, taxable income, regular tax, AMT, credits, and tax after credits

- The value of deleted variables would continue to be included as part of the implied residual variables

New Variables

Demographic information is critical to a wide range of tax research and analysis

Age and gender of taxpayers, and age of dependents (from Social Security) are on the INSOLE, but not the PUF

The proposed design includes the addition of age (in ranges) and gender of taxpayers, and ages (in ranges) of dependents on the PUF

These new variables would only be added to returns in strata 7-18, which are sampled only at the CWHS rate (1 in 1,250)

In addition, as noted above, the number of dependents for which age (in ranges) would be added would be capped

New Variables - Continued

Caps will vary with filing status, and also with other characteristics of the return, to insure nondisclosure

- For some returns (in CWHS-only strata) the cap will be zero

A new variable will also be added to show the split (in ranges) of wage and self-employment earnings on joint returns

Reweighting

The PUF is currently reweighted for deleted “extreme” records and for subsampling

It is not reweighted, however, to take account of the omission of returns filed for years more than three years prior to the current year

Under the proposed design, the omission of these prior year returns would be reflected in the population counts used for reweighting

Tabulations to Accompany the PUF

To help tax analysts use the aggregate record, a table will be included in the PUF documentation with counts for each variable of the number of returns with nonzero entries

SOI also plans to release separate tabulations with information on age, gender, and earnings splits cross tabulated by such variables as AGI and filing status

- These tabulations will be quite useful to all tax analysts
- They will also help PUF users understand and work with the caps on the number and age of dependents, and other missing demographic information

Moving Forward

SOI has solicited comments on the proposed design changes to the PUF from JCT, OTA, and PUF users

After taking into account comments and suggestions from these groups (and you here today!), a provisional 2009 PUF will be produced

Mathematica will analyze this provisional 2009 PUF for disclosure risk

Depending on the results of Mathematica's analysis, refinements may be made to the design before the 2009 PUF is released to the public