**Design Changes to the SOI Public Use File (PUF)**

The Statistics of Income (SOI) Division draws a stratified random sample from the individual income tax returns filed with the IRS. SOI uses all of the microdata records from this sample to produce the INSOLE (INdividual and SOLE proprietor) file, which is used by SOI to produce the tables included in its annual *Individual Complete Report* publication and for other statistical purposes, and by the staffs of the Congressional Joint Committee on Taxation (JCT) and Treasury's Office of Tax Analysis (OTA) to construct microsimulation models and for other tax analyses.

SOI produces a second annual microdata file, the Public Use File (PUF), from the INSOLE file. The high quality of tax return information makes the PUF a critical source of information for researchers and analysts to examine a wide range of tax issues. To insure that the PUF meets the strict protections for taxpayer confidentiality in the Internal Revenue Code, much of the information on the INSOLE file is excluded or altered. Many INSOLE records are excluded from the PUF, primarily through subsampling (particularly for high-income records for which only 1 in 10 INSOLE records are included in the PUF) but also through removal of records with "extreme" amounts prior to subsampling. The PUF also excludes many variables from the INSOLE that would (or might) disclose the identity of a taxpayer directly (e.g., names and addresses), or indirectly using information on the PUF and/or other information. Further, some of the variables on the PUF are modified or statistically "blurred" to protect confidentiality. SOI and its statistical consultant (which is currently Mathematica) perform rigorous checks on the PUF for each year to insure it meets nondisclosure requirements, and as information on individuals has become more accessible to the public have periodically undertaken more in-depth reviews and strengthening of procedures. In consultation with PUF users and the broader research community, SOI also reviews annually, and in more depth periodically, how well the information on the PUF meets the research and analytical needs of users.

Susan Boehmer, Director of the SOI Division, formed a PUF Working Group in the Fall of 2012 to perform an in-depth reassessment of both the disclosure avoidance procedures applied to the PUF and the quality and utility of PUF data to PUF users. Susan asked David Paris, head of the Individual Statistics branch of SOI (which produces the PUF), to chair the Working Group. Members of the group include SOI staff, John Czajka of Mathematica, and Dan Feenberg of NBER and Jim Nunns of the Urban-Brookings Tax Policy Center (who are both members of SOI's PUF Users Group and the SOI Advisory Panel). The Working Group considered a range of possible design changes and refined them to a set of recommended changes that are intended to meet the objectives of reducing disclosure risk and improving PUF data quality and utility, and which can be implemented with available SOI resources. SOI plans to implement the recommendations of the Working Group for the 2009 PUF. The attached table provides a summary of the revised nondisclosure procedures and the new variables for the 2009 PUF.

<u>Disclosure Avoidance Procedures</u>

To strengthen disclosure avoidance procedures, the Working Group recommended the following:

- Aggregation of returns with "large" amount variables. Currently, about 100 returns with amounts that are considered "extreme" are identified by SOI staff and removed from each PUF. In addition, returns in the full INSOLE sample that are sampled at rates above 10% are subsampled to a 10% rate (e.g., only 1 in 10 returns sampled for the INSOLE at 100% is included in the PUF). Several nondisclosure procedures are also applied to these returns.

  To restore the information previously excluded from returns with "extreme" amounts while continuing to protect their confidentiality, the Working Group recommended that all returns with "large" amounts be aggregated into a single record. "Large" amounts would generally be defined as the largest 30 reported for any income item and largest 10 for other items, but other cutoffs would be used for selected variables (e.g., for AGI, the largest 400). For the 2009 PUF, about 1,160 returns would be aggregated, making it impossible to identify any specific return with a "large" amount.[1] The Working Group is exploring the possibility of splitting these returns between those with positive and those with negative AGIs, so there would be two rather than one aggregate record. These two records would provide more useful information for PUF users. However, this splitting will only be done if the reduction in the residual disclosure risk due to aggregation can be maintained, which might require including some additional returns in the aggregated records.

  Subsampling to 10% and current nondisclosure procedures will be retained for returns sampled at a rate above 10% that are not included in the aggregated record(s).

  Base nondisclosure procedures on sample strata. Currently, more extensive nondisclosure procedures are applied to "high-income" returns (generally, returns with AGI of $200,000 or more in absolute value) than to "low-income" returns (returns with AGI under $200,000 in absolute value). The less extensive "low-income" procedures apply to some returns that have AGI below $200,000 but that are sampled at relatively high rates and therefore likely carry a greater disclosure risk.[2]

  To reduce this risk, the Working Group recommended that the more extensive "high-income" nondisclosure procedures (with modifications described below) apply to all returns included in the PUF at a sample rate greater than (approximately) 1 in 1,250 with the "low-income" procedures applied to returns sampled at 1 in 1,250. (The sample rate of 1 in 1,250 is based on 8 SSN 4-digit endings of the Continuous Work History Sample (CWHS) designated by SSA that is part of the INSOLE sample which the Working Group recommended be included in the PUF).

- High-Income Nontaxable Returns (HINTS). HINTS are selected at a 100% rate for the INSOLE sample, and are currently subsampled at a 10% rate for the PUF. HINTS are

[1] The number of returns that would be aggregated would vary somewhat in future PUFs, depending on the correlations among "large" amounts.

[2] Under current procedures, returns sampled at a rate of 10% or higher are considered "high-income" for nondisclosure procedures, regardless of AGI.

inherently different from other high-income returns, so may carry a higher disclosure risk. HINTS are also of little analytical value to PUF users.

To reduce the disclosure risk from HINTS, the Working Group recommended that they be reclassified into the regular sampling strata and then subsampled at the corresponding strata rates for the PUF.

- Further subsampling for certain strata. Currently, the lowest sample rates for the INSOLE are 1 in 1,000 in strata 10 through 16, which covered positive incomes up to about $175,000 in 2009. Strata 7 through 9 cover negative incomes as low as about -$360,000 in 2009 and are sampled at rates up to about 5 in 1,000. Strata 17 and 18 cover positive incomes up to about $360,000 in 2009 and are sampled at about 3 in 1,000.

  The Working Group recommended treating all returns in strata 7 through 18 as "low-income" for purposes of nondisclosure procedures, and also for purposes of adding new variables (see below). For this reason, the Working Group also recommended that the returns in strata 7 through 9 and in strata 17 and 18 be subsampled to the same rate recommended for strata 10 through 16, which is 1 in 1,250, so only CWHS returns in these strata would be included in the PUF.

- Remove state codes. Currently, "low-income" returns on the PUF include a state code. State codes, in combination with other information, may increase disclosure risks, and these risks could increase with the addition of new variables (see below). In addition, the SOI sample is not designed to be representative of each state. State-level estimates produced from the PUF using state codes are therefore subject to high sampling variability, severely limiting the analytical usefulness of state codes.

  To address potential disclosure risks and in recognition of their limited analytical value, the Working Group recommended that state codes be removed from the PUF.

- Changes in groupings for blurring "high-income" returns. Currently, certain variables that are considered to be of highest risk for disclosure, such as wages and deductions for state income taxes, are jointly blurred on "high-income" returns. This multivariate blurring takes returns that have similar values for the high-risk variables and in groups of three returns replaces actual values of each variable with the average over the three returns. Only returns in the same "Category," defined by filing status and number of dependents, are included in the same pool for possible joint blurring.

  The 13 Categories defined for the current multivariate blurring procedure mix filing statuses and in some cases contain so few returns that blurring has to be done on only one or two variables at a time, so is not full multivariate blurring. To improve the disclosure avoidance achieved by the procedure, the Working Group recommended that the Categories be redefined so no filing statuses are mixed and the number of Categories is reduced to 10.

- Cap total number of dependents on "low-income" returns. Currently, the total number of dependents is capped on "high-income" returns, but not on "low-income" returns.[3]

  There may be some disclosure risk associated with providing an uncapped number of dependents on "low-income" returns in the PUF. To address this potential risk, the Working Group recommended that the total number of dependents be capped, with caps varying by filing status.

- Rebalancing returns. Currently, the effects of deleting, modifying and blurring variables appear in the PUF as part of (implied) residual variables.

  The Working Group is evaluating whether gross income, AGI, taxable income, regular tax, AMT and tax after credits should be recomputed after modifying and blurring variables. The value of deleted variables would continue to be included as part of implied residual variables.

Data Quality and Utility

To improve the quality and utility of PUF data for users, the Working Group recommended the following:

- Include age, gender and earnings split variables on "low-income" returns. Demographic information is critical to a wide range of tax research and analysis, but is largely unavailable from tax returns. However, date of birth (age) and gender are supplied to IRS by SSA. Age and gender are included on the INSOLE file, and SOI publishes some tables classified by age and gender, but these variables have not previously been included on the PUF.

  The Working Group recommended that variables for age (in ranges) and gender for taxpayers, age (in ranges) for dependents, and earnings splits (in ranges) on joint returns be included on PUF returns in strata 7 through 18, which are all recommended to be subsampled to a rate of 1 in 1,250. Separate sets of age ranges would be used for taxpayers and dependents, and for joint returns only the age of the primary would be included. The number of dependents for which age is provided would be capped to remove any disclosure risk from the addition of these variables. Caps would be determined by several return variables, and could be zero (i.e., no information on dependent ages would be included). Counts of dependents by age group would be provided in order up to the cap, starting with the youngest dependent.

- Increase subsampling rate for CWHS returns. Currently, the CWHS returns in the INSOLE file, representing a 1 in 1,000 sample from the population, are subsampled to 3 in 10,000 for the PUF (that is, 3 of 10 CWHS endings are retained). This subsampling rate was set when CWHS returns were sampled for the INSOLE at half the current rate, or 5 in 10,000. The Working Group recommended that all returns in strata 7 through 18

[3] The cap is applied first to the number of children at home, and then carries through to other types of dependents.

be subsampled for the PUF to include 8 of 10 CWHS endings, corresponding to a sample rate of approximately 1 in 1,250. This higher sample rate would increase the PUF sample by about 51,000 returns.[4] These returns represent 98 percent of the filing population, so the higher subsampling rate would improve the quality of PUF data for nearly all of the filing population.

- Changes in groupings for blurring "high-income" returns. In addition to reducing disclosure risk, discussed above, these changes would improve data quality by reducing the variance of blurred variables.

- Reweighting. Certain returns (e.g., returns filed for years more than three years prior to the current year) are excluded from the universe of returns included in the PUF, but the PUF weights currently are based on population return counts that include such returns. The Working Group recommended adjusting the population return counts for returns not in the PUF universe.

- Tabulations. To help users use the aggregated record(s) for research and analysis, a simple tabulation of returns included in the aggregation(s) that reported nonzero amounts for each variable, and counts of returns by filing status and of dependents by type would be supplied with the PUF. To help users understand and work with the various caps on dependents and the omission from "high-income" returns of age, gender and earnings split variables on joint returns, cross tabulations of these variables and return characteristics such as AGI and filing status would also be supplied with the PUF.

Prior to release of the 2009 PUF with these changes, a version will be prepared for analysis by Mathematica to determine whether the file, with these changes, raises any new disclosure risks. Depending on the results of that analysis, refinements might be made to the recommendations to insure nondisclosure requirements are met before the 2009 PUF is released to the public.

Schedule for Completing Future PUFs

The Working Group recommends that the lag time between completion of the INSOLE file and completion of the PUF be reduced over the next several years to the time necessary to produce a PUF (about six months). On this schedule, the PUF for a tax year would be released by the end of the second following year.

Attachment

[4] As noted above, non-CWHS returns in strata 7 through 9 and 17 and 18 would be excluded from the PUF under the Working Group's recommendations, whereas all of these returns have been included in previous PUFs.

## Current and Revised Nondisclosure Procedures and New Variables for the 2009 PUF

| | | | Strata[1] | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | 201, 1 & 2, 23 & 24 | | |
| | | | 10 to 16 | 7 to 9, 17 & 18 | | 5 & 6, 19 & 20 | | 3 & 4, 21 & 22 | 101 (HINTS) | No "Large" or "Extreme" Values | "Large" but not "Extreme" Values | "Extreme" Values |
| | | | | AGI<\|$200K\| | AGI >= \|$200K\| | AGI<\|$200K\| | AGI >= \|$200K\| | | | | | |
| Subsampling and Aggregation | Current | CWHS | Subsample to 3 of 10 endings | | | | | | | | | Excluded from PUF |
| | | Other | N/A | No subsampling | | | | Subsampled to achieve a 10% sample rate | | | | |
| | Revised | CWHS | Subsample to 8 of 10 endings | | | | | | | | Aggregate (see 3-15-13 memo on "Large" values for each amount variable); separate tab of return counts for all current and new (see below) variables | |
| | | Other | N/A | Excluded | | No subsampling | | Subsampled to achieve a 10% sample rate | Place in applicable ordinary strata and subsample at strata rates[2] | Subsampled to achieve a 10% sample rate | | |
| 2009 Population and Samples | Population | | 129,594,866 | 6,188,154 | 2,205,762 | 696,759 | 1,588,378 | 262,862 | 35,150 | 26,171 | 1,064 | 100 |
| | INSOLE Sample | | 129,531 | 19,001 | 7,246 | 8,331 | 19,597 | 48,942 | 35,150 | 26,171 | 1,064 | 100 |
| | Current PUF | | 38,882 | 14,669 | 5,702 | 7,843 | 18,485 | 26,286 | 3,515 | 2,617 | 106 | 0 |
| | Revised PUF[3] | | 103,686 | 4,951 | 1,765 | 8,192 | 19,279 | 26,286 | 2,500 | 2,617 | 1 aggregated record | |
| Deleted Variables | Current | | | | State code; state sales tax deduction; alimony paid and received | | State code; state sales tax deduction; alimony paid and received | | | | | |
| | Revised | | State code | | | State code; state sales tax deduction; alimony paid and received (see footnote 2 for HINTS); marital status not provided on aggregate record but counts included in separate tab | | | | | | |
| Modified Variables | Current | | | | Marital status; number of dependents by type; personal exemption amounts | | Marital status; number of dependents by type; personal exemption amts | | | | | |
| | Revised | | Cap total number of dependents by filing status and separately cap number of dependents for which age (in ranges) is provided (see "New Variables" below) with caps based on various return characteristics | | | Marital status; number dependents by type (but vary cap by marital status in same manner as for Categories -- see below); personal exemption amounts (see footnote 2 for HINTS) | | | | Included variables are means for all aggregated returns but not otherwise modified (aggregate record has a weight equal to the number of returns aggregated) | |

| | | Strata[1] | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 201, 1 & 2, 23 & 24 | | |
| | | 10 to 16 | 7 to 9, 17 & 18 | | 5 & 6, 19 & 20 | | 3 & 4, 21 & 22 | 101 (HINTS) | No "Large" or "Extreme" Values | "Large" but not "Extreme" Values | "Extreme" Values |
| | | | AGI<\|$200K\| | AGI >= \|$200K\| | AGI<\|$200K\| | AGI >= \|$200K\| | | | | | |
| New Variables | Age, Gender, Earnings Splits | Age (in ranges) for primary taxpayers and dependents; gender for primary taxpayers, earnings splits (in ranges) for wages and Schedule SE earnings on MFJ | | | | | | | | | |
| | | Tabulations of new variables classified by characteristics of taxpayers such as filing status, number of dependents, and AGI | | | | | | | | | |
| | Other | | | | | | | | | Add separate fields for + and - for variables that can be + or - | |
| Blurring -- Type | Current | Univariate | | Multivariate | Univariate | Multivariate | | | | | |
| | Revised | Univariate | | | Multivariate (see footnote 2 for HINTS) | | | | N/A | | |
| Blurring -- "Categories" (for multivariate) | Current | | | 13 Categories (see box at right) | | 13 Categories based on filing status (MARS) and number of children at home (XOCAH) | | | | | |
| | Revised | | | | 10 Categories based on filing status (MARS) and number of children at home (XOCAH) | | | | | | |
| Blurring -- Distance | Current | | | See box at right | | Distance metric; applied within Categories to normalized variables in subgroup | | | | | |
| | Revised | | | | Distance metric; applied within Categories to normalized variables in subgroup | | | | | | |
| Rebalancing Returns | Current | Returns are rebalanced for deleted, modified and blurred variables by making the change in AGI due to blurring and deletion of alimony paid and received part of an implied residual that includes "other income" and some ATL deductions; total deductions (standard or itemized) and personal exemption amounts are always a combined implied residual so reflect effects of deleting, modifying and blurring component variables | | | | | | | | | |
| | Revised | Rebalance returns and the aggregate record(s) for effects of modifying and blurring variables by recomputing gross income, AGI, taxable income, regular tax, AMT, and tax after credits; keep current procedures for deleted variables | | | | | | | | | |
| Reweighting | Current | Returns are reweighted for subsampling (and removal of returns with "extreme" values), but not for returns excluded from the PUF universe | | | | | | | | | |
| | Revised | Population adjusted for excluded returns prior to reweighting | | | | | | | | Aggregate record will carry weight of returns that are aggregated, so no reweighting | |

[1] Returns that are filed for taxable years more than three years prior to the current year, returns with reject codes greater than zero, and returns filed only for a stimulus payment would continue to be excluded from the universe of returns represented by the PUF. In addition, returns with Forms 2555 that are oversampled in "foreign study" years would be excluded from the PUF.

[2] Subsampled HINT returns would be subject to aggregation, the deletion, modification or addition of new variables, and blurring according to the revised rules for the strata they are reassigned to.

[3] Returns counts for the Revised PUF sample are estimates. Totals for each category:

| | |
|---|---|
| Population | 140,599,266 |
| INSOLE Sample | 295,133 |
| Current PUF | 118,106 |
| Revised PUF | 169,276 |