Design Changes to the SOI Public Use File (PUF)

Victoria L. Bryant, John L. Czajka, Georgia Ivsin, and Jim Nunns

Prepared for the
“New Resources for Microdata-Based Tax Analysis” Session
107th Annual Conference on Taxation
National Tax Association
Santa Fe, New Mexico
November 15, 2014

ABSTRACT

The Statistics of Income (SOI) Division of IRS prepares a publicly available file, the Public Use File, from its annual sample of income tax returns. The PUF is a critical data source for tax policy analysis. To insure taxpayer confidentiality, SOI applies disclosure avoidance procedures to the PUF. In 2012, SOI established a Working Group to perform an in-depth review of these procedures and of the analytical usefulness of the PUF. This paper describes the revised PUF design recommended by the Working Group, and how the design changes improve both disclosure avoidance and the PUF’s analytical usefulness.

Victoria L. Bryant: Statistics of Income Division, Internal Revenue Service, Washington, DC (victoria.l.bryant@irs.gov)
John L. Czajka: Mathematica Policy Research, Washington, DC (jczajka@mathematica-mpr.com)
Georgia Ivsin: Navigant Consulting, Washington, DC (georgia.ivsin@navigant.com)
Jim Nunns: Urban-Brookings Tax Policy Center, Washington, DC (jnunns@urban.org)
Ivsin’s work on this project was performed while she was at the Urban-Brookings Tax Policy Center.
Work by Ivsin and Nunns was supported by a grant from the Alfred P. Sloan Foundation.

Design Changes to the SOI Public Use File (PUF)

I. Introduction

The Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) takes a sample of all federal individual income tax returns filed each year. From this sample, SOI prepares a publicly available file known as the Public Use File, or PUF.[1] The PUF contains information on the income, deductions, exemptions, taxes, credits and characteristics of the taxpayers and dependents reported on the tax returns included in the sample. 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 issues in economics and public policy, such as the effect of taxes on economic behavior and the impact of proposed tax reforms on the distribution of tax burdens.

The Internal Revenue Code provides strict protections for the confidentiality of tax return information.[2] In conformance with these confidentiality provisions, the PUF does not contain any information in a form that would disclose the identity of a specific taxpayer: the PUF does not contain direct identifiers, such as names, addresses, or Social Security numbers, and the information on the PUF is limited or altered in various ways to insure that no one can indirectly identify a taxpayer using that information alone or in combination with information from other sources. SOI and its statistical consultant perform rigorous checks on the PUF for each year to help insure it meets nondisclosure requirements.

As the scope of information on individuals that is publicly accessible has grown over time, especially through the internet, and the power of computers and software to link information has grown, SOI and its statistical consultant have periodically undertaken in-depth reviews of its procedures for avoiding disclosure of taxpayer identities on the PUF.[3] These reviews focus on disclosure avoidance, but also carefully review the impact of existing and proposed disclosure-avoidance procedures on the quality of PUF data for users. The disclosure avoidance procedures for the PUF have been strengthened over time as a result of these reviews, as well as less detailed reviews performed on an annual basis.

In consultation with PUF users and the broader research community, SOI also reviews annually, and in more depth periodically, how well the information included on the PUF meets the research and analytical needs of PUF users. Changes in tax law, for example, often result in additional information reported on income tax returns, and PUF users need at least some of this information to help inform the public about the effects of the new or amended provisions of the law. In addition, as research and analytical methods and models have become more sophisticated and the quality of other publicly available files has deteriorated, PUF users have increasingly relied on the PUF as their primary source of data on U.S. households. However, the PUF has not included basic demographic information, such as on the age and gender of taxpayers

[1] The first PUF, then known as the Tax Model File, covered returns filed for 1960.
[2] See in particular IRC Section 6103 and the associated provisions imposing penalties for breach of confidentiality. See Wilson and Smith (1983) for a history of tax confidentiality provisions and procedures.
[3] See Strudler, Oh and Scheuren (1986), Sailer, Weber and Wong (2001), Winglee, et al. (2002), and Vartivarian, Czajka and Weber (2007).

and dependents, that would make the PUF much more useful for many research and analytical purposes.[4]

Susan Boehmer, then 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 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 Victoria Bryant and other SOI staff, John Czajka of Mathematica Policy Research, 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 intended to meet the objectives of reducing disclosure risk and improving PUF data quality and utility, and that could be implemented with available SOI resources. These recommendations were presented to the SOI Advisory Panel, and circulated for comment to the Treasury's Office of Tax Analysis (OTA), the Congressional Joint Committee on Taxation (JCT), and PUF users. Based on comments from these groups, the Working Group developed a revised design for the PUF.

The remainder of this paper describes the revised PUF design developed by the Working Group. Section II provides background information on the structure of the SOI sample. Subsequent sections cover the current PUF design (i.e., the design used for the 2008 PUF) and the revised design developed by the Working Group for the 2009 and subsequent PUFs. Section III covers returns excluded from the PUF and the subsampling and aggregation of remaining returns; Section IV covers deleted variables, modified variables, blurring of certain variables, and rounding rules; Section V describes the new variables added to the PUF and the procedures used to address potential disclosure risks related to the addition of these variables; and Section VI rebalancing and reweighting returns. The final section describes the planned schedule for completing the 2009 and future PUFs.

II. The SOI Sample

The sample of individual income tax returns used to create the PUF is designed for the larger "INSOLE" file used by SOI in their publications other products, and by OTA and JCT in their microsimulation models and for other analyses.[5] The sample is selected from all individual income tax returns processed during a year by IRS and posted to the Master File except tentative returns, amended returns, and returns that report no income. Selection for the sample is based on the size of "total positive income" or, if larger in absolute value, "total negative income". These two amounts are the sum of nearly all positive and all negative items of income reported on a return. Based on the larger of these two amounts, all returns on the Master File that are eligible for selection are placed in one of 9 negative income strata (strata 1 through 9) or one of 15 positive income strata (strata 10 through 24).[6] Strata boundaries are dollar amounts that are

[4] Age and gender are not reported on tax returns, but are provided to the IRS by the Social Security Administration.
[5] INSOLE is a combination of INdividual and SOLE proprietor, a name that originated when sole proprietorships were the primary form of non-corporate business.
[6] Strata (except two special 100 percent sample rate strata) are identified in this paper by the last two digits of their full three-digit designation by SOI. The INSOLE sample design contains a secondary stratification based on certain

expressed at their original values, established in 1991.[7] However, in selecting the sample the dollar amounts of "total positive income" and "total negative income" on each return are deflated by the change between 1991 and the tax year of the sample in the chained GDP implicit price deflator (which generally grows much more slowly than consumer prices indexes, such as the CPI). Sample rates vary by strata, from a rate of about 0.1 percent (1 in 1,000) in strata 10 through 16, to 100 percent in strata 1, 2, 23 and 24. There are also two special strata for returns sampled at 100 percent, one for returns with gross receipts from one or more nonfarm or farm sole proprietorships reported on Schedules C and F of $50 million or more (stratum 201), and another for "high-income nontaxable returns" or HINTS, which are returns with income of $200,000 or more that report no income tax liability (stratum 101). In addition, periodically sample rates in certain strata are increased to insure an adequate sample of returns that claim an exclusion for foreign earned income on Form 2555.[8]

Returns in the two special 100 percent strata are selected for the INSOLE sample first. Sampling of remaining returns within each stratum is based on SSNs.[9] For returns in all strata, the last four digits of SSNs are examined (there are 9,999 such endings, since SSNs do not end in 0000). Any return with one of 10 specified endings is sampled (making the sample rate 10 in 9,999 or slightly more than 1 in 1,000, or 0.1 percent). The 10 endings used for sampling are part of the Continuous Work History Sample designated by the Social Security Administration (SSA) for research purposes, and are sometimes referred to below as simply CWHS. For returns in strata 10 through 16, only CWHS returns are selected for the sample. For returns in the other strata (1 through 9 and 17 through 24), which have sample rates above the CWHS rate, non-CWHS SSNs are transformed (to correct for slight non-randomness in SSNs), and enough endings of the transform are selected to achieve the sample rate for the stratum (taking into account the CWHS portion of the sample).

Income tax returns contain very limited, and incomplete, demographic information about taxpayers and dependents. However, IRS does have access to the DM-1 file produced by SSA, which contains date of birth (which can be converted directly to age), gender, and (if relevant) date of death information for virtually all taxpayers and dependents.[10] This information is added to returns during IRS processing, and SOI includes it on the INSOLE file.

tax forms and schedules (Form 1116, Form 2555, Schedule C, and Schedule F) filed with a return, which determines the first digit of the three-digit designation. Sample rates are the same across forms within each income strata (i.e., for the same two-digit endings), except in "foreign study years," when a higher sample rate is applied in some income strata to returns that report foreign earned income on Form 2555.

[7] Some of the 1991 strata also separated returns with the same income but different characteristics of more or less interest for tax analysis. Due to subsequent changes in the INSOLE sample, none of these differential sample rates within income strata remain in place.

[8] See footnote 6 above.

[9] SSNs (Social Security Numbers) are issued by the Social Security Administration to every U.S. citizen, every noncitizen who has permission to work in the United States, and certain other noncitizens who require an SSN to receive benefits from the federal government or a state or local government. Noncitizens who do not qualify for an SSN but who are required to file a tax or information return with the IRS are required to have an ITIN (Individual Taxpayer Identification Number), which is a 9-digit number with a first digit of 9 that is otherwise similar to an SSN. IRS also issues an ATIN (Adoption Taxpayer Identification Number) for a child legally placed for adoption with a taxpayer who does not know the child's SSN. Throughout this paper, "SSNs" includes ITINs and ATINs.

[10] SOI also has access to the comparable files produced by IRS that covers individuals who have been issued ITINs and ATINs (see preceding footnote).

As explained more fully in the next section, the INSOLE sample is subsampled for the PUF because subsampling of high-income returns is one of the most effective procedures for reducing disclosure risk on the PUF. The weights on PUF returns are adjusted for this subsampling so that the PUF represents the total filing population (with minor exceptions under the current design described below).

III. Excluded Returns, Subsampling and Aggregation

The revised design modifies which returns in the INSOLE sample are excluded from the PUF, changes the way the INSOLE sample is subsampled for the PUF, and aggregates all returns with a "large" value for any specified amount variable into a single record. The subsampling changes are designed in part to align certain disclosure avoidance procedures with strata boundaries. This alignment improves disclosure avoidance and also data quality. Table 1 provides a summary by strata of subsampling and aggregation, as well as other procedures, under the current and revised PUF designs.[11]

Excluded Returns

Returns Excluded from the PUF Universe. The PUF currently excludes all returns included in the INSOLE sample that are filed for taxable years more than three years prior to the current year and returns filed only for a stimulus payment.[12] These returns are still excluded from the PUF under the revised design, but population counts are now adjusted for the exclusion of these returns from the universe represented by the PUF when the PUF is reweighted.

Returns with "Extreme" Values. Currently, a small number of returns with "extreme" values for certain variables are excluded from the PUF. Under the revised design, these returns are included in the PUF as part of the aggregation into a single record of all returns with a "large" value for any specific amount variable.

Oversampled Foreign Earned Income Returns. Returns with Forms 2555 that are oversampled for the INSOLE in "foreign study" years are currently subsampled for the PUF in the same manner as other returns in the same strata. Under the revised design, oversampled Form 2555 returns will be excluded from the PUF sample. (Note that 2009 was not a "foreign study" year.)

Subsampling

HINTS. The Tax Reform Act of 1976 included a requirement for an annual study of HINTS, and set the limit for "high-income" at $200,000. HINTS are selected for the INSOLE at a 100 percent rate, and in the current PUF design are subsampled at a 10 percent rate (to achieve a 1 in 10 sample rate). HINTS are inherently different from other high-income returns, so may carry a higher risk of disclosure than other returns sampled at high rates. In addition, the $200,000 income threshold for HINTS is not indexed for inflation, so the large increase in the number of HINTS sampled, particularly after 2007, includes many returns of relatively little interest for

[11] For a detailed description of the 2008 PUF and its design, see Bryant (2012).

[12] Stimulus payments were made for tax year 2007, so the related exclusion will become irrelevant starting with the tax year 2011 PUF.

Table 1. Current and Revised Subsampling and Aggregation, Nondisclosure Procedures, New Variables, and Rebalancing and Reweighting for the 2009 PUF

			Strata[1]									
			10 to 16	7 to 9, 17 & 18		5 & 6, 19 & 20		3 & 4, 21 & 22	101 (HINTS)	201, 1 & 2, 23 & 24		
				AGI<$200K	AGI >= $200K	AGI<$200K	AGI >= $200K			No "Large" or "Extreme" Values	"Large" but not "Extreme" Values[2]	"Extreme" Values
Subsampling and Aggregation	Current	CWHS	Subsample to 3 of 10 endings					Subsample to 3 of 10 endings, then subsample like other returns				Excluded from PUF
		Other	N/A	No subsampling				Subsampled to achieve a 10% sample rate				
	Revised	CWHS	Subsample to 7 of 10 endings; 7 endings chosen randomly each year					Subsample to 7 of 10 endings, then like other returns			Aggregated into a single record; loss amounts shown separately; accompanied by tabulation of return counts for all current and new variables	
		Other	N/A	Excluded		No subsampling		Subsampled to achieve a 10% sample rate	Place in applicable strata 1 to 24 and subsample at strata rate[3]	Subsampled to achieve a 10% sample rate		
2009 Population and Samples[4]	Population		129,594,866	6,188,154	2,205,762	696,759	1,588,378	262,862	35,150	26,171	1,065	100
	INSOLE Sample		129,531	19,001	7,246	8,331	19,597	48,942	35,150	26,171	1,065	100
	Current PUF		38,882	14,669	5,702	7,843	18,485	26,286	3,515	2,617	107	0
	Revised PUF		90,725	4,332	1,544	8,122	19,120	26,286	632	2,617	1 aggregate record	
2009 Average Sample Rates	INSOLE Sample		0.10%	0.31%	0.33%	1.20%	1.23%	18.62%	100.00%	100.00%	100.00%	100.00%
	Current PUF		0.03%	0.24%	0.26%	1.13%	1.16%	10.00%	10.00%	10.00%	10.00%	0.00%
	Revised PUF		0.07%	0.07%	0.07%	1.17%	1.20%	10.00%	1.80%	10.00%	100% of returns are aggregated	
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 3 for HINTS); marital (filing) status not provided on aggregate record; some variables with fewer than 10 nonzero entries deleted from the aggregate record						
Modified Variables	Current		Marital (filing) status		Marital status; number of dependents by type; personal exemption amounts; child tax credit	Marital (filing) status	Marital (filing) status; number of dependents by type; personal exemption amounts; child tax credit					
	Revised		See box at right; in addition, separate caps on number of dependents for which age (in ranges) is provided with caps determined by various return characteristics			Marital (filing) status; caps on total number of dependents by marital status, applied sequentially by exemption type and carried through to personal exemption amounts, child tax credit, and education deductions and credits (see footnote 3 for HINTS)					Marital (filing) status modified but not shown on aggregate record; other included variables are unmodified means	

		Strata[1]									
			7 to 9, 17 & 18		5 & 6, 19 & 20		3 & 4, 21 & 22	101 (HINTS)	201, 1 & 2, 23 & 24		
		10 to 16	AGI<|$200K|	AGI >= |$200K|	AGI<|$200K|	AGI >= |$200K|			No "Large" or "Extreme" Values	"Large" but not "Extreme" Values[2]	"Extreme" Values
Blurring -- Type	Current	Univariate		Multivariate	Univariate	Multivariate					
	Revised	Univariate			Multivariate (see footnote 2 for HINTS)					N/A	
Blurring -- Groupings or "Categories"	Current	Marital status (joint/nonjoint) and state (except for alimony)		See box at right	Marital status and state	13 categories based on filing status and number of children at home, then grouped by presence (non zero value) of blurred variables and of Schedule C net receipts					
	Revised	Marital status and prsence of dependents (except for alimony)			10 categories based on filing status and total number of dependents, then grouped by presence (non zero value) of blurred variables and presence of Schedule C					N/A	
Blurring -- Application	Current	Three returns closest returns (based on variable being blurred)		See box at right	Three closest returns	Three most distant returns (measured by a distance metric) within a category; variables normalized in each subgroup					
	Revised	Three to ten closest returns (depending on strata)			Three most distant returns within a category; variables normalized in each subgroup					N/A	
Rounding	Current	Amount fields rounded to four most significant digits; amounts under |$10,000| not rounded									
	Revised	Amounts of |$100,000| or more rounded to four most significant digits; amounts between |$10,000| and |$99,999| rounded to nearest $100; amounts between |$5| and |$9,999| rounded to nearest $10; nonzero amounts under |$5| rounded to $2 (with the sign retained).								Rounding N/A	
New Variables	Age, Gender, Earnings Splits	Age (in ranges) for primary taxpayers and dependents; gender for primary taxpayers; percentage splits (in ranges) for wages plus Schedule SE earnings on joint returns			N/A						
		Tabulations of new variables classified by characteristics of taxpayers such as marital (filing) status, number of dependents, and AGI									
Rebalancing Returns	Current	Returns 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 above-the-line 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; totals rounded independently									
	Revised	Rebalance returns for effects of modifying, blurring and rounding variables by recomputing gross income, education deductions, AGI, personal exemption amounts, itemized deductions, taxable income, regular tax, AMT, credits, and tax after credits; keep current procedures for deleted variables								N/A	
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 returns excluded from the PUF universe prior to reweighting								Aggregate record not reweighted	

[1] Returns that are filed for taxable years more than three years prior to the current year, returns that report no income (i.e., 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] Some of the returns with "large" but not "extreme" values were sampled in other strata in 2009. The total unweighted number of returns included in the aggregate record was 1,144, which weight to 1,319 returns.

[3] 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.

[4] Return counts for the population include 105,139 returns (weighted) that were subsequently excluded from the INSOLE because they were identified as not belonging to the sample universe population (e.g., amended returns) only after sampling. Return counts for the Current and Revised PUF samples are estimates based on PUF design, so do not reflect the effects of exclusions from the PUF universe or the variability in PUF samples due to subsampling. Totals for each category:

Population	140,599,267
INSOLE Sample	295,134
Current PUF	118,106
Revised PUF	153,379

statistical analyses or tax modeling. For both reasons, under the revised PUF design all HINTS are placed in the regular strata they would otherwise have been placed in for sampling purposes and then subsampled at the corresponding PUF rates. This change reduced the 2009 PUF sample by 2,010 returns, leaving only 1,505 HINTS in the sample.

CWHS. All CWHS returns in the INSOLE sample are currently subsampled to three of the 10 CWHS endings for the PUF (i.e., a 30 percent rate). This subsampling rate was established when the INSOLE sample included only five CWHS endings and two endings were set aside for possible use to produce specialized data releases. Under the revised design, the subsampling rate for CWHS returns is increased to seven of 10 endings (70 percent), which will achieve a sample rate for CWHS returns of approximately 1 in 1,430 (0.07 percent). The initial (2009) endings are the seven that had not been used in the 2008 PUF. In addition, starting with the 2010 PUF, the seven endings will be randomly chosen from the 10 available CWHS endings to reduce the potential disclosure risk from longitudinal overlap in this portion of future PUF samples.[13] The higher subsampling rate for CWHS returns under the revised design increased the 2009 PUF sample by about 56,000 returns, most of which are in strata 10 through 16, with some of this increase partially offset by reductions due to the lower sample rates in strata 7 to 9 and 17 and 18 (see below).

Strata 10 through 16. These strata cover positive total income of $1 up to $120,000 (in 1991$; $173,508 in 2009$ based on the chained GDP implicit price deflator used by SOI to index the income stratifier). For the INSOLE sample these strata include only CWHS returns. Approximately 92 percent of the 2009 population of returns is included in these strata, so improving the precision of the PUF in these strata by increasing the subsampling rate for CWHS returns from three to seven endings is important for users. In addition, to limit disclosure risk the new variables added to the PUF in the revised design are only added to returns sampled in these strata and in surrounding strata that under the revised design are subsampled to include only CWHS returns (see next paragraph).

Strata 7 through 9, 17 and 18. Strata 7 through 9 cover returns with total negative income between -$1 and -$250,000 and strata 17 and 18 returns with total positive income between $120,000 and $250,000 ($173,508 and $361,475 in 2009$). These strata cover approximately six percent of the 2009 return population, so all but two percent of the return population is covered by strata 7 through 18.

The revised design balances the analytical value of having new variables added to returns in these strata against the rise in disclosure risk as sample rates increase by reducing the PUF sample in these strata to include only seven of 10 CWHS endings. This change reduced the PUF sample in 2009 by about 14,500 returns. In addition, as discussed below, currently returns in these strata that report AGI of $200,000 or more (in absolute value) are treated as “high-income” returns for purposes of deleting, modifying, and blurring variables, whereas under the revised design all returns in these strata (subsampled to seven CWHS endings) are considered “low-income” for all of these purposes. Consequently, about 1,500 returns classified as “high-

[13] This approach follows the current design for non-CWHS returns subsampled from the INSOLE for the PUF, which each year changes the SSN transforms that are subsampled.

income" under current disclosure procedures are classified as "low-income" under the revised procedures so also have new variables added (see Table 1).

Strata 5 and 6 and 19 and 20. Strata 5 and 6 cover returns with total negative incomes between -$250,000 and -$1 million, and strata 19 and 20 returns with total positive income between $250,000 and $1 million ($361,475 and $1,445,900 in 2009$). Other than CWHS subsampling, returns in these strata are currently not subsampled for the PUF, and are not subsampled under the revised design. However, as discussed below, currently returns in these strata with AGI of under $200,000 (in absolute value) are treated as "low-income" returns for purposes of deleting, modifying and blurring variables, whereas under the revised design all returns in these strata are considered "high-income" for all of these purposes. As a result, about 8,100 returns currently classified as "low-income" are classified as "high-income" under the revised design.

Strata 3 and 4 and 21 and 22. Strata 3 and 4 cover returns with total negative incomes between -$1 million and -$5 million, and strata 21 and 22 returns with total positive income between $1 million and $5 million ($1,445,900 and $7,229,500 in 2009$). Returns in these strata are currently subsampled for the PUF to achieve a 1 in 10 sample rate, and under the revised design they continue to be subsampled in this manner for the PUF.[14] Returns in these strata are also treated as "high-income" returns for purposes of deleting, modifying and blurring variables, which continues under the revised design.

Strata 201, 1 and 2, and 23 and 24. Stratum 201 covers returns with net business receipts from Schedule C or Schedule F of $50 million or more, strata 1 and 2 cover returns with total negative incomes less than -$5 million, and strata 23 and 24 returns with total positive income over $5 million ($7,229,500 in 2009$). Returns in all of these strata are selected at a 100 percent rate for the INSOLE sample, and currently subsampled for the PUF to achieve a 1 in 10 sample rate. In addition, as noted above, a small number of returns with "extreme" values are currently excluded from the PUF.

The Working Group had two main concerns with the current approach for returns in the 100 percent strata (and some other returns sampled at high rates). One is that some of the PUF returns in these strata may be sufficiently unique to allow identification in spite of subsampling and the deletion, modification, blurring and rounding of variables because they report very large values for one or more amount fields. The second is that aggregate values of amount variables are underestimated by the PUF because the exclusion of returns with "extreme" values is not compensated for by reweighting the remaining returns (which report mean values lower than the excluded "extreme" values).

The revised design addresses both concerns in the following way. First, all INSOLE sample returns in these strata are divided into two groups: those with and those without "large" values for any amount field. Returns with large values are aggregated into a single record. Returns with no large values are subsampled as described above.

[14] The 1 in 10 sample rate includes the effects of CWHS subsampling.

Aggregate Record

"Large" Values. Every record in the INSOLE file (except returns excluded from the PUF universe) is tested to see if it contains one or more "large" values, and if so it is included in the aggregate record. In general, the following amount values are considered "large": the 30 highest amounts reported for any income amount included in the sample selection amount "total positive income", the 30 lowest amounts included in "total negative income", and the 10 highest amounts (and 10 lowest, for amounts that can be negative) reported for most other amount fields. Certain amounts have special cutoffs (e.g., for positive AGI, the top 400 returns because SOI has released tabulations covering just these returns). These rules are not applied to amount fields that are statutorily capped, subject to income limits that generally preclude any returns in the very high sampling rate strata, or calculated from other fields that are subject to the rules.

All of the 1,144 returns with large values that are aggregated for the 2009 PUF were sampled in strata 1 through 4, 21 through 24 or 201 (after restratification of HINTS). Because returns in these strata would have been subsampled to a 10 percent rate, aggregation reduced the 2009 PUF sample by about 100 returns.[15]

Aggregation. All returns with "large" values for any variable are included in the aggregation to form the aggregate record. The aggregation rules result, for nearly all variables, in no fewer than 10 returns included in the aggregate record with a nonzero value for any variable considered to carry a potential disclosure risk. Variables that are considered to carry a potential disclosure risk for which fewer than 10 returns report a nonzero value are excluded from the aggregate record.

Only a single weight (the sum of the weights of all aggregated returns) appears on the aggregate record. For nearly all variables, some aggregated returns have no reported amount (i.e., report zero or nothing), so the single weight is not an accurate return count for any amount that is not nonzero on all aggregated returns. To address this issue, a table of return counts for nonzero amounts for each variable will be included in the documentation for the 2009 and future PUFs. This tabulation will also include total return counts for selected categorical variables, such as returns by marital (filing) status, and new variables (age in ranges and gender of the primary taxpayer, and earnings splits on joint returns in ranges) but with no cross-classifications. A second table will show return counts and total amounts for returns reporting positive and separately for returns reporting negative amounts for each variable that can have a positive or negative value.

The net effect of the revised design on excluded returns, subsampling and aggregation increases the total 2009 PUF sample from approximately 118,000 returns to approximately 153,000 returns (and 1 aggregate record). This increase in the size of the sample by approximately 35,000 returns should improve the quality of most variables by reducing variances, while also reducing

[15] The aggregate record covers returns with "extreme" values, which would have been excluded from the PUF under the current design, so less than 10 percent of the returns that are aggregated would have appeared in the PUF. Identifying returns with "extreme" values in the INSOLE sample is a manual process and the number varies somewhat across years, but has typically been about 100 returns.

disclosure risk by removing (through aggregation) all returns with large values reported for any variable that might carry a disclosure risk and also the over sample of HINTS.[16]

IV. Deleting, Modifying, Blurring and Rounding Variables

The nondisclosure procedures for deleting, modifying and blurring variables under the current PUF design differ for "low income" returns, defined as returns with AGI under $200,000 (in absolute value) or with a weight greater than 15 (approximately), and the remaining, "high income" returns, those with AGI of $200,000 or more (in absolute value) or a weight of 15 or less (approximately) after subsampling.[17]

The Working Group was concerned that the current definition of "high-income" omitted some returns with high positive amounts of income but offsetting losses (so had AGI under $200,000) and relatively low weights that carry a higher disclosure risk than other "low-income" returns. Because of this disclosure risk, the Working Group determined that the disclosure avoidance procedures (deleting, modifying, and blurring variables) that apply to "high-income" returns should apply to this group of returns. In addition, to avoid potential new disclosure risks, these returns should not have new variables added to them. The revised design therefore uses strata boundaries alone to define which returns are "low income" and "high income" for purposes of deleting, modifying and blurring variables, and also for the purpose of determining which returns have new variables added. PUF returns (after restratifying HINTS) in strata 7 through 18 are considered "low income" under the revised design, while all remaining returns (those in strata 1 through 6, 19 through 24 and 201) are considered "high income".

Deleted Variables

Currently, state code, state sales tax deduction and alimony paid and received are all deleted from "high income" returns. These variables are deleted because they may be publicly available from other sources so, in combination with other variables on the PUF, could enable identification of a high-income taxpayer and therefore disclose the other items reported on their return.

State code is deleted from all returns in the PUF under the revised design. Two considerations led to this change. One is that state codes, in combination with other information, increase disclosure risks on all returns, and these risks could increase further with the addition of the new demographic and earnings split variables to the PUF described below. The other consideration is that 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.

[16] The overall effects of the revised design on variances and disclosure risk in the 2009 PUF is still being assessed by SOI and Mathematica Policy Research.

[17] The subsampling rate in strata 1 through 4, 21 through 24, 101 (HINTS) and 201 are designed to achieve a 10 percent PUF sample rate, but subsampling variability leads overall to a somewhat lower sample rate (so higher sample weights) for these strata. The PUF sample weight cutoff for "high income" returns is therefore about 15.

To address concerns voiced by some PUF users over the deletion of state codes, SOI is exploring various options and techniques to provide state information to PUF users. In addition, Khitatrakun and Mermin (2014) have extended to the PUF a statistical technique developed by Schirm and Zaslavsky (1997) to improve state-by-state estimates by splitting national weights into weights for each state. Using this technique, PUF users will be able to produce significantly more reliable state-by-state estimates than had been possible using the (limited) state code information that appeared on previous PUFs.

State sales tax deduction and alimony paid and received continue to be deleted from all "high income" returns (as defined for the revised design), and from the aggregate record. These variables are not deleted from the small number of returns with AGI of $200,000 or more in strata 7 through 18 (i.e., that are "low income" returns under the revised design but are "high income" under the current design), but all returns in these strata are now subsampled to seven CWHS endings (i.e., to a sample rate of about 0.07 percent).

Modified Variables[18]

Currently, marital (filing) status is modified on all returns by converting the "surviving spouse" status into "joint" and by collapsing the two "married filing separate" statuses into one.[19] This modification is retained under the revised design. In addition, under the revised design "high income" returns claiming "head of household" filing status and no dependents are converted to "single" filing status.[20] Marital status does not appear on the aggregate record.

Under the current design, the number of dependents by type, personal exemption amounts, and the number of children for which the child tax credit is claimed are modified on "high income" returns. The number of dependents is modified only if the number of children living at home is three or more, in which case the number of children living at home is capped at three and the number of all other types of dependents is set to zero.[21] If the number of children living at home is less than three, the number of dependents of other types is not modified. Personal exemption amounts are modified to reflect any change in the total number of dependents, and also for the effect of the personal exemption phaseout (PEP). Finally, if the number of children living at home is three or more, the number of children claimed for the child tax credit is also capped at three.

The Working Group identified a number of potential additional disclosure risks related to number of dependents by type. One is that the current procedures apply only to "high-income" returns, but in some circumstances there may be disclosure risks for these variables on "low-income" returns. Second, if the number of children living at home is less than three the number

[18] Returns in the PUF that cover fiscal (i.e., non-calendar) years are converted to reflect the parameters of the most recent year-end tax year. Although these conversions may require modifying at least some variables, these modifications are not considered here.

[19] On the INSOLE, married filing separate returns receive different marital status codes depending on whether the spouse also filed a return or was claimed as a dependent (which requires that the spouse had no income, did not file a return, and could not be claimed as a dependent by another taxpayer).

[20] This change is related to the reduction in the number of "categories" for purposes of multivariate blurring discussed below.

[21] Other dependent types are children living away from home, parents, and "other".

of dependents of other types is not modified so the total number of dependents can exceed three. A relatively infrequent number for total dependents or for dependent types other than children living at home carry potential disclosure risk. Third, currently the same procedures apply irrespective of marital (filing) status, even though the frequency of the total number of dependents and number of dependents by type differs significantly by marital status. And fourth, any cap on the number of children living at home or on other types of dependents should be taken into account in the number of students that qualify for, and the amount of, education deductions and credits and the amount of the child tax credit as well as the personal exemption amount and the number of children that qualify for the child tax credit.

The revised design addresses these disclosure risks with a new set of procedures for modifying number of dependents. New caps that vary with filing status apply to the total number of dependents on all returns, regardless of income or sample weight. These new caps are: three for joint and head of household returns, two for single returns, and one for married filing separate returns. These caps are applied sequentially by type of dependent, starting with children living at home, then children living away from home, etc. This sequential application helps insure consistency with the age of dependents (in ranges), one of the new variables added to the PUF starting with the youngest dependent. This consistency helps reduce potential disclosure risk because the youngest dependents are in most cases children, and their age (in ranges) can sometimes be inferred from other variables on the PUF. The new caps are also taken into account for the education deductions and credits that are claimed on a return, so that the combination of the number of taxpayer(s) and dependents and the amount claimed for education benefits is consistent with the capped number of dependents.

Blurring

Blurring (also referred to as masking and microaggregation) is a statistical technique for reducing disclosure risk by replacing the value of one or more variables on a group of returns with the average value(s) for the variable(s) for returns in the group. If a single variable is blurred across returns in each group the blurring is “univariate”, whereas if two or more variables are blurred across the returns in each group the blurring is “multivariate”. Under the current PUF design, multivariate blurring is applied to three variables (salaries and wages, state and local income taxes, and real estate taxes) on “high-income” returns. Univariate blurring applies on all other (i.e., “low-income”) returns to six variables (alimony paid and received (which are deleted from “high-income” returns), salaries and wages, medical and dental expenses, state and local income taxes (in Wisconsin only),[22] and real estate taxes).

The Working Group identified several issues with the way returns are grouped into “categories” for purposes of multivariate blurring of “high income” returns. Currently, 13 combinations of marital (filing) status and the number of children at home are used to define categories. Only returns in the same category, and with the same combination of the presence (zero and nonzero

[22] Wisconsin publishes income tax payments, which might allow identification of a return if amounts were not blurred.

amounts) of variables to be blurred and the presence of Schedule C net receipts, can be blurred.[23] Within each of these subgroups, variables to be blurred are normalized, a distance metric among all returns is calculated, and then the returns are selected in groups of three, starting with the returns that are furthest apart, and the variables are multivariate blurred. One issue is that some of these subgroups contain few returns so multivariate blurring cannot be applied (i.e., blurring is univariate) or is done only for two of the three variables. Small subgroups can also affect data quality because returns are not "close" so blurring significantly reduces variances. Another issue is that some of the categories cover more than one marital (filing) status, which might allow the effects of blurring to be partially reversed. To address these issues, under the revised design the definition of categories is changed to remove mixing of marital statuses and to reduce the number of categories to 10.[24] The new categories are defined by marital status and total number of dependents (rather than number of children at home), and use (a subset of) the revised caps on number of dependents. With fewer categories, there are more returns in the smaller subgroups used for blurring, so more returns will be subject to full multivariate blurring and we expect variances to be less affected.[25] Aside from the new definition of "high income" and new categories, the process of multivariate blurring is unchanged.

All returns subject to univariate blurring (all returns in strata 7 through 18 under the revised design), continue to be grouped together (i.e., nationally) for blurring of the alimony variables. For univariate blurring of other variables, returns are grouped by filing status, presence of dependents and strata groups, but are no longer grouped by state code (which, as noted above, is removed from the PUF). In addition, the state and local income tax variable is blurred on all returns, and univariate blurring of some of the six variables is now performed across up to 10 closest returns (ranked by the size of the blurred variable), rather than only the closest three.

Rounding

Rounding reduces disclosure risk in a manner similar to blurring, and can be more effective than univariate blurring if variable amounts are clustered. Under the current design, all dollar amounts are rounded to the four most significant digits (e.g., $14,371 would be rounded to $14,370 and $228,867 would be rounded to $228,900). Under this rounding rule, amounts under $10,000 (in absolute value) are not rounded at all, and amounts between $10,000 and $99,999 (in absolute value) are only rounded to the nearest $10. Totals are rounded independently, so sums of components may differ from totals due only to rounding.

Although the disclosure risk of reporting small amounts is generally quite low, because of the range of variables included in the PUF the Working Group determined that the rounding rules should be strengthened for amounts under $100,000 (in absolute value). The revised design therefore rounds amounts (in absolute value) between $10,000 and $99,999 to the nearest $100, amounts between $5 and $9,999 to the nearest $10 and (nonzero) amounts under $5 to $2 (with

[23] Schedule C net receipts does not appear on the PUF, so is currently used only to determine which returns are eligible for multivariate blurring. Some of the combinations of presence of variables are condensed for blurring to avoid very small subgroups.

[24] In addition, for placing returns within each category into subgroups the presence of Schedule C net receipts variable is replaced by an indicator variable for whether a Schedule C was filed (a variable that appears on the PUF).

[25] Mathematica Policy Research, as SOI's statistical consultant, performs an analysis of the effects of multivariate blurring for each PUF, but has not yet completed that analysis for the 2009 PUF.

the sign retained). Amounts of $100,000 or more (in absolute value) continue to be rounded to the four most significant digits. Total amounts are recomputed (see next section), so are not independently rounded.

V. New Variables

Although age (year of birth) and gender information for taxpayers and dependents is included on the INSOLE file, only limited tabulations of tax return information classified by age and gender are published by SOI, and these variables have never appeared on the PUF.[26] The Working Group recognized that adding information on age and gender to the PUF would be extremely valuable for analysis of a range of issues, such as how the work, investment, and retirement decisions of men and women differ, and how those decisions vary over the life cycle and with the ages of children. Similar to age and gender information, W-2 wages for each spouse are included on the INSOLE file,[27] but only limited tabulations on earnings splits of spouses on joint returns have been published by SOI[28] and no information on wage splits is included on the PUF.[29] The Working Group recognized that earnings splits are critical for a range of analyses, including proposals to change the wage cap for Social Security retirement (OASDI) purposes, proposals to change the limits on retirement contributions, and the size and distribution of marriage penalties and bonuses.

The revised design enhances the research and analytical usefulness of the PUF by adding variables for age (in ranges) for primary taxpayers and dependents, gender for primary taxpayers, and earnings splits (in ranges) for joint filers. As shown in Table 2, the age ranges differ for primary taxpayers who are not dependents of another taxpayer, primary taxpayers who are dependents of another taxpayer, and dependents. The age ranges for non-dependent primary taxpayers are the same as those used in SOI's annual *Individual Income Tax Returns* publication.[30] Age ranges for dependent primary taxpayers were chosen to reflect the age distribution of these returns in the population while assuring a sufficient number in each range to avoid disclosure concerns. For dependents, age ranges reflect a combination of their age distribution, relevance for analysis of current or proposed policies, and, to reduce disclosure risk, consistency with other information on returns that may indicate the age range of certain dependents.[31] The new age variables for dependents simply provide the age range for each

[26] Beginning for tax year 2006, SOI has included tables based on the age (in ranges) of primary taxpayers in the annual *Individual Income Tax Returns* publication. Prior publications with tabulations by age and/or gender of taxpayers include Sailer, Yau and Rehula (2002), Yau, Gurka, and Sailer (2003), and Bryant (2008).

[27] The inclusion of W-2 information on the INSOLE is relatively recent, but previously W-2 information had periodically been linked to the INSOLE from information return samples.

[28] SOI recently began publishing detailed tables based on W-2 information that include wage splits on joint returns, but these tables do not cover earnings from self-employment (see Pierce and Gober, 2013).

[29] The PUF does contain information on the split of earnings from self-employment.

[30] Table 1.5 in the *Individual Income Tax Returns* publication includes age categories of under 18 and 18 under 26. The age ranges used on the PUF for non-dependent primary taxpayers combine these two ranges, while the age ranges for dependent primary taxpayers combine all the ranges above these two ranges.

[31] Other information includes, in particular, the child and dependent care tax credit, which typically applies to children under age 13; the child tax credit, which applies to children under age 17; the dependent exemption and EITC for children who are not students, which generally apply only to children under age 19; and the dependent exemption and EITC for children who are students, which is under age 24. Note that student status might be inferred from education variables.

dependent, starting with the youngest, up to the cap on the number of dependents for the marital status; these variables are not associated with specific dependent types.

Table 2: Age Ranges for Non-Dependent Filers, Dependent Filers, and Dependents

Non-Dependent Filers		Dependent Filers		Dependents	
1	Under 26	1	Under 18	1	Under 5
2	26 under 35	2	18 under 26	2	5 under 13
3	35 under 45	3	26 and over	3	13 under 17
4	45 under 55			4	17 under 19
5	55 under 65			5	19 under 24
6	65 and over			6	24 and over

The ranges for splits of earnings (wages plus self-employment income) on joint returns are limited to three: 100 percent primary, 100 percent secondary, and all splits between these two (i.e., for two-earner couples). This limited set of ranges reflects the distribution of earnings splits in the population, while assuring a sufficient number in each range, in combination with other variables, to avoid disclosure concerns.

The new variables for age, gender and earnings splits are added only to PUF records in strata 7 through 18 (including HINTS re-stratified into these strata). The PUF sample rate for returns in these strata, roughly 1 in 1,430 (0.07 percent), is very low and much lower than in any of the remaining strata, so this restriction partially addresses the potential increase in disclosure risk due to the addition of these new variables. To further address the remaining increase in potential disclosure risk, we extracted a set of variables that might be directly observable or might be obtained or inferred from publicly-available sources for returns on the IRS Master File that were processed in 2010 (i.e., the period covered by the tax year 2009 INSOLE sample). SOI sampling strata is not one of the variables included on the Master File, so we used proxy variables to limit the extract to returns that would very likely be in strata 7 through 18. The variables obtained for each extracted return included marital (filing) status, age of primary (converted to the ranges describe above), gender of primary, number of dependents (with caps described above and alternatives applied), ages of each dependent (converted to the ranges described above) in ascending order, wages (in ranges that increased with the level of wages), earnings splits on joint returns (calculated in more ranges than those described above), and the presence of unemployment benefits, a farm net income or loss, or a first-time homebuyers credit.

Most returns do not report dependents, so any increase in potential disclosure risk would likely arise from the addition of ages (in ranges) of dependents (up to the cap for each marital status). Because relatively few returns report farm net income or loss or a first-time homebuyers credit, to avoid disclosure risk the age of dependents variables do not appear on returns reporting these items.

To assess the potential disclosure risk of including the age of dependents variables on the remaining returns, we cross-tabulated these returns by all of the remaining variables extracted from the Master File. For returns reporting any wages, for each combination of other variables (except gender, but including earnings splits, for joint returns) a search across wage ranges was made to determine the lowest wage range with fewer than 100 returns. The bottom wage for this range was then set as a tentative cap for the level of wages for returns with that combination of other variables for which age(s) of dependent(s) appear on the PUF. For single, head of household, and married filing separately returns, the minimum of the tentative caps for male and female primaries (for each combination of other variables) was set as the final wage cap. For joint returns, the minimum across the tentative caps for earnings splits for two-earner couples (for each combination of other variables) was set as the final wage cap for all two-earner couples. (As noted above, only a single wage split category is included on the PUF for two-earner couples.) The caps on the total number of dependents by filing status were also based on these cross-tabulations. For returns reporting no wages, a similar set of cross-tabulations was prepared using all of the remaining variables except wages, earning splits, and presence of unemployment benefits.[32] Cells with fewer than 100 returns were identified, and age(s) of dependent(s) do not appear on any PUF return with the characteristics that define one of these cells.

Two considerations lead us to set the quite stringent requirement of at least 100 returns with a given set of characteristics in the filing population in order for the age of dependent variables to appear on the 2009 PUF. One is that the process of preparing the Master File extract, preparing the cross-tabulations, identifying cells for which age of dependents can and cannot be included on the PUF, coding and testing these rules, and then applying them in the production of the PUF requires significant staff time. If the requirement had been set just high enough to address nondisclosure concerns in 2009 alone, the process (and cost in staff time) would need to be repeated for each PUF. By setting the requirement at 100 returns, the rules determined from the 2009 population provide a large margin of safety against any shifts in disclosure risk due to the relatively small shifts in the filing population covered by the PUFs for the next several years. The other consideration is that the PUF contains many variables, so a high requirement is a safeguard against the possibility that a PUF variable that was not included in the cross tabulations might pose a disclosure risk.

Because the new age, gender and earnings split variables will not appear on all returns in strata 7 through 18 due to the various nondisclosure procedures applied and will never appear on returns in the other strata, and because the age and earnings split variables are in fairly broad ranges, SOI will produce a set of tabulations of these new variables to accompany the PUF each year, starting in 2009. The tabulations will provide some cross-classifications of the new variables by AGI, marital (filing) status, and number of dependents, and also tabulate some of the new variables in additional detail. These new tabulations will provide very useful information that can be used for separate analyses as well as to aid PUF users in the imputation of missing data and detail on these variables.

As noted above, under the new design a tabulation of new variables for returns included in the aggregate record is also added to the PUF documentation. A separate tabulation in the documentation provides the total of positive and total of negative amounts from aggregated returns for each variable that can be positive or negative to help PUF users perform analyses using the aggregate record.

[32] Very few returns report unemployment benefits but no wages.

VI. Rebalancing and Reweighting Returns

Rebalancing Returns

Deleting, modifying, blurring and rounding variables changes relationships among some of the variables on a tax return, making them out of balance. Currently, the change in AGI due to blurring of wages and deletion of alimony paid and received is not carried through as a change in AGI, but rather is included in an implied residual variable that includes "other income" (line 21 of Form 1040) and some above-the-line deductions. Total deductions (standard or itemized) and personal exemption amounts are always a combined implied residual variable which includes the effects of deleting, modifying and blurring component variables, so no other rebalancing is required.

A concern with the current approach is that the resulting implied residual variables may increase disclosure risk. The Working Group asked Mathematica Policy Research, SOI's statistical consultant, to assess this risk, which it did. Based on that assessment, under the revised design returns are rebalanced for the effects of blurring, modifying and rounding variables by re-computing gross income, education deductions, AGI, taxable income, regular tax, AMT, the child tax credit, the education credits, and tax after credits. Current procedures are retained for rebalancing necessary for deleted variables, i.e., the effects show up in implied residual variables.

Reweighting

Currently, the population counts are not adjusted for returns that are excluded from the PUF universe prior to reweighting the PUF to adjust for subsampling (see above). Under the revised design, the population counts are adjusted for these excluded returns. In addition, because the aggregate record represents all of the aggregated returns and therefore is not reweighted, these returns are also removed from the population counts.

VII. Completing the 2009 and Future PUFs

A preliminary version of the 2009 PUF file has been produced using all of the design changes describe above. Prior to its release with these changes, SOI and Mathematica Policy Research must complete an analysis to determine whether the file, with these changes, raises any new disclosure risks and the effect of the revisions on variances. Depending on the results of these analyses, refinements might be made to elements of the revised design to insure nondisclosure and data quality requirements are met before the final 2009 PUF is publicly released.

The 2010 and subsequent PUFs will use the same design as the final 2009 PUF. A goal is to gradually reduce the lag time between completion of the INSOLE file and completion of the corresponding PUF to the time necessary to produce a PUF, which is about six months. On this schedule, the PUF for a tax year will be released by the end of the second following year.

References

Bryant, Victoria, 2012. *General Description Booklet for the 2008 Public Use File*. Statistics of Income Division, Internal Revenue Service, Washington, DC.

Bryant, Victoria, 2008. “Accumulation and Distribution of Individual Retirement Arrangements, 2004.” *SOI Bulletin* (Spring), 90-101.

Joulfaian, David and David Richardson, 2001. “Who Takes Advantage of Tax-Deferred Saving Programs? Evidence from Federal Income Tax Data.” *National Tax Journal* 54 (3), 669-688.

Khitatrakun, Surachai and Gordon Mermin, 2014. “Re-Weighting a Model Based on IRS Statistics of Income Public Use Tax File for State-Level Analysis.” Paper to be presented at the 107th Annual Conference on Taxation of the National Tax Association.

Pierce, Kevin and Jon Gober (2013), ‘‘Wage Income and Elective Retirement Contributions From Form W-2, 2008-2010,’’ *SOI Bulletin* (Summer), 5-21.

Sailer, Peter, Victoria Bryant and Sarah Holden, 2005. “Trends in 401(k) and IRA Contribution Activity, 1999-2002 – Results from a Panel of Matched Tax Returns and Information Documents.” Paper presented at the American Statistical Association Meetings.

Sailer, Peter and Kurt Gurka, 2003. “Accumulation and Distributions of Retirement Assets, 1996-2000 – Results from a Matched File of Tax Returns and Information Returns.” Paper presented at the American Statistical Association Meetings.

Sailer, Peter and Sarah Holden, 2004. “Use of Individual Retirement Arrangements to Save for Retirement – Results from a Matched File of Tax Returns and Information Documents for Tax Year 2001.” Paper presented at the American Statistical Association Meetings.

Sailer, Peter and Sarah Nutter, 2004. “Accumulation and Distribution of Individual Retirement Arrangements, 2000.” *SOI Bulletin* (Spring), 121-134.

Sailer, Peter, Michael Weber and William Wong, 2001. “Disclosure-Proofing the 1996 Individual Tax Return Public use File.” Paper presented at the American Statistical Association Meetings.

Sailer, Peter, Ellen Yau, Kurt Gurka and Michael Weber, 2002. “Salaries and Wages and Deferred Income, 1989-1999.” Paper presented at the American Statistical Association Meetings.

Sailer, Peter, Ellen Yau and Victor Rehula, 2002. “Income by Gender and Age from Information Returns.” *SOI Bulletin* (Winter), 83-102.

Schirm, Allen and Alan Zaslavsky, 1997. “Reweighting Households to Develop Microsimulation Estimates for States.” *1997 Proceedings of the Section on Survey Research Methods*. American Statistical Association, Alexandria, VA.

Statistics of Income Division, Internal Revenue Service, Annual. *Individual Income Tax Returns*. Washington, DC.

Strudler, Michael, H. Lock Oh and Fritz Scheuren, 1986. “Protection of Taxpayer Confidentiality with Respect to the Tax Model.” Paper presented at the American Statistical Association Meetings.

Vartivarian, Sonya, John L. Czajka and Michael Weber, 2007. “Measuring Disclosure Risk and an Examination of the Possibilities of Using Synthetic Data in the Individual Income Tax Return Public Use File.” Paper presented at the American Statistical Association Meetings.

Wilson, Oliver H. and William J. Smith, Jr., 1983. “Access to Tax Records for Statistical Purposes.” Paper presented at the American Statistical Association Meetings.

Winglee, Marianne, Richard Valiant, Jay Clark, Yunhee Lim, Michael Weber and Michael Strudler, 2002. “Assessing Disclosure Protection for a SOI Public Use File.” Paper presented at the American Statistical Association Meetings.

Yau, Ellen, Kurt Gurka and Peter Sailer, 2003. “Comparing Salaries and Wages of Women Shown on Forms W-2 to Those of Men, 1969-1999.” *SOI Bulletin* (Fall), 274-283.