Redesign of SOI's Individual Income Tax Return Edited Panel Sample

Yan K. Liu[1], Gerald Auten[2], Valerie Testa[1] and Michael Strudler[1]

Abstract

The Statistics of Income (SOI) Division of the Internal Revenue Service started the Individual Income Tax Return Edited Panel Sample in 1999. It has been used for both longitudinal and cross-sectional purposes. Because the panel drifted over time, the efficiency of the panel was reduced in terms of the longitudinal study of taxpayers' behavioral changes. Further, the yearly refreshment sample is unable to capture enough new high-income returns to support cross-sectional estimation. Therefore, it was decided to start a new panel in 2007. The new panel follows the same stratified random sample design as the old panel. The key is the sample size reduction and allocation. Using the information from the current panel, the stratum sample sizes are determined by balancing the cost and precision and factoring in the needs of panel users. In addition, a yearly refreshment sample is designed to better represent out-year populations.

Key words: Cross-sectional estimation, longitudinal analysis, panel sample, sample size allocation, stratified sample

1. Introduction

The Statistics of Income Division of the Internal Revenue Service started a panel sample in tax year 1999 called the Individual Income Tax Return Edited or High Income Cohort Panel Sample. One use of this panel sample is to study the Sales of Capital Assets (SOCA) and therefore it is also sometimes called the SOCA panel sample. Sales of capital assets include such things as stocks, bonds, mutual funds, and property as well as other assets. The SOCA panel sample is used for both longitudinal analysis and cross-sectional estimation.

There are two panel-related issues limiting the analysis of future years with the 1999 panel. First, many panel members were lost over time through attrition. Some of this was natural, such as death, while other causes were related to filing behaviour (Bryant, 2008). This reduces the efficiency of the longitudinal study. Second, because the panel is based on the population from 1999, it is no longer representative of the current population because the population changed while the panel did not. Further, the filers that started with low-income in 1999 and then became rich are not well represented. We call these newly rich 'poppers.' This also limits the study of the effects of the tax system on the rise and fall of taxpayer's income. Even though there is a small refreshment sample every year, it is a simple random sample, and far from enough to keep the panel sample representative of the cross-section over time. Of particular concern is that high-income returns are not well represented by the refreshment sample. We are interested in adequate coverage of both the new filers and the poppers.

Because of these issues, it was decided to start a new SOCA panel sample in tax year 2007. The decision to start a new panel in 2007 is also supported by the fact that there is a new administration in 2009. The goal is to have a new panel sample smaller and lower cost than the 1999 panel sample with a yearly refreshment sample that can support cross-sectional estimates.

[1] Statistics of Income Division, IRS, P.O. Box 2608, Washington, DC 20013. Email:yan.k.liu@irs.gov.
[2] Office of Tax Analysis, U.S. Department of the Treasury, Washington, DC.

2. SOI Yearly Cross-Sectional Sample of Individual Returns

SOI selects a cross-sectional sample of individual returns every year. However, it provides capital gains data only at the return level, not at the capital asset transaction level because of the high processing cost associated with editing the more detailed data. Therefore, periodic smaller cross-sectional samples and a still smaller panel sample are used to collect SOCA data at the transaction level, for both longitudinal and cross-sectional analysis. In fact, the 1999 panel sample was a subset of the 1999 yearly cross-sectional sample.

The cross-sectional sample is a stratified sample where strata are determined primarily by income range but with some oversampling of certain returns with a high degree of interest. The high degree of interest is based on certain relatively rare characteristics where a larger sample is needed. Returns of high interest include returns with high income and no tax after credits, or returns with large business or farm receipts respectively and are sampled with certainty, regardless of the income amount. The rest of the returns are divided into 24 strata within each tax return type by a combination of income and degree of interest. The boundaries of these strata are basically the same as those shown in Table 1 of Section 3 except that the top positive and negative strata are further split in Table 1 for panel purpose.

A Bernoulli sample is selected independently from each stratum. The sample selection is done in two independent parts. First, the Continuous Work History Sample (CWHS) is selected based on the last four digits of the primary taxpayer's Social Security Number (SSN) (Weber, 2005). Five ending digit combinations were selected during tax years 1999 through 2004 giving a sample selection rate of 0.05%. Beginning in 2005, ten ending digit combinations were included in the sample thereby increasing the sample selection rate to 0.10%. In addition to the CWHS sample, all returns are also subject to sampling based on a permanent random number that uses the primary taxpayer's SSN as the seed. This permanent random number is referred to as the Transformed Taxpayer Identification Number (TTIN). If the last five digits of the TTIN, which range from 0 to 99,999 are less than the sample selection rate for the strata into which the tax return falls multiplied by 100,000, then it is selected for the sample.

3. 1999 Panel Sample Design

The new panel design is a modification of the 1999 panel design. The 1999 panel sample design was a stratified sample and a subsample of the 1999 cross-sectional sample. Unlike the yearly cross-sectional sample design that is stratified by return type, income, and the degree of interest, the panel sample stratification is by income, as shown in Table 1.

Table 1. Sample Design of TY1999 Panel Sample and TY1999 Cross-sectional Sample

Stratum	Income Range	Degree of Interest	Expected Sampling Rate (%)	
			Cross-sectional Sample*	Panel Sample
	NEGATIVE INCOME			
1	$20,000,000 and over	All	100.00	100.00
1	$10,000,000 - under $20,000,000	All	100.00	48.47
2	$5,000,000 - under $10,000,000	All	100.00	22.05
3	$2,000,000 - under $5,000,000	All	33.42	4.20
4	$1,000,000 - under $2,000,000	All	16.03	1.42
5	$500,000 - under $1,000,000	All	3.36	0.05
6	$250,000 - under $500,000	All	0.94	0.05
7	$120,000 - under $250,000	All	0.46	0.05
8	$60,000 - under $120,000	All	0.26	0.05
9	Under $60,000	All	0.14	0.05
	POSITIVE INCOME			
10	Under $30,000	1	0.05	0.05
11	Under $30,000	2	0.05	0.05
12	Under $30,000	3, 4	0.10	0.05
13	$30,000 - under $60,000	1, 2	0.05	0.05
14	$30,000 - under $60,000	3, 4	0.11	0.05
15	$60,000 - under $120,000	1,2,3	0.05	0.05
16	$60,000 - under $120,000	4	0.10	0.05
17	$120,000 - under $250,000	1,2,3	0.14	0.05
18	$120,000 - under $250,000	4	0.28	0.05
19	$250,000 - under $500,000	All	0.67	0.18
20	$500,000 - under $1,000,000	All	2.43	0.59
21	$1,000,000 - under $2,000,000	All	12.14	1.72
22	$2,000,000 - under $5,000,000	All	32.42	5.73
23	$5,000,000 - under $10,000,000	All	100.00	18.88
24	$10,000,000 - under $20,000,000	All	100.00	57.62
24	$20,000,000 and over	All	100.00	100.00

**Returns of high interest are selected with certainty, regardless of the income amount.*

The panel is also a Bernoulli sample selected in two independent parts. First, returns having one of the five ending digit combinations for the primary taxpayer's SSN are selected. These are the same CWHS returns in the 1999 cross-sectional sample, which represent 0.05% of the population. In addition to the CWHS sample, all returns are also subjected to sampling based on the Transformed Taxpayer Identification Number (TTIN). Table 1 gives the summary of the sample designs of the 1999 panel sample and 1999 cross-sectional sample. The panel sample selection rate is not greater than the cross-sectional sample selection rate for any stratum.

The panel sample is used not only for longitudinal analysis, but also for cross-sectional estimation. While the yearly SOI cross-sectional sample produces adequate estimates of capital gains and losses on a tax return basis, the panel study provides much more detailed information about each transaction reported on the tax returns using Schedule D as well as other forms. There is a high cost associated with processing a large number of individual transactions. Therefore it is more cost effective to obtain these cross-sectional estimates from the smaller panel sample than it would be to obtain these same estimates from the larger cross-sectional sample. To make the cross-sectional estimates in out-years using the panel sample, additional returns need to be added to the panel sample in order to reflect the population change. Due to various resource and planning constraints, only a small refreshment sample is added to the panel sample every year. This small refreshment sample includes CWHS returns having one of the five specific ending digit combinations of the primary filer's SSN. It is a small simple random sample. Because the distribution of income is highly skewed, this refreshment sample cannot capture enough new high-income returns. Therefore, the refreshment sample design has been revised in the new design, as described in Section 4.2.

4. New Panel Sample Design

The new tax year 2007 panel design is a modification of the old tax year 1999 panel sample design (Mathematica, 2006). It is also a stratified sample design where stratum boundaries are basically the same as the old panel except that the two certainty strata are further divided based on income. Because the high-income returns in the certainty strata are very expensive to process, we decided to make the certainty strata smaller to cut down the cost of the panel. Each certainty stratum is split into two strata – one certainty stratum selected at the rate of 100% and one stratum sampled at a rate of 50%. The key to the new panel design is the stratum sample size allocation to balance the precision of the estimates, the SOCA return processing cost, and client's needs. Like the old panel sample and the yearly cross-sectional sample, the Tax Year 2007 panel sample selection is a two-step procedure. The first step is to select CWHS returns. A specified percent of returns are also randomly selected within each stratum independent of the CWHS selection. In each out-year, a base-year panel return stays in the panel if either the primary or secondary filer files in that year, regardless of the marital status. In addition, a refreshment sample is added to keep the panel representative of the cross-section population by accounting for new population entrants and newly rich filers. Since approximately two percent of returns each year are prior year tax returns, late-filed tax year 2007 returns are included in the panel sample if they are filed in the next two years as well.

Key features of the new panel sample design include a larger CWHS sample, the stratum sample size reallocation in the base-year and cost-saving sample design features, as well as the stratified sample design for the out-year refreshment samples.

4.1 Base-Year Panel Sample Size Allocation

First, the new panel sample increases the sample size of CWHS returns, primarily to accommodate the need to examine subsets of the lower income population. The new panel sample includes ten specific ending digit combinations for both the primary and secondary taxpayers' SSN. These CWHS returns from primary tax filers do not add much cost to the panel because they are already processed in the cross-sectional sample. The primary reason that the secondary taxpayers were added to the CWHS sample is to have a complete history of prior filings. For example, if the secondary taxpayers in later years become primary taxpayers due to change in marital status, we will have their returns from previous years. In addition, while the panels are used to study the effects of tax policy on reported income, the sample is largely chosen on the basis of current year reported income, thereby creating endogeneity issues. Adding secondary taxpayers mitigates, but doesn't solve this problem. The CWHS returns in the new panel sample represent 0.14%[3] of the entire population, while CWHS returns in the old panel sample represent

[3] About 0.1% is from the primary SSN and 0.04% from secondary SSN.

0.05% of the entire population. The CWHS sample is broadly representative of the lower and middle-income ranges in the population. The stratified portion of the sample has the same design as the 1999 panel with the selection being based on the TTIN, however the rates changed.

The major changes in stratum sample sizes are intended to balance the cost and precision and also to factor in the analytical needs of the panel users. To reduce the cost, we first reduced the sample sizes of the top strata. For strata 5 – 19, we use only the CWHS returns. For the rest of the strata, we started by using an initial stratum sample sizes based on the idea that panel costs could be reduced by using sampling rates from two strata down in the 2006 cross-sectional sample so as to reduce the non-overlap of the panel with the cross-sectional sample. For example, the panel sampling rate for stratum 23 could use the cross-sectional sample rate for stratum 21. This is based on the observation that most of the tax returns in the 1999 panel did not drop more than 2 strata. Thus, with the proposed sampling scheme, a high percentage of the panel should be in the yearly cross-sectional sample anyway and SOI would save the costs of processing tax returns that are not already included in another SOI sample. We then adjusted those stratum sample sizes based on the cost and precision, as well as the input of the users of the panel data. Section 5 gives details on the evaluation. The final sampling rates are shown in Table 2.

Table 2. Selection Criteria of the Base-Year Panel Sample

Stratum	Income Range	New Panel		Old Panel	
		Expected Sampling Rate %	TTIN Cutoff	Expected Sampling Rate %	TTIN Cutoff
	NEGATIVE INCOME				
1	$150,000,000 and over	100.000	99,999	100.000	99,999
1	$40,000,000 - under $150,000,000	100.000	99,999	100.000	99,999
1	$20,000,000 - under $40,000,000	50.000	49,929	100.000	99,999
1	$10,000,000 - under $20,000,000	50.000	49,929	48.470	48,443
2	$5,000,000 - under $10,000,000	22.510	22,400	22.050	22,010
3	$2,000,000 - under $5,000,000	3.410	3,274	4.200	4,151
4	$1,000,000 - under $2,000,000	2.000	1,862	1.420	1,370
5-9	Under $1,000,000	0.140	*	0.050	*
	POSITIVE INCOME				
10-18	Under $250,000	0.140	*	0.050	*
19	$250,000 - under $500,000	0.140	*	0.180	129
20	$500,000 - under $1,000,000	0.335	194	0.590	539
21	$1,000,000 - under $2,000,000	1.900	1,761	1.720	1,670
22	$2,000,000 - under $5,000,000	2.480	2,342	5.730	5,682
23	$5,000,000 - under $10,000,000	12.200	12,076	18.880	18,838
24	$10,000,000 - under $20,000,000	28.600	28,499	57.620	57,598
24	$20,000,000 - under $40,000,000	50.000	49,929	100.000	99,999
24	$40,000,000 - under $150,000,000	100.000	99,999	100.000	99,999
24	$150,000,000 and over	100.000	99,999	100.000	99,999

* No selections based on TTIN.

Table 2 gives the selection criteria of the base-year panel sample for each stratum. For comparison, the selection criteria of the 1999 panel sample are also given in Table 2. Since the Bernoulli sample selection method is used in selecting the sample of returns, the expected sampling rates can be slightly different from the actual sampling rates. For the random number selection portion of the sample, the TTIN cutoff is calculated using the following formula

$$\text{TTIN Cutoff} = \frac{100{,}000 \times (r\% - CWHS\%)}{(1 - CWHS\%)} - 1, \qquad (2.1)$$

where $r\%$ is the expected sampling rate that includes the random number selection and the CWHS selection; and $CWHS\%$ is the expected percent of CWHS returns. $CWHS\%$ is 0.14% in the new 2007 panel sample and 0.05% in the old 1999 panel sample.

4.2 Refreshment Sample in Out-Years

Each year after the base-year, a refreshment sample is selected to support cross-sectional estimation. The refreshment sample should capture enough new entrants and high-income returns. In the old panel, the yearly refreshment included only the CWHS selection that includes returns having one of the five specific SSN ending digit combinations of the primary filer. Together with the surviving panel returns, this sample provides a good representation of the lower- and middle-income population (including new entrants), but not of the high-income population (including new filers and old filers whose income increased dramatically). In the new refreshment sample design, we made two changes. First, we increased the CWHS selection to include new returns having one of ten specific SSN ending digit combinations of both the primary filer and the secondary filer. Second, we selected a small ‘popper’ sample that includes newly ‘popped’ high-income returns that are neither already in the panel nor in the CWHS part of the refreshment.

To include enough high-income returns, the popper sample design is stratified by income with higher rates for higher income strata. The stratum boundaries are the same as in the new base-year panel sample. Tax returns with income of $150 million or more of either positive or negative income are selected with certainty. This guarantees that roughly 200 of the top returns each year will be included in the SOCA data, which would be extremely valuable for the analysis of major data users – the Office of Tax Analysis (OTA) and the Congressional Joint Committee on Taxation (JCT). The users also want to include smaller representative samples of the newly rich in other high-income strata in their analysis as well as valid cross-section samples for SOCA purposes.

Once the final decision on the stratum sampling rates was made, the TTIN cutoffs are calculated from the sampling rates using formula (2.1). The design of yearly refreshment sample is summarized in Table 3. The selection criterion of the refreshment sample for the old panel is also given in Table 3 for comparison. The purpose of the refreshment sample is to guarantee that there are enough returns for out-year cross-sectional estimation. It is expected that about 200 to 400 high-income returns would be selected for the refreshment sample each year. This would roughly compensate for those that drop-out of the panel each year as well as providing systematic high-income refreshment of returns for SOCA and other high-income analysis. So as to increase the likelihood that prior year returns could be found for the "popper" refresh sample, the sampling rates were generally three strata down in the 2006 cross-sectional sample scheme. Thus, prior year returns should be present unless selection income had increased more than eight-fold. Because of questions about the proper weighting of popper returns, these will be kept separate from the main panel until weighting and other issues are resolved.

Table 3. Selection Criteria of the Out-Year Refreshment Sample

Stratum	Income Range	New Panel		Old Panel	
		Expected Sampling Rate %	TTIN Cutoff	Expected Sampling Rate %	TTIN Cutoff
	NEGATIVE INCOME				
1	$150,000,000 and over	100.000	99,999	0.050	*
1	$40,000,000 - under $150,000,000	22.512	22,402	0.050	*
1	$20,000,000 - under $40,000,000	3.410	3,274	0.050	*
1	$10,000,000 - under $20,000,000	3.410	3,274	0.050	*
2	$5,000,000 - under $10,000,000	2.480	2,342	0.050	*
3	$2,000,000 - under $5,000,000	0.963	823	0.050	*
4	$1,000,000 - under $2,000,000	0.140	*	0.050	*
5-9	Under $1,000,000	0.140	*	0.050	*
	POSITIVE INCOME				
10-18	Under $250,000	0.140	*	0.050	
19	$250,000 - under $500,000	0.140	*	0.050	*
20	$500,000 - under $1,000,000	0.140	*	0.050	*
21	$1,000,000 - under $2,000,000	0.140	*	0.050	*
22	$2,000,000 - under $5,000,000	0.335	194	0.050	*
23	$5,000,000 - under $10,000,000	1.200	1,060	0.050	*
24	$10,000,000 - under $20,000,000	2.480	2,342	0.050	*
24	$20,000,000 - under $40,000,000	6.071	5,938	0.050	*
24	$40,000,000 - under $150,000,000	22.512	22,402	0.050	*
24	$150,000,000 and over	100.000	99,999	0.050	*

** No selections based on TTIN.*

5. Evaluations of the New Panel Sample

As described in Section 4, before determining the final stratum sample sizes we chose a few sample design options based on the consideration of many factors. We then evaluated them, adjusted the stratum sample sizes based on the evaluation, and decided on the final sample design described above. We conducted both a direct analysis and a dynamic analysis in the evaluation.

5.1 Direct Analysis - Precision and Cost in the Base-Year

The direct analysis is based on the precision and the cost of the cross-sectional estimates using the information from the 2006 cross-sectional sample data since the panel is a subset of the cross-sectional sample. The key variables we evaluated were Adjusted Gross Income, Wages, Taxes, Positive Capital Gains, Negative Capital Gains, Partnership Capital Gains, Positive Business Income, Negative Business Income, Total itemized Deductions, Charitable Deductions, and the design variable Selection Income. These variables were chosen because they are often the subject of research or tax policy analyses. We calculated stratum variance estimates of these key variables under the proposed sample design outlined in Table 3. We obtained stratum sample counts since the proposed panel sample design is a subset of the yearly cross-sectional sample. Then we calculated the processing cost and the Coefficients of Variation (CVs) and the standard errors of the totals at the stratum level. The stratum sampling rates were adjusted so as to reduce costs while also producing reasonable CVs. The standard errors of the total were also compared across strata for each key variable. This is because strata with similar CVs may have much different standard errors of the total due to different income levels. Therefore, we looked at this measure to make sure only a reasonable contribution to the overall standard error comes from each stratum. The higher costs were allowed for some negative income strata because many returns in those strata do not owe any tax, but they may later become high positive income cases. In addition, there are only a few hundred of these returns included in the panel sample.

CWHS returns identified based on both the primary and secondary taxpayers' SSNs are included in the new panel sample. However, the 2006 cross-sectional sample data include only CWHS returns identified based on the primary taxpayer's SSN. That is, it does not include CWHS returns indentified by the secondary taxpayer's SSN. To simulate these returns, we used returns identified by the primary taxpayer's SSN whose marital status indicated 'married' as an approximation. The selections based on the TTIN were obtained applying the sampling rates in Table 3 to the 2006 cross-sectional sample data. Table 4 summarizes the simulated base-year panel sample size of CWHS and non-CWHS returns[4]. Also given in Table 4 are the processing cost and population counts. The processing cost measure is the average processing time per panel return that is derived from 2006 data. The average cost per return is the total processing cost of SOCA returns divided by the number of returns, including SOCA and non-SOCA returns, within each stratum. Here, the SOCA processing time excludes the processing time from the cross sectional sample. Since some small SOCA returns are already processed in the cross-sectional sample, their SOCA processing time is already counted so their SOCA processing time is treated as zero. As shown in Table 4, the estimated total cost for the new panel is 2,160,048 minutes, approximately the same as the 1999 panel cost of 2,138,428 minutes. However, this estimated cost is based on the 2006 population. We expect a larger initial cost because the 2007 population is expected to be larger than the 2006 population, but out-year costs should be reduced because of the expected greater overlap with the cross-section sample.

In addition to the processing cost, the evaluation used the coefficient of variation (CV) and the standard error of the total for each of the key variables. We adjusted the starting stratum sample sizes based on the CVs, standard errors of the totals and the clients' input. Specifically, CVs are used to measure the precision within each stratum, while the standard errors of the total are used to compare the relative contributions across strata. Table 5 gives the evaluation results of the final sample design for 5 of the key variables. In adjusting the stratum sample sizes, we focused on the cost and the contribution to the standard error of the total. For example, if the cost is too large in a stratum, we reduced the sample size to a level where the CV and the standard error of the total were reasonable, while the expected sample size would still meet clients' needs. The CV of Adjusted Gross Income (AGI) in stratum 9 is extremely large due to the small average AGI and large variation. However, its standard error of the total is not over the limit compared to other strata. The last row of Table 4 shows that the overall CV of the total is reasonable for each variable. The final sample size allocation at the base-year is based on the comprehensive evaluation of cost and CV for each key variable, within each stratum and across all strata, as well as the clients' request that certain numbers of high-income returns be included.

[4] Twenty-one 'outlier' returns from this simulation were removed as they are extremely large or small within their strata in terms of key variables other than income amount

Table 4. Summary Information for the Evaluation of Base-Year Panel (Simulated Based on TY2006 Data)

Stratum	Income Range	Population Size	Average Cost Per return (Minutes)	# Sample Returns		Total Cost (Minutes)
				Non-CWHS	CWHS	
	NEGATIVE INCOME					
1	$100,000,000 and over	56	55	56	0	3,083
1	$90,000,000 - under $100,000,000	14	64	14	0	893
1	$80,000,000 - under $90,000,000	15	233	15	0	3,498
1	$70,000,000 - under $80,000,000	15	66	15	0	985
1	$60,000,000 - under $70,000,000	16	209	16	0	3,338
1	$50,000,000 - under $60,000,000	36	142	36	0	5,113
1	$40,000,000 - under $50,000,000	51	73	51	0	3,745
1	$30,000,000 - under $40,000,000	93	121	47	1	5,808
1	$20,000,000 - under $30,000,000	201	75	91	2	7,008
1	$10,000,000 - under $20,000,000	790	88	393	2	34,653
2	$5,000,000 - under $10,000,000	2,400	59	561	1	33,188
3	$2,000,000 - under $5,000,000	10,963	81	312	9	26,140
4	$1,000,000 - under $2,000,000	23,667	70	374	40	28,988
5	$500,000 - under $1,000,000	60,204	49	0	70	3,434
6	$250,000 - under $500,000	141,125	37	0	241	8,817
7	$120,000 - under $250,000	299,998	19	0	432	8,020
8	$60,000 - under $120,000	414,106	6	0	575	3,696
9	Under $60,000	1,136,234	3	0	1,274	4,200
	POSITIVE INCOME					
10	Under $30,000	31,663,929	0	0	28,631	56
11	Under $30,000	29,365,466	0	0	36,869	6,069
12	Under $30,000	10,829,551	1	0	12,978	11,177
13	$30,000 - under $60,000	24,185,285	0	0	35,780	11,323
14	$30,000 - under $60,000	10,749,781	1	0	17,016	25,170
15	$60,000 - under $120,000	14,420,675	1	0	25,555	30,628
16	$60,000 - under $120,000	6,372,224	4	0	10,773	44,129
17	$120,000 - under $250,000	1,964,476	3	0	3,687	10,570
18	$120,000 - under $250,000	4,206,510	12	0	7,145	83,233
19	$250,000 - under $500,000	1,723,453	29	0	3,047	87,842
20	$500,000 - under $1,000,000	590,710	48	1,119	1,002	101,163
21	$1,000,000 - under $2,000,000	202,454	71	3,422	337	268,687
22	$2,000,000 - under $5,000,000	86,073	96	1,985	172	207,368
23	$5,000,000 - under $10,000,000	20,414	135	2,523	51	347,191
24	$10,000,000 - under $20,000,000	7,849	138	2,171	26	303,211
24	$20,000,000 - under $30,000,000	1,421	158	657	3	104,395
24	$30,000,000 - under $40,000,000	944	169	476	0	80,284
24	$40,000,000 - under $50,000,000	465	160	465	0	74,514
24	$50,000,000 - under $60,000,000	289	165	289	0	47,731
24	$60,000,000 - under $70,000,000	186	168	186	0	31,320
24	$70,000,000 - under $80,000,000	133	231	133	0	30,737
24	$80,000,000 - under $90,000,000	74	126	74	0	9,293
24	$90,000,000 -under $100,000,000	71	182	71	0	12,890
24	$100,000,00 and over	334	139	334	0	46,459
OVERALL		**138,482,751**		**15,886**	**185,719**	**2,160,048**

Table 5. Evaluation Results for Selected Key Variables at the Base-Year

Stratum	Income Range	Coefficient of Variation (CV) (%)				Standard Error of the Total (Million $)			
		Selection Income	AGI	Wages	Taxes	Selection Income	AGI	Wages	Taxes
	NEGATIVE INCOME								
1	$40,000,000 - under $50,000,000					0	0	0	0
1	$30,000,000 - under $40,000,000					19.9	144.9	5.2	0.3
1	$20,000,000 - under $30,000,000					43	182	16	14
1	$10,000,000 - under $20,000,000					81	206	26	11
2	$5,000,000 - under $10,000,000					120	323	70	25
3	$2,000,000 - under $5,000,000					501	1,027	211	85
4	$1,000,000 - under $2,000,000					332	830	231	79
5	$500,000 - under $1,000,000					1,062	2,154	1,044	200
6	$250,000 - under $500,000					617	1,530	554	198
7	$120,000 - under $250,000	-1.01	40.29	6.98	9.25	527	1,429	726	118
8	$60,000 - under $120,000	-0.85	43.25	6.09	9.79	305	906	515	56
9	Under $60,000	-1.91	-496.52	6.01	14.38	554	719	360	48
	POSITIVE INCOME								
10	Under $30,000	0.45	0.45	0.45	0.77	1,562	2,086	2,091	216
11	Under $30,000	0.24	0.27	0.39	0.99	1,194	1,623	1,961	135
12	Under $30,000	0.42	0.59	0.91	1.81	774	949	912	76
13	$30,000 - under $60,000	0.1	0.14	0.27	0.38	1,093	1,803	3,063	357
14	$30,000 - under $60,000	0.15	0.32	0.56	0.75	700	1,520	1,649	217
15	$60,000 - under $120,000					1,421	2,086	3,507	513
16	$60,000 - under $120,000					1,033	2,148	2,223	384
17	$120,000 - under $250,000					892	1,654	2,538	407
18	$120,000 - under $250,000					1,742	3,665	4,479	912
19	$250,000 - under $500,000					2,129	4,721	5,527	1,367
20	$500,000 - under $1,000,000					1,756	4,081	4,255	1,220
21	$1,000,000 - under $2,000,000					902	2,024	2,119	634
22	$2,000,000 - under $5,000,000					1,449	2,532	2,551	790
23	$5,000,000 - under $10,000,000					523	1,007	1,149	341
24	$10,000,000 - under $20,000,000					392	709	834	247
24	$20,000,000 - under $30,000,000					116	311	164	64
24	$30,000,000 - under $40,000,000					88	298	406	116
24	$40,000,000 - under $50,000,000					0	0	0	0
OVERALL		**0.07**	**0.13**	**0.2**	**0.25**	**5,158**	**10,102**	**11,365**	**2,519**

*For disclosure reason, we only show CVs for a few strata where the population sizes are large.

5.2 Dynamic Analysis - Precision and Cost in Out-Years

We conducted a dynamic analysis to evaluate the refreshment sample planned for the out-years of the new 2007 cohort panel. To assess the precision and cost in each out-year, we applied the sampling rates of the new panel sample to the 1999 data and selected the base-year panel as it would have been selected had the panel design been in place. The design for the new refreshment sample was applied to the out-year data for 2000 through 2006. The cross-sectional estimates were made for each out-year based on using the surviving panel returns as well as the refreshment returns. These returns were also used to assess the precision and cost in each out-year.

First, we simulated the base-year of the panel sample using the 1999 panel sample data. It includes CWHS selections and TTIN selections. For the CWHS selections, the new design includes returns having ten specific last-four digits of the primary SSN and secondary SSN. However, the 1999 sample data included only CWHS returns for the primary taxpayer and included only five of the ten SSN ending digit combinations. To account for the number of CWHS returns in the new panel, we used an approximation. The number of CWHS returns from the 1999 panel is a good approximation of the number of additional returns from the new panel for the additional five SSN ending digit combinations. Also CWHS returns from the 1999 panel that were married filing joint returns were used to approximate the secondary returns that will be included in the new panel. These same returns can be used to approximate the secondary CWHS selections for the additional five SSN ending digit combinations.

For the TTIN selections, we selected the TTIN selections from the 1999 cross-sectional sample[5] using the criterion in Table 2. Then we matched returns in this simulated panel to the returns in the actual 1999 panel data using unique identifiers. The matched returns are in the actual 1999 panel sample and the surviving returns are linked across years in the panel sample across years (1999 – 2006)[6]. There are a small number of returns that are not included in the actual 1999 panel. These returns are from four strata that have a TTIN cut-off larger than the cut-off in the old panel sample. These returns were not included in the simulated panel data. To account for them in the analysis, we adjusted their stratum sample sizes proportionally in out-years. For example, as shown in Table 6, in stratum 1, there are 11 returns that should be included, but were not included in the simulated panel sample. This accounts for 4.49% of the included returns. Then the sample size was adjusted up by 4.49% in stratum 1. The adjusted sample size of the simulated base-year panel is given in Table 7. Also given in Table 7 is the base-year 1999 population size. For the out-years, the simulated panel includes surviving returns that are linked to the base-year simulated panel by either primary SSN or secondary SSN or both. The sample sizes are adjusted using the same adjustment in Table 6. The adjusted simulated sample size of surviving panel returns in out-years is summarized in Table 8. The estimated panel sample sizes in Table 8 are the surviving returns from the simulated 2007 panel sample (simulated based on 1999 cross-sectional sample data) that are still in years 2008 – 2014 (simulated based on 2000 – 2006 cross-sectional sample data). Refreshment returns are not included in Table 8.

[5] The cut-off of the panel sample is smaller than the yearly cut-off of the cross-sectional sample in each stratum.

[6] A panel return stays in the panel in out-years if it is linked to a base-year return by either primary SSN or secondary SSN.

Table 6. Adjustment % for Returns Not Covered in the Simulated Panel

Stratum	Number of returns that should be in the simulated sample, Non-CWHS		Adjustment % = # Not included over # Included
	Included	Not included	
1	245	11	4.49%
2	291	3	1.03%
4	173	62	35.84%
21	2,433	118	4.85%

Table 7. Summary of the Number of Returns in Simulated Panel (Base-Year 1999)

Stratum (a)	Income Range (b)	Population (c)	CWHS (5 endings) (d)	CWHS (5 endings & married) (e)	TTIN Selections (Non-CWHS Returns) (f)	% not Included (g)	Adjusted Sample Size 2d+2e +(1+g)*f
	NEGATIVE INCOME						
1	$150,000,000 and over	16	0	0	16		16
1	$40,000,000, - under $150,000,000	90	0	0	90		90
1	$20,000,000 - under $40,000,000	239	0	0	123		123
1	$10,000,000 - under $20,000,000	535	0	0	245	4.49%	256
2	$5,000,000 - under $10,000,000	1,399	0	0	291	1.03%	294
3	$2,000,000 - under $5,000,000	5,633	4	2	192		204
4	$1,000,000 - under $2,000,000	12,216	6	1	173	35.84%	249
5-9	Under $1,000,000	1,147,293	561	250	0		1,622
	POSITIVE INCOME						
10-19	Under $500,000	125,144,751	62,399	24,320	0		173,438
20	$500,000 - under $1,000,000	435,344	224	188	841		1,665
21	$1,000,000 - under $2,000,000	141,595	82	67	2,433	4.85%	2,849
22	$2,000,000 - under $5,000,000	59,284	29	25	1,425		1,533
23	$5,000,000 - under $10,000,000	14,307	8	8	1,745		1,777
24	$10,000,000 - under $20,000,000	5,391	1	1	1,509		1,513
24	$20,000,000 - under $40,000,000	1,984	2	2	982		990
24	$40,000,000 - under $150,000,000	850	0	0	850		850
24	$150,000,000 and over	86	0	0	86		86
	Total	**126,971,013**	**63,316**	**24,864**	**11,001**		**187,555**

Table 8. Adjusted Panel Sample Sizes of Out-years 2008 – 2014 from Simulations on 2000 – 2006 Cross-Sectional Sample Data (Refreshment Returns Not Included)

Stratum	Income Range	2000 (2008)	2001 (2009)	2002 (2010)	2003 (2011)	2004 (2012)	2005 (2013)	2006 (2014)
	NEGATIVE INCOME							
1	150,000,000 and over	14	14	12	15	15	12	6
1	$40,000,000 - under $150,000,000	84	95	102	95	90	73	68
1	$20,000,000 - under $40,000,000	127	155	163	153	138	113	80
1	$10,000,000 - under $20,000,000	205	232	277	259	202	170	146
2	$5,000,000 - under $10,000,000	235	309	398	379	329	280	225
3	$2,000,000 - under $5,000,000	298	463	656	657	546	459	341
4	$1,000,000 - under $2,000,000	314	560	765	753	652	598	446
5-9	Under $1,000,000	2,204	3,224	4,658	4,812	4,564	3,997	3,316
	POSITIVE INCOME							
10-19	Under $500,000	179,655	181,962	182,444	183,755	185,583	188,357	190,553
20	$500,000 - under $1,000,000	2,244	2,208	2,024	2,114	2,280	2,386	2,354
21	$1,000,000 - under $2,000,000	2,190	2,029	1,749	1,722	1,775	1,749	1,788
22	$2,000,000 - under $5,000,000	2,016	1,710	1,454	1,417	1,560	1,725	1,724
23	$5,000,000 - under $10,000,000	1,272	1,055	885	915	912	1,000	1,049
24	$10,000,000 - under $20,000,000	953	713	525	552	684	730	754
24	$20,000,000 - under $40,000,000	618	416	357	337	423	447	482
24	$40,000,000 - under $150,000,000	527	281	210	262	352	378	339
24	$150,000,000 and over	92	47	38	37	64	81	109
	Overall	**193,048**	**195,473**	**196,717**	**198,234**	**200,169**	**202,555**	**203,780**

**Income is measured in each year rather than in the base year.*

Next, we selected the refreshment sample for each out-year. The TTIN selections are based on the criteria in Table 3. For the CWHS selections, although the new panel sample design included ten SSN ending digit combinations for both the primary and secondary taxpayers, we only counted returns with the original five SSN ending digit combinations for the primary, since the sample data before tax year 2005 only included these returns. Starting from tax year 2005, returns with all ten CWHS SSN ending digit combinations are included in the cross-sectional sample. The secondary CWHS selections have not been included in any yearly cross-sectional samples. Since there are very few returns of these returns in the high-income strata, ignoring returns with them should not have much impact on the dynamic analysis. The number of returns selected for the refreshment sample by tax year and stratum is given in Table 9.

Finally, we pooled the panel data and the refreshment data together and estimate precision measures for each year and each key variable. Technically, the weights of refreshment returns are different from the weights of panel returns. However, for the design purpose, we simply treat all the returns within the same stratum with equal weight. To avoid the possible large influence of outliers, two largest returns and one smallest return in each of strata 1 – 24 are not included in calculating the stratum variance of each variable. Table 10 gives the overall coefficient of variation (CV) for each tax year and each key variable. These two tables show that if the refreshment panel had been in place for the 1999 panel, it would have provided good cross sectional estimates. The overall CV's for AGI, taxes, capital gains and losses, partnership gains, total and charitable itemized deductions are 1% or less in all years through 2006. The individual cells by income class are almost all under 10% (most are under 5%) with the exception of some negative income class cells where one would expect a large variance because such returns can bounce between positive and negative income over time.

Table 9. Number of Refreshment Returns – Random Selection and CWHSI=1 (Not overlap with the panel sample)

Stratum	Income Range	2000	2001	2002	2003	2004	2005	2006
	NEGATIVE INCOME							
1	$150,000,000 and over	7	17	18	12	12	21	24
1	$40,000,000 - under $150,000,000	8	10	14	13	17	8	17
1	$20,000,000 - under $40,000,000	3	3	0	1	1	3	3
1	$10,000,000 - under $20,000,000	6	6	4	6	8	7	6
2	$5,000,000 - under $10,000,000	11	9	15	17	15	17	24
3	$2,000,000 - under $5,000,000	12	20	26	31	34	46	45
4	$1,000,000 - under $2,000,000	0	1	0	0	0	0	0
5 - 9	Under $1,000,000	71	121	167	187	200	209	193
	NEGATIVE INCOME							
10 - 19	Under $500,000	5,869	8,458	10,458	12,355	14,320	16,505	18,670
20	$500,000 - under $1,000,000	2	3	4	4	9	15	10
21	$1,000,000 - under $2,000,000	0	0	1	2	1	2	2
22	$2,000,000 - under $5,000,000	25	22	20	23	32	59	55
23	$5,000,000 - under $10,000,000	41	23	20	32	56	67	75
24	$10,000,000 - under $20,000,000	27	12	9	9	35	57	73
24	$20,000,000 - under $40,000,000	31	19	12	20	42	64	77
24	$40,000,000 - under $150,000,000	85	41	31	39	69	119	139
24	$150,000,000 and over	52	25	13	22	33	57	77

Table 10. Estimated Overall Coefficient of Variation

	Tax Year (%)							
Variable	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Selection Income	0.08	0.07	0.07	0.08	0.08	0.07	0.07	0.07
AGI	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
Wages	0.20	0.19	0.19	0.19	0.19	0.19	0.20	0.20
Taxes	0.23	0.23	0.23	0.25	0.26	0.26	0.26	0.25
Positive Capital Gain	0.78	0.71	0.99	1.15	1.07	0.90	0.81	0.73
Negative Capital Gain	-1.78	-1.15	-0.84	-0.70	-0.73	-0.79	-0.87	-0.95
Partnership Capital Gain	1.77	1.70	2.09	2.27	1.98	1.60	1.55	1.43
Pos Business Income	0.93	0.89	0.88	0.87	0.87	0.86	0.86	0.88
Neg Business Income	-1.80	-1.72	-1.55	-1.49	-1.44	-1.45	-1.46	-1.47
Total itemized	0.30	0.44	0.34	0.32	0.32	0.31	0.33	0.30
Charitable deduct	0.76	0.78	0.75	0.72	0.73	0.74	0.85	0.79

6. Conclusions

The Statistics of Income Division of the Internal Revenue Service is starting in new cohort panel in 2007 as part of a long-run plan of periodically beginning new cohort panels as the prior ones become dated. Challenges in designing a new panel included increasing the size of the random CHWS sample and expanding it to include secondary taxpayers, choosing a new high-income sample, making the panel more useful for later cross-section analysis, and providing an improved refreshment sample while at the same time reducing costs of the new panel. This paper explains the sample procedures adopted for the new panel and presents simulations of how well the new procedures would work using the prior 1999 panel and other available data. The results of the simulations suggest that the new refreshment panel is likely to provide good representations of cross-sections as well as being useful for panel analysis.

7. References

Bryant, V (2008), “Attrition in the Tax Years 1999 - 2005 Individual Income Tax Return Panel,” *Proceedings of the Survey Methodology Section, American Statistical Association, 2008.*

Mathematica Policy Research (2006), “Final Weighting of the Edited Panel, Years One through Five,” *Internal Memo*.

Testa, V. and Scali, J (2005), “Description of the Sample,” *Statistics of Income – 2005, Individual Income Tax Returns, Internal Revenue Service, Washington, DC.*

Weber, M. (2001), “The Statistics of Income 1979-2002 Continuous Work History Sample Individual Income Tax Return Panel,” *Proceedings of the Survey Methodology Section, American Statistical Association, 2001.*