The 1999 Individual Income Tax Return Edited Panel

Michael E. Weber and Victoria L. Bryant, Internal Revenue Service

The primary product of the Statistics of Income Division's Individual Statistics Branch is an annual cross-sectional sample of individual income tax returns. Some form of this annual cross section, also known as the Individual Complete Report File, has been produced every year since 1916. These annual cross sections provide the basis for most Federal tax policy analysis and research as they are consistently and reliably produced with well-known statistical properties. Longitudinal or panel samples of individual income tax returns, however, have a much shorter history. This has been largely due to their statistical and operational complexity relative to cross-sectional samples, and the added cost of producing panels given limited budgets. SOI produced a few small panels in the mid-to-late seventies and the early eighties, but all of these panels were focused on capital gains and losses. They were not meant to provide longitudinal information on other types of income, deductions, or credits. Beginning with Tax Year 1979, SOI incorporated a few Continuous Work History Sample (CWHS) Social Security Number (SSN) endings as part of the annual Individual Income Tax Return Cross Sectional Sample. These CWHS cross-sectional samples can be used to form a panel as the name implies and have been used for tax policy analysis by researchers both inside and outside the Government.[1] But, while the SOI CWHS has many wonderful longitudinal aspects, it lacks the ability to provide statistically reliable data for high-income taxpayers. For example, in 1999, taxpayers reporting over $1,000,000 in Adjusted Gross Income (AGI) accounted for 11 percent of all reported AGI and 20 percent of all income taxes. In the annual cross-section file, which utilizes a highly stratified sample design based on income, there were 53,587 returns with $1,000,000 or more in AGI but only 123 CWHS returns, a statistically inadequate sample for tax policy analysis.[2]

The first panel that attempted to use a stratified sample design that adequately sampled high-income returns and also represented the underlying annual cross-section or Complete Report File was the 1987-based Family Panel. This panel followed all of the primary and secondary taxpayers shown on nondependent tax returns found in the 1987 Complete Report. The panel continued until 1996.

▶ Why the 1987 Family Panel was Terminated

Financial considerations were paramount in the decision to end the panel in 1996. As noted above, the 1987 Family Panel was drawn from the nondependent returns found in the 1987 Complete Report File. So, initially, the Complete Report and the Family Panel samples overlapped. However, since there is great volatility in the reported incomes of taxpayers in the upper income strata, many taxpayers sampled for SOI's Complete Report File at rates of 100 percent in a given year fall into strata with sampling rates of 25 percent or even 10 percent in subsequent years. These original 100-percent strata returns, once selected for the panel, must be processed in subsequent years even though they are not needed for the annual cross-sectional sampling. In addition, in 1991 the Treasury Department's Office of Tax Analysis (OTA) and SOI jointly redesigned the annual cross-sectional sample and thereby shifted the entire underlying sample structure, further reducing the overlap of the two samples. As can be seen from Table 1, in 1988, some 56 percent of the returns sampled for the Complete Report were also used in the 1987 Family Panel. By 1993, that percentage had dropped to 33 percent. If dependent returns, which are usually simple returns, are removed, the comparable figures are 71 percent and 39 percent, respectively (Table 2). If only returns selected for the panel with a 100-percent probability of selection are examined, the comparable figures are 62 percent and 28 percent, respectively (Table 3). This diminishing overlap in the high-income returns is, therefore, very problematic from a cost perspective. In terms of manual processing time, returns in the various 100-percent strata take over 26 minutes on average to process, almost 5 times the amount of time it takes to process returns with AGI under $100,000. During preparations for processing Tax

Year 1997 returns, it became apparent that, due to the diminishing overlap, SOI would not have enough funds available to complete the processing of both the 1987 Family Panel and the 1997 Complete Report File.

A second reason for ending the 1987 Family Panel was its age. The longer any panel continues, the less its usefulness for the analysis of current issues. For example, assume the 1987 Panel had continued through 2005 and an analysis was performed on the Bush 2001

Table 1.--Overlap between the 1987 Family Panel and the 1987-1993 Complete Reports (CR)

SOIYR	87 Panel	CR	Both	Panel Overlap with CR
1987	86,975	125,788	86,907	99.9%
1988	116,342	110,495	65,385	56.2%
1989	120,803	110,566	59,077	48.9%
1990	124,087	104,277	55,791	45.0%
1991	123,295	125,756	49,494	40.1%
1992	125,228	103,190	45,479	36.3%
1993	132,583	104,357	44,283	33.4%

Table 2.--Overlap between the 1987 Family Panel (nondependent returns) and the 1987-1993 Complete Reports (nondependent returns)

SOIYR	87 Panel	CR	Both	Panel Overlap with CR
1987	86,950	120,520	86,883	99.9%
1988	92,363	106,876	65,109	70.5%
1989	97,207	106,836	58,882	60.6%
1990	101,839	101,512	55,650	54.6%
1991	104,154	123,094	49,385	47.4%
1992	107,917	100,589	45,388	42.1%
1993	112,951	101,779	44,221	39.2%

Table 3.--1987 Panel Returns sampled at 100 percent rate and overlap with SOI cross-section*

1987 100% panel rate = 12,411

SOIYR	Both	Panel overlap with CR
1987	12,411	100%
1988	7,642	62%
1989	6,301	51%
1990	5,480	44%
1991	4,096	33%
1992	3,571	29%
1993	3,422	28%

* Obtained by matching the 1987 panel 100 percent sample returns in each year with the 100 percent returns in the CR for each year. This is an overestimate as the number of 100 percent records in the panel grows each year due to divorce and dependents filing their own return.

Tax Cuts. The results would not have provided an analysis of how American taxpayers of year 2000 responded to the tax cuts over the next 5 years. It would have provided an analysis of how individual taxpayers who filed a return in the panel base year of 1987 responded to the 2001 tax cuts. Those populations of taxpayers almost certainly were very different. This is not to say that long-lived panels are useless; indeed, long-lived panels are highly valued by researchers, but, as they age, the nature of the analysis that can be performed upon them changes. Given limited resources, there is a tradeoff between the longevity of a panel and the age of its underlying base year data. As any panel ages, it loses its ability to speak to the issues of the current day. Most researchers and analysts find that the most pressing issues, usually defined by their job requirements, are those of the current day.

Thus, given the resource concerns and the age of the panel, a decision was made jointly between SOI and OTA to end the 1987 panel after processing of the 1996 data was complete.

▶ The 1999 Edited Panel--The Beginning

The planning process for the next panel began in the fall of 1997. Consultants from Westat were contracted to moderate the process and to provide statistical guidance and sample design recommendations. Over the next year, Westat met extensively with members of SOI and also moderated several meetings between members of SOI and individuals from OTA.[3] The wide-ranging discussions covered such topics as greater utilization of the CWHS concept to completely integrating the cross-section and panel studies into one sample.[4] In January 1999, Westat produced a report entitled "Issues in the Design of a New Panel of Individual Tax Returns" which provided the basic contours of the sample design for the Tax Year 1999 Edited Panel that was put into operation in May 2001.[5]

▶ Basics of the Individual Cross-Section Sample

Before discussing the specifics of the Edited Panel sample design, the basics of the Complete Report sample design should be discussed. Table 4 shows the final weighting stratifications for the 1999 Complete Report. The stratifications are based on a tabulated income amount, which is indexed to the GDP each year, and the inclusion of various IRS forms and schedules. For certain income strata, a few additional substrata are created based on a "Degree of Interest" variable. This variable is derived from various components on the tax return such as filing status and the number of dependents.[6] Prior to the planning and implantation of the 1999 Edited Panel, the prescribed sampling rates ranged from a low of 1 to a high of approximately 1-in-5,000. When ranking the cost of processing returns for the SOI program by stratification, the lower income stratifications (which are dominated by CWHS returns) are the cheapest to process, and the 100-percent stratifications are the most expensive. [7]

▶ The 1999 Edited Panel Sample Design

One of the key Westat panel design recommendations, and one that was readily accepted and implemented, was that the 1999 Edited Panel should make greater use of the CWHS concept and thus contain a larger sample of CWHS returns. This would produce many analytical benefits but would also help SOI to maintain a more constant cost structure over time since CWHS returns could be readily used in the annual cross-sectional file as well as in the 1999 Edited Panel. Consequently, the SOI Complete Report sample design was changed to include five CWHS endings.[8] Table 5 shows the various Complete Report strata for 1997 and 1999, as well as the percentage of returns found in each stratum that were selected due to their membership in the SOI CWHS sample. As can be seen, some strata now consist entirely of CWHS returns. Indeed, if the "Degree of Interest" stratifications, which require a larger sample size than that generated by five CWHS endings, were eliminated, the CWHS sample would provide all returns required for the Complete Report for returns showing $120,000 or less of positive income and about one third of the required sample for returns between $120,001 and $250,000. In fact, it was decided that the "Degree of Interest" stratifications were not needed for the panel and that a roughly 33-percent subsample of the returns between $120,000 and $250,000 of positive income would be adequate as well. Thus, the CWHS sample accounts for all sampled records in the panel with

Table 4.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 1999

Description of the sample strata	Degree of interest [1]	Number of Returns by type of form attached: Form 1040, with Form 1116 or Form 2555 — Population counts	Sample counts	Sampling Rate	Form 1040, with Schedule C but without Form 1116 or Form 2555 — Population counts	Sample counts	Sampling Rate	Form 1040, with Schedule F but without Schedule C, Form 1116 or Form 2555 — Population counts	Sample counts	Sampling Rate	All other forms — Population counts	Sample counts	Sampling Rate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Total		2,698,596	36,528		17,272,967	36,746		1,521,415	4,470		105,825,250	95,824	
Indexed Negative Income [2]													
$10,000,000 or more	All	101	101	100.00	504	504	100.00	65	65	100.00	586	586	100.00
$5,000,000 under $10,000,000	All	86	86	100.00	609	609	100.00	121	121	100.00	750	750	100.00
$2,000,000 under $5,000,000	All	346	103	29.77	2,349	741	31.55	533	190	35.65	2,673	862	32.25
$1,000,000 under $2,000,000	All	703	100	14.22	5,188	818	15.77	1,312	214	16.31	5,192	847	16.31
$500,000 under $1,000,000	All	1,472	54	3.67	14,089	498	3.53	3,990	123	3.08	12,007	401	3.34
$250,000 under $500,000	All	3,007	35	1.16	34,810	310	0.89	9,768	78	0.80	27,489	258	0.94
$120,000 under $250,000	All	5,467	34	0.62	75,090	352	0.47	17,257	89	0.52	58,046	267	0.46
$60,000 under $120,000	All	**	**	**	117,062	292	0.25	17,810	36	0.20	87,367	224	0.26
Under $60,000	All	**	**	**	321,426	425	0.13	33,741	52	0.15	327,804	446	0.14
Indexed Positive Income [2]													
Under $30,000	1										27,809,524	13,804	0.05
Under $30,000	2	143,649	65	0.05	1,874,895	973	0.05	108,513	62	0.06	29,242,683	14,749	0.05
Under $30,000	3-4	199,772	223	0.11	3,464,052	3,586	0.10	172,357	188	0.11	6,205,425	6,492	0.10
$30,000 under $60,000	1-2	198,137	101	0.05	1,686,282	787	0.05	184,402	83	0.05	20,613,240	10,179	0.05
$30,000 under $60,000	3-4	314,375	373	0.12	3,351,363	3,562	0.11	281,068	299	0.11	5,618,229	6,224	0.11
$60,000 under $120,000	1-3	408,896	191	0.05	1,874,804	959	0.05	232,413	120	0.05	10,025,047	4,905	0.05
$60,000 under $120,000	4	350,365	355	0.10	2,274,376	2,361	0.10	190,886	161	0.08	2,374,629	2,408	0.10
$120,000 under $250,000	1-3	243,101	367	0.15	466,388	680	0.15	106,656	139	0.13	1,584,226	2,346	0.15
$120,000 under $250,000	4	328,531	958	0.29	1,085,930	3,115	0.29	76,074	198	0.26	1,017,036	2,910	0.29
$250,000 under $500,000	All	277,335	1,849	0.67	454,376	3,100	0.68	61,525	371	0.60	567,361	3,727	0.66
$500,000 under $1,000,000	All	128,630	3,105	2.41	125,068	2,979	2.38	16,675	404	2.42	166,746	4,029	2.42
$1,000,000 under $2,000,000	All	54,290	6,581	12.12	31,129	3,767	12.10	4,280	542	12.66	52,437	6,447	12.29
$2,000,000 under $5,000,000	All	27,424	8,938	32.59	10,170	3,321	32.65	1,532	498	32.51	20,333	6,545	32.19
$5,000,000 under $10,000,000	All	7,813	7,813	100.00	2,015	2,015	100.00	302	302	100.00	4,273	4,273	100.00
$10,000,000 or more	All	5,096	5,096	100.00	992	992	100.00	135	135	100.00	2,145	2,145	100.00

1 Each population member is assigned a degree of interest based on how useful it is for tax modeling purposes. Degree of interest ranges from one (1) to four (4), with a one being assigned to returns that are the least interesting, and a four being assigned to those that are the most interesting. 'All' refers to income classes for which returns with all four degrees of interest are assigned.

2 Positive and Negative Income classes are divided by a Chain-Type Price Index for the Gross Domestic Product of 1.1480 to represent a base year of 1991.

** Sampling Strata Collapsed.

Table 5.—CWHS Selection as Percentage of Cross-sectional Sample Stratifications, 1997 and 1999 SOI Samples

		Stratification by type of form attached							
		Form 1040, with Form 1116 or Form 2555		Form 1040, with Schedule C but without Form 1116 or Form 2555		Form 1040, with Schedule F but without Schedule C, Form 1116 or Form 2555		All other forms	
Description of the sample strata	Degree of interest [3]	1997 CWHS %	1999 CWHS %	1997 CWHS %	1999 CWHS %	1997 CWHS %	1999 CWHS %	1997 CWHS %	1999 CWHS %
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Indexed Negative Income [4]									
$10,000,000 or more	All								
$5,000,000 under $10,000,000	All								
$2,000,000 under $5,000,000	All		0.97%		0.13%				
$1,000,000 under $2,000,000	All	1.41%			0.24%		0.93%		
$500,000 under $1,000,000	All		1.85%	0.51%	1.00%	0.88%		1.67%	2.24%
$250,000 under $500,000	All		11.43%	4.35%	5.16%	2.25%	6.41%	4.95%	6.20%
$120,000 under $250,000	All		14.71%	3.70%	11.93%	4.29%	5.62%	5.77%	12.36%
$60,000 under $120,000	All	**	**	7.84%	20.21%	5.77%	11.11%	8.76%	18.30%
Under $60,000	All	**	**	24.47%	35.14%		25.00%	19.52%	32.81%
Indexed Positive Income [4]									
Under $30,000	1							90.93%	100.00%
Under $30,000	2	0%	100.00%	61.42%	100.00%	66.67%	100.00%	61.96%	100.00%
Under $30,000	3-4	24.14%	53.36%	23.70%	47.52%	23.35%	51.06%	24.73%	48.54%
$30,000 under $60,000	1-2	56.76%	100.00%	62.00%	100.00%	59.72%	100.00%	61.79%	100.00%
$30,000 under $60,000	3-4	20.59%	46.38%	21.81%	46.50%	20.39%	39.46%	22.96%	45.76%
$60,000 under $120,000	1-3	54.08%	100.00%	55.87%	100.00%	52.05%	100.00%	57.05%	100.00%
$60,000 under $120,000	4	19.92%	50.70%	19.49%	49.98%	21.88%	50.93%	20.51%	50.00%
$120,000 under $250,000	1-3	12.56%	33.79%	16.12%	33.97%	14.09%	28.78%	14.89%	34.65%
$120,000 under $250,000	4	6.84%	18.16%	7.04%	16.18%	6.71%	16.67%	7.73%	17.05%
$250,000 under $500,000	All	3.84%	7.95%	2.67%	8.10%	2.30%	7.01%	3.09%	8.48%
$500,000 under $1,000,000	All	0.93%	2.19%	0.76%	2.32%	1.76%	1.98%	0.76%	1.99%
$1,000,000 under $2,000,000	All	0.23%	0.43%	0.10%	0.61%	0.39%	0.74%	0.26%	0.37%
$2,000,000 under $5,000,000	All	0.05%	0.13%	0.08%	0.18%	0.00%	0.20%	0.09%	0.15%
$5,000,000 under $10,000,000	All	0.04%	0.05%	0%	0.05%	0.00%	0.33%	0.04%	0.07%
$10,000,000 or more	All	0%	0.04%	0%	0.10%	0.00%	0%	0%	0.00%

positive income up to $250,000. It was also determined that the additional stratifications by form type would not be needed either. Consequently, the lowest sampling rate in each income strata sampling group (determined by the type of forms and schedules attached to the return) became the maximum sampling rate for that income stratum.

Another recommendation of the Westat consultant's was to design a targeted high-income cohort. The 1987 Family Panel design essentially selected all 1987 cross-section high-income returns for inclusion in the panel, and, in the end, the costs associated with that decision forced the termination of the panel after 10 years. As a general rule, the larger the selection probability, the more expensive the return is to process; therefore, decisions about sample size for high-income returns, particularly those with over $2,000,000 of positive income, are crucial in determining project costs. A smaller high-income sample would create the possibility of a longer lived panel and/or the possibility of multiple high-income waves starting perhaps every 5 years. The first step in subsampling high-income returns was to determine how much if any of the 100-percent stratum should be subsampled. A Westat report confirmed OTA's initial opinion that returns above $20,000,000 of positive in-

come should not be subsampled but rather included in the panel at 100 percent.[9] Consequently, returns below $20,000,000 and above $250,000 would be subjected to subsampling. To that end, analysts from Westat, in conjunction with SOI and OTA, analyzed over 30 potential subsampling schemes using a linked version (or panel) of the 1996 and 1997 Complete Report files.[10,11] This intensive process required Westat to evaluate each scheme in terms of coefficients of variation (CV) for various items in 1996 and also to compute the CVs for the differences in totals for the various items between 1996 and 1997. To quote from the report: "The primary goal was to select a panel that had acceptably low CV's for cross-sectional estimates and estimates of change…In addition, a secondary consideration was how the distribution of the sample among income classes would change over time ..(as).. one of OTA's desires was to avoid allocations that would become too thin at the tails of the income distribution as incomes changed over time." As various designs were discarded, others were refined, and, in the end, Design 16A was chosen. (See Table 6)

▶ The Issue of Late Filed Returns

A subtlety of the annual cross-section must be addressed at this point: Not all Tax Year 1999 returns are filed by the end of Calendar Year 2000. A significant portion of Tax Year 1999 returns were filed in Calendar Years 2001 and 2002. Keeping the sample open for an additional 2 years in order to obtain these returns would force policymakers to use outdated data for decision-making. For instance, sampling for the Tax Year 1999 file would not be complete until as late as December 31, 2002. Therefore, in order to provide more timely statistics, SOI produces a sample of tax returns filed during each calendar year. Approximately 97 percent of the returns received in a given calendar year are for the preceding tax year. For example, in Calendar Year 2000, some 97 percent of taxpayers filed their Tax Year 1999 returns. The remaining 3 percent of the returns filed in a given calendar year are generally for the preceding 2 tax years. In our example, these would be Tax Years 1997 and 1998. These "prior year" returns are used as proxies for the Tax Year 1999 returns that were not filed timely during Calendar Year 2000.

When creating panels, however, we have the luxury of time and are thus able to create a sample from a virtually complete set of returns for a given tax year. The Tax Year 1999 Edited Panel is a sample of Tax Year 1999 returns. Since each calendar year was sampled independently, it would be appropriate, when combining all 3 years of Tax Year 1999 sampling, to treat each year as a separate level of stratification. But as can be seen from Table 6, the sample sizes for most of the stratifications for Calendar Years 2001 and 2002 are rather small. This would cause a proliferation of weights. Consequently, a decision was made not to stratify on Tax Year but to treat the 3 years as one sample with one set of stratifications and thereby reduce the variability in the weights.

▶ Linking Individuals and Tax Returns Over Time

In order to link tax returns and individuals over time, a unique identifier is required. Fortunately, taxpayers are required to provide their Social Security numbers on their tax forms. However, sometimes the SSN's that are shown on the tax forms are incorrect, and, sometimes IRS transcribes them incorrectly. So, in order to prevent billionaires and millionaires from either disappearing or being linked to Earned Income Tax Credit recipients, SOI performs a review of panel member SSN's. The 1999 Edited Panel contains 125,108 unique panel member SSN's. This is simply the number of base year returns in the sample plus the number of spouses on joint returns. Of the 125,108 panel members, only 456 SSN's (44 for the primary taxpayers and 412 secondary taxpayers) were determined to be incorrect. For 392, a correction was obtained. A total of 29 returns were deleted because the primary SSN's on these nonjoint returns were determined to be incorrect and no correction could be obtained. Note that this is not a confirmation that the remaining SSN's are correct. Frequently, invalid SSN's are not detectable for many years until some point in the future, often when multiple individuals use the same SSN. In addition, many corrections are made to nonpanel member individuals who accidentally, or perhaps intentionally, use an SSN that does not belong to them and thus cause an incorrect linkage to a panel member. While these figures paint a positive picture for the quality of the SSN linkages, one area of concern is

Table 6.—1999 Complete Report and 1999 Edited Panel Sampling Rates, Tax Year 1999 Population and Sample Counts by Calender Year

Description of the sample strata	Complete Report Sampling Rate [1]	Edited Panel Sampling Rate	Estimated Population and Tax Year 1999 Edited Panel Sample Counts							
			Calendar Year 2000		Calendar Year 2001		Calendar Year 2002		Tax Year 1999	
			Population	Sample	Population	Sample	Population	Sample	Population	Sample
Indexed Negative Income										
$20,000,000 or more	100.00	100.00	329	329	14	14	5	5	348	348
$10,000,000 under $20,000,000	100.00	48.47	498	232	29	11	3	3	530	246
$5,000,000 under $10,000,000	100.00	22.05	1,276	267	98	19	7	7	1,381	293
$2,000,000 under $5,000,000	29.77	4.20	5,140	212	396	28	19	6	5,555	246
$1,000,000 under $2,000,000	14.22	1.42	11,149	164	875	13	13	2	12,037	179
$500,000 under $1,000,000	3.08	0.58	27,742	176	2,409	12	116	4	30,267	192
$250,000 under $500,000	0.80	0.12	67,633	93	4,564	8	443	2	72,640	103
$120,000 under $250,000	0.46	0.05	146,165	84	9,646	4	358	1	156,169	89
$60,000 under $120,000	0.20	0.05	199,848	90	11,885	9	5,857	8	217,590	107
Under $60,000	0.13	0.05	617,324	282	35,784	21	227,466	113	880,574	416
Indexed Positive Income										
Under $30,000	0.05	0.05	67,044,058	33,563	1,228,584	625	334,912	196	68,607,554	34,384
$30,000 under $60,000	0.05	0.05	31,733,460	15,663	400,357	201	47,455	41	32,181,272	15,905
$60,000 under $120,000	0.05	0.05	17,505,122	8,694	241,313	129	23,691	25	17,770,126	8,848
$120,000 under $250,000	0.13	0.05	4,832,584	2,378	88,301	36	1,745	5	4,922,630	2,419
$250,000 under $500,000	0.66	0.18	1,333,893	2,410	29,466	46	749	5	1,364,108	2,461
$500,000 under $1,000,000	2.38	0.59	427,468	2,521	6,915	36	206	5	434,589	2,562
$1,000,000 under $2,000,000	12.10	1.72	138,498	2,449	2,684	58	77	10	141,259	2,517
$2,000,000 under $5,000,000	32.19	5.73	58,147	3,369	1,009	54	26	9	59,182	3,432
$5,000,000 under $10,000,000	100.00	18.88	14,037	2,680	245	43	8	8	14,290	2,731
$10,000,000 under $20,000,000	100.00	57.62	5,291	2,994	91	52	7	7	5,389	3,053
$20,000,000 or more	100.00	100.00	2,876	2,876	45	45	3	3	2,924	2,924
Total			124,172,538	81,526	2,064,710	1,464	643,163	465	126,880,411	83,455

1 - Lowest sampling rate found within collapsed strata

with the use of IRS-generated Taxpayer Identification Numbers or ITIN's which are provided to individuals who are required to file a return but who have not been issued an SSN. Quite often, these individuals will, in time, obtain an SSN from the Social Security Administration and then file using it in subsequent years. This breaks the link to the previous set of returns and, if not caught prior to sampling, will cause the loss of valid sample units.

▶ Future Plans

The 1999 Individual Income Tax Return Panel is currently being weighted and will include data from 1999 through 2003. Subsequent years of data will be appended to the panel as they become available. Our attention now turns to learning how to use the panel and the publication of tabulations and analysis, hopefully the subject of many future papers.

▶ Endnotes

[1] For more information on the CWHS panel, see Weber, Michael (2004), "The Statistics of Income 1979-2002 Continuous Work History Sample Individual Income Tax Return Panel," 2004 Prceedings of the American Statistical Association, Social Statistics Section.

[2] For example, the estimated amount of AGI, using the full sample of returns with a reported AGI of $1,000,000 or more, was $653,184,370,292. The coefficient of variation for this amount is .19. Using the 123 CWHS returns and applying a weight of 2,000 (5 different endings were used in 1999, thus producing a 1-in-2000 sampling rate) produced an estimate of $696,643,752,000. The specific coefficient of variation for this amount has not been calculated, but can be assumed to be significantly larger than .19.

[3] Notes from these meetings are found in an unpublished Westat document entitled "Meeting Minutes For Task Order #13 Under Contract No. TIRNO-96-D-00030.0005."

[4] More information on this topic is found in an unpublished Westat document entitled "Integrated versus Separate Panel and Cross-Sectional Sample Designs," September 1999.

[5] Tax Year 1999 returns were generally filed in Calendar Year 2000. As the Tax Year 1999 Based Edited Panel was defined as a subsample of the 1999 Complete Report File, panel membership did not need to be defined for sampling purposes until Tax Year 2000 returns, which were generally filed in Calendar Year 2001, were received by IRS and ready for SOI sampling in May 2001. As is often the case, final sample decisions were not finalized until the last possible moment.

[6] For additional information on the sample design of the annual Complete Report sample, see Internal Revenue Service, Statistics of Income--Individual Income Tax Returns, Publication 1304, 1999, "Section 2: Description of Sample."

[7] It should be noted that SOI processes many CWHS returns without any manual processing costs.

[8] This change was actually instituted for Tax Year 1998. The sample design for Tax Year 1999 is identical to Tax Year 1998. Consequently, a table showing the Tax Year 1998 stratifications has been omitted.

[9] Westat unpublished memo, "Report on Substrata for Strata 1 and 24," October 9, 2000.

[10] Unpublished Westat report "Design of a Panel Sample of Tax Returns--Final Report," May 2001.

[11] The 1997 file was augmented by data from the IRS Individual Returns Transaction File when a 1996 Complete Report SSN did not appear in the 1997 Complete Report.