The Statistics of Income 1979-2002 Continuous Work History Sample Individual Income Tax Return Panel

Michael Weber, Internal Revenue Service

Since 1979, the annual SOI Individual Income Tax Return Cross-Sectional Sample has had at least one Continuous Work History Sample (CWHS) Social Security Number (SSN) ending embedded in the sampling framework. The CWHS utilizes a feature of the SSN numbering system where the last four digits of the number have the properties of a random number. Thus, by sampling on the last four digits, a random sample can be obtained.[1] The CWHS sample was embedded in the sample in order to create an occupational match study with the Social Security Administration. It was envisioned that the study would be repeated, and, eventually, longitudinal occupation data could be produced.[2] The match study never came to fruition, but the CWHS sample remained embedded in the annual SOI cross-sectional sample. Limited use was made of the longitudinal aspects of the CWHS portion of the SOI sample until the mid-1980's when a tax return panel was created. This panel, which began with 1979 data, was then made into a public-use file. Data for the public-use file was released through Tax Year 1990. SOI stopped the public release of data beyond 1990 due to disclosure concerns. Subsequently, SOI turned its attention to the creation of stratified panels: a 1985-2001 Sales of Capital Assets Panel, a 1987-1996-based Family Panel, and two ongoing 1999 based panels. The 1979-1990 CWHS panel was never updated until now.

▶ The Cross-Sectional Problem

Before turning to the 1979-2002 CWHS Panel, one cross-sectional issue needs to be addressed. Table 1 shows the weighted total return counts for the SOI Individual Income Tax Return Cross-Sectional Sample and the CWHS subsample for the SOI years 1979-2002.[3] An interesting feature of this table is that the CWHS cross-sectional totals are always less than the full SOI cross-sectional totals. One would expect some random differences between two samples with the CWHS cross-section sometimes estimating more and sometimes fewer tax returns than the full SOI cross-section. But one would not expect the CWHS cross-section total to always be less than the full SOI cross-section. However, since the CWHS sample involves the same SSN's each year, and since the SOI sample is based on a transformed SSN, both of these samples in a given year have a high degree of overlap with the samples in all other years. Thus, if there were some systemic error involved with the CWHS sample as compared to the full SOI sample, one would expect that same error and the sign of the error to occur from year to year. The question then becomes what is the source of this consistent shortfall in the CWHS. One source of this shortfall may reside in the IRS issuance of Individual Taxpayer Identification Numbers or ITIN's to individuals who do not have SSN's. The IRS system of issuance for ITIN's may not produce numbers where the last four digits are random numbers. While this is a possible explanation, the issue merits further research. Interestingly, the difference between the CWHS and the full SOI cross-section significantly diminished as the CWHS sample was expanded to five endings for SOI Year 1998.

▶ The Use of the Primary SSN as the Unique Panel Identifier

Panel files require the use of a unique control number to identify the temporal observations associated with the unit that is being followed. Since taxpayers are required to provide their SSN's on their tax returns, the SSN is a good candidate for this unique person control number. In IRS computer systems, the use of the primary taxpayer SSN as a unique individual identifier is generally very accurate. This is due to fact that IRS returns processing rules do not permit duplicate primary SSN's to be "posted" or moved onto the IRS Individual Master File after the initial tax return transcription process has been completed. Thus, for a given tax year, there is one tax return for each primary SSN and one primary SSN for each tax return. No system, however, is error-free, and duplicate primary SSN's do slip in.[4] But for the entire 24 years of the panel, there were, approximately, only

700 cases or less than 0.1 percent of the sample where a deletion was required due to multiple returns using the same primary SSN for the same tax year.

▶ Eliminating Tax Returns Incorrectly Linked to an SSN

If one assumes that any taxpayer or IRS transcription errors found with primary SSN's are random, then each tax return found in the SOI CWHS sample is a valid sample record for cross-sectional purposes. Some returns in the sample should not be there, but a like number of returns that should be there are not. Longitudinally, however, sampled returns bearing the same primary SSN are useful only if they actually represent the same individual. Mistakes, intentional and unintentional, do occur in the use of SSN's as unique personal identifiers on tax returns. In a longitudinal, sample, this situation must be corrected. The question then becomes how to identify these situations. The most easily identifiable situation is where multiple returns show the same primary SSN for the same tax year. Fortunately, as mentioned earlier, this problem accounts for only a very small portion of the sample.

The next step is to identify and separate the true owner of the SSN from the incorrect user(s) of that SSN. Fortunately, SOI has a few tools at its disposal for this purpose. First, taxpayers are required to list their full names on the tax return; thus, a simple comparison of taxpayer names solves many problems. Unfortunately, SOI did not retain the full name listed on the tax return until 1988, and then only for special studies. For the CWHS panel, the full names for all members of the panel exist only for returns filed for SOI Year 1999 and later. What has been retained for all years is the IRS-generated name control, which is derived from the full name listed on the return. A name control is the first four digits of an individual's last name.

Second, IRS has access to an extract of the SSA's Numident file, which contains information on all of the name controls legally used with a given SSN. This file also contains a date of birth, gender, and, if applicable, a date of death.[5] As a general rule, for this paper, a taxpayer incorrectly uses an SSN when the name control listed for that SSN by SSA does not correspond to the name shown on the tax return, while a taxpayer correctly uses an SSN when the name control listed for that SSN by SSA does correspond to that shown on the return. In most SSN multiple-use cases, the taxpayer who incorrectly uses the SSN is readily identifiable. For example, for a year where two returns were filed using the same primary SSN, one return will have a name control that does not correspond to any of the valid SSA name controls, while the name control listed on the other return does correspond with a valid SSA name control.

Once the duplicate return situation has been resolved for the particular tax year in question, the rest of the returns for the remaining tax years in that SSN sequence need to be checked because a taxpayer may have filed using an incorrect SSN for years without causing a multiple return problem. This is most likely to occur due to one of three situations. The first situation occurs when the age associated with the SSN is under 21. If a taxpayer incorrectly uses a given SSN for a number of years, and then the true owner of the SSN enters the workforce after high school or college and begins to file returns as a primary taxpayer, multiple returns appear. The returns in the sequence filed prior to the first filing by the true owner must be removed. This can also happen in reverse when a taxpayer retires and perhaps is no longer required to file a tax return. A third situation occurs when a single woman files as an unmarried person and thus reports her SSN in the primary position and then marries and files returns as the secondary taxpayer. While she is married, another taxpayer incorrectly uses her SSN. If the woman subsequently divorces and again files as unmarried, a multiple return situation occurs. Approximately 75 returns were removed from the sample because, after finding at least one duplicate situation in a particular year, other returns in other years were found to have been filed by the same "incorrect" taxpayer but without causing a duplicate SSN problem.

In all of these cases, multiple returns using the same SSN within a tax year trigger the review process. A more difficult problem arises when multiples are not present but two different taxpayers are represented within the same longitudinal sequence of tax returns. This situation can be found by examining a sequence of returns using the same primary SSN but where the IRS name controls differ between years. First, let us examine the

case of males. Generally, males have only one SSA name control since men seldom change their last names. Consequently, all CWHS SSN's listed as Males were checked if the IRS name controls changed between any combinations of years. Once again, using the SSA name controls and the full name found on the tax return, this problem can be readily resolved. Approximately 225 returns were removed as a result of this test.

Returns where a woman is the owner of the primary SSN are more complicated because additional name controls are added to the SSA name control list when a woman changes her name due to marriage. Therefore, these returns were reviewed for name control problems only when an IRS name control did not match any of the valid SSA name controls. Approximately 500 returns were removed due to this check.

In the end, as shown in table 2, 1,517 records were removed from the sample, or 0.23 percent of all sampled returns.

▶ An Implication of Removing "Bad" Returns

As noted above, some returns selected for the SOI CWHS sample were selected because the SSN's listed on the returns were incorrect. In other words, the SSN's were SOI CWHS SSN's but they did not belong to the taxpayers who used them on the tax returns. Over time, as taxpayers resolve these SSN problems and begin to use their correct SSN's, they disappear from the CWHS sample. If SOI was able to perform real-time SSN resolution, SOI could continue sampling those taxpayers using their correct SSN's. Since this is not currently possible, these individuals were removed from the sample since, at a minimum, their longitudinal observations are incomplete. Conversely, individuals whose true SSN's are SOI CWHS SSN's but who filed returns using incorrect SSN's are not included in the SOI CWHS sample, and no realistic attempt could have been made to find them and follow them as they continued to use incorrect SSN's. The net result of these two situations is that the weighted totals generated by the CWHS panel sample, when weighted using the inverse of the sampling rate, are shy of the true population totals. It is possible that a post sampling weighting adjustment could be made for each possible base year of the panel, but such an attempt will require more research.

▶ The Gender Bias Problem

A very unfortunate implication of a panel based on sampling primary SSN's is that it produces a profound gender bias. Table 3 shows the gender of the primary taxpayers in the SOI CWHS and of the spouses listed as secondary taxpayers on those CWHS returns that show a joint filing status. Table 4 shows the gender of just the primary taxpayers. The source of the difference between table 3 and 4 is shown in table 5; Over 95 percent of joint returns are filed with the male listed as the primary taxpayer. This does not create a cross-sectional problem, as the total number of taxpayers (primary and secondary) by gender will still be correctly represented as shown in Table 3.

Longitudinally, however, this is an enormous problem because taxpayers are followed solely on the basis of the primary SSN. If taxpayers never change their marital status from an initial base year state the gender bias problem would not exist. However, people do get married and divorced. Thus, from a panel perspective, if one wishes to study individuals who never get married or who are married to the same person for the period under study, the gender bias created by sampling on primary SSN's is not a problem. For all other situations, the problem is inescapable.

▶ From Filer to Nonfiler to Death

When analyzing a longitudinal sample, a user must always be aware of, and have a strategy for, dealing with missing observations and panel attrition. The most important piece of information a user needs in order to develop such a strategy is an explanation of what happened to the missing observations.[6] Suppose a taxpayer files returns for 3 years then vanishes never to file again; what happened to this individual? Did the individual die, retire, or marry? The answer to these questions affects the meaning of any analysis developed using a panel.

One possible explanation is that the taxpayer was a woman who married and subsequently filed as the secondary taxpayer on a joint return. As a result, she

disappears from a panel of primary taxpayers. This is the gender bias problem discussed above. Fortunately, for 2 base years, we are able to solve this problem. In 1987 and in 1999, SOI began panels where the base year primary SSN's were followed in future years whenever they appeared in either the primary or secondary positions. But a limitation of these two panels is that, unlike the primary SSN-based CWHS panel, in which any year from 1979 to 2002 can be used as a base year, the beginning, or base year, is limited to 1987 and 1999. In addition, the 1987 panel ended in 1996.

There are other legitimate reasons why a taxpayer may disappear from the CWHS primary SSN panel, or any other tax return panel for that matter. Two primary reasons are: an income insufficient to require the filing of a tax return; and, death. Fortunately, we have some tools to help with these situations. Someone once said there were only two things certain in life--death and taxes--but our income tax system provides a third possibility. It is possible to be alive and be the recipient of income and not be required to file a tax return or pay income tax. This situation occurs most often with individuals living on Social Security whose incomes are below the filing thresholds for the income tax system. But for purposes of tax return panels, these individuals disappear. Fortunately, IRS creates something called the Information Returns Master File, which contains information documents (Form W-2, Form 1099, Form 1098, etc.) that show whether an individual received any income from a variety of sources during a given year. So, for individuals whose only source of income is Social Security Benefits, and who thus do not file tax returns, SOI has evidence that they are alive and receiving income. Unfortunately, such data are only available for the years 1989, 1993, and 1996 to the present. The use of the IRMF has been the subject of previous ASA papers.[7] Finally, the same SSA files that provide information on name control and gender also provide us with dates of death.

▶ The 1979-2002 SOI CWHS Primary SSN Panel -- The Conclusion

To summarize:

- SOI has created a panel of primary taxpayers that begins in 1979 and continues to the present.
- Duplicate returns and erroneous returns have been removed to the extent possible.
- Age, gender, and date of death information are available for these individuals.
- Base year 1987 primary taxpayers are followed even if they file as secondary taxpayers through 1996.
- Base year 1999 primary taxpayers are followed in future years even if they file as secondary taxpayers.
- Information Returns data are available for all individuals in this panel for the years 1989, 1993, and 1996 through the current year.

▶ Footnotes

[1] Smith, Creston M., "The Social Security Administration's Continuous Work History Sample," *Social Security Bulletin*, Social Security Administration, Office of Research and Statistics, October 1989, Volume 52, Number 10.

[2] Sailer, Peter; Orcutt, Harriet; and Clark, Phil (1980), "Coming Soon: Taxpayer Data Classified by Occupation," *1980 Proceedings of the American Statistical Association, Government Statistics Section*, 1981.

[3] The SOI year is one less than the calendar year or processing year. For example, taxpayers generally filed their Tax Year 2003 returns during Calendar Year 2004. Thus, the returns filed in Calendar Year 2004 would be included in the 2003 SOI file. Over 97 percent of the returns sampled for the 2003 SOI file will be for Tax Year 2003.

[4] It is possible that the source of many of these primary SSN duplicates is the SOI sampling process itself. SOI samples tax returns on a weekly basis throughout a given processing year. It does not receive later IRS corrections to those weekly sample extracts. Thus, if, in January, a taxpayer uses a specific primary SSN, and, at a later date, another taxpayer lists the same primary SSN, IRS

will resolve this situation. For example, if the second occurrence of the SSN was determined to be incorrect, the return would not be posted to the IRS master file, and that return would never be subject to SOI sampling. But if the first occurrence of the SSN was determined to be wrong, SOI would still have the tax return listing the first occurrence in its sample, as well as the second tax return. This would produce a duplicate use of a primary SSN in SOI files.

[5] IRS does not receive all of the death information contained on the NUMIDENT file. The death information SSA obtains from approximately half the States, and for which SSA cannot independently verify the date of death, cannot be shared with IRS due to restrictions placed on that information by these States. Fortunately, SSA is able to independently verify a significant number of the deaths in these States due to the administrative process of stopping Social Security Benefit payments for the deceased individuals. At this time, SSA is not able to provide an estimate of the number of missing entries for date of death, but a reasonable guess would place it below 5 percent.

[6] For some data on CWHS panel attrition and ideas on how to use a panel of tax returns, see Sailer, Peter; Weber, Michael; and Wong, William, "Attrition in a Panel of Individual Income Tax Returns, 1992-1997," 2000 *Proceedings of the American Statistical Association, Government Statistics Section*, 2001.

[7] Sailer, Peter; Weber, Michael; and Yau, Ellen, "How Well Can IRS Count the Population," 1993 *Proceedings of the American Statistical Association, Government Statistics Section*, 1994.

Sailer, Peter and Weber, Michael, "The IRS Population Count: An Update," 1998 *Proceedings of the American Statistical Association, Government Statistics Section*, 1999.

Table 1

All Records						
SOIYR *	CWHS Endings in SOI	Unweighted Count	Weighted Total	SOI Complete Report (CR)	SOI CR less CWHS Total	SOI CR less CWHS Total %
1979	3	27,248	90,826,576	92,694,302	1,867,726	2.01%
1980	3	27,684	92,279,908	93,902,469	1,622,561	1.73%
1981	3	27,799	92,663,241	95,396,123	2,732,882	2.86%
1982	1	9,353	93,530,000	95,337,432	1,807,432	1.90%
1983	2	19,155	95,775,000	96,321,310	546,310	0.57%
1984	1	9,752	97,520,000	99,438,708	1,918,708	1.93%
1985	2	20,207	101,035,000	101,660,287	625,287	0.62%
1986	1	10,138	101,380,000	103,045,170	1,665,170	1.62%
1987	2	21,238	106,190,000	106,996,270	806,270	0.75%
1988	2	21,718	108,590,000	109,708,280	1,118,280	1.02%
1989	2	22,379	111,895,000	112,136,673	241,673	0.22%
1990	2	22,694	113,470,000	113,717,138	247,138	0.22%
1991	2	22,759	113,795,000	114,730,123	935,123	0.82%
1992	2	22,609	113,045,000	113,604,503	559,503	0.49%
1993	2	22,730	113,650,000	114,601,819	951,819	0.83%
1994	2	22,965	114,825,000	115,943,131	1,118,131	0.96%
1995	2	23,469	117,345,000	118,218,327	873,327	0.74%
1996	2	23,878	119,390,000	120,351,208	961,208	0.80%
1997	2	24,172	120,860,000	122,421,991	1,561,991	1.28%
1998	5	62,318	124,636,000	124,770,662	134,662	0.11%
1999	5	63,435	126,870,000	127,075,145	205,145	0.16%
2000	5	64,677	129,354,000	129,373,500	19,500	0.02%
2001	5	64,910	129,820,000	130,255,237	435,237	0.33%
2002	5	64,858	129,716,000	130,076,443	360,443	0.28%

* SOIYR is defined as the Calendar Year of IRS Processing minus one. Thus, the returns filed and sampled in 1980, of which most are for Tax Year 1979, are found in the SOIYR 1979 Individual Income Tax Return File.

Table 2

	All Records less Deleted Records					Deleted Records	
SOIYR *	Unweighted Count	Weighted Total	SOI Complete Report (CR)	SOI CR less CWHS Total	SOI CR less CWHS Total %	Records Deleted	Weighted
1979	27,162	90,539,909	92,694,302	2,154,393	2.32%	86	430,000
1980	27,566	91,886,575	93,902,469	2,015,894	2.15%	118	590,000
1981	27,720	92,399,908	95,396,123	2,996,215	3.14%	79	395,000
1982	9,303	93,030,000	95,337,432	2,307,432	2.42%	50	250,000
1983	19,078	95,390,000	96,321,310	931,310	0.97%	77	385,000
1984	9,694	96,940,000	99,438,708	2,498,708	2.51%	58	580,000
1985	20,118	100,590,000	101,660,287	1,070,287	1.05%	89	445,000
1986	10,084	100,840,000	103,045,170	2,205,170	2.14%	54	540,000
1987	21,119	105,595,000	106,996,270	1,401,270	1.31%	119	595,000
1988	21,634	108,170,000	109,708,280	1,538,280	1.40%	84	420,000
1989	22,314	111,570,000	112,136,673	566,673	0.51%	65	325,000
1990	22,641	113,205,000	113,717,138	512,138	0.45%	53	265,000
1991	22,688	113,440,000	114,730,123	1,290,123	1.12%	71	355,000
1992	22,537	112,685,000	113,604,503	919,503	0.81%	72	360,000
1993	22,658	113,290,000	114,601,819	1,311,819	1.14%	72	360,000
1994	22,906	114,530,000	115,943,131	1,413,131	1.22%	59	295,000
1995	23,411	117,055,000	118,218,327	1,163,327	0.98%	58	290,000
1996	23,835	119,175,000	120,351,208	1,176,208	0.98%	43	215,000
1997	24,146	120,730,000	122,421,991	1,691,991	1.38%	26	130,000
1998	62,269	124,538,000	124,770,662	232,662	0.19%	49	98,000
1999	63,389	126,778,000	127,075,145	297,145	0.23%	46	92,000
2000	64,645	129,290,000	129,373,500	83,500	0.06%	32	64,000
2001	64,879	129,758,000	130,255,237	497,237	0.38%	31	62,000
2002	64,835	129,670,000	130,076,443	406,443	0.31%	23	46,000

* SOIYR is defined as the Calendar Year of IRS Processing minus one. Thus, the returns filed and sampled in 1980, of which most are for Tax Year 1979, are found in the SOIYR 1979 Individual Income Tax Return File.

Table 3
SOI CWHS - Unweighted Taxpayer Counts by Gender

SOI Year	All Taxpayers	Male	Female	Percent Male
1979	40,434	20,137	20,131	49.8%
1980	40,852	20,276	20,427	49.6%
1981	41,071	20,316	20,602	49.5%
1982	13,839	6,773	7,023	48.9%
1983	28,259	13,842	14,316	49.0%
1984	14,385	7,046	7,305	49.0%
1985	29,591	14,516	14,992	49.1%
1986	14,800	7,235	7,530	48.9%
1987	30,592	15,042	15,496	49.2%
1988	31,184	15,336	15,792	49.2%
1989	31,944	15,766	16,138	49.4%
1990	32,284	15,916	16,304	49.3%
1991	32,342	15,939	16,340	49.3%
1992	32,092	15,786	16,238	49.2%
1993	32,187	15,797	16,305	49.1%
1994	32,474	15,980	16,424	49.2%
1995	33,108	16,205	16,826	48.9%
1996	33,490	16,448	16,997	49.1%
1997	33,840	16,596	17,220	49.0%
1998	87,035	42,509	44,485	48.8%
1999	88,233	42,998	45,208	48.7%
2000	89,707	43,777	45,902	48.8%
2001	90,216	44,034	46,158	48.8%
2002	90,399	43,917	46,461	48.6%

Table 4
SOI CWHS - Primary Taxpayer Unweighted Counts by Gender

SOI Year	All Returns	Male	Female	Unclassified	Percent Male
1979	27,162	19,899	7,097	166	73.3%
1980	27,566	20,058	7,359	149	72.8%
1981	27,720	20,080	7,487	153	72.4%
1982	9,303	6,686	2,574	43	71.9%
1983	19,078	13,660	5,317	101	71.6%
1984	9,694	6,957	2,703	34	71.8%
1985	20,118	14,331	5,704	83	71.2%
1986	10,084	7,149	2,900	35	70.9%
1987	21,119	14,852	6,213	54	70.3%
1988	21,634	15,154	6,424	56	70.0%
1989	22,314	15,567	6,707	40	69.8%
1990	22,641	15,700	6,877	64	69.3%
1991	22,688	15,723	6,902	63	69.3%
1992	22,537	15,561	6,908	68	69.0%
1993	22,658	15,541	7,032	85	68.6%
1994	22,906	15,722	7,114	70	68.6%
1995	23,411	15,898	7,436	77	67.9%
1996	23,835	16,145	7,645	45	67.7%
1997	24,146	16,298	7,824	24	67.5%
1998	62,269	41,719	20,509	41	67.0%
1999	63,389	42,190	21,172	27	66.6%
2000	64,645	42,900	21,717	28	66.4%
2001	64,879	43,076	21,779	24	66.4%
2002	64,835	42,860	21,954	21	66.1%

Table 5
SOI CWHS Joint Returns - Unweighted Counts by Gender

SOI Year	All Returns	Male	Female	Unclassified	Percent Male
1979	13,272	13,034	188	50	98.2%
1980	13,286	13,068	170	48	98.4%
1981	13,351	13,115	190	46	98.2%
1982	4,536	4,449	77	10	98.1%
1983	9,181	8,999	156	26	98.0%
1984	4,691	4,602	82	7	98.1%
1985	9,473	9,288	164	21	98.0%
1986	4,716	4,630	77	9	98.2%
1987	9,473	9,283	177	13	98.0%
1988	9,550	9,368	173	9	98.1%
1989	9,630	9,431	193	6	97.9%
1990	9,643	9,427	202	14	97.8%
1991	9,654	9,438	204	12	97.8%
1992	9,555	9,330	211	14	97.6%
1993	9,529	9,273	235	21	97.3%
1994	9,568	9,310	248	10	97.3%
1995	9,697	9,390	290	17	96.8%
1996	9,655	9,352	295	8	96.9%
1997	9,694	9,396	295	3	96.9%
1998	24,766	23,976	783	7	96.8%
1999	24,844	24,036	807	1	96.7%
2000	25,062	24,185	875	2	96.5%
2001	25,337	24,379	954	4	96.2%
2002	25,564	24,507	1,054	3	95.9%