Attrition in a Panel of Individual Income Tax Returns, 1992-1997

Peter Sailer, Michael Weber, and William Wong, Internal Revenue Service

One of the main functions of the Statistics of Income Division of the Internal Revenue Service is to provide files for the Office of Tax Analysis (OTA) at the Treasury Department, so that they can analyze not only how the income tax system is working, but also project how it might work under many different proposals for tax law changes. Longitudinal files can help accomplish some of this analysis. With longitudinal files, one can study how the same group of taxpayers reacted to certain tax law changes; and one can see how the tax system affected this group over a number of years, as their incomes rose or fell, they married or divorced, had children, and retired. Over the years, SOI has produced a number of panels. Each successive panel incorporated many improvements, and each, in its own way, somehow managed to be more difficult to use than had been anticipated.

With this paper, the authors propose to summarize some of the problems involved in putting together a panel of tax returns. For this purpose, we are using a very simple panel which was created by incorporating two four-digit Social Security Number endings, taken from the Continuous Work History Sample (CWHS), in the design of our annual cross-sectional sample. Any 1992 return with a primary Social Security Number ending in either of these four-digit combinations was included in the Individual Statistics of Income file, without regard to income level. Using a weight of 5,000 (these two endings represent two of 10,000 possible endings), this file weighted up very nicely to the 1992 population of individual income tax returns filed. By selecting returns with the same two four-digit Primary SSN endings in subsequent years, we created an embedded panel of tax returns that could be used to do longitudinal analysis for any series of years desired. This kind of an unstratified sample is much easier to use than a panel highly stratified by size of income. With this sample, we need not worry about variability increasing at the upper income levels as taxpayers migrate from lower to higher income levels, and vice versa. With this sample, our income estimates for the very rich do not deteriorate-they are just not very good from the beginning.

Nonetheless, this sample has in common with all other panels we have devised two major problems inherent in the study of income tax returns: incomplete panel units and panel units that change composition.

In this paper, we will attempt to quantify the magnitude of these problems, determine the reasons they exist, and suggest some strategies for constructing panels that are both more complete and more comparable. This research has been made even more important by the fact that we are in the midst of designing another large panel: a panel which is scheduled to begin with Tax Year 1999.

◆ Construction of the 1992 Base Year Sample

According to published data (Internal Revenue Service, 1994), there were 113,604,503 returns for Tax Year 1992. SOI's 1992 CWHS sample contained 22,609 returns. Using the theoretical weight of 5,000, they weighted up to 113,045,000 returns, a very good estimate. Unfortunately, we immediately had to throw out 513 of these returns before we could even start forming a panel. This is because they were either prior-year returns, duplicate returns for the same SSN, or, in many cases, both. The standard SOI procedure of using late-filed prior-year returns as stand-ins for current-year returns to be received after the close of the processing year works well for the cross-section, but makes little sense for a panel. So, we were only able to use the 22,096 returns for 1992 that were timely filed, giving us a population estimate that fell short by about 3 percent.

However, it was not necessary to settle for this shortfall. Presumably, those late Tax Year 1992 filers would file in some subsequent year. Indeed, within the SOI cross-section files for Tax Years 1993 through 1998, we found enough returns to make up the shortfall. Our final count of Tax Year 1992 returns (filed between Calendar Years 1993 and 1999) with the two SSN endings was 22,739, giving us a weighted estimate of 113,695,000.

◆ Status of the Sample in 1993

Column 1 of table 1 shows where the individuals from 1992 ended up in 1993. The initial match of SSN's in the CWHS sample for 1992 to those in the 1993 file yielded a match rate of only 90 percent. We were able to match an additional 2 percent of the SSN's when we checked the 1993 master file of tax returns for the unmatched 1992 SSN's in the secondary SSN slot. Most of these individuals were women who got married and switched from single primary filers to married secondary filers.

Table 1: 1992-base panel, 1 & 5 years later

	1993	1997
Base year sample	22,739	22,739
Ending year percentages:	100	100
Match to primary taxpayer	90	79
Match to secondary taxpayer	2	6
Match to late-filed return	2	1
Match to info. doc. only	4	7
Deceased	1	4
Unmatched remainder	1	3

An additional 2 percent of the 1992 individuals did eventually file tax returns for 1993--but they did not do so until much later. Two-thirds of these late filers filed a year late, but IRS was still receiving Tax Year 1993 returns as late as Processing Year 1999.

About 1 percent of the individuals for Tax Year 1992 had died before the year was over, and thus could not be expected to file for 1993. We identified these individuals by matching to the "Numident" file IRS receives from the Social Security Administration each year.

Finally, about 4 of the remaining missing 5 percent were found in a match to the 1993 Information Returns Master File (IRMF), leaving only 1 percent of the 1992 individuals unaccounted for in 1993. While we never got a 1993 record for this 1 percent, over two-thirds resurfaced again with documents for one of the next 5 years.

◆ Characteristics of Taxpayers Who Were Not Timely Primary Filers in 1993

It would be simplest to confine any analysis of changes from 1992 to 1993 to the 90 percent of all taxpayers who remained primary filers and filed timely returns for 1993. However, such analysis would be valid only if the characteristics of the taxpayers dropped from the study were similar to those who remained in the sample. This is not the case.

Overall, in the population as a whole, 69 percent of primary taxpayers were male; this is a reflection of the fact that married couples filing jointly generally look upon the husband as being the primary taxpayer. Not surprisingly, the taxpayers who switched from primary to secondary were overwhelmingly female--only 13 percent were male.

The most noticeable fact about the late filers is that they are even more predominantly male than the population of primary filers as a whole: about 76 percent, as opposed to 69 percent for the whole population.

Taxpayers who died after filing their 1992 returns were, on average, older than the population as a whole: their average age was about 74, as opposed to an average age of 41 for all taxpayers. On the other hand, those who dropped out of the IRS system entirely without dying were quite a bit younger than the population as a whole--about 28 years old, on average. Their ranks may include students who had held part-time or summer jobs, but who returned to college full time for the year. And it may include young mothers who went on Aid to Families with Dependent Children.

Looking at income: The individuals who went from tax return filers to information document recipients between 1992 and 1993 tended to have low incomes, even in 1992. Their mean adjusted gross income--the bottom-line figure on page 1 of the income tax return--was $8,569, as opposed to $33,159 for the population as a whole.

The average 1992 income of the taxpayers who dropped completely out of all IRS systems for 1993 was even lower: $6,421. So, this is definitely a group of poor, young people.

In summary, if SOI removed all these individuals from the sample, it would change the demographic and economic mix of the panel. That is why SOI plans to check the secondary taxpayer SSN's on the Master File of all tax returns; to check the information documents for non-filers; to keep the file open for a year or more to bring in the late filers; and to check out the Numident file to make sure that the missing are truly deceased.

◆ Status of the Sample in 1997

Column 2 of table 1 shows the status of the 1992 sample in 1997. Obviously, more of our 1992 cohort has died, gotten married, gotten divorced, moved into the information documents only group, or simply dropped out. Only 79 percent are left as primary filers, as opposed to 90 percent for 1993. And there are further complications.

Table 2: 1992 sample of 22,739 returns: % of primary taxpayers by status, 1997

		Solution				
	All	1	2	3	4	5
Total used, by solution	100	63	81	86	96	81
Non-joint, all years	32	32	32	32	32	32
Same couple, all years	31	31	31	31	31	31
Non-joint to joint	12		12	12	12	12
Jnt to non-jnt/diff.spouse	6		6	6	6	6
Missing intervening years	5			5	5	
Match to info. doc. only	7				7	
Unmatched remainder	3				3	
Deceased	4					
Additional sample						18

Column 1 of table 2 breaks the data in column 2 of table 1 into greater detail. The last three lines--deceased, unmatched, and matched to information documents only--remain the same. However, the 86 percent of the 1992 population who filed for 1997 are divided up differently. Five percent of the panel units are missing returns for some year from 1993 through 1996, making a complete historical analysis difficult. Six percent of the 1992 population started off as joint return filers, but ended up either as non-joint filers or as married to somebody other than their 1992 spouses. So, the data of the former spouses are now no longer in our sample. Twelve percent of the sample started off as non-joint, but became joint filers. So, their income, deduction, and tax information has now become intertwined with that of a taxpayer not present in the base year. Only the remaining 63 percent of the file are easy to analyze--the 31 percent who remained joint return filers married to the same spouses for all 6 years, and 32 who remained non-joint filers for all 6 years.

◆ Using the Panel File

How should a researcher use a panel file with so many unstable units? The answer may depend on the type of research being conducted. We will suggest five possible solutions to this dilemma. We are greatly indebted to John Czajka and Larry Radbill for getting us started in thinking about strategies for analyzing multiperson units (see Czajka and Radbill, 1995). As a generalization, the simpler the solution, the smaller the proportion of the file that can be used.

Solution 1--use unchanged filing units only

Solution 1 is the very simplest--you use just those returns that represent the same taxpayers for all 6 years. Czajka and Radbill dismiss this solution as ignoring the most interesting returns, but we think we can make a case for using it in some limited situations. Let us say we want to test the hypothesis that taxpayers who have sole proprietorship farm income, as well as income from other sources, tend to time their farm net profits and/or losses in such a way as to even out their taxable incomes over the years. Since marriages or the dissolution thereof may have a major effect on the relationships between income types and amounts, it is probably best to go with only the consistent family units. The 63 percent that file every year and do not change filing units are usable, and receive a weight of 5,000.

Solution 2--follow the primary taxpayer only

Our second solution is the equivalent of Czajka and Radbill's solution 3: choose one taxpayer and follow that person throughout the 6-year period. For example, if you wanted to test how many base-year tax return filers remained in, fell into, or got out of poverty throughout the length of the study, you would just recompute the poverty level each year based on family size, but follow only those returns containing the selected 1992 taxpayer. Under this solution, 81 percent of the base-year sample would be usable-all panel members for whom we have a return every year; and the weight remains 5,000. It should be noted that, while secondary taxpayers are not followed under this solution, their presence and their incomes still form an important part of the analysis.

Solution 3--follow primary taxpayers using tax returns and information documents

Solution 3 is a variation on solution 2. It still demands a tax return in the beginning and ending year, but is content with information documents in the intermediate years. Of course, we would have to make the assumption that the demographics on marital status and family size remained the same over the years for which no tax return data were available (which is probably safe if they are the same at the beginning and the end of the non-filing period). In the case of joint returns, information documents for both taxpayers would be used. More about the use of information documents later; but we are making the assumption that they can be used to compute the unit's income level where no return is present. The panel weight remains at 5,000 for solution 3, and 86 percent of the file is usable.

Solution 4--follow all taxpayers individually

A fourth solution is to keep all the returns reflecting changed marital status in the sample, and express the results in terms of numbers of taxpayers. As a matter of fact, for some types of analysis, the individual taxpayer may be the only logical unit to follow. Let us assume you want to follow the wages of men and women separately, along with the marginal tax rates to which they are subjected. Obviously, you do not want to drop any individuals from the sample just because they get married or divorced--these may in fact be the most interesting individuals to study. What you can do, thanks to the availability of information documents, is get separate earnings data for each taxpayer on a joint return from the appropriate Forms W-2. Each taxpayer in 1992 gets a weight of 5,000 and is followed through all subsequent years. Tax returns are used strictly for the purpose of obtaining marginal tax rates, which are the same for the primary and the secondary taxpayer. But the absence of a tax return--assuming enough time has been left for one to come in--does not interrupt the series. No tax return means a marginal rate of zero. For that matter, you could argue that you do not need a W-2 either. No W-2 simply means no salaries and wages. We will call this extreme solution 4; it includes everybody who has not died, and weights up to the total survivors of 1992 taxpayers.

Solution 5--obtain additional records to complete panel units

There are many analyses for which it is essential to have tax return data for every year in the series, and for which the tax return is the only logical unit of analysis. For example, to what extent did taxpayers use the various tax law breaks on capital gains in successive years, and what effect did they have on capital gain realizations? Here is our fifth solution: a way you could build complete panel units for a sample of base-year tax returns, even if they include taxpayers who got married or divorced. You start from your base-year sample chosen on primary SSN with a weight of 5,000. If this is a joint return, and the two taxpayers split, you duplicate the returns for the previous year or years, then attach one set to each of the following years' separate returns. For this, you have to cut the weight in half, to 2,500, so that you will be weighting up to the correct number of returns.

More complicated to deal with is the situation where tax units are formed in an out-year through the joining of two single taxpayers into joint filing status, and one (but not the other) is a panel member. If we are to constitute two complete series of tax returns, we need to get the earlier years' returns for the "visitor" to the panel. They would then get weights of 2,500, and the weight for the

pre-marriage years of the panel member would also be reduced to 2,500. After the marriage, the joint return would be duplicated, and each copy assigned a weight of 2,500.

In order to do this for future panels--we have not done so for the one currently under discussion--we will be accumulating master files of all individual income tax returns for every year of our next panel. While the Master Files do not contain all of the information in our typical SOI samples, it will be a start. In solution 5, you get to use the 81 percent of the base-year filers who filed a return in all subsequent years, and get a bonus of bringing in matching spousal return for the 18 percent who changed marital status during the duration of the panel.

◆ Using Information Documents

In two of the above solutions (3 and 4), we recommended the use of information documents (such as Forms W-2, 1099-INT, 1099-DIV) to augment data from the tax return (see Sailer and Weber, 1998). In one case, they were to stand in for tax returns when none were filed. In the other, they were to be used to divide up income amounts between taxpayers on joint returns. It is, therefore, advisable to discuss what information documents can and cannot do. In table 3, data are shown for joint returns in the 1993 SOI sample.

The clearest one-for-one substitution, at least in theory, comes from salaries and wages. In actuality, about 97 percent of 1040 salaries and wages on Form 1040 appear on Form W-2, allowing you to neatly divide income amounts between husband and wife. The remaining 3 percent of salaries and wages are frequently documented on Forms 1099-MISC, although this form also feeds into other parts of the 1040. Unemployment compensation is also pretty much an identical item on the 1040 and information document side, although 5 percent of it appears not to have made it to the 1040 side.

Investment income is a bit of a problem when it comes to dividing income between husband and wife, since much of it may, in fact, belong to both taxpayers, and be filed using only the primary taxpayer's SSN. However, when used as a substitute for 1040 data, divi-

Table 3: 1040-Information Records Comparisons
(Estimates based on a sample of TY 1993 matched joint returns and information documents)
Money amounts in thousands of dollars

	Form 1040 Amount	Info. documents Amount	Info. documents as % of 1040	Source of info. documents
Employee compensation	1,916,612,019	1,992,645,840	103.97%	
Salaries & wages	1,916,612,019	1,855,387,621	96.81%	W-2
Awards	*	2,570,019	*	1099-MISC
Non-wage compensation	*	134,688,200	*	1099-MISC
Unemployment compensation	15,933,502	16,813,345	105.52%	1099-G
Interest	80,886,025	51,463,388	63.62%	1099-INT
OID interest	**	1,119,698	**	1099-OID
Dividends	49,611,822	51,129,369	103.06%	1099-DIV
State income tax refunds	8,506,263	10,132,863	119.12%	1099-G
Social Security income	76,773,806	116,787,936	152.12%	SSA-1099
Pensions	194,545,359	222,085,011	114.16%	1099-R
Rents and royalty gains	31,108,715	33,140,785	106.53%	1099-MISC
Rents & royalties + farm rental	33,347,744	33,140,785	99.38%	1099-MISC
Business income	137,193,782	137,215,048	100.02%	Sch. C
Other self-employment income	***	32,236,312	***	Sch.SE - Sch.C

* Some taxpayers show these amounts as salaries and wages on Form 1040.
** Generally reported on the "Interest" line of Form 1040.
*** On Form 1040, most of these amounts appear as farm or partnership income.

dends appear to be a rough equivalent, whereas interest, even when the amounts reported on Form 1099-OID (original issue discounts) are included, falls short of the 1040 total. Social Security income is actually more complete on the information document side, since taxpayers with no taxable benefits do not have to show total benefits on the 1040. Pensions on the information document side include some rollovers, which are frequently not reported on the 1040. Rents and royalties on the information document side compare very nicely with rents and royalties plus farm rental income on the tax return side, but this is largely a coincidence. Many rents are not covered by 1099 forms, while many rents reported on Form 1099 are actually reported as business income on the 1040. Royalties from Form 1099 are also frequently reported as business income. It is in large part serendipity that the two figures both came out to $33 billion.

Generally, there are no information documents that show the business income of non-filers, although most business people do have to file for self-employment tax purposes. If a return is filed, there are two ways to separate the business income of husbands and wives. Each Schedule C can be gender-coded, based on the name shown on that schedule, as is done in the SOI program. Or Schedule SE, which must be submitted separately for a husband and a wife on a joint return, can be used to obtain total self-employment income (including that from farms and partnerships where the taxpayer is a working partner).

◆ Future Plans

The next panel SOI prepares will include a larger CWHS component (five 4-digit endings) plus a stratified high-income cohort. The stratified nature of this panel will add a whole new set of technical difficulties to the weighting and interpretation of the data (see Czajka and Schirm, 1992). We will collect information document data for the same five 4-digit endings throughout the period of the panel, whether the individual files a tax return or not. Data for base-year panel members (primary and secondary) will be collected from both tax returns and information documents. We are also planning to keep copies of selected items from all IRS Master File systems, so that data on visitors can be traced back to the base year.

◆ References

Czajka, John L. and Radbill, Larry M. (1995), "Weighting Panel Data for Longitudinal Analysis," *1995 Proceedings of the Section on Survey Research Methods*, Volume I, American Statistical Association, Alexandria, VA.

Czajka, John L. and Schirm, Allen L. (1992), "Enhancing the Representativeness of a Longitudinal Sample of Individual Income Tax Returns: Weighting and Sample Supplementation," *Proceedings of the 1992 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC.

Internal Revenue Service (1994), *Statistics of Income-1992, Individual Income Tax Returns*, U.S. Government Printing Office, Washington, DC.

Sailer, Peter and Weber, Michael (1998), "The IRS Population Count: An Update," *1998 Proceedings of the Section on Government Statistics*, American Statistical Association, Alexandria, VA.■