Weight Development for the SOI Individual Income Tax Return Panel

Yan K. Liu[1], Victoria Bryant[1] and Fritz Scheuren[2]

Abstract

Statistics of Income of IRS uses a panel sample for longitudinal analyses and cross-sectional estimations. Each year, a small refreshment sample is also added for the cross-sectional estimation purpose. The panel sample was selected as a stratified sample from the tax year 1999 population of individual income tax returns, where stratum boundaries were formed using the income of return. Because of the much skewed income distribution, the sampling rates across strata were quite different. This poses a particular problem for returns whose income grows dramatically such that its associated base weight is no longer appropriate for either longitudinal analyses or cross-sectional estimations. In this paper, we look into the treatment for the influential movers. We also introduce the production of cross-sectional weights from the combined sample of panel returns and refreshment returns.

Key Words: Cross-sectional weight, Longitudinal weight, Panel sample, Poststratification, Stratified sample.

1. Background of the Panel Sample

The Statistics of Income (SOI) of the Internal Revenue Service (IRS) selects a cross-sectional stratified random sample of individual returns from the population of all U.S. individual tax returns filed to the IRS every year. This yearly cross-sectional sample provides a broad scope of individual return data for federal tax policy analysis and research. However, it does not provide detailed information on certain subjects, such as transaction-level data for taxpayers' sales of capital assets. In addition, it has very limited use in analyzing the behavior of taxpayers over time. Therefore, panel samples, such as the tax year 1999 Individual Income Tax Return Edited Panel sample, have been introduced.

Since the 1999 Individual Income Tax Return Edited Panel sample was selected from SOI's Tax Year (TY) 1999 cross-sectional sample, they are closely related, but their target populations, stratum definitions, and sample sizes are different. The target population of the cross-sectional sample for year *t* includes all returns filed in calendar year *t*+1, where the majority of the returns are for Tax Year *t*, with a small number of returns for other tax years (mainly late returns for TY *t*-1 and TY *t*-2). The target population for the 1999 panel sample includes all returns filed for TY1999. The 1999 panel sample returns were drawn from the 1999 cross-sectional sample, supplemented with the 2000 and 2001 cross-sectional samples to include returns that were filed up to two years later.

The 1999 cross-sectional sample and the 1999 panel sample are both stratified random samples, but their strata boundaries are different. For the panel sample, the stratification is achieved by size of the income of return (either positive or negative) and the 'degree of interest' of return for tax modeling purposes (see Table 1 below). The 'degree of interest' is a four-level categorical variable, where '1' is assigned to returns that are least interesting and '4' to those most interesting. For the cross-sectional sample, stratification is achieved by using the same stratification variables of income amount and 'degree of interest' combined with a 'return type code' that puts returns in different categories based on the presence/absence of particular schedules attached to the taxpayer's return. In comparison, the panel sample has 25 strata, while the cross-sectional sample has 208. We will utilize the fact that returns with an income of $5,000,000 or more are selected with certainty in the cross-sectional sample (for more information, see Testa and Scali, 2005), which correspond to panel sample returns in current-year strata[2] 0, 1, 2, 23, 24 and 25.

When a return was selected for the 1999 panel sample, both the primary filer and the secondary filer became members of the longitudinal sample, while other members in the family were not included. These panel member returns have been followed through subsequent tax years. The first year, Tax year 1999, is termed the *base-year* and the subsequent years are *out-years*. Due to the changes in family compositions, it is possible that two out-year panel returns are linked to one base-year panel return (e.g., divorced couple who filed jointly in TY1999) or one out-year panel return is linked to two base-year panel returns (e.g., married couple who filed separately in TY1999).

This panel also provides extra detailed information about sales of capital assets at the transaction level that the yearly cross-sectional samples do not. For this level of information, the panel sample can be used for both longitudinal analysis and yearly cross-sectional estimation. However, the panel becomes less representative of each out-year population over time. Adding refreshment is one way to capture new filers, but due to various resource constraints, only a small number of refreshment returns are added every year for the purpose of cross-sectional estimation and are not followed afterwards. The refreshment sample is a simple random sample and the returns are selected by having one of the five specific last four digits of the nonpanel filer's primary SSN and called Continuous Work History Sample (CWHS) returns (Weber, 2005). The CWHS returns are considered randomly selected since the four-digit SSN endings are approximately random. The refreshments include two sources of returns: (1) new returns that were not in the 1999 population; and (2) nonpanel returns that filed in 1999 as a secondary filer.

[1] Statistics of Income, IRS, P.O.Box 2608, Washington, DC 20013.
[2] NORC, University of Chicago, 1402 Ruffner Rd., Alexandria, VA 22302.
[2] The current-year stratum is defined using the current-year income and same boundaries as the panel sample as shown in Table 1.

Table 1. Sample Design of the 1999 Panel Sample

Stratum	Income Range	Degree of Interest	Specified Sampling Rate (%)	Specified Weight[3]	Actual Weight[4]
	NEGATIVE INCOME				
0	$20,000,000 or more	All	100	1	1
1	$10,000,000 - under $20,000,000	All	48.47	2	2.18
2	$5,000,000 - under $10,000,000	All	22.05	5	4.81
3	$2,000,000 - under $5,000,000	All	4.20	24	22.9
4	$1,000,000 - under $2,000,000	All	1.42	70	68.25
5	$500,000 - under $1,000,000	All	0.58	172	160.34
6	$250,000 - under $500,000	All	0.12	833	727.85
7	$120,000 - under $250,000	All	0.05	2000	1790.99
8	$60,000 - under $120,000	All	0.05	2000	2146.78
9	Under $60,000	All	0.05	2000	2138.21
	POSITIVE INCOME				
10	Under $30,000	1	0.05	2000	2017.37
11	Under $30,000	2	0.05	2000	1979.98
12	Under $30,000	3-4	0.05	2000	1998.22
13	$30,000 - under $60,000	1-2	0.05	2000	2034.27
14	$30,000 - under $60,000	3-4	0.05	2000	2006.87
15	$60,000 - under $120,000	1-3	0.05	2000	2029.59
16	$60,000 - under $120,000	4	0.05	2000	1969.88
17	$120,000 - under $250,000	1-3	0.05	2000	2003.24
18	$120,000 - under $250,000	4	0.05	2000	2081.19
19	$250,000 - under $500,000	All	0.18	556	556.38
20	$500,000 - under $1,000,000	All	0.59	169	169.92
21	$1,000,000 - under $2,000,000	All	1.72	58	56.3
22	$2,000,000 - under $5,000,000	All	5.73	17	17.28
23	$5,000,000 - under $10,000,000	All	18.88	5	5.24
24	$10,000,000 - under $20,000,000	All	57.62	2	1.77
25	$20,000,000 or more	All	100	1	1

Two independent uses are derived from this panel sample and yearly refreshment sample. The panel sample alone is used for longitudinal analyses. The panel sample and the refreshment sample together are used for cross-sectional estimations. Our primary interest is thus to develop appropriate return weights for both longitudinal analyses and cross-sectional estimations. We will discuss two sets of weights: one for longitudinal analyses and the other for cross-sectional out-year estimates. The longitudinal weight is representative of the base-year population, while the cross-sectional weight is intended to be representative of each out-year population. Both weights should be adjusted for a small number of influential 'outliers,' but in different ways. We focus on the income change from base-year to current-year when adjusting the panel weights and focus on current-year income when adjusting the cross-sectional weight. Further, we use poststratification within each panel sample stratum to calibrate weights panel returns and refreshment returns such that they produce each out-year's stratum population sizes, as estimated from the yearly cross-sectional samples. It is important to note that the cross-sectional weight that we derive for the combined sample of panel returns and refreshment returns is different from the cross-sectional weight of the yearly cross-section sample. Through this text, we refer to the former, unless it otherwise specified.

2. Droppers, Poppers and Outliers

Our sample was selected on return income in the base-year. However, income changes over time due to economic success/failure or return composition change (e.g., marriage or devoice), which leads to units shifting to different strata where the

[3] The actual weights are slightly different from the specified weights since the Bernoulli sample selection was used.
[4] Since we use Bernoulli sampling to select returns, the actual sample size is slightly different from the expected sample size in each stratum.

probability of selection is different. We consider two such cases, termed 'poppers' and 'droppers.' Some returns with a very high income in the base-year have a very low income in the out-year. These returns are called 'droppers.' On the other hand, some returns that started with a very low income in base-year may end up with an extremely high income in the out-year. These returns are called 'poppers.' Recall, the sampling rates across strata differed dramatically, resulting in base weights that ranging from 1 to over 2,000. A problem arises when returns shift between strata for different years, due to dramatic changes in yearly income[5]. This particularly affects 'poppers' that were originally selected with low sampling rates and then move into a high-income[6] stratum; thus assigned a higher base weight. The associated base weights for such returns, which are the inverse of the selection probabilities, are too large and no longer appropriate for longitudinal analyses and cross-sectional estimations.

For example, assume that 10 returns in the TY1999 population had an income of $50,000 that increased to $25,000,000 in TY2004. If only one such return was selected in the panel sample, their weight for the base-year would be 2000 and also for any out-years. We would overestimate the number of total returns by 1990. Conversely, if none of those returns were included in the panel sample, we would underestimate the number of such returns by 10. In terms of the dollar amount, if we sampled one such return, we would overestimate the income by $49,750,000,000; but if we selected no such returns we would underestimate the income by $250,000,000.

In summary, the 'popper' returns that experience very large growth in income (the absolute value) along with their large weights will exert an unduly large influence upon the estimates of income and tax variables at income levels where most panel members have much smaller weights. The extreme 'poppers' and 'droppers' are also called 'outliers' and their weights should be adjusted. We will focus our discussion on the 'poppers' for this discussion.

3. Original Base Weights and Adjustments

Before we discuss the weight adjustments, we will describe the original weights calculated by the Mathematica Policy Research Inc. (Mathematica Policy Research, 2006). Our final longitudinal weights and cross-sectional weights are developed based on these original weights. First, in the base-year, the person weight for each social security number (SSN) was equal to the sample weight of its return. In out-years, each matched primary or secondary SSN received a longitudinal person weight from the matching base-year panel member. These person weights are used to calculate return weights for each out-year file. If an out-year return includes only panel members as filers, the return weight on the out-year return is set equal to the primary filer's person weight, with one exception. If a joint out-year record included two panel members drawn from different base-year panel returns, then the return weight is calculated as a function of the primary and secondary person weights:

$$W_R = \frac{1}{(1/W_P)+(1/W_S)-(1/W_P)(1/W_S)} \qquad (3.1)$$

where W_R is the return weight; W_P is the person weight of the primary filer and W_S is the person weight of the second filer. In this weight calculation, the inverse of each person weight is treated as a selection probability.

If a married joint out-year return contains one panel member and a nonpanel spouse, then W_R is calculated by taking into account the selection probability of the nonpanel spouse. The SSN of the nonpanel spouse is matched against the TY1999 population file[7]. If identified as a TY1999 filer, then the nonpanel spouse is assigned a pseudo-person weight based on its 1999 stratum membership. W_R is then calculated as above, substituting the nonpanel spouse's pseudo-person weight for W_P or W_S, as appropriate. If a nonpanel spouse did not file a 1999 return, then it is assumed that the spouse's panel selection probability be zero. In that case, return weight W_R becomes the panel member's person weight.

In terms of trimming the weights for potential outliers, Mathematica Policy Research (MPR) proposed the ad hoc method that includes two treatments. An influential 'popper' could move to a higher absolute income stratum by "marrying up," i.e., a taxpayer who filed a lower income return in base-year married a taxpayer with very high income in base-year, resulting in its out-year return with high income. In this case, MPR used the new spouse's base-year selection probability in (3.1) to derive the new weight. If "marrying up" did not occur, then they derived the trimmed weight as the geometric mean of the original base-year weight and the weight associated with the high-income stratum to which a panel member migrated into during the out-year.

Offsetting adjustments were also applied to the remaining returns in the stratum that the popper migrated from, to preserve the departed stratum population total in the base-year. For example, suppose that a return from base-year stratum 19 had a weight of 556.38. Then this return achieved a high income associated with stratum 24, which had a weight of 1.77. If the income jump of this return was not the result of the taxpayer 'marrying up,' then the trimmed weight was $\sqrt{556.38 \times 1.77}$ = 31.38. The difference,

[5] All incomes in out-years are adjusted so that they are comparable to the tax year 1999 income.
[6] We talk about income in the absolute value in the rest of the paper.
[7] The 1999, 2001 and 2002 population files were searched for nonpanel spouses who might have been late filers. But the 2000 population file was not searched because it was not available. If a nonpanel spouse had filed a late 1999 return in 2000, then the assumption that the panel selection probability was zero is incorrect, and return weight would be biased upward. But this should be a minor issue the number of missed matches should be small.

556.38-31.38 = 525 would be offset (i.e., redistributed) among the remaining returns in the base-year stratum 19. This was done by apportioning an offsetting weight increase among the remaining returns and adding this increment to each original weight.

The trimmed base-year return weight is also equal to the trimmed base-year person-level longitudinal weights. Then these trimmed weights are used to derive the trimmed return weights in each out-year, which replaced the base-year return-level weights on the panel sample's data file as the longitudinal weights and used with the unadjusted spouse weights to derive a trimmed return-level weight for each of the four remaining out-years. MPR recommended that this weight trimming be performed only on influential returns whose base-year weights were severely different from the weight strata they were associated with, either if their weights were much larger than those of the units within the strata to which they migrated into or that their weights accounted for a substantial share of the estimated population total of their new strata. MPR identified 69 such returns from the out-years 1999-2003 as candidates (i.e., potential outliers) for the associated trimming procedures described above.

4. Proposed Weights

We produce two separate sets of longitudinal and out-year cross-sectional weights. The longitudinal weight is essentially the same as MPR's sampling weight (described in Section 3), but we used a different trimming procedure. With access to more data sources, we used the large number of certainty returns from SOI's yearly cross-sectional sample to trim the influential panel outliers. The trimming is intended to reflect the true number of influential panel returns that went through the similar income changes across years. The cross-sectional weight is produced so that the combined sample of panel returns and refreshment returns is representative of the current population and produces cross-sectional estimates that are consistent to those produced from SOI's cross-sectional sample for each out-year. The weights are adjusted every year to reflect the current population distribution (especially for high income returns) as much as possible.

The adjustment of longitudinal weights is mainly focused on trimming the weights for these influential 'poppers' or outliers We adjust the longitudinal weights based on the known number of influential 'poppers' in the panel sample and the number of similar 'poppers' in the population. However, some influential 'poppers' in the population have no similar poppers in the panel sample; and are not represented by any panel returns. Therefore, no weighting adjustment can compensate for these omitted units. Conversely, the cross-sectional weights of the combined sample are intended to represent the snapshot of each out-year population, estimated from the SOI's larger cross-sectional sample. We first produced the return weights using selection probability. Then we adjust the return weights using poststratification.

4.1. Modified Weights for Longitudinal Analysis

The target population for longitudinal analysis is the cohort of base-year filers, which includes all returns filing in subsequent tax years with at least one SSN (primary or secondary) appearing in the TY1999 population. In other words, the target population in each out-year excludes the returns where neither primary filer nor secondary filer was in the TY1999 population. The longitudinal weight is intended to represent the base-year population and the income-change over time. Thus, we start with the MPR weights calculated using joint probability (see equation (3.1)) and then adjust weights for rare cases where the weights could result in a significant bias in the longitudinal analysis. We only consider trimming for returns that are in low-income strata in the base-year and are in high-income strata in out-years. We restrict the potential trimming candidates to the returns that were not selected with certainty in the base-year and have out-year income of $5 million or more. These returns are in out-year strata 0, 1, 2, 23, 24 and 25 and are sampled with certainty in the yearly cross-sectional samples. Therefore, we know the associated stratum population sizes. Further, we divide these returns by size of out-year income (over $500 million, $100 million to $500 million, or $5 million to $100 million) for different treatments that are described as follows:

1. The weights for base-year certainty returns (i.e., returns in base-year strata 0 and 25) are not adjusted. Thus, these returns retain their weight of 1, regardless of their out-year income.
2. Returns with an out-year income of over $500 million are, and should be, treated as certainty returns since the income is so large that these returns are arguably self-representing. Four such panel returns with a weight from 1.77 to 2.18 have their weight adjusted to 1. However, another 32 such returns in the cohort population were not in the panel sample, which means they are not represented in the panel sample.
3. For returns with an out-year income between $100 million to $500 million, we performed case-by-case trimming due to the possible large impact of the weight of these returns on longitudinal analysis. We first summarize the returns by ***weight adjustment cell*** that is defined by the base-year stratum group and the out-year income class. Base-year stratum groups are strata 3-6, 7-18 and 19-22 and the rest of strata and the out-year income is classified into four classes each by $100 million increments. As shown in Table 2, we use the difference between the sum of panel weights and the number of population returns and the ratio and of the total panel weights over the true number of population returns in each cell to identify outliers. Returns with a large difference or large ratio are potential outliers. For example, if a panel return started with a base-year income in range (-$250,000, $250,000) and then moved to a much higher income in the range of ($100 million, $200 million), while there are only two population returns in this cell. This return has a base-year weight of 2,006.9. Therefore, this return is clearly a 'popper' and its weight is adjusted to 2. Other returns in the table are treated as outliers and therefore are not subject to weight adjusting.

Table 2. An illustration example

Weight Adjustment Cell		Number of Returns in the Panel Sample	Number of Returns in the Population	Sum of Panel Return Weights	Ratio	Difference
Base-Year Stratum (a)	Out-Year Income (b)	(c)	(d)	(e)	(f)=(e)/(d)	(g)=(e)-(d)
1	(100, 200)	2	3	4.4	1.5	1.4
2	(100, 200)		1			
3-6	(100, 200)		2			
7-18	(100, 200)	1	2	2006.9	1003.4	2,004.90
19-22	(100, 200)	1	15	17.3	1.2	2.3
23	(100, 200)		17			
24	(100, 200)	15	21	26.6	1.3	5.6

4. For noncertainty panel returns with an out-year income between $5 million to $100 million, we define weight adjustment by the base-year stratum group (1, 2, 3-6, 7-18, 19-22, 23 and 24) and the out-year stratum group (0, 1-2, 23-24, 25). Then we identify outliers using the difference or ratio as illustrated above. The adjusted weight is simply the number of population returns divided by the number of panel returns in those cells.

In summary, there are a total of 17 returns whose weight was trimmed. After we trimmed the weights for outliers, we checked if any individual filer on a return was adjusted over multiple years. In such cases, we used the smallest adjusted weight. The 17 returns whose weights were adjusted were associated with 11 unique taxpayers. One return gets adjusted in two tax years and we take the smaller adjusted weight for this return in the two tax years. The same procedure was applied for the one return whose weight was trimmed in three tax years and one whose weight was adjusted in four tax years.

Offsetting is performed using MPR's method. The amount of total weight adjustment was offset among the remaining returns in the same base-year stratum by apportioning an offsetting weight increase or decrease among the remaining returns. The trimmed base-year person-level weight was set to equal to the trimmed return-level weight. Overall, the trimming is performed only for rare cases. There are many cells that had influential poppers in the population, but no panel return to represent them, but there is no alternative data available such that weight adjustments could be used to compensate for them in the longitudinal weighting.

4.2. Weights for Cross-sectional Analysis

For cross-sectional estimations, a small supplemental sample is added to the panel sample every year. It is called CWHS refreshment sample and includes returns in the Continuous Work History Sample (CWHS) that are not in the panel. CWHS returns are returns with five specific 4-digit endings of the taxpayers' primary SSN. Therefore, CWHS returns are selected with a probability of 5/9999 in SOI's yearly cross-sectional sample. CWHS refreshment sample includes all CWHS returns that are ***not*** in the panel. The additional information captured for panel-level analysis (e.g., the taxpayer's transaction-level sales of capital assets) was obtained for the CWHS refreshment returns each year so they could be combined with the panel sample to make cross-sectional estimates on certain subjects that are not available in the yearly cross-sectional samples.

The cross-sectional weights were adjusted so the combined sample of panel returns and refreshment returns represents each out-year's population. The combined sample is divided into three non-overlapping segments for weighting purposes: (1) panel returns; (2) nonpanel CWHS returns that were also in the TY1999 population; and (3) nonpanel CWHS returns that were not in the TY1999 population. These three segments are shown in Table 3. To separate refreshment returns into segment (2) and segment (3), we searched the TY1999 population returns[8]. If a match was found, the stratum identification 'PSAMP' on the population file was used to determine the return's selection probability for the panel sample. High income returns, especially new filers, are often missed in the CWHS refreshment sample due to the extremely skewed income distribution. In fact, no returns over $2,000,000 were selected in the CWHS refreshment sample[9] for any tax year. Therefore, we can only compensate to a certain degree by adjusting weights using poststratification.

[8] We first searched TY1999 returns from the yearly cross-sectional samples of SOIYR 1999, 2000 and 2001 where the panel sample was selected since the cross-sectional samples are edited and more accurate than the population. We used EPSSN and ESSSN for the matching. Then, for the unmatched returns, we searched TY1999 returns from SOIYR 1999 and SOIYR 2001 populations using PSSN and SSSN. The SOIYR 2000 population was not available and therefore was not used. However, the number of TY1999 returns in the SOI year 2000 population was small.

[9] There are a small number of CWHS returns with an income over $2,000,000, but they are already in the panel and therefore, not in the refreshment sample.

Table 3. Number of Returns by Tax Year and Weighting Segment for the Combined Sample

Tax Year	Segment of the Combined Sample			Total
	1.Panel Returns	2. Nonpanel CWHS Returns, and 1999 filers	3. Nonpanel CWHS Returns, new filers	
1999	83,432	0	93[10]	**83,525**
2000	80,952	1,076	4,866	**86,894**
2001	80,044	1,702	6,881	**88,627**
2002	79,045	2,274	8,356	**89,675**
2003	78,464	2,719	9,829	**91,012**
2004	77,814	3,065	11,463	**92,342**
2005	77,043	3,387	13,343	**93,773**
2006	75,351	3,797	15,077	**94,225**

We begin by looking at the weighting adjustment procedure for high-income panel returns that are different from other returns. The panel sample does not include all of the large current-year high-income returns, but we know how many there are in the population because returns in out-year strata 0, 1, 2, 23, 24 and 25 are certainty returns in the yearly cross-sectional samples. Thus we have stratum population totals for each out-year, estimated from SOI's cross-sectional sample.

Since income follows a highly skewed distribution, we adjust those large returns within different income ranges. First, for returns with an income of $1 billion or more, we give them a weight of 1 due to the large amount of income. Then for returns in with an income between $200 million and $1 billion, we divide each weighting adjustment cell into finer weight adjustment cells by each $100 million income increment. The adjusted weight is simply the ratio of the number population returns over the number of combined sample returns. In the case where there are no sample returns to represent the associated ones in the population, these returns are ignored from the estimation. Table 4 gives illustrated examples.

Table 4. An Illustrated example

Out-Year Stratum	Income (in Million $)	Number of Returns		Adjusted Weight for Panel Returns	Number of returns not represented
		Population	Panel Sample		
25	[200, 300)	28	17	1.65	
25	[300, 400)	18	14	1.29	
25	[400, 500)	10	9	1.11	
25	[500, 600)	2			1
25	[600, 700)				
25	[700, 800)				
25	[800, 900)	2	2	1.00	
25	[900, 1000)				
25	1000 and over	5	2	1.00	3

For returns with an income between $20 million and $200 million, along with returns with an income between $5 million and $20 million (in out-year strata 1, 2, 23 and 24), we define the weight adjustment cell as the combination of base-year stratum and current-year stratum in order to factor in the income change from base-year to current-year as much as possible. These returns are all taken with certainty in each yearly cross-sectional sample and we know how many there are in each out-year population. The initial weight within the cell is simply the ratio of the number of population returns over the number of panel sample returns. Then we use poststratification within each current-year stratum to adjust for the population returns that have no panel returns to represent. The details are given next.

Let n_i be the number of panel sample returns and N_i be the number of population returns in cell i. Let M_i be the number of non-represented population returns and $M_i = \begin{cases} N_i, & \text{if } n_i > 0 \\ 0, & \text{if } n_i = 0 \end{cases}$ in cell i. The initial weight for panel returns in cell i is $W_{0i} = N_i / n_i$, for $n_i > 0$. We proportionally allocated the total number of non-represented returns $\sum_i M_i$ to other cells. So the

[10] These are the CWHS returns that were filed after the panel sample was selected. They were in SOIYR 2004, 2005 and 2006 cross-sectional samples.

adjusted weight for panel returns in cell i is $W_i = aW_{0i}$, where a is the adjustment factor and defined as $a = \sum_i N_i \Big/ \left(\sum_i N_i - \sum_i M_i \right)$. This is done for each out-year stratum. Table 5 illustrates the weight adjusting procedure. There are a total of $\sum_i N_i$ =572 population returns of which $\sum_i M_i$ =195 are non-represented by panel sample returns. Of these, twenty were not in the base-year population and 175 were in the base-year population and 'popped' in from base-year strata 3-22. The adjustment factor of a =572/(572-195) was applied to the initial weight for every return in this adjustment cell, resulting in the adjusted weights that are given in Table 5.

Table 5. Illustration of Weight Adjusting Procedure

Initial Adjustment Cell i		**Number of Returns**		**Number of non-represented Population Returns** M_i	**Initial Weight** W_{0i}	**Adjusted Weight** W_i
Base-Year Stratum	**Current-Year Stratum**	**Panel Sample** n_i	**Population** N_i			
New filers	0		20	20		
0	0	120	120	0	1.00	1.52
1	0	32	50	0	1.56	2.37
2	0	6	28	0	4.67	7.08
3-22	0		175	175		
23	0	7	34	0	4.86	7.37
24	0	28	50	0	1.79	2.71
25	0	95	95	0	1.00	1.52
Total		**288**	**572**	**195**		

One advantage of the above weighting procedures for high-income returns is that the possible bias from large panel weights of influential 'popper' returns is avoided. This is because the weighting procedure is exclusive of the panel weights. Instead, we make use of the known number of returns there should be in the high-income certainty strata.

Next, we move to returns with an income under $5 million. These are the non-certainty returns in each out-year. The weighting procedure here is to first calculate the initial weights using returns' selection probabilities and then adjust them using poststratification within each out-year stratum. The initial weights for returns in each of the three segments are straightforward. For panel returns in segment (1), the initial weights are the same return weights calculated using the joint probability approach described in section 3. In segment (2), the CWHS returns that were also in the TY1999 population were subject to the sample selection in both base-year and current-year. Therefore, their weights are calculated using the joint probability approach in equation (3.1). Specifically, we use the equation:

$$W_{CWHS99} = \frac{1}{(1/2000) + (1/W_{99}) - (1/2000)(1/W_{99})}, \qquad (4.1)$$

where W_{CWHS99} is the weight of CWHS returns that were also in the base-year population; W_{99} is the weight associated with the return's TY1999 panel selection, not the weight of cross-sectional sample. This is determined by looking up the weight table (Table 1) using the last two digits of the associated 'PSAMP.' The values of W_{CWHS99} are smaller than and dominated by the values of W_{99}. In segment (3), or CWHS returns that were not in the TY1999 population and represent new filers, the initial weight is the inverse of the selection probability of CWHS refreshment returns, i.e., 2000. Thus, the initial weights for return in non-certainty out-year strata are

$$W_0 = \begin{cases} W_R, & \text{if the return is in segment 1} \\ W_{CWHS99}, & \text{if the return is in segment 2} \\ 2000, & \text{if the return is in segment 3} \end{cases}. \qquad (4.2)$$

The initial weights are calculated using equations (3.1), (4.1) and (4.2).

After we obtain the initial weights, we apply poststratification within each out-year stratum. Since we make estimations for tax years, not filing years, constructing the population totals for poststratification would require much effort pulling out returns for the tax year from multiple population files. In addition, because we need to use income to classify the returns to out-year strata, the uncorrected income in the population could introduce errors. Therefore, we decide to use the estimates from cross-sectional samples.

We apply the poststratification such that the total adjusted weight is the same as the total cross-sectional weight in each out-year stratum. Note that the cross-sectional weight is the one used in the cross-sectional sample selection and different from the panel return weight in Table 1. Here are the details to apply poststratification. Let W_{0ij} be the initial weight and W_{ij}^* be the cross-sectional weight for return j in cell i. Note that initial weights W_{0ij} can differ for returns even within the same cell; and so too can the cross-sectional weights W_{ij}^*. Let M_{ij}^* be the non-represented cross-sectional weight for return j in cell i, $M_{ij}^* = \begin{cases} W_{ij}^*, & \text{if } W_{0ij} > 0 \\ 0, & \text{if } W_{0ij} = 0 \end{cases}$ for return j in cell i. The adjustment factor is defined as $a^* = \sum_{i,j} W_{ij}^* \Big/ \left(\sum_{i,j} W_{ij}^* - \sum_{i,j} M_{ij}^* \right)$. Therefore, we get the adjusted weight $W_{ij} = a^* W_{0ij}$. An example of the weight adjusting procedure using postatrtification is outlined in Table 6. The total cross-sectional weight in out-year stratum 3 is 11,917. The total unrepresented cross-section weight is 2,856 that include 458 from new filers and 2,398 from base-year strata 6-19 where there is no panel return. The adjustment factor is $a^* = 11{,}917/(11{,}917 - 2856)$. Using $W_{ij} = a^* W_{0ij}$, we take care of the unrepresented portion. Therefore the total of cross-sectional weights of combined sample returns is the same as the estimated population size within each out-year stratum.

Table 6. Illustration of Poststratification

Initial Adjustment Cell i		**Total Weight**		**Total Non-represented Cross-sectional weight** $\sum_j M_{ij}^*$
Base-Year Stratum	**Current-Year Stratum**	**Initial Weight** $\sum_j W_{0ij}$	**Cross-Sectional Weight** $\sum_j W_{ij}^*$	
New filers	3	0	458	458
6-19	3	0	2,398	2,398
0-5, 20-25	3	9,628	9,062	0
Total		**9,628**	**11,917**	**2,856**

We have thus been able to produce two separate sets of longitudinal and out-year cross-sectional weights. We were able to use the large number of certainty returns from SOI's yearly cross-sectional sample to trim the influential panel outliers. The trimming is intended to reflect the true number of influential panel returns that went through the similar income changes across years. The cross-sectional weight is produced so that the combined sample of panel returns and refreshment returns is representative of the current population and produces cross-sectional estimates that are consistent to those produced from SOI's cross-sectional sample for each out-year. The weights are adjusted every year to reflect the current population distribution (especially for high income returns) as much as possible.

As we have shown, the adjustment of longitudinal weights is mainly focused on trimming the weights for these influential 'poppers', or outliers We adjust the longitudinal weights based on the known number of influential 'poppers' in the panel sample and the number of similar 'poppers' in the population. However, no weighting adjustment can compensate for population 'poppers' not present in the panel sample. Cross-sectional weights of the combined sample are intended to represent the snapshot of each out-year population, estimated from the SOI's larger cross-sectional sample. Here, we have introduced adjusting return weights using poststratification.

5. References

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

Mathematica Policy Research (2001), "A Review of the Design of a Panel Sample of Individual Income Tax Return filers," *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.*