An Application of Calibration Approach in Weight Trimming for Stratum Jumpers

Yan K. Liu[1], Phillip S. Kott[2], Lance Harris[1]
[1] Statistics of Income division, IRS, P.O. Box 2608, Washington, DC 20013,
yan.k.liu@irs.gov
[2] RTI International, 6110 Executive Blvd., Rockville, MD 20852

Abstract
The Statistics of Income division of the IRS started a panel sample of individual returns in tax year 1999. This panel sample is also used for cross-sectional estimations, with a small supplemental sample added to it each year. The base-year panel sample was a stratified sample, where stratum boundaries were formed using the return income. Because the income distribution is highly skewed, base weights vary dramatically. This poses a particular problem for returns whose income grows so dramatically that its associated base weight is no longer appropriate for cross-sectional estimations. In addition to stratum jumpers, high-income returns from both new filers and surviving filers in out-years are not well represented by weights based on selection probabilities. In this paper, we consider a calibration approach to adjusting return weights, including trimming for influential stratum jumpers, so that the estimates conform to the out-year population. The adjusted weights are for multipurpose estimations and a few key variables are used in the calibration

Key Words: Calibration, Outlier, Stratum Jumper, Weight Trimming.

1. Introduction

The Statistics of Income division (SOI) started a panel sample of individual returns (Forms 1040, 1040A, and 1040EZ) in 1999, and the sampled returns have been followed ever since. When a return was selected for the 1999 sample, both the primary filer and the secondary filer became permanent panel members, while other members in the family were not included. These panel returns have been followed through subsequent tax years 2000 - 2008. The panel sample is used to study the Sales of Capital Assets such as stocks, bonds, mutual funds, property and other assets. There are two types of analyses using the panel sample; the longitudinal and the cross-sectional. The longitudinal analysis studies the behavior of taxpayers over time, which is the main purpose of the panel sample. However, it is not the topic of this paper. The cross-sectional analysis gives yearly cross-sectional estimates such as totals, means and percents. Proper weighting is important for cross-sectional estimates from the panel sample. Therefore, in this paper, we look at the weighting issue for the cross-sectional estimation.

1.1 Panel Sample Design

The panel sample started from Tax Year (TY) 1999, with the first year termed the *base-year*. The base-year panel sample was a stratified sample where the stratification was achieved by a return's income (either positive or negative) and 'degree of interest'. The 'degree of interest' is a four-level categorical variable for the tax modeling purpose. '1' is assigned to returns that are least interesting, with '4' assigned to those most interesting. Table 1 gives the summary of the panel sample design. The specified sample weight for

each return is the inverse of its selection probability. The actual weight is the actual sample size divided by the population size within each stratum. Because the Bernoulli method is used in selecting sample returns, the actual sample size is slightly different from the expected sample size in each stratum and therefore the actual weights are slightly different from the specified weights.

Table 1. Sample Design of the 1999 Panel Sample of Individual Returns

Stratum	Income Range	Degree of Interest	Specified Sampling Rate (%)	Specified Return Weight	Actual Return Weight
	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

1.2 Panel Sample Selection

First, it is important to know that SOI selects a special sample every year that includes returns with specific 4-digit endings of the taxpayers' primary social security numbers, or PSSN. It is called the **Continuous Work History Sample (CWHS)**. It is approximately a simple random sample since the 4-digit endings of an SSN can be considered random. Five specific final four digits are used for the CWHS[1], which represents 5 in 9,999 (the

[1] Starting from 2005, additional five sets of final four digits were added to selecting CWHS samples for other purposes. The panel sample analysis was not impacted by this change.

sequence 0000 is not used in assigning SSNs) population returns. In other words, the CWHS constitutes 0.05% random returns of the entire population of returns.

The base-year panel sample selection was a two-step procedure. First, all the CWHS returns were included in the base-year panel sample. Then, for returns not selected into the CWHS, the sample selection utilized the taxpayer's PSSN. An integer function of the PSSN, called the Transformed Taxpayer Identification Number (TTIN), was computed. The last five digits of the TTIN was a pseudo-random number. A return for which the pseudo-random number was less than the cutoff sampling rate multiplied by 100,000 was selected in the sample. The cutoff sampling rate needed to account for the approximately 5% CWHS returns already included in the sample. For example, if the desired sampling rate is 20%, then the cutoff sampling rate a is calculated by $a+(1-a)*0.05=0.20$, which results in $a=0.1579$.

1.3 Panel Member Following Rule

The first year is termed the *base-year*, while subsequent years are *out-years*. To identify panel returns to follow in out-years, both the primary social security numbers (PSSNs) and secondary social security numbers (SSSNs**)** in the base-year sample are matched against the PSSNs and SSSNs on all returns that post to the individual master file in each out-year. In other words, an out-year return is identified if its PSSN is matched to either the PSSN or SSSN of a base-year panel return; or if its SSSN is matched to either the PSSN or SSSN of a base-year panel return. 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).

1.4 The Yearly Supplemental Sample (CWHS) and Combined Sample

In addition to longitudinal analyses, the panel sample is also used for cross-sectional estimations. However, the panel sample becomes less representative of each out-year population over time. In particular, the panel sample does not include new filers, or enough newly-rich filers that are due to economic change. Adding appropriate supplemental returns is one way to deal with this. It would be desired that the supplemental sample can supplement with out-year high-income returns. But due to various resource constraints, the yearly supplemental sample includes only CWHS returns that are not already in the panel sample. In other words, returns whose PSSNs have the five specific 4-digit endings are selected in the supplemental sample, if they are not already in the panel sample. The yearly supplemental sample does not provide good representation for high-income returns. For example, no returns with an income over $2,000,000 were included in the TY2004 supplemental sample (See the following Table 3). The yearly supplemental sample does provide enough additional returns, including new filers, for low-income and middle-income strata. The CWHS supplemental sample and the panel sample together, called the **combined sample**, are used to make cross-sectional estimations for out-years. The combined sample includes surviving panel returns and CWHS supplemental returns.

1.5 Weighting Issues

The purpose of this paper is to develop the appropriate cross-sectional weights for the combined sample returns, in order to represent the out-year population. There are two major issues we hope to deal with through trimming and adjusting return weights for stratum jumpers and other high-income returns.

Weights of Stratum Jumpers**.** The panel sample was selected based on the 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 selection probabilities were different in the base-year sample. Returns with very low income in the base-year may end up with extremely high income in the out-year. Those are called 'poppers'. On the other hand, returns with a very high income in base-year may end up with an extremely low income in the out-year. These are called 'droppers'. Both poppers and droppers are stratum jumpers because they would have been assigned to another stratum had the out-year income been used for stratification. A problem arises when returns shift strata due to dramatic changes in yearly income[2]. Some stratum jumpers 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. Those returns with both large weights and large incomes can inflate the variance and cause estimation bias. Therefore, appropriate weight trimming for the extreme stratum jumpers should be considered.

Representation of High-Income Returns**.** The representation of high-income returns diminishes with time because many newly-rich are not included in the combined sample, especially in later years. The newly-rich could be either new filers that were not in the base-year population, or filers that were in the base-year population. For high-income new filers, their chances of being selected in the CWHS supplemental sample are very small, as the CWHS sample is a small simple random sample from the highly skewed population of out-year returns. For newly-rich filers that were in the base-year population, they had very small selection probabilities as their incomes were not high at that time. As a result, the combined sample of panel returns and CWHS supplemental returns does not provide enough representation to the tails of the out-year income distribution. Therefore, appropriate weight adjusting is used to compensate for the loss of the panel sample strength to represent high-income returns.

1.6 Goal of This Paper

In this paper, we try to adjust return weights to best represent the out-year population as much as possible. We first develop the weights for the combined sample of panel returns and supplemental returns based on their selection probabilities. We then consider weight trimming for the extreme stratum jumpers (poppers) and adjust weights for other returns, especially those with high out-year income. In trimming/adjusting weights, we throw in additional population benchmarks, and take into consideration the multi-purpose use of the weights. We are interested in the estimation for a few key variables such as AGI, Short Term Gain/Loss, Long Term Gain/Loss, Income, and Itemized Deductions. These key variables are not all highly correlated with each others; therefore, a set of weights that works well for the estimation of one variable may not be good for other variables. In order to find a set of compromised weights that balance all the key variables, we make use of the resources of population control totals on those key variables and apply the calibration approach. In the end, we have a weighted sample that reflects both the situations in the base-year and the current-year, as well as the balance of the estimates of a few key variables.

Organizationally, this paper is divided into six parts, with this introduction as Section 1. Section 2 looks at the calculation of cross-sectional weights of surviving panel returns for

[2] All incomes in out-years are adjusted so that they are comparable to the tax year 1999 income.

the cohort populations in out-years. This weight calculation procedure was developed by Mathematica Policy Research (2006) and based on the selection probabilities. Section 3 extends to the calculation of weights for combined sample returns based on selection probabilities, including how to incorporate the supplemental sample into the panel sample for cross-sectional estimation purposes. Section 4 provides details of the yearly cross-sectional sample of individual returns that is the source of auxiliary information for trimming and adjusting weights. Section 5 looks at the weight calibration approach and SUDAAN's WTADJUST procedure, as well as our applications. Finally, Section 6 gives the summary and discussion.

2. Cross-Sectional Weights of Panel Sample Returns Based on Selection Probabilities

The cross-sectional weights of out-year panel sample returns were constructed by the Mathematica Policy Research Inc. (2006). The weights of surviving panel returns in out-years were intended for the estimations of out-year cohort populations. The base-year return weight is simply the number of population returns divided by the number of sample returns within each stratum, which is the "Actual Return Weight" column in Table 1. The longitudinal person weight of each filer is set equal to its base-year return weight. By the panel following rule, an out-year panel return is linked to the base-year panel return by at least one SSN (either primary or secondary). Each matched primary or secondary filer on an out-year record received a longitudinal person weight from the matching base-year panel member. We used these person weights to calculate return weights for each out-year file.

Let W_1 denote the longitudinal person weight of the PSSN of an out-year return and W_2 denote the longitudinal person weight of the SSSN of the out-year return. Then W_1 is equal to the weight of the base-year return that is linked to the out-year PSSN; and W_2 is equal to the weight of the base-year return that is linked to the out-year SSSN. W_1 and W_2 may be different if they are from two different returns, but are the same if they are from the same return. Let W_R be the return weight of an out-year return. There are different scenarios in calculating W_R:

(1) If an out-year return is a single filer, the return weight is set equal to the primary filer's person weight, that is, $W_R = W_1$.
(2) If a joint out-year return includes two panel members from one base-year panel return, then $W_R = W_1$.
(3) If a joint out-year return includes two panel members from two different base-year returns, then the return weight W_R is calculated as a function of the primary and secondary person weights:

$$W_R = \frac{1}{(1/W_1) + (1/W_2) - (1/W_1)(1/W_2)} \tag{2.1}$$

In this weight calculation, the inverse of each person weight is treated as a selection probability. In other words, $(1/W_1)$ is treated as the selection probability of PSSN and $(1/W_2)$ is treated as the selection probability of SSSN.

(4) If a joint out-year return contains a panel member and a nonpanel spouse who did file a return in the base-year, then the return weight is still calculated using equation (2.1). The personal weight for the nonpanel spouse is assigned by taking into account its selection probability in the base-year. The SSN of the nonpanel spouse is first matched against the base-year population file[3]. Then the nonpanel spouse gets a weight based on its stratum membership in the base-year.
(5) If the out-year return contains a panel member and a nonpanel spouse who did not file a return in the base-year, then this spouse has a panel selection probability of zero. In that case, the return weight is equal to the personal weight of the panel member.

The above out-year cross-sectional weights of panel returns are developed based on their base-year selection probabilities. Technically, they can be used for estimations of the cohort population, which includes all the out-year returns that filed in TY1999. Practically, these weights become less and less representative of the cohort population over time because return income changes over time. Also, the returns may not be representative of the base-year sample design strata in which they fall, especially for some influential stratum jumpers. Therefore, weight trimming and adjusting should be performed, which is discussed in Section 4. But first, we will develop the weights for the combined sample of panel returns and supplemental returns based on selection probabilities since our interest is the estimate for the complete out-year population (not just the cohort population).

3. Cross-Sectional Weights of Combined Sample Returns Based on Selection Probabilities

A supplemental sample is added to the panel sample every year to support representative cross-sectional estimates. It simply includes all the out-year CWHS returns that are not already in the sample. The combined sample of surviving panel returns and supplemental sample returns is used to make cross-sectional estimations in each out-year. Developing an appropriate weighting scheme is the key to make the combined sample representative of the out-year population. In this section, we develop the selection probability of each out-year return in the combined sample and take the inverse as its return weight. We take into account the selection probability of the panel sample in the base-year and the selection probability of the supplemental sample in the out-year.

For a new return that did not file in the base-year, the selection probability is 0.0005. For a return that did file in the base-year, we need to count its selection probabilities in the base-year and in the out-year. Each primary or secondary filer on an out-year record receives a person weight. We then used these person weights to calculate return weights for each out-year return. The return probability is calculated by

$$P = P_1 + P_2 - P_{12}, \tag{3.1}$$

[3] The 1999, 2001 and 2002 population files were searched for non-panel spouses who might have been late filers. But the 2000 population file was not searched because it was not available. If a non-panel 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 since the number of missed matches should be small.

where P_1 is the probability of selecting the primary filer (PSSN), P_2 is the probability of selecting the secondary filer (SSSN), and P_{12} is the probability of choosing both. Primary and secondary refer to the status in the out-year return and in the combined sample. If an out-year filer in the combined sample also filed in the base-year, it carries a personal longitudinal weight from its base-year return, denoted by W_1 for the PSSN and W_2 for SSSN separately. Recall, W_1 is the weight of the base-year return whose SSN (either primary or secondary) is matched to an out-year primary filer (PSSN); and W_2 is the weight of the base-year return whose SSN (either primary or secondary) is matched to an out-year secondary filer (SSSN). Note that if the return is not a panel return, W_1 and W_2 must be determined by first searching the base-year population file to locate this filer and identify the SOI stratum. Then, a weight associated with each stratum can be obtained from Table 1. W_1 and W_2 are the same if they are from the same base-year return.

An out-year return is included in the combined sample either through the base-year panel sample selection or through the out-year supplemental sample selection. When calculating P_1, the selection probability of an out-year PSSN in the combined sample, we need to consider the conditional selection probability. At the base-year panel sample selection stage, a return in the base-year population had a probability of $1/W_1$ to be selected, with a probability of $(1\text{-}1/W_1)$ to not be selected. For an out-year return, the selection probability of a PSSN is 0.0005 if it is not matched to any return in the base-year population and $1/W_1$ if it is matched to a PSSN in the base-year population. If the PSSN is matched to a SSSN in the base-year population, its selection probability is 1 if it was in the base-year panel sample; and 0.0005 if it was not in the base-year panel sample. That is,

$$P_1 = \begin{cases} 1/W_1, \textit{ if it is matched to a base-year PSSN} \\ 1/W_1 + 0.0005(1-1/W_1), \textit{ if it is matched to a base-year SSSN} \\ 0.0005, \textit{ if it is a new filer} \end{cases} \quad (3.2)$$

The selection probability of an SSSN in the combined sample is only through the base-year panel selection because SSSN is not used for the out-year supplemental sample selection. That is,

$$P_2 = \begin{cases} 1/W_2, \textit{ if it is matched to a base-year SSN (either PSSN or SSSN)} \\ 0, \quad \textit{if it is a new filer} \end{cases} . \quad (3.3)$$

The joint probability of selecting both the PSSN and SSSN of an out-year return is

$$P_{12} = \begin{cases} 0, & \textit{if it is a new return} \\ 1/W_1, & \textit{if both are from one base-year return} \\ P_1 \times P_2, & \textit{if they are from two different returns} \end{cases} . \quad (3.4)$$

Let d denote the weight of returns in the combined sample based on selection probability, then $d = 1/P$. We call this the ***initial weight*** to distinguish it from the ***final weight,*** w, after trimming/adjusting.

4. Auxiliary Information from the Yearly Cross-Sectional Sample of Individual Returns

We choose the calibration approach as our re-weighting method because we have the rich information of auxiliary variables. SOI selects a cross-sectional 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 is much larger than the panel sample and provides a good representation of the current population. While this yearly sample is used for various cross-sectional studies, including the study of items on Form 1040 Sales of Capital Assets (SOCA) on a tax return basis, it does not provide detailed information about each transaction reported on the tax returns using Schedule D and other forms. This is due to the high processing cost associated with the editing. That is why the surviving panel sample returns are used to study the cross-sectional SOCA at the transaction level (in addition to longitudinal analysis). The weights are based on the return level though and are adjusted to represent the out-year population.

The yearly cross-sectional individual return sample is also a stratified random sample, but the stratum definition is different from that of the panel sample. The stratification is achieved by the return type code, as shown in Table 2, and the same income range of the panel sample stratification, as shown in Table 1. The final stratification is achieved by the combination of return type code and income code. The sample consists of two parts: a CWHS, and a Bernoulli sample that are selected using the same two-step method as that of panel sample. The sampling rates are much larger than those of the panel sample. In fact, all returns with income $5 million or more (panel sample strata 0, 1, 2, 23, 24 and 25 in Table 1) are taken with certainty. Therefore, we have the known control totals for high-income strata. The known totals of the number of returns and key variables are used to calibrate original weights to account for all high-income returns, including stratum jumpers and the newly-rich. For the returns with income under $5 million, the estimated stratum totals are used as control totals. These estimated stratum totals are considered stable because of the high sampling rates. The following Table 3 gives the comparison of sample sizes by stratum between the cross-sectional sample and the combined sample. The total weights of the cross-sectional sample match the numbers of population returns.

Table 2. Return Type Code

Return Type Code	Special Category
1	High income nontaxable returns
2	Large Business Receipts
3	Form 2555 (foreign earned income)
4	Form 1116 + Schedule C or F
5	Form 1116 (foreign tax credit)
6	Schedule C and Schedule F
7	Schedule C (nonfarm sole proprietors)
8	Schedule F (farm sole proprietors)
0	All Others

Table 3. Number of Returns – Comparison between the Cross-Sectional Sample and the Combined Sample (TY2004)

Out-Year Stratum	Out-Year Income Range	# Returns, Cross-Sec Sample	# Returns Combined Sample		Total Weight, Combined Sample	Total Weight, Cross-Sample (Pop Size)
			Panel	Supp		
	NEGATIVE INCOME					
0	$20,000,000 or more	619	317	0	431	619
1	$10,000,000 - under $20,000,000	998	264	0	753	998
2	$5,000,000 - under $10,000,000	2,901	457	0	2,269	2,907
3	$2,000,000 - under $5,000,000	4,094	864	0	9,631	11,917
4	$1,000,000 - under $2,000,000	4,285	776	0	26,134	25,665
5	$500,000 - under $1,000,000	2,456	729	0	69,949	65,886
6	$250,000 - under $500,000	1,738	681	3	154,346	153,838
7	$120,000 - under $250,000	1,600	496	16	338,766	317,817
8	$60,000 - under $120,000	1,257	363	21	454,975	460,419
9	Under $60,000	2,245	740	163	1,554,877	1,525,071
	POSITIVE INCOME					
10	Under $30,000	13,710	6,878	6,755	26,558,141	26,561,252
11	Under $30,000	16,490	13,300	4,184	31,683,780	31,637,170
12	Under $30,000	10,702	4,358	1,283	10,134,274	10,132,485
13	$30,000 - under $60,000	11,989	12,269	1,154	23,595,204	23,700,300
14	$30,000 - under $60,000	10,978	5,293	445	10,060,981	10,009,692
15	$60,000 - under $120,000	7,133	7,959	315	14,059,717	13,977,947
16	$60,000 - under $120,000	6,375	3,513	155	5,915,021	5,930,782
17	$120,000 - under $250,000	4,069	1,312	50	1,993,213	1,913,195
18	$120,000 - under $250,000	11,737	3,070	79	3,580,654	3,621,478
19	$250,000 - under $500,000	11,992	2,910	29	1,402,370	1,473,758
20	$500,000 - under $1,000,000	12,906	2,691	9	520,102	482,282
21	$1,000,000 - under $2,000,000	19,475	2,350	1	159,685	156,515
22	$2,000,000 - under $5,000,000	20,734	2,507	0	50,822	63,472
23	$5,000,000 - under $10,000,000	14,799	1,529	0	14,723	14,824
24	$10,000,000 - under $20,000,000	5,712	1,159	0	8,590	5,712
25	$20,000,000 or more	3,031	1,256	0	5,078	3,031
Total		**204,025**	**78,041**	**14,662**	**132,354,488**	**132,249,031**

5. Calibrating Weights Using WTADJUST procedure

Weight calibration is a method to adjust sampling weights using auxiliary information. Let d_k be the initial weight of return k in the combined sample, and w_k be the calibration weight of return k. The calculation of d_k in our application is based on the selection probabilities of return k and discussed in Section 3. The calculation of w_k is through the weight calibration procedure (see, e.g., Särndal, 2007; Kott, 2009). The weights go through an iterative process of adjustments until convergence at the predefined population totals. We use SUDAAN's WTADJUST procedure to calculate calibration weight, w_k. In this section, we first briefly describe the weight calibration procedure. Then we look at our application WTADJUST procedure and discuss the results.

Calibration is a weight-adjustment method that creates a set of weights, $\{w_k\}$, such that (1) they are close to the original design weights, d_k (as the sample size grows arbitrarily large, w_k converges to d_k), and are therefore nearly unbiased under the randomization distribution; and (2) satisfy a set of calibration equations:

$$\begin{aligned} \sum_S w_k &= N \\ \sum_S w_k \mathbf{x}_k &= \sum_U \mathbf{x}_k \end{aligned}, \tag{5.1}$$

where N and $\sum_U \mathbf{x}_k$ are the known control totals. There is one calibration equation for each auxiliary variable.

To calculate calibration weight, w_k, we use SUDAAN's WTADJUST procedure that is based on a generalized exponential model (SUDAAN Language Manual, Release 10.0). This procedure allows separate weight adjustment factor α_k for each k such that

$$w_k = \alpha_k d_k . \tag{5.2}$$

In our application, we use the default of SUDAAN's WTADJUST procedure that reduces the generalized exponential model to the following standard exponential model:

$$\alpha_k = \exp(\mathbf{x}_k^{'} \beta) . \tag{5.3}$$

Here, $\mathbf{x}_k^{'}$ is the design matrix that includes intercept and auxiliary variables; and β is the vector of model parameters that will be estimated within the procedure.

Five auxiliary variables are of interest and listed in the order of their importance: AGI (x_1), Short Term Gain/Loss (x_2), Long Term Gain/Loss (x_3), Income (x_4), and Itemized Deduction (x_5). However, the calibration model treats all the variables in the model as equally important. AGI and Income are moderately correlated, while the other variables are not correlated with each other. As shown in Table 3, we have the reliable population totals in each out-year stratum and would like to apply calibration method within each out-year stratum. In other words, for surviving panel returns and supplemental returns that fall in the same out-year stratum, we adjust their initial weights so that the estimated totals match the known population totals within the out-year stratum. The initial weights within each out-year stratum could vary a lot since those returns could be from different design strata and the supplemental sample. If we use the term **Calibration Group** to denote the group within which the weight calibration is applied, each of out-year strata 1 – 24 is a calibration group. For out-year strata 0 and 25, because the income distributions are highly skewed, we further divided each into 10 calibration groups based on the value of return income, as shown in Table 4. Then we applied calibration method within each calibration group except for two open-ended groups where absolute value of income was over \$400 million[4].

[4] We do not adjust weights for these two groups because of the extremely large incomes. Data users need to deal with them using subject knowledge or making use other data sources.

Table 4. Calibration Groups of Top Two Out-Year Strata (TY2004)

Absolute Value of Income Range (Million $)	Negative Income (Out-Year Stratum 0)			Positive Income (Out-Year Stratum 25)		
	Number of Returns in Combined Sample	Total Weight of Returns in the Combined Sample	Number of Returns in the Population	Number of Returns in Combined Sample	Total Weight of Returns in the Combined Sample	Number of Returns in the Population
Over 400	*Masked* *			*Masked* *		
350 - 400						
300 - 350						
250 - 300						
200 - 250						
150 - 200						
100 - 150				70	2103	110
80 - 100				61	124	102
60 - 80	32	39	60	109	145	195
40 - 60	55	78	107	245	565	510
20 - 40	187	265	381	698	2061	2017

** Numbers are masked to protect taxpayers' information.*

SUDAAN's WTADJUST procedure allows the trimming of initial weights (d_k) before the calibration adjustment using the option WTMAX and WTMIN. We first looked at the weight distribution within calibration group and truncated the initial weights for a couple of outliers. For example, there are 70 returns in the calibration group where the return income is in the range of ($100, $150) million (see Table 4); two returns have significantly larger weights than the rest, as shown in Table 5. The total of initial weights is 2103, while the known total number of returns is 110. Therefore, we set their initial value to be 2.18, the largest weight after removing the two top weights, before the calibration (i.e., WTMAX=2.18) is applied. The choice of the initial weight cutoff was ad hoc here. A similar ad hoc procedure was used to other calibration groups. Although calibration would adjust weights to conform to control totals, different choices on the initial weight cutoff would result in different calibration weights as the calibration procedure strives to achieve the minimum distance between w_k and d_k.

Table 5. Initial Weight Summary of the Calibration Group ($100 Million, $150 Million)

Value of Initial Weight	**Number of Sample Returns**	**Total Initial Weight**
1.00	54	54.00
1.77	13	23.01
2.18	1	2.18
17.28	1	17.28
2006.87	1	2006.87
Total	**70**	**2103.34**

In addition to putting a cap on outliers of initial weights, we also want to limit the final weights in a reasonable range, especially a lower bound to be at least 1. SUDAAN's WTADJUST procedure does not have the option to directly set lower and upper bounds on final weights (w_k), though there are indirect ways to handle this. We started out by letting the calibration weights go as low as they want, and as high as they want. The calibration did not converge if all five variables were included in the model, but did converge for the first three variables. In other words, the weighted totals on the Number of Returns (n), AGI (x_1), Short Term Gain/Loss (x_2),, and Long Term Gain/Loss (x_3), matched the control totals.

Then we turned to the issue on bounds of calibration weights. That is, some calibration weights were smaller than 1, while a few were undesirably larger than the maximum initial weight of 2,150 (all in low income groups). To solve this problem, we chose to perform the second round of calibration on the already calibrated weights and only included the one most important variable (x_1) in the calibration model. We used the option WTMIN=1 and WTMAX=2,150 in each calibration group. The second-round calibration weights were further tuned and only a few of them were slightly smaller than 1. The few large first-round calibration weights were also brought down to near 2,150. Since we wanted the final weights to be at least 1, for those few weights that were still below 1, we forced them to be 1 and generated an offsetting adjustment among the remaining returns in order to preserve the group population total. We did this by apportioning an offsetting weight decrease among the remaining returns in the same calibration group. Note that by cycling back to the highest ranked variable (x_1), the weighted totals on the other two variables (x_2 and x_3) no longer match their control totals, but are still much closer than those using the initial weights.

The following Figure 1 gives the scatterplot of final calibration weights versus the initial weights for calibration group 1 where the out-year return income is between $10,000,000 and $20,000,000. Two weights have major changes, while others are slightly adjusted upwards (above the 45-degree line). The largest weight of 67 is trimmed to 32, while another weight of 23 is increased to 45. Figure 2 is the scatterplot for the calibration group 5 where the out-year return income range is ($500,000, $1,000,000). The calibration weights in this group are adjusted around their initial weights with no major decrease or increase.

Finally, we look at comparisons of relative errors of totals estimated from initial weights and from calibration weights for a few key variables. The numbers in the Table 6 and Table 7 are the error of estimates in the percent of the population total, where the error is the difference between the estimated total and the population total. Table 6 includes the three variables in the first round calibration model. Table 6 shows that errors of estimates from the calibration weights are significantly smaller than those from the initial weights. Since we did the second round calibration on AGI, the calibration AGI totals by stratum are extremely accurate, while the minor difference from the population in some strata are caused by forcing the final weights to be at least 1. Also note that two open groups where the return incomes are too large, we did not apply any weight adjustment. Table 7 gives the same comparison for two variables that were not included in calibration model. Again, the numbers show that there is a significant improvement in the estimations of totals using calibration weights over using initial weights, especially in some strata that have outliers.

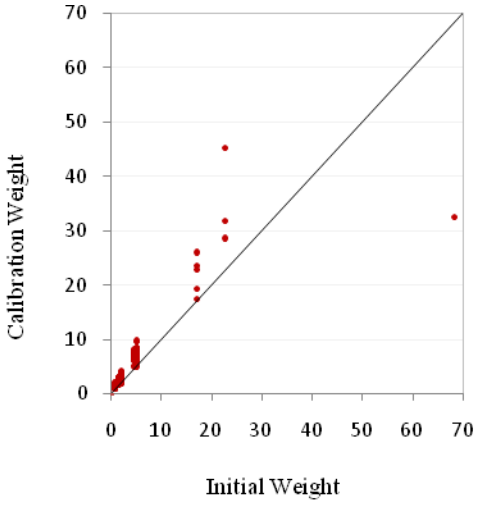

Figure 1. Scatterplot of Final Weight Vs. Initial Weight for Calibration Group **1**

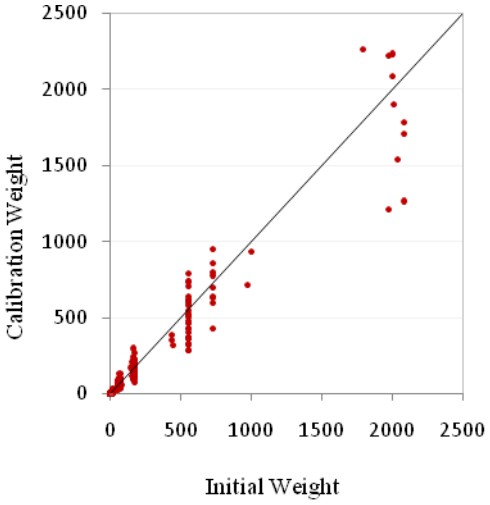

Figure 2. Scatterplot of Final Weight Vs. Initial Weight for Calibration Group **5**

6. Summary

This paper introduces the use of weight calibration on the panel data for the cross-sectional purpose. The accuracy of estimates is greatly improved using a weight calibration that borrows strength of auxiliary information. Since we have the rich information on population benchmarks, we are comfortable about how the outlier weights are trimmed. The estimates using the calibration weights have much less errors, and are especially good for the calibration variables.

There are some limitations. While extensive work was done to produce the most accurate cross-sectional tabulations, little effort was put into making them longitudinally consistent. The weighting process is not related to returns' behavior over time. We just look at a snapshot of the population. But again, this is mainly for the cross-sectional purpose. The other limitation is that for extremely large newly-rich returns in the current population, there may not be comparable returns to represent them, either from the panel sample or the supplemental sample. For example, the returns with an income over $400 millions in tax year 2004 belong to this category and are put in two top groups in Table 6 and Table 7. In this situation, we cannot make it up through adjusting weights. Even though the number of such returns is small, we need to be cautious since a few top returns can be so large and so influential, and can badly bias the estimates if not treated correctly.

References

Kott, Phillip. "Calibration Weighting: Combining Probability Samples and Linear Prediction Models." *Handbook of Statistics Sample Surveys: Inference and Analysis* 29B (2009): 55-82.

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

Särndal, Carl-Erik. "The Calibration Apporach in Survey Theory and Practice." *Survey Methodology* 33 No. 2 (December, 2007): 99-119.

SUDAAN Language Manual, Release 10.0

Table 6. Comparison of Relative Errors of Totals Estimated by Initial Weights and by Calibration Weights for Calibration Variables (%)

Out Year Stratum	Out-Year Income Range ($1000)	AGI		Short Term Gain/Loss		Long Term Gain/Loss	
		Initial	Calibrat	Initial	Calibrat	Initial	Calibrat
	NEGATIVE INCOME						
0	$400,000 or more	NA					
0	$40,000 - under $400,000	-1.0	1.7	-49.1	-0.6	-61.3	-1.0
0	$20,000 - under $40,000	-23.4	0.0	-39.6	0.0	-59.0	-0.1
1	$10,000 - under $20,000	-32.5	0.0	-12.8	0.2	-44.3	-0.2
2	$5,000 - under $10,000	-12.1	0.0	-54.7	-0.1	-5.1	0.0
3	$2,000 - under $5,000	-26.8	0.0	-14.2	0.0	-19.1	0.0
4	$1,000 - under $2,000	40.9	0.0	7.9	0.0	-41.3	0.0
5	$500 - under $1,000	-198.4	0.0	30.7	0.2	-13.6	-0.5
6	$250 - under $500	-615.8	-0.1	-9.8	-5.0	18.6	0.2
7	$120 - under $250	-125.2	0.0	-12.6	-2.1	-0.2	0.9
8	$60 - under $120	18.8	0.0	-9.2	-2.7	8.6	1.1
9	Under $60	-65.5	0.0	6.4	1.4	12.8	1.9
	POSITIVE INCOME						
10	Under $30	0.2	0.0	0.0	0.0	0.0	0.0
11	Under $30	0.1	0.0	11.4	0.0	5.9	0.0
12	Under $30	-0.4	0.0	3.5	0.0	18.9	0.0
13	$30 - under $60	-0.5	0.0	-0.4	0.0	-5.7	-0.6
14	$30 - under $60	0.4	0.0	25.3	1.5	-186.1	-11.2
15	$60 - under $120	0.7	0.0	12.1	0.1	-2.7	-0.1
16	$60 - under $120	-0.9	0.0	-12.0	-4.2	31.0	4.9
17	$120 - under $250	4.9	0.0	48.5	1.3	-1.7	-0.2
18	$120 - under $250	-1.3	0.0	2.8	0.1	4.4	0.1
19	$250 - under $500	-4.7	0.0	-12.7	-0.1	-10.9	-2.5
20	$500 - under $1,000	4.2	0.0	-5.2	0.0	4.0	0.0
21	$1,000 - under $2,000	6.6	0.0	34.1	0.9	9.7	0.0
22	$2,000 - under $5,000	-19.8	0.0	-880.7	-1.5	-29.8	0.0
23	$5,000 - under $10,000	0.6	0.0	227.4	1.0	11.8	0.0
24	$10,000 - under $20,000	44.9	0.0	7.2	-3.0	48.6	-0.5
25	$20,000 - under $40,000	2.4	0.0	-40.2	-3.2	6.3	-0.1
25	$40,000 - under $400,000	327.6	-0.1	-92.2	-17.9	9.3	-0.4
25	$400,000 or more	NA					

Table 7. Comparison of Relative Errors of Totals Estimated by Initial Weights and by Calibration Weights for Key Variables That Were not in the Calibration Model

Out Year Stratum	Out-Year Income Range ($1000)	Income		Itemized Deduction	
		Initial	Calibration	Initial	Calibration
	NEGATIVE INCOME				
0	$400,000 or more	NA			
0	$40,000 - under $400,000	-28.1	1.6	-12.2	52.6
0	$20,000 - under $40,000	-31.1	-0.7	-20.3	32.0
1	$10,000 - under $20,000	-25.9	-0.4	10.9	44.9
2	$5,000 - under $10,000	-22.0	0.6	-19.1	24.7
3	$2,000 - under $5,000	-16.4	3.7	-13.8	2.4
4	$1,000 - under $2,000	3.8	1.0	-1.6	5.5
5	$500 - under $1,000	7.0	1.3	42.3	24.3
6	$250 - under $500	-0.3	-0.4	5.4	-7.9
7	$120 - under $250	6.6	-0.1	-8.0	-4.4
8	$60 - under $120	-1.0	0.0	17.1	14.6
9	Under $60	6.3	-0.4	9.8	9.5
	POSITIVE INCOME				
10	Under $30	0.2	0.0	0.0	0.0
11	Under $30	0.1	0.0	5.1	4.9
12	Under $30	-1.0	-0.8	-3.5	-3.1
13	$30 - under $60	-0.4	0.0	-0.4	0.1
14	$30 - under $60	0.4	-0.1	-2.9	-3.4
15	$60 - under $120	0.6	0.0	0.4	-0.2
16	$60 - under $120	-0.2	0.2	-2.6	-1.6
17	$120 - under $250	3.9	-0.4	6.3	1.5
18	$120 - under $250	-1.4	-0.3	-0.4	1.0
19	$250 - under $500	-4.9	-0.1	-3.6	1.2
20	$500 - under $1,000	6.5	-0.7	3.3	-1.8
21	$1,000 - under $2,000	4.9	1.8	-1.4	-5.2
22	$2,000 - under $5,000	-20.5	0.1	-11.1	13.6
23	$5,000 - under $10,000	-2.1	-1.4	-1.3	1.9
24	$10,000 - under $20,000	40.9	-1.5	21.5	3.3
25	$20,000 - under $40,000	2.0	0.4	31.0	34.7
25	$40,000 - under $400,000	283.9	0.1	74.0	17.7
25	$400,000 or more	NA			