Improving Techniques to Use Panel Data to Produce Cross-Sectional Estimates

Yan K. Liu, Michael Strudler, Janette Wilson, Young Lim
Statistics of Income, Internal Revenue Service
77 K Street, NE, Washington, DC 20002.

Abstract

The Statistics of Income Division (SOI) of the IRS started a panel sample of individual returns in 2007 for longitudinal analyses. SOI edits Sales of Capital Assets (SOCA) transactions reported on individual tax returns for this panel. This panel sample, together with a small yearly refreshment sample has also been used for cross-sectional SOCA estimations. This has allowed us to get annual SOCA estimates with a small expense every year. The cross-sectional weights are derived using weight calibration method. The budget allows SOI to select and edit a SOCA cross-sectional sample once every five years, most recently in 2012. This paper compares the 2012 SOCA estimations based on the panel sample to those based on the 2012 SOCA cross-sectional sample and learns how to better weight the panel returns for SOCA cross-section purposes.

Key words: calibration weighting, cross-sectional estimation, panel sample, R.

1. Introduction

The Statistics of Income Division (SOI) of the IRS started a new panel sample of individual returns in Tax Year 2007. This panel sample is a stratified random sample selected from the population of returns filed for Tax Year 2007. The longitudinal weights refer to the 2007 population and are adjusted to take into account the attrition of the sample over time. The longitudinal weights are used when performing analyses of the longitudinal data through years. An important part of analyses is around the Sales of Capital Assets (SOCA) data. For this, SOI records and edits Sales of Capital Assets (SOCA) transactions reported on different tax forms. There are seven different tax forms that need to be looked at in order to compile SOCA data per return. Obviously, the cost to process SOCA data is very high, especially for large returns that tend to have many transactions.

SOI also produces SOCA cross-sectional estimates every year. But due to the high cost, a cross-sectional SOCA sample was selected only once every five years in 2007 and 2012. For years 2008 – 2011, SOI used the panel sample with a small yearly supplemental sample in order to count for new returns and high-income returns. In other words, the surviving panel sample and yearly supplemental sample returns form the sample that is used to make SOCA cross-sectional estimations every year. We call this sample the **panel-converted SOCA sample**. The selection probabilities of returns could not be calculated in a strict mathematical form and therefore the weights could not be defined as the inverse of selection probability. Instead, an ad hoc weighting scheme is used to produce initial weights and then the weight calibration method is applied to adjust initial weights such that the cross-sectional weights can reflect the current year population and

some known SOCA benchmarks. How good are the estimates from panel-converted samples? This paper will compare estimates from the 2012 panel-converted SOCA sample to those from the 2012 cross-sectional SOCA sample, learn how to better weight the panel returns for SOCA cross-section purposes, and modify the future selection of the yearly supplemental sample.

2. 2012 SOCA Samples

The 2007 panel sample was a stratified random sample and the stratum was defined by the selection income[1] of tax return (See Liu *et al*, 2009). A base-year panel return stays in the panel in the following years if either its primary SSN, secondary SSN, or both file tax returns in the out-year. These are referred as "surviving panel returns".

The panel-converted SOCA sample consists of surviving returns of the 2007 SOCA panel sample and the out-year refreshment sample returns. The out-year refreshment sample includes three parts. The first part is a very small, stratified random sample with the same stratum definition as in the 2007 base-year panel sample. The purpose of this part is to add high-income returns (especially stratum jumpers) and new filers that were not already in the panel sample. The second and third parts[2] are added at no cost as these returns are being edited for another SOI sample. However, these returns do not have much impact on the SOCA estimation as they are small returns and barely have SOCA activities. The panel-converted SOCA sample was used to make SOCA estimations for years 2008 – 2011.

The 2012 **SOCA cross-sectional sample** is a stratified random sample and selected once every five years due to the cost constraint. It includes much more large returns than the panel-converted SOCA sample. The stratum boundary was the same as the panel sample, using the 2012 selection income (not the 2007 selection income used in the panel sample), with an additional stratifier: the means of filing (electronic filing and paper filing). Two SOCA cross-sectional samples have been selected, one in 2007 and one in 2012.

Table 2.1 compares the number of returns by stratum among the panel sample, the panel-converted sample (i.e., panel returns plus refreshment returns), the SOCA cross-sectional sample and the population for 2012. Column E is the number of refreshment sample returns. Column G is the difference in sample size between the SOCA cross-sectional sample and the panel-converted sample. Because not all of panel-converted sample returns overlap with the SOCA cross-sectional sample, the extra number of returns that need to be edited for the SOCA cross-sectional sample is actually more that in Column G. Strata with selection income under $500,000 in both positive and negative are classified as less important in the analysis as small returns have very small contribution to SOCA estimates,. For high-income strata with the income over $500,000 in both positive and negative, as shown in column G, the SOCA cross-sectional sample is much larger than the panel-converted sample, while the refreshment sample is very small as shown in column E.

[1] It is the size of total gross income that is calculated from multiple tax items.
[2] The second part is a simple random sample of primary taxpayers who did not file tax returns for Tax Year 2007 but did for the current study year. The third part is a simple random sample of secondary taxpayers (i.e., married and filing jointly) and was not part of the base-year panel.

Table 2.1 Number of Returns of the 2012 Samples and the Population

Selection Income ($1000)	Population	Panel Sample	Panel Converted SOCA Sample	Refreshment Sample	SOCA Cross Sectional Sample	Difference in Size between SOCA cross sample and panel converted sample
A*	B	C	D	E=D-C	F	G=F-D
Negative Income						
$150,000 or more	58	41	58	17	58	0
$40,000 to < $150,000	322	186	196	10	322	126
$20,000 to < $40,000	624	242	257	15	618	361
$10,000 to < $20,000	1,808	454	480	26	1,095	615
$5,000 to < $10,000	4,941	633	694	61	1,461	767
$2,000 to < $5,000	18,996	912	1,009	97	5,220	4,211
$1,000 to < $2,000	39,168	738	786	48	5,254	4,468
$500 to < $1,000	90,305	668	718	50	3,307	2,589
$250 to < $500	193,736	666	708	42	2,083	1,375
$120 to < $250	376,018	768	815	47	1,940	1,125
$60 to < $120	500,951	827	896	69	1,549	653
Under $60	1,168,982	1,392	1,624	232	2,146	522
Positive Income						
Under $30	28,893,838	17,144	31,964	14,820	29,111	-2,853
Under $30	35,899,253	38,373	48,177	9,804	35,972	-12,205
Under $30	12,762,955	14,597	17,694	3,097	12,943	-4,751
$30 to < $60	24,096,932	35,881	38,785	2,904	24,100	-14,685
$30 to < $60	11,153,134	18,297	19,438	1,141	11,434	-8,004
$60 to < $120	14,585,202	26,531	27,596	1,065	15,404	-12,192
$60 to < $120	6,259,443	11,165	11,695	530	6,854	-4,841
$120 to < $250	1,999,423	3,821	3,987	166	6,802	2,815
$120 to < $250	4,283,865	8,528	8,873	345	15,086	6,213
$250 to < $500	1,707,012	4,537	4,730	193	13,278	8,548
$500 to < $1,000	579,818	3,220	3,331	111	13,814	10,483
$1,000 to < $2,000	197,265	2,676	2,783	107	10,983	8,200
$2,000 to < $5,000	82,501	2,659	2,835	176	8,991	6,156
$5,000 to < $10,000	19,722	1,649	1,777	128	5,573	3,796
$10,000 to < $20,000	7,550	1,224	1,330	106	5,948	4,618
$20,000 to < $40,000	2,914	795	883	88	2,492	1,609
$40,000 to < $150,000	1,502	723	845	122	1,502	657
$150,000 or more	228	164	228	64	228	0
Total	**144,928,465**	**199,511**	**235,192**	**35,681**	**245,568**	**10,376**

**. The grouping by the selection income is taken from the annual SOI Individual Cross-Sectional Sample that is stratified by the selection income and an indicator. The multiple lines of same income group in Column A is due to the indicator.*

3. Weighting

The 2012 population includes all individual returns filed in 2012. The panel-converted sample is used to make cross-sectional estimates of population totals on SOCA measures. The proper weighting of panel-converted sample is necessary to produce good estimates. In this section, we briefly outline the weight development for returns of the 2012 panel-converted sample. The procedure is similar every year. We make use of the limited SOCA information from another SOI sample, the yearly cross-sectional individual return sample. This sample includes some SOCA measures from the Schedule D at the return

level. Population totals on a few SOCA measures such as the net long-term capital gain and loss, the net short-term capital gain and loss, the net long-term capital gain/loss from sales of capital assets, and etc. are estimated from the 2012 individual sample of 338,350 returns. The initial weights are calculated using an ad hoc method and then adjusted using the calibration method. The final weights are calibration weights such that weighted estimates from the panel-converted sample match these estimated population totals on a few key SOCA measures.

Note that the 2012 SOCA cross-sectional sample plays no role in developing weights for the panel-converted sample returns. The SOCA estimates from this sample are used for comparison purpose here.

First, the initial weights are calculated using an ad hoc method. The initial weights cannot be the inverse of selection probabilities because the selection probabilities of returns in the panel-converted SOCA sample could not be calculated in a strict mathematical form. The ad hoc method is to pool all sample returns and post-stratify them into groups and then take the average weight for every return within the group. Because the panel-converted sample includes secondary CWHS returns that are married and jointly filing returns (Liu *et al*, 2009), married and jointly filing returns are over-represented. Therefore, initial sample weights are calculated separately for married and jointly filing returns and other returns. Specifically, within each income group shown in Table 2.1, returns are further divided by EMARS=2 (married and jointly filing returns) and EMARS≠2 (other filing statuses). The ad hoc initial weight within each post-stratified cell h is the population size (N_h) divided by the number of sample returns (n_h), that is (N_h/n_h).

Then, the initial weights are adjusted using the weight calibration approach such that the sample estimates from calibration weights are close to the population totals of auxiliary variables. The benchmark population totals are estimated from the yearly individual return cross-sectional sample. This yearly individual return sample is large and has the SOCA information at the return-level from the IRS Form 1040 Schedule D, while the SOCA estimates are calculated from the detailed SOCA Data Compiled from 7 Tax Forms (Schedule D / Form 8949, Forms 4797, 6252, 6781, 8824, 4684, & 2439) and at the transactional-level. The auxiliary information considered in the weight calibration procedure includes a few key SOCA measures, while the final weights are used to estimate population totals on thousands of SOCA measures. However, adjusting initial weights and matching totals of some a few key SOCA measures can still help making SOCA better estimates than other choices (Liu, Henry and Strudler, 2012). The goal here is to produce reasonable SOCA estimates that represent the current year population and reflect multiple variables of interest.

Let d_k be the initial weight of return k in the panel-converted sample; and w_k be the calibration weight of return k. 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 R software to calculate calibration weight w_k. 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:

$$\begin{aligned} \sum_S w_k &= N \\ \sum_S w_k \mathbf{x}_k &= \sum_U \mathbf{x}_k \end{aligned}, \tag{3.1}$$

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

The SOCA data are highly dispersed, so the SOCA sample data are divided into calibration groups to reduce the variability of key variables, while group sample sizes are large enough for calibration. The above calibration weight procedure is applied within each calibration group. The variables on schedule D are SOCA variables used for calibration, including Net short-term gain/loss from Sales of Capital Assets, Net short-term capital gain/loss, Net long-term gain/loss from Sales of Capital Assets, Net long-term capital gain/loss, etc.. The benchmark totals $\mathbf{x}_k$ are estimated from the yearly individual return sample that includes return-level SOCA information on Form 1040 Schedule D. More details on weighting the panel-converted sample returns can be found in Liu, Henry and Strudler (2012).

4. SOCA Estimates

The SOCA estimates include capital gains, losses and net gains, as well as number of returns and number of transactions, overall and by different categories. These estimates are summarized in the SOI's Sales of Capital Assets Data Report (Wilson & Liddell, 2016) in multiple tables: Tables 1A-1C, Tables 2A-2E, Tables 3A-3E and Tables 4A-4E. These are the key tables in the SOCA report and the numbering of those tables is not related to the numbering of tables in this paper. The following snapshots are part of these SOCA report tables. Table1A gives estimates of the long-term and short-term capital gain/loss overall and for different asset types. Table 1B and Table 1C are estimates for the short-term only and for the long-term only, broken down by the same asset types. Tables 2A-2E are estimates for selected asset types and broken down by the AGI group. Tables 3A-3E are estimates for the same selected asset types as Table 2A-2E, but broken down by the transaction's month of sale. Finally, Tables 4A-4E are also estimates for the same selected asset types as Table 2A-2E, but broken down by the transaction's length of time held.

There are 3,854 cells from SOCA report Tables 1A-1C, 2A-2E, 3A-3E and 4A-4E). The estimate and the *CV* in each cell are calculated using the full SOCA cross-sectional sample of 245,568 returns. Except that fifteen cells in the SOI report Table 1A (the first row where asset type='Total') are the estimates of population totals, the other 3,839 cells are domain estimates. Some cells are very small domains as they have a very small number of returns that have a non-zero SOCA value. For example, Column (1) in SOCA report Table 1A is the estimated number of transactions (in thousands). The number is 80 for Asset Type='Timber' and 247,937 for Asset Type='Total.' In other words, 'Timber' category only counts for 0.03% of the total in the number of transactions. Also can be seen from the snapshot of SOCA report Table 1A is that domains are even smaller when Timber category is further divided into gain transaction, loss transactions and transactions with no gain/loss. All estimates are calculated using the *extended domain variable of interest*, defined as $y_{di} = y_i$ for return *i* in the domain and $y_{di} = 0$ for return *i* not in the

domain. The extended domain variables are useful for estimations here because a single set of calibrated weights is attained for all domains and domain sample sizes are random. The contribution of extra variance caused by random domain sample sizes is incorporated in the variance expressions using the full sample and extended domain variables (Lehtonen, R. & Veijanen, A., 2009). For very small cells such as the Timber category, the construction of confidence interval using the assumption of asymptotic normality may not be appropriate sometimes, especially when the study variable is highly skewed.

Sales of Capital Assets Data Reported on Individual Tax Returns, 2007–2012

Table 1A. Short-Term and Long-Term Capital Gains and Losses, by Asset Type, Tax Year 2012

[All figures are estimates based on samples—number of transactions is in thousands, money amounts are in thousands of dollars]

Asset type	All transactions				Gain transactions			
	Number	Sales price	Basis	Net gain/loss	Number	Sales price	Basis	Gain
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total	**247,937**	**4,443,556,071**	**4,174,521,721**	**639,876,890**	**142,709**	**2,405,033,175**	**1,902,580,613**	**844,423,487**
Corporate stock	112,550	1,696,286,995	1,574,128,882	159,602,976	64,405	963,863,494	737,995,878	229,430,517
U.S. Government obligations	2,077	84,828,599	84,536,900	266,927	604	46,479,477	45,723,994	756,237
State and local government obligations	3,759	169,453,672	166,009,413	3,381,784	1,370	68,644,431	63,935,273	4,570,193
Other bonds, notes and debentures	2,795	64,788,208	64,352,421	682,998	1,408	29,830,801	26,814,897	3,062,120
Put and call options	9,666	53,275,093	54,830,421	1,096,805	5,444	31,721,643	22,676,743	10,156,673
Futures contracts	998	19,744,357	14,756,473	4,976,942	412	13,312,273	3,227,566	10,099,331
Mutual funds, except tax-exempt bond funds	74,189	600,939,523	602,669,778	17,110,263	40,687	331,539,537	296,795,368	35,426,266
Tax-exempt bond mutual funds	4,646	51,536,355	49,930,658	1,605,814	3,218	39,534,402	37,373,148	2,180,988
Partnership, S corporation, and estate or trust interests	5,519	170,566,604	123,763,913	48,149,424	3,367	118,935,159	59,113,994	61,997,223
Livestock	583	5,572,825	2,156,439	2,561,121	342	4,721,041	1,544,785	2,816,129
Timber	80	1,785,921	766,981	992,570	72	1,603,791	496,571	1,090,196
Involuntary conversions	425	1,085,636	94,924	344,752	37	887,690	136,971	746,679
Residential rental property	1,084	152,543,413	136,532,830	11,627,126	599	77,876,066	49,791,060	25,064,350
Depreciable business personal property	2,131	20,203,795	14,114,510	1,733,322	267	5,685,223	1,956,635	2,943,075
Depreciable business real property	464	61,507,402	47,850,899	12,167,901	270	41,649,606	22,443,584	18,446,399
Farmland	76	10,966,726	5,957,564	4,470,710	60	9,116,220	4,225,119	4,776,830
Other land	439	39,471,021	26,864,335	11,699,151	296	29,146,440	13,595,345	15,185,487
All residences	445	136,558,360	123,929,890	5,404,328	96	35,374,467	22,342,774	7,058,554
Residences	92	17,693,774	15,814,389	1,125,709	43	8,702,425	5,989,076	1,841,414
Principal residences	353	118,864,586	108,115,502	4,278,619	53	26,672,042	16,353,698	5,217,139
Other assets	8,097	936,559,534	914,721,103	38,557,461	4,921	480,008,839	424,631,165	61,073,035
Unidentifiable	1,515	165,882,033	166,553,387	4,372,659	865	75,102,577	67,759,743	8,845,104
Passthrough gains or losses	6,000	N/A	N/A	291,290,976	3,571	N/A	N/A	320,917,218
Capital gain distributions	10,399	N/A	N/A	17,780,882	10,399	N/A	N/A	17,780,882

Sales of Capital Assets Data Reported on Individual Tax Returns, 2007–2012

Table 1A. Short-Term and Long-Term Capital Gains and Losses, by Asset Type, Tax Year 2012—Continued

[All figures are estimates based on samples—number of transactions is in thousands, money amounts are in thousands of dollars]

Asset type	Loss transactions				Transactions with no gain or loss		
	Number	Sales price	Basis	Loss	Number	Sales price	Basis
	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Total	**74,662**	**1,495,545,720**	**1,713,757,831**	**204,546,597**	**30,566**	**542,977,176**	**558,183,277**
Corporate stock	39,755	581,697,884	666,563,320	69,827,542	8,390	150,725,617	169,569,684
U.S. Government obligations	413	15,641,935	16,118,810	489,310	1,059	22,707,187	22,694,097
State and local government obligations	737	22,497,484	23,665,255	1,188,409	1,653	78,311,757	78,408,885
Other bonds, notes and debentures	681	15,928,952	18,328,014	2,379,122	706	19,028,455	19,209,509
Put and call options	3,733	18,366,651	27,840,952	9,059,868	489	3,186,799	4,312,725
Futures contracts	571	5,667,926	10,792,210	5,122,389	15	764,158	736,696
Mutual funds, except tax-exempt bond funds	20,671	232,009,573	265,135,051	18,316,003	12,831	37,390,414	40,739,359
Tax-exempt bond mutual funds	445	8,492,768	9,046,584	575,174	983	3,509,185	3,510,926
Partnership, S corporation, and estate or trust interests	1,726	41,105,266	54,916,629	13,847,799	427	10,526,180	9,733,289
Livestock	79	190,690	445,912	255,008	162	661,094	165,742
Timber	4	142,664	231,036	97,626	4	39,466	39,374
Involuntary conversions	53	17,654	-221,406	401,928	335	180,292	179,360
Residential rental property	285	35,830,686	49,234,161	13,437,224	200	38,836,660	37,507,609
Depreciable business personal property	294	2,397,065	3,603,500	1,209,753	1,570	12,121,507	8,554,375
Depreciable business real property	119	15,024,802	21,462,543	6,278,498	75	4,832,994	3,944,771
Farmland	6	471,629	749,763	306,120	10	1,378,878	982,682
Other land	97	5,653,516	9,028,125	3,486,336	46	4,671,064	4,240,865
All residences	37	4,678,902	6,191,439	1,654,226	313	96,504,991	95,395,677
Residences	19	1,815,961	2,510,277	715,706	30	7,175,389	7,315,035
Principal residences	17	2,862,942	3,681,161	938,520	283	89,329,602	88,080,642
Other assets	2,429	401,975,321	434,973,459	22,515,575	747	54,575,375	55,116,479
Unidentifiable	474	87,754,353	95,652,474	4,472,446	176	3,025,103	3,141,170
Passthrough gains or losses	2,053	N/A	N/A	29,626,242	377	N/A	N/A
Capital gain distributions	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Sales of Capital Assets Data Reported on Individual Tax Returns, 2007–2012

Table 2A. Returns with Short-Term and Long-Term Capital Gains and Losses, by Size of Adjusted Gross Income and Selected Asset Type, Tax Year 2012

[All figures are estimates based on samples—money amounts are in thousands of dollars]

Selected asset type and size of adjusted gross income	Returns with short-term capital gain or loss						
	Number of returns	Returns with short-term gain transactions [1]			Returns with short-term loss transactions		
		Number of returns	Number of transactions	Gain	Number of returns	Number of transactions	Loss
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All asset types							
All returns	**10,856,594**	**8,825,193**	**82,810,766**	**98,678,301**	**6,554,647**	**41,997,439**	**73,708,875**
Adjusted gross deficit	327,897	252,730	2,500,392	4,623,130	212,292	1,572,025	7,209,589
Under $20,000	1,196,014	959,243	8,886,094	2,066,888	666,638	3,904,652	3,260,489
$20,000 under $50,000	1,665,807	1,326,258	10,970,885	3,188,480	954,839	4,652,497	6,047,018
$50,000 under $100,000	2,602,352	2,115,202	14,974,287	6,692,021	1,479,696	7,052,327	7,744,399
$100,000 under $200,000	2,791,078	2,264,568	17,127,754	10,313,381	1,668,180	8,871,059	10,521,296
$200,000 under $500,000	1,578,047	1,297,758	15,080,343	14,100,639	1,056,180	7,852,174	11,566,120
$500,000 under $1,000,000	406,844	350,455	5,919,890	8,042,943	292,063	3,314,880	5,974,127
$1,000,000 or more	288,555	258,978	7,351,121	49,650,819	224,758	4,777,824	21,385,837

Selected asset type and size of adjusted gross income	Returns with long-term capital gain or loss						
	Number of returns	Returns with long-term gain transactions [1]			Returns with long-term loss transactions		
		Number of returns	Number of transactions	Gain	Number of returns	Number of transactions	Loss
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
All asset types							
All returns	**19,888,075**	**17,966,529**	**90,463,866**	**745,745,186**	**8,242,192**	**32,664,415**	**130,837,722**
Adjusted gross deficit	633,988	485,944	2,219,587	25,270,895	378,387	1,169,896	25,783,229
Under $20,000	2,695,598	2,430,891	7,813,131	7,449,876	946,177	2,826,339	6,560,605
$20,000 under $50,000	3,466,065	3,092,546	11,574,533	13,977,528	1,277,184	3,885,014	10,106,122
$50,000 under $100,000	5,193,706	4,674,512	18,759,976	31,207,920	1,986,529	6,462,065	16,489,188
$100,000 under $200,000	4,740,758	4,314,502	19,611,584	51,300,935	2,000,802	6,937,133	18,133,647
$200,000 under $500,000	2,265,549	2,113,925	16,246,212	78,837,107	1,100,242	5,858,621	18,549,387
$500,000 under $1,000,000	543,411	516,574	6,057,670	62,298,650	315,863	2,307,970	9,270,092
$1,000,000 or more	349,001	337,636	8,181,173	475,402,276	237,007	3,217,377	25,945,452

[1] Transactions with no gain or loss are included with gain transactions.
NOTES: Number of returns with gain plus number of returns with loss does not add to the total column because some returns show both. Detail may not add to totals because of rounding.
SOURCE: IRS, Statistics of Income Division, Sales of Capital Assets Data, Tax Years 2007–2012, February 2016.

Sales of Capital Assets Data Reported on Individual Tax Returns, 2007–2012

Table 3A. Capital Gains and Losses, by Selected Asset Type and Month of Sale, Tax Year 2012

[All figures are estimates based on samples—transactions are in thousands, money amounts are in thousands of dollars]

Type of transaction, month of sale	All asset types							
	Gain transactions [1]				Loss transactions			
	Number of transactions	Sales price	Basis	Gain	Number of transactions	Sales price	Basis	Loss
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Short-term transactions								
Total	**82,811**	**1,420,384,356**	**1,379,242,467**	**98,678,301**	**41,997**	**1,090,036,048**	**1,185,338,652**	**73,708,875**
January	6,624	71,648,269	68,549,680	4,312,190	3,320	35,175,647	38,597,814	3,358,493
February	6,631	79,235,837	75,391,884	4,681,721	2,785	32,451,365	35,091,396	2,673,216
March	6,724	89,133,313	85,479,185	4,652,695	3,096	37,883,608	41,189,164	3,400,950
April	6,401	76,513,597	73,977,199	4,125,571	3,218	34,042,840	37,122,202	3,191,961
May	5,668	61,893,903	59,978,302	3,443,931	4,349	41,995,236	46,735,479	4,674,246
June	5,138	63,141,558	60,562,231	3,452,892	3,776	43,371,042	48,071,293	3,933,003
July	5,826	59,072,291	56,750,776	3,358,546	3,187	32,427,765	36,546,358	3,555,304
August	6,190	61,597,564	58,376,271	3,558,219	2,754	28,388,858	31,692,847	2,874,568
September	5,587	58,466,566	55,421,570	3,686,198	2,133	24,171,815	27,255,870	2,393,339
October	6,381	67,425,719	65,361,181	3,121,246	2,532	26,874,837	30,002,726	2,805,050
November	5,394	60,450,602	59,073,927	3,263,111	3,266	38,858,402	43,029,403	3,667,598
December	6,884	272,909,467	269,539,765	9,995,231	3,472	314,085,137	339,616,792	11,573,064
Not determinable	9,363	398,895,670	390,780,495	47,026,752	4,108	400,309,696	430,387,308	25,608,081
Long-term transactions								
Total	**90,464**	**1,527,625,995**	**1,081,519,311**	**745,745,186**	**32,664**	**405,509,672**	**528,419,179**	**130,837,722**
January	5,545	77,235,559	54,257,209	22,854,073	2,221	18,660,572	25,310,917	6,661,710
February	5,509	74,666,922	56,038,401	18,235,325	1,938	81,532,134	101,876,940	5,241,530
March	5,979	83,537,243	61,083,849	21,475,586	2,013	19,350,747	24,944,673	5,575,055
April	5,741	82,536,363	57,407,021	23,515,259	2,246	20,881,047	27,368,376	6,511,748
May	5,132	81,516,119	59,926,969	20,207,918	2,710	20,988,376	27,356,421	6,412,050
June	5,822	107,827,810	76,503,895	29,119,106	2,898	27,410,919	34,786,600	7,420,262
July	6,201	97,530,330	73,049,548	22,529,652	2,782	24,818,955	31,932,618	7,046,229
August	6,292	100,362,318	75,841,651	22,536,872	2,633	22,947,018	30,070,685	7,177,279
September	5,734	86,059,895	61,029,919	23,633,149	2,031	19,179,495	24,941,321	5,842,028
October	6,134	93,912,900	65,393,387	25,877,672	2,024	19,337,648	25,596,955	6,463,596
November	5,837	109,937,408	76,989,161	29,789,803	2,460	24,460,307	31,957,178	7,541,922
December	9,851	317,085,179	204,845,708	110,098,376	3,816	60,850,442	84,308,834	23,132,600
Not determinable	16,685	215,417,951	159,152,593	375,872,395	2,892	45,092,013	57,967,662	35,811,714

[1] Transactions with no gain or loss are included with gain transactions.

Sales of Capital Assets Data Reported on Individual Tax Returns, 2007–2012

Table 4A. Capital Gains and Losses, by Selected Asset Type and Length of Time Held, Tax Year 2012

[All figures are estimates based on samples—transactions are in thousands, money amounts are in thousands of dollars]

Type of transaction, length of time held	All asset types							
	Gain transactions [1]				Loss transactions			
	Number of transactions	Sales price	Basis	Gain	Number of transactions	Sales price	Basis	Loss
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Short-term transactions								
Total	**82,811**	**1,420,384,356**	**1,379,242,467**	**98,678,301**	**41,997**	**1,090,036,048**	**1,185,338,652**	**73,708,875**
Under 1 month	20,430	354,647,148	349,630,564	12,185,718	10,600	176,122,756	190,819,963	11,687,897
1 month under 2 months	6,332	56,907,199	54,917,385	3,921,290	3,575	26,804,793	30,394,217	3,505,566
2 months under 3 months	4,894	38,101,062	35,829,753	2,725,686	2,958	19,810,813	22,547,719	2,716,974
3 months under 4 months	4,652	32,914,928	31,095,221	2,417,294	2,447	18,600,464	20,954,193	2,304,712
4 months under 5 months	3,805	26,007,005	24,007,687	2,202,328	2,072	14,168,498	16,195,441	2,038,509
5 months under 6 months	3,864	27,463,224	25,484,972	2,225,939	1,773	20,047,636	21,856,500	1,712,341
6 months under 7 months	3,710	30,069,829	28,024,394	2,343,819	1,745	11,871,104	13,546,056	1,731,322
7 months under 8 months	3,365	21,198,372	19,631,747	1,790,477	1,625	12,331,371	14,009,630	1,620,727
8 months under 9 months	2,755	18,217,622	16,613,062	1,573,507	1,499	10,165,001	11,700,527	1,524,554
9 months under 10 months	2,873	22,151,704	20,743,641	1,683,607	1,499	9,145,429	10,786,104	1,626,478
10 months under 11 months	2,702	18,959,091	17,195,715	1,638,795	1,415	7,938,127	9,410,604	1,460,897
11 months under 12 months	2,795	20,376,120	18,907,751	1,643,463	1,507	12,699,503	14,563,751	1,825,552
1 year or more [2]	1,051	24,372,763	21,580,749	3,201,332	520	100,216,992	109,565,063	2,857,706
Period not determinable	19,583	728,998,289	715,579,825	59,125,045	8,763	650,113,561	698,988,884	37,095,641
Long-term transactions								
Total	**90,464**	**1,527,625,995**	**1,081,519,311**	**745,745,186**	**32,664**	**405,509,672**	**528,419,179**	**130,837,722**
Under 18 months [3]	13,588	115,255,122	98,755,562	16,842,311	7,338	60,665,158	70,078,031	8,930,511
18 months under 2 years	9,429	72,063,018	60,475,980	11,405,257	3,902	74,175,186	94,656,583	5,940,314
2 years under 3 years	13,118	120,451,758	97,824,715	22,085,514	3,282	24,951,785	31,151,501	6,230,884
3 years under 4 years	8,343	96,348,437	73,872,289	22,317,229	1,786	16,173,630	21,568,558	5,362,988
4 years under 5 years	3,572	64,934,123	52,414,517	13,564,392	2,343	20,164,455	28,538,634	8,444,958
5 years under 10 years	7,410	198,269,353	156,168,658	45,903,334	3,782	45,673,556	69,663,083	24,037,682
10 years under 15 years	2,692	102,823,875	61,637,598	34,272,119	1,349	13,452,562	19,815,312	6,348,252
15 years under 20 years	995	50,317,334	23,436,171	21,657,870	241	2,572,601	4,633,703	2,070,525
20 years or more	1,174	109,094,859	37,071,263	61,005,280	111	2,611,392	4,155,932	1,560,222
Period not determinable	30,143	598,068,115	419,862,557	496,691,880	8,530	145,069,348	184,157,841	61,911,387

[1] Transactions with no gain or loss are included with gain transactions.

Next, we compare the estimates based on the panel-converted sample to the estimates based on the SOCA cross-sectional sample. The estimates in SOCA report tables outlined above are calculated using the SOCA cross-sectional sample. Their CVs and confidence intervals are also calculated. All estimates shown in SOCA report tables are calculated using the panel-converted sample. To see how close the two sets of estimates are in general, we summarize the number of cells where estimates from the panel-converted sample fall within 95% confidence intervals from SOCA cross-sectional sample for each of SOCA report tables in the SOI report. As mentioned above, the assumption of normal distribution in constructing confidence intervals may not be appropriate for very small domains. But as a rough measure, all are included in the summary. The 'coverage rate' is the number of domain estimates (cells) from the panel-converted sample falling within their 95% confidence intervals divided by the total number of domain estimates (cells) in the table. The following Tables 4.1A -4.1D give the 'coverage' rate of panel-converted sample estimates for each of those SOI SOCA report tables. For example, in Table 4.1, there are 356 cell estimates in SOCA report Table 1A, 185 of panel-converted sample estimates are within the 95% confidence interval, which is 52% in coverage.

Tables 4.1A-4.1D show the overall picture of how close domain estimates from the panel-converted sample to those from the SOCA cross-sectional sample, where some domains are large and some are small. We further look at those estimates in more detail. First, all 3,854 estimates (or cells) from SOCA report tables are grouped into seven categories, as shown in Table 4.2. For each category, cells are from different SOCA report tables. There are 1,012 cells are domain estimates of 'Number of Transaction,' 816 cells are domain estimates of 'Basis,' and so on. The Estimated population total is taken from the SOCA report Table 1A (the first snapshot above) for all categories except for the Number of Returns. The population total number of returns that are associated to any SOCA activities is estimated using the SOCA cross-sectional sample data. For each category, estimates are then grouped by cell size. Here, the cell size is relative to the population total, defined as the cell estimate divided by the estimated population total. The cell estimate is simply the number in each cell of SOCA report tables and population total estimate is the number in the appropriate category in Table 4.2. For example, the first cell of Asset Type='Total' in column (1) of the SOCA report Table 1A (the first snapshot above) has a cell size of 247,397/247,397=100% and the second cell of Asset Type='Corporate Stock' in column (1) has a cell size of 112,550//247,397=45.5%. The cell size ranges from 0.00001% to 100%, where the 100% is the population total and others are domains. Cells in each category are divided into 10 groups by the cell size.

Table 4.1A. "Coverage" Rate of Capital Gain/Loss Estimates by Asset Type

SOI SOCA Report Table 1A -1C Capital Gains and Losses by Asset Type	Number of cells in the Table	Number of Cells where Panel-Converted Sample Estimates Fall within the 95% Confidence Interval	"Coverage" Rate
Table 1A. Short-Term and Long-Term	356	185	52.0%
Table 1B. Short-Term	352	204	58.0%
Table 1C. Long-Term	356	192	53.9%

Table 4.1B. "Coverage" Rate of Capital Gain/Loss Estimates for Selected Asset Type, by AGI Group

SOI SOCA Report Table 2A -2E Short-Term and Long-Term Capital Gains and Losses by AGI Groups	Number of cells in the Table	Number of Cells where Panel-Converted Sample Estimates Fall within the 95% Confidence Interval	"Coverage" Rate
Table 2A. For All Asset Types	126	67	53.2%
Table 2B. For Asset Type=Corporate stock	126	69	54.8%
Table 2C. For Asset Type=Bonds and other securities	126	46	36.5%
Table 2D. For Asset Type=Real Estate	126	105	83.3%
Table 2E. For Asset Type=Others	126	75	59.5%

Table 4.1C. "Coverage" Rate of Capital Gain/Loss Estimates for Selected Asset Type, by Short-Term and long-Term Capital Asset Transactions and Month of Sale

SOI SOCA Report Table 3A -3E Short-Term and Long-Term Capital Gains and Losses by Short-Term and Long-Term Capital Asset Transactions and Month of Sale	Number of cells in the Table	Number of Cells where Panel-Converted Sample Estimates Fall within the 95% Confidence Interval	"Coverage" Rate
Table 3A. For All Asset Types	224	55	24.6%
Table 3B. For Asset Type=Corporate stock	224	74	33.0%
Table 3C. For Asset Type=Bonds and other securities	224	27	12.1%
Table 3D. For Asset Type=Real Estate	224	191	85.3%
Table 3E. For Asset Type=Others	224	96	42.9%

Table 4.1D. "Coverage" Rate of Capital Gain/Loss Estimates for Selected Asset Type, by Short-Term and Long-Term Capital Asset Transactions and Length of Time Held

SOI SOCA Report Table 4A -4E Short-Term and Long-Term Capital Gains and Losses by Short-Term and Long-Term Capital Asset Transactions and Length of Time Held	Number of cells in the Table	Number of Cells where Panel-Converted Sample Estimates Fall within the 95% Confidence Interval	"Coverage" Rate
Table 4A. For All Asset Types	208	50	24.0%
Table 4B. For Asset Type=Corporate stock	208	75	36.1%
Table 4C. For Asset Type=Bonds and other securities	208	37	17.8%
Table 4D. For Asset Type=Real Estate	208	173	83.2%
Table 4E. For Asset Type=Others	208	102	49.0%

Table 4.2. SOCA Subject Category, Estimated Population Total and Number of Estimates

SOCA Subject Category	Estimated Population Total Associated to SOCA Activities	Number of Cells from SOI Report Tables
Number of Transactions	247,937	1,012
Basis	4,174,521,721	816
Sales Price	4,443,556,071	816
Number of Returns	22,096,082	270
Net Gain/Loss	639,876,890	74
Gain	844,423,487	434
Loss	(204,546,597)	432
Total		**3,854**

For each SOCA subject category, cells are divided into eight groups by the cell size with unequal increments 0-0.1%, 0.1-0.5%, 0.5-1%, 1-2%, 2-4%, 4-10%, 10-50% and 50-100%. This is because there are a lot more small cells than large cells. The following Table 4.3A looks at the distribution of relative error by size group for the category of 'Number of Transactions.' The 1,012 cell estimates of the number of transactions are divided into eight groups by the cell size and then within each cell size group, estimates are distributed across 10 groups of relative error. The relative error is defined as $|(E_p/E_s)-1|$, where E_p is the cell estimate from the panel-converted sample and E_s is the cell estimate from the SOCA cross-sectional sample. The first nine relative error groups, ranging from 0% to 90%, have the same increment of 10%. The last group is open ended group of 'more than 90%.' As shown in Table 4.3A, the first column has a total of 355 cell estimates of number of transactions that are smaller than 0.1% of the total number of transactions. Among 355 small cells, 118 have a relative error in absolute value under 10%, 76 cells between 10-20%, and so forth. The last column under 'Total' shows the distribution of the relative error of all estimates. Overall, 368 cells (36.4%) whose estimated number of transactions from the panel-converted sample is within 10% of the estimated number of transactions from the SOCA cross-sectional sample regardless of cell size, 263 cells (26.0%) have a relative of 10-20%, and so forth. Table 4.3A also shows 19 cells have a large relative error of over 90% and all are small cells with a cell size under 1%. Tables 4.3B-G are for other six categories.

Table 4.3A. The Distribution of Relative Error of *Number of Transactions* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	118	78	43	52	35	27	15		**368**	**36.4%**
10 - 20	76	38	31	57	30	9	19	3	**263**	**26.0%**
20 - 30	44	30	41	16	9	10	9		**159**	**15.7%**
30 - 40	33	18	9	9	3				**72**	**7.1%**
40 - 50	28	22	4	7	2				**63**	**6.2%**
50 - 60	21	7	5	2					**35**	**3.5%**
60 - 70	10	6							**16**	**1.6%**
70 - 80	7	2							**9**	**0.9%**
80 - 90	5	2		1					**8**	**0.8%**
More than 90%	15	3	1						**19**	**1.9%**
Total	**357**	**206**	**134**	**144**	**79**	**46**	**43**	**3**	**1,012**	**100.0%**

Table 4.3B. The Distribution of Relative Error of *Basis* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	31	45	18	7	6	4	7		**118**	**14.5%**
10 - 20	31	46	20	15	6	5	9		**132**	**16.2%**
20 - 30	22	47	18	5	13	5	7		**117**	**14.3%**
30 - 40	25	35	19	13	6	8	6	2	**114**	**14.0%**
40 - 50	17	11	18	9	4	1	9		**69**	**8.5%**
50 - 60	22	10	5	6	3	6			**52**	**6.4%**
60 - 70	10	8	7	5	4	2	4		**40**	**4.9%**
70 - 80	8	2	3	3	4				**20**	**2.5%**
80 - 90	9	6	2	1					**18**	**2.2%**
More than 90%	54	55	14	5	8				**136**	**16.7%**
Total	**229**	**265**	**124**	**69**	**54**	**31**	**42**	**2**	**816**	**100.0%**

Table 4.3C. The Distribution of Relative Error of *Sales Price* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	33	49	29	11	9	9	11		**151**	**18.5%**
10 - 20	36	33	26	8	10	10	1		**124**	**15.2%**
20 - 30	29	35	17	10	14	4	11	3	**123**	**15.1%**
30 - 40	20	42	14	9	6	3	6		**100**	**12.3%**
40 - 50	15	15	13	4	5	5	5		**62**	**7.6%**
50 - 60	17	7	6	7	4	4	1		**46**	**5.6%**
60 - 70	12	9	6	3	1	2	2		**35**	**4.3%**
70 - 80	11	5	3	4					**23**	**2.8%**
80 - 90	8	5	2	1					**16**	**2.0%**
More than 90%	61	52	10	6	7				**136**	**16.7%**
Total	**242**	**252**	**126**	**63**	**56**	**37**	**37**	**3**	**816**	**100.0%**

Table 4.3D. The Distribution of Relative Error of *Number of Returns* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	10	16	27	34	29	46	32	4	**198**	**73.3%**
10 - 20	3	12	5	5	3	2			**30**	**11.1%**
20 - 30	5	7	3	2	1				**18**	**6.7%**
30 - 40	5	7							**12**	**4.4%**
40 - 50	2	6							**8**	**3.0%**
50 - 60	1	1							**2**	**0.7%**
60 - 70	1								**1**	**0.4%**
70 - 80										**0.0%**
80 - 90	1								**1**	**0.4%**
More than 90%										**0.0%**
Total	**28**	**49**	**35**	**41**	**33**	**48**	**32**	**4**	**270**	**100.0%**

Table 4.3E. The Distribution of Relative Error of *Net Gain/Loss* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10		3	6		4	2	4	2	**21**	**28.4%**
10 - 20	3	5	1	2	2	2			**15**	**20.3%**
20 - 30	2	4	2	4					**12**	**16.2%**
30 - 40	1	4	1						**6**	**8.1%**
40 - 50	1	1							**2**	**2.7%**
50 - 60	1								**1**	**1.4%**
60 - 70		1							**1**	**1.4%**
70 - 80	1								**1**	**1.4%**
80 - 90		1							**1**	**1.4%**
More than 90%	13		1						**14**	**18.9%**
Total	**22**	**19**	**11**	**6**	**6**	**4**	**4**	**2**	**74**	**100.0%**

Table 4.3F. The Distribution of Relative Error of *Gain* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	21	48	28	20	21	10	13	10	**171**	**39.4%**
10 - 20	16	28	12	17	14	18	4		**109**	**25.1%**
20 - 30	12	20	9	9	2				**52**	**12.0%**
30 - 40	16	6	2	1					**25**	**5.8%**
40 - 50	8	2							**10**	**2.3%**
50 - 60	7	1	3						**11**	**2.5%**
60 - 70	10	1							**11**	**2.5%**
70 - 80	10	2	1						**13**	**3.0%**
80 - 90	3	5							**8**	**1.8%**
More than 90%	21	3							**24**	**5.5%**
Total	124	116	55	47	37	28	17	10	**434**	**100.0%**

Table 4.3G. The Distribution of Relative Error of *Loss* by Cell Size Group

Relative Error (Absolute Value, %)	Cell Size (%)								Total	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	9	14	22	19	20	11	13		**108**	**25.0%**
10 - 20	14	18	27	24	15	6	12	5	**121**	**28.0%**
20 - 30	15	17	19	10	9	9	11		**90**	**20.8%**
30 - 40	5	7	4	7	3	5	1		**32**	**7.4%**
40 - 50	6	2	6	5	2	2			**23**	**5.3%**
50 - 60	9	5	1						**15**	**3.5%**
60 - 70	3	2							**5**	**1.2%**
70 - 80	3	1							**4**	**0.9%**
80 - 90	3								**3**	**0.7%**
More than 90%	25	6							**31**	**7.2%**
Total	**92**	**72**	**79**	**65**	**49**	**33**	**37**	**5**	**432**	**100.0%**

5. Considerations for Future Design

The above comparison of two sets of 3,854 estimates from two samples indicates that most estimates are reasonably close between the panel-converted sample and the SOCA cross-sectional sample. To increase precision, a larger yearly refreshment sample is needed. There is a balance between the cost-saving and the precision-losing when the panel sample is used for out-year cross-sectional estimations. The panel sample is free as panel returns are edited anyway for the panel study. The number of additional returns selected as the refreshment sample is summarized in the following Table 5.1. The non-overlap between SOCA cross-sectional sample and the panel sample, also summarized in Table 5.1, is the extra cost (in terms of number of returns) to do the SOCA cross-sectional sample. Note that in some strata of low-income returns, there are more returns in the panel-converted sample. This is due to returns linked to the secondary CWHS returns selected in the panel sample. Small returns are very cheap but have basically insignificant SOCA values. Therefore, we only look at returns with the selection income over $500,000. Column D of Table 5.1 is the ratio of the extra number of returns of the 2012 panel-converted sample to the extra number of returns of the 2012 SOCA cross-sectional sample. It shows that the extra cost for the panel-converted sample is a very small portion of the extra cost for the SOCA cross-sectional sample, especially in strata of high income returns.

To see how much the precision level can be gained using a larger panel-converted sample if we have the budget, say 20% of cost of the 2012 SOCA cross-sectional sample, we look at a simulation scenario where number of returns in the panel-converted sample is about 20% of returns in the SOCA cross-sectional sample, especially for high-income strata. Table 5.2 outlines the sample size by stratum. Column B is the SOCA cross-sectional sample size after taking out the ones already in the panel sample and Column C is the panel-converted sample size after taking out the ones already in the panel sample. Column D is the number of extra returns needed so that the cost of the panel-converted sample is about 20% of the SOCA cross-sectional sample. Few adjustments are made. First, certainty strata are larger and all returns with a selection income over $100,000,000 in both positive and negative are selected. Second, low income strata with a selection income of $0-$500,000 are set to have no extra returns because they have very small impact in SOCA estimates. Third, for the stratum with the selection income of $500,000-$1,000,000, the number is reduced to 1000 because of its small contribution to SOCA estimates. Column E is the final number of extra returns to be drawn. An additional 9,232 returns are selected from 87,271 that are in the SOCA cross-sectional sample, but not in the panel-converted sample. In summary, the simulation sample includes a total 244,424 returns, 235,192 of them in the panel-converted returns and 9,232 are additional returns.

Table 5.1 Number of Additional Sample Returns After Removing Panel Returns, 2012

Selection Income ($1000)	SOCA Cross-Sectional Sample	Panel-Converted Sample	Ratio
A	B	C	D=C/B
Negative			
$150,000 or more	17	17	1.00
$40,000 to < $150,000	136	10	0.07
$20,000 to < $40,000	378	16	0.04
$10,000 to < $20,000	682	26	0.04
$5,000 to < $10,000	979	61	0.06
$2,000 to < $5,000	4,504	97	0.02
$1,000 to < $2,000	4,732	48	0.01
$500 to < $1,000	2,892	50	0.02
$250 to < $500	1,768	42	0.02
$120 to < $250	1,540	47	0.03
$60 to < $120	1,071	69	0.06
Under $60	1,124	232	0.21
Positive			
Under $30	13,632	14,820	1.09
Under $30	6,370	9,804	1.54
Under $30	2,307	3,097	1.34
$30 to < $60	1,293	2,904	2.25
$30 to < $60	735	1,141	1.55
$60 to < $120	1,057	1,065	1.01
$60 to < $120	845	530	0.63
$120 to < $250	4,876	166	0.03
$120 to < $250	10,734	345	0.03
$250 to < $500	11,049	193	0.02
$500 to < $1,000	11,795	111	0.01
$1,000 to < $2,000	9,132	107	0.01
$2,000 to < $5,000	7,207	176	0.02
$5,000 to < $10,000	4,322	128	0.03
$10,000 to < $20,000	4,884	106	0.02
$20,000 to < $40,000	1,761	88	0.05
$40,000 to < $150,000	779	122	0.16
$150,000 or more	64	64	1.00

Table 5.2. Number of Additional Sample Returns for the Simulation

Selection Income ($1000)	Number of Returns after removing the overlap with the panel sample		Additional Number of Returns to be Drawn for the Simulation		Number of Returns to be Drawn From
	SOCA Cross-Sectional Sample	Panel-Converted Sample	Calculated	Actual	
A	B	C	D=0.2B-C	E	F*
Negative					
$150,000 or more	17	17	0	0	0
$100,000 to < $150,000	14	2	1	12	12
$40,000 to < $100,000	122	8	16	16	114
$20,000 to < $40,000	378	16	60	60	362
$10,000 to < $20,000	682	26	110	110	661
$5,000 to < $10,000	979	61	135	135	924
$2,000 to < $5,000	4,504	97	804	804	4,419
$1,000 to < $2,000	4,732	48	898	898	4,698
$500 to < $1,000	2,892	50	528	528	2,862
$250 to < $500	1,768	42	312	312	1,750
$120 to < $250	1,540	47	261	261	1,530
$60 to < $120	1,071	69	145	145	1,052
Under $60	1,124	232	-7	0	970
Positive					
Under $30	13,632	14,820	-12,094	0	36
Under $30	6,370	9,804	-8,530	0	100
Under $30	2,307	3,097	-2,636	0	169
$30 to < $60	1,293	2,904	-2,645	0	179
$30 to < $60	735	1,141	-994	0	331
$60 to < $120	1,057	1,065	-854	0	755
$60 to < $120	845	530	-361	0	673
$120 to < $250	4,876	166	809	0	4,828
$120 to < $250	10,734	345	1,802	0	10,634
$250 to < $500	11,049	193	2,017	0	10,967
$500 to < $1,000	11,795	111	2,248	1,000	11,736
$1,000 to < $2,000	9,132	107	1,719	1,719	9,068
$2,000 to < $5,000	7,207	176	1,265	1,265	7,094
$5,000 to < $10,000	4,322	128	736	736	4,222
$10,000 to < $20,000	4,884	106	871	871	4,791
$20,000 to < $40,000	1,761	88	264	264	1,677
$40,000 to < $100,000	703	109	32	32	594
$100,000 to < $150,000	76	13	2	63	63
$150,000 or more	64	64	0	0	0
Total				**9,232**	**87,271**

** Column F is the number of SOCA cross-sectional sample after removing the overlap with the panel converted sample.*

The same weighting procedure and estimation procedure are applied to the larger simulated panel-converted sample. Estimates from the simulated sample and the panel-converted sample are compared and summarized in Table 5.3 and Tables 5.4A-G.

In Table 5.3, numbers of column B are 'coverage rates' calculated from the panel-converted sample, taken from Tables 4.1A-D in Section 4 and numbers in column C are 'coverage' rates calculated from the simulated sample. Column D is the difference between two samples and shows that the larger simulated sample results in higher 'coverage rate' for all SOCA report tables except for SOCA report Table 2D. Note that a decrease in 'coverage' rate for Table 2D is not surprising because this table is about

estimates for real estates that is a very small subdomain and estimates of small domains can be unstable.

Table 5.3. "Coverage" Rate - Percent of Panel-converted Sample Estimates Falling within the 95% Confidence Interval

SOI SOCA Report Table	Total Number of Cell Estimates	Coverage Rate		
		Panel Converted Sample	Simulated Sample	Difference
	A	B	C	D=C-B
Table 1A. Short-Term and Long-Term	356	52%	57%	5%
Table 1B. Short-Term	352	58%	70%	12%
Table 1C. Long-Term	356	54%	57%	3%
Table 2A. For All Asset Types	126	53%	56%	2%
Table 2B. For Asset Type=Corporate stock	126	55%	60%	5%
Table 2C. For Asset Type=Bonds and other securities	126	37%	52%	16%
Table 2D. For Asset Type=Real Estate	126	83%	74%	-10%
Table 2E. For Asset Type=Others	126	60%	61%	2%
Table 3A. For All Asset Types	224	25%	52%	27%
Table 3B. For Asset Type=Corporate stock	224	33%	52%	19%
Table 3C. For Asset Type=Bonds and other securities	224	12%	24%	12%
Table 3D. For Asset Type=Real Estate	224	85%	88%	2%
Table 3E. For Asset Type=Others	224	43%	67%	24%
Table 4A. For All Asset Types	208	24%	58%	34%
Table 4B. For Asset Type=Corporate stock	208	36%	57%	21%
Table 4C. For Asset Type=Bonds and other securities	208	18%	26%	9%
Table 4D. For Asset Type=Real Estate	208	83%	86%	3%
Table 4E. For Asset Type=Others	208	49%	66%	17%

Tables 5.4A-G look at the change of using the simulation sample versus using the panel-converted sample for each category of SOCA estimations (as shown in Table 5.2). Table structures are the same as those of Table 4.3A-G, where cells are defined by the cell size group and the relative error group. The tables present the **difference** in the number of domain estimates between two samples, one table for each SOCA estimation category. In other words, they give the change in the number of domain estimates due to the use of simulation sample, compared to the use of panel-converted sample. Table 5.4A is for the category of Number of Transactions. Of the total of 1,012 domain estimates of Number of Transactions, the simulation sample results in 184 or 18.2% more domain estimates in the smallest relative error group of 0-10% and 40 less in the second smallest relative error group of 10-20%, and so forth. Another way to analyze the table is to fix each cell size group and look at the numbers across relative groups. For the smallest size group 0-0.1%, more estimates are in small relative error groups and less estimates are in large relative error groups due to the larger sample size of the simulation sample. For the largest cell size group of 50-100%, three estimates move from the relative error group of 10-20% to a smaller relative error group of 0-20%. Other six tables are for other six SOCA estimation categories. In some tables, we see the simulation tables actually results in less small relative error estimates, such as Table 5.4D for estimates of the Number of Returns. But numbers of negative change are small though. Overall, the simulation sample has a lot more small-error estimates and a lot less large-error estimates than the panel-converted sample in general, which is the improvement due to a larger sample with about 20% the cost of the SOCA sample.

Table 5.4A. The Change in the Number of Domain Estimates of *Number of Transactions* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	9	35	31	51	19	10	26	3	**184**	**18.2%**
10 - 20	5	6	17	-35	-13	0	-17	-3	**-40**	**-4.0%**
20 - 30	15	-12	-38	-1	-3	-10	-9		**-58**	**-5.7%**
30 - 40	-4	1	-1	-7	-1				**-12**	**-1.2%**
40 - 50	5	-17	-4	-7	-2				**-25**	**-2.5%**
50 - 60	-10	-5	-5						**-20**	**-2.0%**
60 - 70	-7	-5							**-12**	**-1.2%**
70 - 80	-4	0							**-4**	**-0.4%**
80 - 90	-2	-2	1	-1					**-4**	**-0.4%**
More than 90%	-7	-1	-1						**-9**	**-0.9%**
Total	0	0	0	0	0	0	0	0	**0**	**0.0%**

** The % is based on the total of 1,012 domain estimates of the Number of Transactions*

Table 5.4B. The Change in the Number of Domain Estimates of *Basis* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	6	55	16	16	5	4	15	1	**118**	**14.5%**
10 - 20	4	3	18	-3	6	4	-2	1	**31**	**3.8%**
20 - 30	2	-14	2	7	-3	3	4		**1**	**0.1%**
30 - 40	2	-22	-10	-6	-1	-5	-4	-2	**-48**	**-5.9%**
40 - 50	1	-9	-12	-3		2	-9		**-30**	**-3.7%**
50 - 60	-2	1	-3	-5	-1	-6			**-16**	**-2.0%**
60 - 70	9	-4	-5	-3	-3	-2	-4		**-12**	**-1.5%**
70 - 80	-1	6	-3	-1	1				**2**	**0.2%**
80 - 90		-1	3	2	1				**5**	**0.6%**
More than 90%	-21	-15	-6	-4	-5				**-51**	**-6.3%**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.0%**

** The % is based on the total of 816 domain estimates of Basis*

Table 5.4C. The Change in the Number of Domain Estimates of *Sales Price* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	2	23	13	5	9	13	5	3	**73**	**8.9%**
10 - 20	4	36	10	10	3	-9	16		**70**	**8.6%**
20 - 30	10	-8	3	-5	-2	7	-8	-3	**-6**	**-0.7%**
30 - 40	-6	-29	-4	0	-1		-5		**-45**	**-5.5%**
40 - 50	7	-10	-8	-2	-3	-5	-5		**-26**	**-3.2%**
50 - 60	1	4	-4	-4	-4	-4	-1		**-12**	**-1.5%**
60 - 70	5	-2	-4	-2	-1	-2	-2		**-8**	**-1.0%**
70 - 80	0	1	-3		4				**2**	**0.2%**
80 - 90	1	2		2					**5**	**0.6%**
More than 90%	-24	-17	-3	-4	-5				**-53**	**-6.5%**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.0%**

** The % is based on the total of 816 domain estimates of Sales Price*

Table 5.4D. The Change in the Number of Domain Estimates of the *Number of Returns* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	-1	7	-14	-17	-3	2			**-26**	**-9.6%**
10 - 20	1	1	9	10	4	-2			**23**	**8.5%**
20 - 30	-2	1	3	2	-1				**3**	**1.1%**
30 - 40	3	-4	2	5					**6**	**2.2%**
40 - 50	-1	-4							**-5**	**-1.9%**
50 - 60		-1							**-1**	**-0.4%**
60 - 70	1								**1**	**0.4%**
70 - 80										**0.0%**
80 - 90	-1								**-1**	**-0.4%**
More than 90%										**0.0%**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.0%**

** The % is based on the total of 270 domain estimates of the Number of Returns*

Table 5.4E. The Change in the Number of Domain Estimates of the *Net Gain/Loss* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	1	1	-5				-4	-2	**-9**	**-12.2%**
10 - 20	-1		2	-2		-2		1	**-2**	**-2.7%**
20 - 30	-2	1	2	-4			4	1	**2**	**2.7%**
30 - 40	2	-2	1	6		2			**9**	**12.2%**
40 - 50	1	-1							**0**	**0.0%**
50 - 60	-1	2							**1**	**1.4%**
60 - 70	1		1						**2**	**2.7%**
70 - 80									**0**	**0.0%**
80 - 90	4	-1							**3**	**4.1%**
More than 90%	-5		-1						**-6**	**-8.1%**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.0%**

** The % is based on the total of 74 domain estimates of Net Gan/Loss*

Table 5.4F. The Change in the Number of Domain Estimates of the *Gain* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10	4	18	-7	5	-6	4	-9	-10	**-1**	**-0.2%**
10 - 20	-3	5	-3	-12		-14	5	9	**-13**	**-3.0%**
20 - 30	13	-13	9	3	5	10	4	1	**32**	**7.4%**
30 - 40	-1	-4	4	4					**3**	**0.7%**
40 - 50	4								**4**	**0.9%**
50 - 60	4	4	-2		1				**7**	**1.6%**
60 - 70	-5								**-5**	**-1.2%**
70 - 80	-5	-2	-1						**-8**	**-1.8%**
80 - 90	1	-5							**-4**	**-0.9%**
More than 90%	-12	-3							**-15**	**-3.5%**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.0%**

** The % is based on the total of 434 domain estimates of Gain*

Table 5.4G. The Change in the Number of Domain Estimates of the *Loss* by Cell Size Group and Relative Error Group Due to Using the Simulation Sample

Relative Error (Absolute Value, %)	Cell Size (%)								Total*	
	0 - 0.1	0.1 - 0.5	0.5 - 1	1 - 2	2 - 4	4 - 10	10 - 50	50 - 100		
0 - 10		10	22	13	9	5	13	4	**76**	**17.6%**
10 - 20	2	-4	-11	-5	-4	2	-3	-4	**-27**	**-6.3%**
20 - 30	1	-5	-10	-2	-3	-8	-10		**-37**	**-8.6%**
30 - 40	5	7	3	-3	-2	1			**11**	**2.5%**
40 - 50	-4	1	-4	-3					**-10**	**-2.3%**
50 - 60	-1	-5							**-6**	**-1.4%**
60 - 70									**0**	**0.0%**
70 - 80	2								**2**	**0.5%**
80 - 90	4								**4**	**0.9%**
More than 90%	-9	-4							**-13**	**-3.0%**
Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0.0%**

** The % is based on the total of 432 domain estimates of Loss*

References

Lehtonen, R. & Veijanen, A. (2009), "Design-based Methods of Estimation for Domains and Small Areas," Handbook of Statistics Vol. 29B. Sample Surveys. Inference and Analysis. Elsevier Scientific Publ. Co, (C.R. Rao and Danny Pfeffermann Eds.), p. 219–249

Liu, Y. K., Auten, G., Testa, V. and Strudler, M. (2009), "Redesign of SOI's Individual Income Tax Return Edited Panel Sample," *Proceedings of the Section on Government Statistics, American Statistical Association*, p. 3129 - 3143.

Liu, Y. K., Henry, K. A., Strudler, M. (2012), "Practical Issues When Calibrating Weights for Multiple Skewed Variables," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 4619-4632.

Wilson, J. and Liddell, P. (2016), "Sales of Capital Assets Data Reported on Individual Tax Returns, 2007–2012," *SOI Bulletin*, p. 2-93. http://soi.soi.irs.gov/publications.html,