WEIGHTING A PANEL OF INDIVIDUAL TAX RETURNS FOR CROSS-SECTIONAL ESTIMATION

Allen L. Schirm and John L. Czajka, Mathematica Policy Research, Inc.

Keywords: Income; Longitudinal; Stratification

1. INTRODUCTION

The question of how and even whether to use weights in analyzing sample data has been the subject of several papers in recent years. This paper explores issues related to the use of weights with panel data--specifically, the relevance of weights reflecting the base year sample design for estimates applying to years after the base year. The dataset we employ has such a high degree of stratification clearly relevant to the major variables of interest that it is hard to question the need for some type of differential weighting of observations. What is not so clear is whether weighting the observations according to their probabilities of selection under the base year design continues to provide satisfactory results for analyses using data well after the base year.

Section 2 of this paper discusses several preliminary issues, including the design of a panel of individual tax returns, while Section 3 describes the construction of design-based weights. Section 4 discusses the implications of panel dynamic behavior, and Section 5 examines alternative methods of weighting panel returns for cross-sectional estimation. Section 6 describes other methods of weighting that will be considered more fully in future work.

2. THE STATISTICS OF INCOME INDIVIDUAL PANEL

Each year the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) draws a sample of individual (i.e., Form 1040) tax returns from the population of returns processed in that calendar year. The sample is large -- frequently exceeding 100,000 returns -- and the design employs stratification with sharply differential sampling rates but essentially simple random sampling within strata.

Starting with the 1987 tax year sample (which includes returns processed in calendar year 1988), the IRS began a large panel. The panel was selected from the 1987 cross-sectional sample and is representative of nondependent tax returns processed in 1988. For the most part these returns have 1987 reference periods (tax years) -- hence the 1987 designation for the sample. The returns filed by panel members have been captured in every year since the initiation of the panel. While the returns selected for the panel were limited to nondependent returns (i.e., returns whose filers were not claimed as dependents on other filers' returns), the dependents claimed on these nondependent returns are panel members.

2.1 Design of the SOI Individual Panel

The panel was selected as a stratified random sample. Stratum definitions and sampling rates are provided in Table 1. The design included 39 strata based on a combination of income, return type, and total receipts (the latter from sole proprietorship business and farm returns). The sampling rates varied from just above .02 percent to 100 percent. Within each stratum the sample returns were selected on the basis of a transformation of the primary filer's social security number (SSN), as described in Harte (1986).

The base year sample includes some returns with tax periods prior to 1987. These "prior year" returns include late-processed returns for the immediately preceding tax year (1986) as well as returns for taxpayers who are catching up on their filing by filing more than one return. The presence of prior year returns in the base year sample (and what they imply about both filing behavior and sample selection) contribute to panel coverage deficiencies in future years (Czajka and Schirm, 1992).

One of the concerns regarding a panel sample is its continued representativeness over time. The SOI individual panel sample fully represents population growth attributable to the separation of joint filers and the

Table 1. Selection of the 1987 Individual Panel

Description of the sample strata		Stratum number	Number of returns: Nondependent population	Number of returns: Sample size	Sampling rate	Implied weight
Grand total			97,813,147	89,755		
Form 1040 returns only with adjusted gross income of $200,000 and over with no income tax after credits and no additional tax for tax preferences, total		28	873	873	100.000	1.00
Form 1040 returns only with combined Schedule C (business or professional) net profit or net loss of $350,000 and over, total		38	9,590	9,590	100.000	1.00
Larger of total income amounts or total loss amounts	Size of business receipts plus farm receipts					
Forms 1040 only with Form 2555						
Under $50,000	Under $500,000	80	95,382	29	0.030	3289.03
$50,000 under $100,000	Under $500,000					
Under $50,000	$500,000 under $1,000,000	81	43,514	7	0.016	6216.29
$100,000 under $500,000	Under $1,000,000					
Under $100,000	$1,000,000 under $10,000,000	82	30,736	120	0.390	256.13
$500,000 under $2,000,000	Under $10,000,000					
Under $500,000	$10,000,000 under $30,000,000	83	825	167	20.242	4.94
$2,000,000 and over	Any amount					
Under $2,000,000	$30,000,000 and over	84	39	39	100.000	1.00
Forms 1040 only with Form 1116, but without Form 2555						
Under $50,000	Under $500,000	90	203,433	50	0.025	4068.66
$50,000 under $100,000	Under $500,000					
Under $50,000	$500,000 under $1,000,000	91	153,037	55	0.037	2782.49
$100,000 under $500,000	Under $1,000,000					
Under $100,000	$1,000,000 under $10,000,000	92	132,523	531	0.401	249.57
$500,000 under $2,000,000	Under $10,000,000					
Under $500,000	$10,000,000 under $30,000,000	93	18,004	957	5.315	18.81
$2,000,000 and over	Any amount					
Under $2,000,000	$30,000,000 and over	94	3,031	747	24.645	4.06
Forms 1040 only with Schedule C, but without a Form 2555 or Form 1116						
Under $25,000	Under $200,000	60	5,730,379	3,089	0.054	1855.09
$25,000 under $50,000	Under $500,000					
Under $25,000	$200,000 under $500,000	61	4,320,756	3,527	0.082	1225.05
$50,000 under $100,000	Under $1,000,000					
Under $50,000	$500,000 under $1,000,000	62	2,246,815	3,763	0.167	597.08
$100,000 under $200,000	Under $5,000,000					
Under $100,000	$1,000,000 under $5,000,000	63	567,340	2,291	0.404	247.64
$200,000 under $500,000	Under $10,000,000					
Under $200,000	$5,000,000 under $10,000,000	64	162,976	1,896	1.163	85.96
$500,000 under $1,000,000	Under $20,000,000					
Under $500,000	$10,000,000 under $20,000,000	65	26,950	1,078	4.000	25.00
$1,000,000 under $2,000,000	Under $30,000,000					
Under $1,000,000	$20,000,000 under $30,000,000	66	8,794	1,732	19.695	5.08
$2,000,000 under $5,000,000	Under $50,000,000					
Under $2,000,000	$30,000,000 under $50,000,000	67	3,368	1,684	50.000	2.00
$5,000,000 and over	Any amount					
Under $5,000,000	$50,000,000 and over	68	985	985	100.000	1.00
Forms 1040 only with Schedule F, but without Form 2555, Form 1116, or Schedule C						
Under $25,000	Under $200,000	50	807,378	259	0.032	3117.29
$25,000 under $50,000	Under $500,000					
Under $25,000	$200,000 under $500,000	51	657,766	493	0.075	1334.21
$50,000 under $100,000	Under $1,000,000					
Under $50,000	$500,000 under $1,000,000	52	291,152	374	0.128	778.48
$100,000 under $200,000	Under $5,000,000					
Under $100,000	$1,000,000 under $5,000,000	53	63,417	177	0.279	358.29
$200,000 under $500,000	Under $10,000,00					
Under $200,000	$5,000,000 under $10,000,000	54	24,096	337	1.399	71.50
$500,000 under $1,000,000	Under $20,000,000					
Under $500,000	$10,000,000 under $20,000,000	55	5,150	198	3.845	26.01
$1,000,000 under $2,000,000	Under $30,000,000					
Under $1,000,000	$20,000,000 under $30,000,000	56	1,623	554	34.134	2.93
$2,000,000 under $5,000,000	Under $50,000,000					
Under $2,000,000	$30,000,000 under $50,000,000	57	626	626	100.000	1.00
$5,000,000 and over	Any amount					
Under $5,000,000	$50,000,000 and over	58	176	176	100.000	1.00
Forms 1040, 1040A and 1040EZ without a Form 2555, Form 1116, Schedule C or F						
Under $25,000		40	50,054,822	19,548	0.039	2560.61
$25,000 under $50,000		41	22,372,524	10,757	0.048	2079.81
$50,000 under $100,000		42	8,277,243	7,909	0.096	1046.56
$100,000 under $200,000		43	1,108,456	2,559	0.231	433.16
$200,000 under $500,000	Not applicable	44	311,216	3,176	1.021	97.99
$500,000 under $1,000,000		45	54,251	1,415	2.608	38.34
$1,000,000 under $2,000,000		46	16,310	3,343	20.497	4.88
$2,000,000 under $5,000,000		47	5,886	2,939	49.932	2.00
$5,000,000 and over		48	1,705	1,705	100.000	1.00

"graduation" of dependents into nondependent filers. The panel sample does not represent growth attributable to filing by persons who did not file as nondependents in 1987 and were not claimed as dependents in 1987 -- unless they file jointly with persons who are represented by the panel. For example, a new filer filing as a single person in 1989 is not represented by the panel, but a new filer filing as the spouse of someone who filed in 1987 is represented.

2.2 Change in Panel Composition over Time

An implication of differentiating sampling rates so sharply by income is that downward movement is represented by many more observations than upward movement. Hence, as shown in Czajka and Schirm (1992), there is a sizable net decline in sample returns falling into the higher income strata (66-68, 56-58, 46-48) and special strata (28 and 38) between 1987 and 1988 or, even more so, 1989 [1]. For example, panel returns falling into the two highest income current year strata declined by more than 40 percent between 1987 and 1989.

There is also a sizable net increase in sample returns falling into the lower income strata (60-61, 50-51, 40-41) between 1987 and 1988 or 1989. This is partly due to downward movement over time. It is also partly due to the sizable growth in the total sample size -- 5.4 percent between 1987 and 1988 and 4.7 percent between 1988 and 1989 -- from nondependent returns filed by panel members who were selected in 1987 as dependents claimed on their parents' returns.

2.3 Types of Estimation with Panel Data

Panel data can be used for both longitudinal and cross-sectional estimation. Longitudinal estimation would focus on the population represented in the base year, so the diminishing representativeness over time need not be at issue. However, a data user might also be interested in studying change for short intervals over the duration of the panel, in which case the incomplete representation of the population at the start of each interval would be of concern.

Cross-sectional estimation with panel data would ideally reflect the entire population in a given year. To achieve this would require some method of compensating for units joining the population since the panel was initiated (i.e., after the 1988 processing year). While cross-sectional estimation may seem an odd use of panel data, a panel may capture data that are not available from a cross-sectional survey. This is true, for example, of the Census Bureau's Survey of Income and Program Participation and the IRS's Sale of Capital Assets (SOCA) panel, which supplements the SOI data with capital transactions data. This paper focuses on weighting issues related to cross-sectional estimation.

3. DESIGN-BASED WEIGHTING

The construction of design-based cross-sectional weights for the SOI panel for years after the base year begins with the panel base year weight -- i.e., the inverse of the selection probability, calculated using final sample counts and true base year population totals by stratum.

This preliminary weight must be adjusted for any new persons added to panel filing units from other units that were eligible for selection in the base year. Such an adjustment is needed to avoid double-counting persons exposed to selection twice. The adjusted panel filing unit weight is the average base weight of the primary and secondary filers.

If a new member of a filing unit was not included in the base year population, no adjustment is necessary. Unfortunately, we cannot determine definitively whether a new member was in the base year population. With an available population-level data file, we can identify 1987 filers who were in the primary or secondary position or who filed as dependents, but we cannot identify nonfilers who were claimed as dependents and thus included in the base year population. Therefore, we adjust the weights of all units with new members and thus tend to overadjust.

The final step in constructing cross-sectional weights consists of an adjustment for coverage deficiencies. Such adjustment goes beyond the sample design, so it is not, strictly speaking, design-based. Czajka and Schirm (1992) discuss alternative methods of adjusting panel weights for coverage deficiencies arising after the base year. Adjustments would not be required for longitudinal analysis of the 1987 filing population, as noted above, but they become more important for cross-sectional or short-term longitudinal estimates as distance from the base year increases.

Czajka and Schirm (1992) compared estimates of the numbers of returns in the population derived from the panel and cross-sectional SOI samples for 1988 and 1989. With no coverage adjustment, the panel estimates fall 6 percent or 6 million returns short of the complete population in 1988 and 8 percent or 8 million returns short in 1989. The greatest deficiencies in percentage terms occur at high income levels. For some form types there are large deficiencies at the lowest income levels as well -- for example, strata 40 and 60. In some strata, such as strata 8, 51, 42, and 44, the panel estimates exceed the complete population [2]. To a large degree this may reflect nonsampling error -- in particular, falses elections attributable to erroneously recorded SSNs of panel members (especially dependents) in the base year or nonmembers in subsequent years. Overestimates may also reflect sampling error. This is clearly true in stratum 8 in 1988, where taxpayers whose returns were selected at very low probabilities (and thus received high weights) in the base year filed returns that would have been selected with certainty in 1988.

4. IMPLICATIONS OF DYNAMIC BEHAVIOR

This section examines the implications of panel dynamic behavior for cross-sectional estimation with the panel data. The net movement of the panel sample with respect to characteristics governing the current year stratum assignment is the result of substantially greater gross movement. Table 2 displays the distribution of design-based weights among panel returns that would have been assigned to one of three strata -- 43, 44, and 45 -- in 1988. These strata include nonbusiness, nonfarm returns with incomes ranging from $100,000 to $199,999 (stratum 43), $200,000 to $499,999 (stratum 44), and $500,000 to $999,999 (stratum 45). The weights assigned to returns originating in these three strata in 1987 are underscored. All of the returns in a given column, had they been selected into the 1988 cross-sectional sample, would have been weighted equally (with weights close to the underscored weights). Instead, the design-based weights range from 1 to over 1,000 (2,000 in strata 43 and 44) with substantial dispersion. In no stratum does the original base year weight account for more than half of all panel returns, and in stratum 45 the original base year weight is not even the modal value and accounts for fewer than one quarter of the panel returns.

Table 2. Distribution of Panel Base Weights for Selected Cross-Sectional Strata, 1988

Panel Base Weight	1988 Cross-Sectional Stratum		
	43	44	45
1.00	264	352	319
2.00	172	280	247
2.44	6	6	12
2.93	8	11	7
4.06	17	13	13
4.88	181	331	566
5.08	46	31	35
18.81	11	41	65
19.17	*	8	8
25.00	16	37	32
26.01	*	7	*
38.34	76	301	547
42.98	*	7	*
49.00	10	30	6
71.50	10	12	*
85.96	47	119	20
97.99	572	1,639	280
123.82	4	4	0
124.79	4	0	0
216.58	20	5	4
247.64	135	30	*
249.57	47	36	8
358.29	5	*	0
433.16	1,351	303	18
523.28	29	*	0
597.08	44	*	0
778.48	4	*	0
1046.56	390	33	*
1280.31	9	*	0
2079.81	21	7	0
2560.61	6	0	0
2782.49	6	0	0
Total	3,547	3,667	2,210

*Cell value suppressed because the sample count is less than 4. The column totals include weight classes not shown here because no cell count was 4 or greater.

This wide variability in the design-based weights may introduce substantial imprecision into sample estimates unless it is strongly related to variability in the characteristics being estimated. We investigated the relationships between several income and tax variables

and the design-based weights within these three strata and found the relationships to be weak.

Table 3 reports Pearson product-moment correlations between the design-based weights and each of 18 items within the three strata in 1988. For the most part the correlations are very low. Particularly notable are the correlations for adjusted gross income (AGI), which for these strata is closely related to the measure of income used for stratification. The within-stratum correlations are effectively zero, suggesting that within these 1988 strata, at least, the wide range of design-based weights does little but increase the variability of estimates of aggregate AGI. Even for items with larger correlations the range of weights seems likely to increase variance more than it reduces bias and thus to induce a net increase in the mean squared error of cross-sectional estimates.

Table 3. Correlations Between Design-Based Weights and Selected Income and Tax Items Within Three Strata, 1988

Item	Stratum		
	43	44	45
AGI or Deficit	-0.01	-0.04	0.03
Salaries & Wages	0.17	0.08	0.07
Taxable Interest	-0.27	-0.21	-0.18
Dividends	-0.17	-0.12	-0.05
Pensions/Annuities in AGI	0.01	0.03	0.01
Net Capital Gain or Loss	-0.08	-0.07	-0.02
Supplemental Gain or Loss	-0.02	-0.05	-0.01
Net Schedule E Income or Loss	-0.16	-0.10	-0.08
Gross Short-Term Capital Gain	-0.12	-0.13	-0.11
Gross Short-Term Capital Loss	-0.11	-0.09	-0.08
Gross Long-Term Capital Gain	-0.07	-0.06	-0.04
Gross Long-Term Capital Loss	-0.13	-0.09	-0.10
Partnership Passive Income	-0.12	-0.06	-0.08
Partnership Passive Loss	-0.17	-0.08	-0.12
Partnership Nonpassive Income	-0.05	-0.01	-0.04
Partnership Nonpassive Loss	-0.09	-0.07	-0.09
Total Itemized Deductions	-0.06	-0.04	0.01
Total Tax Liability	0.11	0.07	0.08

NOTE: Income items are transformed to logs of one plus their absolute values.

Across the entire sample, of course, the relationships between the design-based weights and these same characteristics are much stronger because of the association between the measures of income used for stratification in the base year and subsequent years. Table 4 reports correlations between the design-based weights

Table 4. Correlations Between Design-Based Weights and Selected Income and Tax Items by Year, 1987-1989

Item	1987	1988	1989
AGI or Deficit	-0.84	-0.77	-0.74
Salaries & Wages	-0.07	-0.04	-0.01
Taxable Interest	-0.72	-0.70	-0.69
Dividends	-0.60	-0.59	-0.59
Pensions/Annuities in AGI	0.01	-0.01	-0.00
Business Net Profit or Loss	-0.48	-0.44	-0.41
Net Capital Gain or Loss	-0.68	-0.64	-0.63
Supplemental Gain or Loss	-0.33	-0.32	-0.31
Net Schedule E Income or Loss	-0.71	-0.68	-0.66
Farm Net Profit or Loss	-0.19	-0.17	-0.16
Gross Short-Term Capital Gain	-0.41	-0.40	-0.42
Gross Short-Term Capital Loss	-0.43	-0.39	-0.40
Gross Long-Term Capital Gain	-0.65	-0.59	-0.58
Gross Long-Term Capital Loss	-0.46	-0.46	-0.45
Partnership Nonpassive Income	-0.29	-0.30	-0.30
Partnership Nonpassive Loss	-0.35	-0.35	-0.35
Total Itemized Deductions	-0.65	-0.65	-0.61
Total Tax Liability	-0.56	-0.53	-0.48

NOTE: Income items are transformed to logs of one plus their absolute values.

and the selected income and tax items across all strata in 1987, 1988, and 1989. Although the correlations decline over time for almost every item, the reductions are modest, providing evidence of the continuing relevance of the base year stratification to current year cross-sectional estimation.

Nevertheless, in light of Tables 2 and 3 we must ask if an alternative to the design-based weights will support better cross-sectional estimates. Ignoring the question of coverage adjustment, for which there is no truly design-based methodology, makes this strictly an issue of variance, as the design-based weights support unbiased estimates.

The large size of the SOI panel may seem to make moot any concerns about the variance introduced by the design-based weights for all but rare items. However, items for which the SOI sample sizes are not particularly large include some of the most policy-relevant fields on the tax return. Furthermore, the SOCA panel, with its much smaller sample size, presents the same issues with respect to weightingfor cross-sectional estimation as does the SOI panel. With SOCA, concerns about the variances of estimates of even relatively com-

mon items are paramount. The SOI panel pro-vides a data base for research on panel weighting. Alternative methods can be tested on small subsamples, and variances can be estimated from multiple replications. In our continuing research we hope to use the SOI panel for such studies.

5. ALTERNATIVE METHODS OF WEIGHTING

To explore the broader implications of what we observed for selected current year strata, we prepared estimates of 1988 income and tax items for the entire panel sample, using three alternative methods of weighting:

1) design-based weighting with a one-cell post-stratification to the 1988 complete population (specifically, a uniform 6 percent upward adjustment);
2) post-stratification to the 1988 stratum totals, ignoring the design-based weights entirely;
3) design-based weighting with post-stratification to the 1988 stratum totals.

Method (3) is a Horvitz-Thompson estimator within each 1988 stratum whereas method (2) utilizes uniform weighting within each 1988 stratum,treating panel returns as a simple random sample. Method (1) ignores the 1988 stratification.

For each estimator, post-stratification serves two functions: variance reduction and coverage adjustment. We have not produced estimates of the variance reduction resulting from post-stratification, but we can draw some inferences about it. With respect to coverage adjustment, greater use of the 1988 strata in the second and third estimators compared to the first estimator will generate better results, given the evidence of differential coverage described earlier. Comparison of the first two estimators will show the trade-off between this better coverage adjustment and the elimination of bias for the substantial portion of the population fully covered by the panel.

Table 5 presents cross-sectional sample estimates of aggregate amounts of selected income and tax items reported on nondependent returns in 1988 along with percentage deviations for estimates from the three alternative panel-based estimators. Comparing the final two columns we find the following. For AGI/D, ignoring the 1987 design within the 1988 strata does not seem to hurt the estimate. The deviation from the cross-sectional sample estimate is actually smaller when the 1987 base weights are ignored. Since the 1988 stratification is closely related to the level of AGI/D, this is where we would expect the 1988 post-stratification to perform best. Nevertheless, with respect to the deficit component,post-stratifying on the 1988 stratification without regard to the initial design, as with the second estimator, yields a very poor estimate.

For some items -- particularly the gross capital gains components -- weighting on the 1987 design with only a crude coverage adjustment is no worse than using the 1987 weights with full post-stratification on the 1988 strata. For many critical items, using the 1987 weights with a crude post-stratification produces larger deviations than using the 1987 weights with a full post-stratification, but the results from the former are often far superior to what is achieved by weighting on the 1988 stratification alone.

Clearly we cannot dispense with the 1987 weights in favor of the 1988 stratification. If we weight on the basis of the 1988 stratification alone -- without differentially weighting the returns within strata -- we appear to pay a high price in bias for most items.

What can we say about the variance impact of the variability of the 1987 design weights within the 1988 strata? Table 6 presents estimates of coefficients of variation (CVs) for the alternative estimators in Table 5. The estimates of CVs are based on the assumptions of each estimator. For example, it is assumed for the second estimator that the panel returns within a 1988 stratum are a simple random sample.

Comparing the two alternative methods of post-stratifying the design-weighted panel sample, we see the benefits of using the full 1988 stratification. Most strikingly, the estimated CVs of AGI/D and AGI are halved. While this may not be too surprising, given that the 1988 stratification is strongly related to AGI (more specifically, the absolute value of AGI/D), we find that the CV of total tax liability is also halved while that of dividends is reduced by almost as much. We find smaller reductions for other items.

Weighting on the 1988 stratification alone often produces a lower estimated CV than design-based weighting within the 1988 stratification. We suspect that

Table 5. Percentage Deviations from Cross-Sectional Estimates of Aggregate Amounts of Selected Income and Tax Items, 1988 Nondependent Returns

Item	Cross-Sectional Estimate ($1,000,000)	Panel Sample Stratification		
		1987 Strata	1988 Strata	
			Unweighted	Horvitz-Thompson
AGI or Deficit	3,048,956	6.0	0.9	1.3
Income	3,089,954	5.8	1.7	1.2
Deficit	40,998	-12.3	62.4	-6.4
Salaries & Wages	2,311,428	6.5	-2.1	1.7
Taxable Interest	183,520	2.0	41.5	1.4
Dividends	75,977	3.4	45.9	1.0
Pensions/Annuities in AGI	138,694	4.2	-3.1	1.5
Business Net Profit or Loss	126,122	14.6	7.9	-0.8
Profit	145,272	13.6	12.4	-1.5
Loss	19,150	6.8	42.1	-5.8
Net Capital Gain or Loss	151,845	-2.9	-7.0	-5.8
Gain	159,761	-2.4	-5.0	-5.4
Loss	7,916	8.6	34.6	2.2
Supplemental Gain or Loss	1,843	-15.4	-57.0	-26.6
Gain	6,260	-4.8	3.5	-10.8
Loss	4,417	-0.3	28.8	-4.2
Net Schedule E Income or Loss	58,919	5.0	-19.4	3.8
Income	126,695	1.8	17.7	1.0
Loss	67,776	-1.0	50.0	-1.5
Farm Net Profit or Loss	-1,246	-18.0	286.9	-12.6
Profit	11,136	-2.9	-1.5	-5.6
Loss	12,382	-4.4	27.5	-6.3
Gross Capital Gain or Loss				
Short-Term Gain	14,000	-1.1	93.8	0.9
Short-Term Loss	26,737	-2.2	124.3	-4.5
Long-Term Gain	169,953	-2.3	3.6	-5.3
Long-Term Loss	45,519	1.5	119.5	-1.5
Partnership				
Passive Income	18,530	-0.2	51.3	2.9
Passive Loss	27,844	-6.8	70.8	-0.7
Nonpassive Income	45,932	1.7	9.4	1.5
Nonpassive Loss	13,704	4.2	75.3	9.0
Total Itemized Deductions	394,912	7.8	17.3	0.8
Total Tax Liability	428,092	5.8	1.8	0.3

Table 6. Coefficients of Variation (Percent) for Estimates of Aggregate Amounts of Selected Income and Tax Items, 1988 Nondependent Returns

Item	Cross-Sectional Sample	Panel Sample Stratification		
		1987 Strata	1988 Strata	
			Unweighted	Horvitz-Thompson
AGI or Deficit	0.15	0.34	0.18	0.17
Income	0.14	0.33	0.16	0.15
Deficit	2.27	3.34	2.45	3.09
Salaries & Wages	0.25	0.30	0.22	0.24
Taxable Interest	1.10	1.01	0.80	0.98
Dividends	1.64	2.82	1.32	1.82
Pensions/Annuities in AGI	1.57	1.46	1.31	1.41
Business Net Profit or Loss	1.58	1.84	1.46	1.60
Profit	1.24	1.52	1.09	1.27
Loss	2.97	3.48	2.37	3.21
Net Capital Gain or Loss	1.18	3.65	1.91	2.25
Gain	1.11	3.45	1.77	2.13
Loss	2.54	2.13	1.77	2.18
Supplemental Gain or Loss	24.52	36.84	51.64	38.51
Gain	4.91	6.74	4.00	6.14
Loss	7.52	9.30	5.54	9.21
Net Schedule E Income or Loss	3.74	6.10	5.14	4.55
Income	1.44	2.73	1.26	1.88
Loss	1.61	1.87	1.37	1.78
Farm Net Profit or Loss	66.87	79.37	15.21	67.91
Profit	5.26	5.75	4.46	5.28
Loss	3.82	3.83	2.93	3.36
Gross Capital Gain or Loss				
Short-Term Gain	3.54	3.89	3.71	3.83
Short-Term Loss	8.56	5.53	5.94	5.45
Long-Term Gain	1.15	3.31	1.98	2.06
Long-Term Loss	5.49	4.15	6.79	3.92
Partnership				
Passive Income	3.08	3.46	2.39	3.28
Passive Loss	1.94	2.17	1.81	2.05
Nonpassive Income	2.99	3.62	2.55	3.38
Nonpassive Loss	3.32	4.30	2.66	4.05
Total Itemized Deductions	0.55	0.55	0.53	0.51
Total Tax Liability	0.23	0.60	0.40	0.28

this result may reflect mainly the erroneousness of the assumption that the returns in each 1988 stratum are a simple random sample rather than the adverse effects of the variability of the base year weights. Comparing the last column of panel CVs with the 1988 cross-sectional CVs, for which the assumption of simple random sampling within the 1988 strata is correct, provides some sense of the impact of variability in the base year weights on the estimates. We see little impact for AGI, which surprises us, as we have seen how in at least three strata the variability of the base weights is unrelated to AGI, suggesting that the weights should do little but reduce precision. Similarly, on most other items the 1987 weighting combined with post-stratification on the 1988 stratification yields CVs roughly equal to CVs for estimates from the cross-sectional sample.

In short, the estimated CVs do not give much evidence of the adverse effects of the variability of the 1987 weights among returns that are similar with respect to their 1988 stratification. Does our direct calculation of estimated variances for the estimators that use both the 1987 design and the 1988 cross-sectional stratification understate those variances? We plan further evaluation using resampling as a basis for estimating variances.

6. ADDITIONAL ESTIMATORS

We considered the use of shrinkage estimators to reduce the variability of weights while retaining some of the design-based differential weighting within the 1988 strata. Within each stratum we define a shrinkage weight as a weighted sum of an observation's own 1987 weight and that of the filing units that remained in that stratum from 1987 (usually this is the modal 1987 weight). Thus we shrink the weights to the weight of the "stayers."

We specified and applied a number of alternative schemes for defining the fractional shares assigned to the two weights. These included uniform shares within each 1988 stratum and more elaborate schemes based on the relative precision of the population estimate of the number of returns exhibiting a given transition (and thus having a given base weight) versus the population estimate of the number of stayers. We have not evaluated the variances of estimates based on these shrinkage weights -- which, of course, is what we are seeking to reduce. The point estimates (see Czajka and Schirm, 1992) suggest that for many items the bias introduced by the best shrinkage estimators is not large. Clearly, though, there was need for improvement.

We are also considering the specification of shrinkage estimators that differentiate among returns with the same 1987 weight such that the relative weighting assigned to the 1987 design weight versus the weight of the stayers depends on characteristics of each observation. One method of achieving this is to utilize propensity scores to differentiate among observations with respect to their resemblance to stayers, but we have not developed an approach for implementing this method. A principal problem is that giving the greatest weight to the base weights of observations that differ most from stayers seems to enhance rather than reduce the impact of variable base weights.

ACKNOWLEDGMENTS

This research was performed under contract to the Board of Governors of the Federal Reserve System. We are grateful for this support. We also thank the Statistics of Income Division of the Internal Revenue Service for supplying the data and providing other research support, and Donald B. Rubin and Roderick J. A. Little for encouraging us to pursue some of the issues addressed in this work. Finally, we are deeply grateful to Bob Cohen and Randy Hirscher of Mathematica Policy Research, Inc. for exceptionally skillful programming. Any errors are our own.

NOTES

[1] The current year "stratum" of a panel return indicates the stratum to which that return would be assigned if selected into the cross-sectional sample in that year. In fact, many -- up to about two-thirds (Czajka and Schirm, 1990) -- of the panel returns are selected into the cross-sectional sample in subsequent years, owing to the use of the SSN in selecting the cross-sectional sample. However, with respect to the panel sample design, the stratum of each panel member remains fixed over time.

[2] Stratum 8 in the current year cross-sectional sample includes returns that editing revealed had been incorrectly assigned to and selected from strata sampled with certainty. Returns in stratum 8 receive current year cross-sectional weights of one, but their base year weights (if they are panel members) may differ.

REFERENCES

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

Czajka, John L. and Schirm, Allen L. (1990). "Overlapping Membership in Annual Samples of Individual Tax Returns." *1990 Proceedings of the Section on Survey Research Methods*. Alexandria, VA: American Statistical Association.

Harte, James M. (1986). "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS." *1986 Proceedings of the Section on Survey Research Methods*. Washington, DC: American Statistical Association.