COMBINING PANEL AND CROSS-SECTIONAL SELECTION IN AN ANNUAL SAMPLE OF TAX RETURNS

John L. Czajka, Mathematica Policy Research, Inc.; and Bonnye Walker, Internal Revenue Service
John L. Czajka, 600 Maryland Ave., S.W., Suite 550, Washington, D.C. 20024

1. INTRODUCTION

Each year the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) draws a sample of individual tax returns filed during that year. The principal usage of the individual sample lies in the production of aggregate statistics. However, policy analysts in the Treasury Department, Congressional agencies, and elsewhere also use SOI microdata--primarily for research on the operation of the tax system. Applications include simulation of prospective changes to the tax code in order to estimate their revenue implications and distributional impact.

The broad scope of tax legislation in recent years has placed extreme demands upon the microdata. One area in which these demands have been strongly felt is in the estimation of behavioral responses to hypothetical tax law changes. In response to these demands, the SOI Division has undertaken a major redesign of the individual returns sample. The new design will introduce the following elements: (1) the collection of data for family units, (2) a revision of the income stratification, and (3) the inclusion of a large panel representative of the total population of individual taxpayers. This paper focuses upon the design of the panel sample and, in particular, its relationship to the cross-sectional selection that will continue to be the major focus of the SOI individual returns program.

2. THE CURRENT SAMPLE DESIGN

Under the current SOI sample design, the sampling unit is the individual tax return. Each tax return processed by the IRS during a given calendar year is assigned to a stratum and then subjected to SOI selection with a probability that varies widely by stratum. Weights for sampled units are derived from counts of the population of all filed returns, classified by the same strata. This system, with the same sampling unit and weighting methodology, albeit with a different stratification scheme, will remain the heart of the individual sample design. Currently there are 39 strata, based on nine income classes and other characteristics. The sampling rates utilized in selecting the SOI sample range from about .03 percent in the lowest income strata to 100 percent in two of the specialized strata and in the highest income strata for all types of returns.

2.1 Sample Selection

Within each stratum the sample selection utilizes the taxpayer's social security number (SSN). On joint returns, only the first listed or primary taxpayer's SSN is used for selection, which is a two step procedure. First, returns with a specific set of final four digits in the taxpayer's SSN are selected into a special subsample, the Continuous Work History Sample (CWHS), which represents a one in ten thousand or .01 percent random sample of the entire filing population. This subsample, which is included each year, contains about 10,000 returns in even numbered tax years and about 20,000 returns in odd numbered years, when a second set of ending digits is added to supplement the SOI sample.

For returns not selected into this subsample, selection is based upon a transformation of the SSN. Truncation of the transformed value yields a five-digit pseudo-random number which is compared to a target number for that return's stratum. Returns with transforms below the target number are selected into the sample.

The transformation algorithm remains constant from year to year, so that a given SSN always produces the same transform. Consequently, a particular SSN, once selected, will continue to be selected as long as the taxpayer's return falls into a stratum with the same or higher sampling rate. A taxpayer who drops into a lower stratum will face a reduced probability of selection.

2.1 Longitudinal Aspects of the Current Design

We have described two longitudinal features of the current SOI sample design: the CWHS subsample, and a selection mechanism that utilizes a taxpayer characteristic that remains fixed over time. In both cases sample selection is based upon the primary taxpayer's SSN, and this has implications for the continuity of sample membership over time.

A taxpayer with an SSN designated for the CWHS subsample will not be selected for the subsample in any year in which his or her SSN is recorded in the secondary position on the tax return. About half of the previously single filers who marry will disappear from the CWHS subsample for this reason, as will joint filers who reverse the listing of their SSNs on their returns. In this respect the CWHS tracking is incomplete. Of the nine percent of CWHS subsample members who leave the sample between consecutive years, perhaps as many as one-third are lost because of changes in SSN position. This loss could be eliminated by extending the CWHS selection to the secondary SSN. However, this would increase the size of the CWHS subsample

by about one-half and would require that the weighting of CWHS returns take into account the dual selection probability of joint returns.

The selection of returns on the basis of an SSN transform accounts for a substantial year-to-year overlap even among the non-CWHS portion of the sample. The overlap for this portion exceeds two-thirds. For any given sample member, the probability of reselection in the next year depends on the relative sampling rates in the strata in which that individual falls in the two years. The gradient in the sampling rates is such that the likelihood of remaining in the sample diminishes rapidly with downward mobility between strata. Even maintaining stratum membership does not guarantee selection if the sampling rate for that stratum is reduced.

For the non-CWHS portion of the sample, changes in sampling stratum among filers account for the great majority of sample losses. For the CWHS portion, changes in stratum do not affect selection, so nonfiling is the major reason for sample loss.

Because the probability of sample continuity is affected by changes in income, the existing year-to-year overlap in the SOI sample, while substantial, is inadequate for most longitudinal research. The fact that the reason for a sample member's disappearance from the sample is not known with certainty adds to the difficulty of interpreting longitudinal relationships in the current SOI sample.

3. THE NEW SAMPLE DESIGN

The interest in a panel component of the SOI sample is driven by the need for better longitudinal data generally, and by the periodic need for specialized studies utilizing multi-year panels (currently a panel study of capital assets sales is underway). A large, continuing panel will be able to provide the sampling frame for such specialized studies in the future.

3.1 Design Considerations

To address these needs, the new panel component must satisfy a number of requirements. First, to provide adequate precision the panel must be stratified in the same general manner as the cross-sectional sample. This implies sharply differentiated sampling rates for strata defined along some dimension of income. For this reason, expanding the CWHS subsample could not answer the panel sample needs. Second, the panel must include all members of tax families. The importance of this requirement is increased by the fact that the panel sample will permit more complete capture of tax family members than will the cross-sectional sample. Third, the anticipated usage of the panel as a sampling frame for specialized studies requires that the panel size be several times that of most special study panels, which typically focus on subsets of the return population. The implication is a sample size of perhaps 40,000 to 60,000 returns. Finally, since the underlying objective in drawing a panel sample is to create a data base suitable for studying change, the panel should perhaps overrepresent tax filing units that experience change. For this to be possible, panel selection must be based on at least two years of data.

Cross-sectional estimation will continue to be the primary use of the SOI individual returns sample. For this reason making the entire sample a panel, with additions each year to represent new and returning filers, was not an acceptable option. The expected mobility of sample members among strata with highly differentiated sampling rates meant that the distribution of sampled returns by strata would drift substantially from the initial allocation within as little as one year. To maintain the precision of cross-sectional estimates would require supplementation of the sample to compensate for this drift. Moreover, panel selection is significantly more difficult operationally than the current method of cross-sectional selection. Even in the absence of the statistical concerns, the risk of unanticipated problems was too great to allow the SOI Division to rely exclusively on a panel sample to support annual cross-sectional statistics.

The cross-sectional needs would have to be met by a stand alone, cross-sectional sample, probably no smaller than the current lean year sample size of about 83,000 returns. Only if the new stratification produced significantly greater efficiencies than the current design could this sample size requirement be reduced. In fact, however, a cross-sectional sample of 93,000 returns is under consideration. This includes the permanent addition of a second CWHS subsample.

Independent selection of the cross-sectional and panel returns would imply a total first year sample size of 153,000 plus approximately 14,000 additional returns for spouses and dependents filing separately. Moreover, the panel sample size would grow from year to year as couples filed separately or divorced, and as dependents became filers. Independent selection was out of the question, therefore. Instead, we had to devise a sample design that would provide significant overlap between the cross-sectional and panel components.

3.2 Elements of the New Design

As we have noted, there is substantial overlap in the SOI sample membership between consecutive years, primarily because of the manner in which the taxpayer's SSN is employed to determine selection. Selecting the panel from one year's cross-sectional sample will result in

substantial overlap between the panel and cross-sectional samples in subsequent years even if the panel design is not optimal in this regard. The overlap will diminish over time, of course, but substantial savings can be realized in the early years. The revised stratification could conceivably increase the year-to-year overlap, as well.

The SOI Division has taken advantage of the current year-to-year overlap in developing the panel sample design. Specifically, a base year panel has been designated from the TY 1987 SOI sample. Selecting a large base year panel provided a means of starting panel sampling early while deferring more specific decisions about the composition of a 40,000 to 60,000 member panel. The base year panel included more than 89,000 returns for TY 1987. This number will grow with the introduction of family sampling for TY 1988, with future filing on the part of currently nonfiling dependents, and with separations by married couples (sample size projections are provided in the next section).

A tax family includes a primary taxpayer and spouse, plus all dependents claimed by either taxpayer. The collection of family unit data for both the cross-sectional and panel samples will be accomplished by supplementing the regular sample selection. The SOI sample will continue to be a sample of filing units. However, the returns selected on this basis will be supplemented by the identification and collection of the returns of dependents and separately filing spouses of all nondependent sample members. Separate filing unit and family unit weights will be constructed, the principal differences being that dependents selected into the initial sample will not get family weights, while the family weights of couples filing separately will reflect their dual exposure to selection.

The base year panel includes all members of the tax families of the individuals whose returns were designated for the panel. All base year panel members will be "followed"--i.e., their returns will continue to be selected for SOI processing for the duration of the panel (or until panel reduction, if they are removed at that point), regardless of their tax family membership. The panel will remain representative of the cohort of tax families filing tax returns in 1988. The returns of new family members (e.g., new spouses or new dependents) will be selected for inclusion in the sample, but only for as long as these individuals remain associated with base year panel members. They will not be followed if they terminate their family memberships (e.g., through divorce or becoming nondependent).

3.3 Sample Size Projections

Preliminary projections of the SOI sample size through the first five years of the redesign are reported in Table 1. Separate projections are provided for the cross-section sample, the additional family members whose returns will be selected beginning in 1988, the base year panel sample (including family members), and the family members who joined panel families after the base year and who are not recognized as panel members themselves (i.e., they will not remain in the sample if they separate from their panel families). We also project the overlap between panel and cross-section sample membership.

Within each sample component we distinguish between nondependent and dependent returns, and we extend this breakdown to the projected number of additional returns beyond the basic cross-section sample. This distinction is important because dependent returns require minimal SOI editing in most cases and thus cost less to add to the sample than do nondependent returns.

The projections assume no reduction in the panel size, although one is planned. Without a reduction the panel sample size is projected to reach 100,755 nondependent returns and 7,785 dependent returns by 1990, with 433 additional nonpanel spouse returns and 519 nonpanel dependent returns. By 1991 more than 4,000 additional panel returns will have been added.

The overlap between the panel and cross-section samples is projected to decline from 68,890 in 1988 (the first year of distinct panel and cross-section sampling) to 40,177 by 1991, with the rate of decline diminishing over time. If panel reduction were to be implemented in 1990, and the number of nondependent panel member returns reduced to 60,000, the net impact on the total SOI sample size would not be as large as the reduction in the size of the panel. Because of the overlap between the two samples, about 40 percent of the panel members eliminated from the panel sample would still be selected into the cross-sectional sample. Thus the net reduction in sample size would be about 60 percent of the reduction in the size of the panel sample.

The total size of the SOI sample is projected to climb from 120,000 in 1987 to 176,330 by 1991, if no panel reduction occurs. Moreover, there is preliminary evidence from the TY 1988 sample suggesting that the actual number of returns filed by dependents of sample members may be double the projected number. This appears to be due to higher rates of filing by dependents of sample members than by dependents as a whole--a possibility about which we could only speculate until suitable data on tax families became available.

4. CROSS-SECTIONAL ESTIMATION

For the near-term, the SOI Division will continue to base its published income statistics on the cross-sectional portion of the sample--i.e.,

excluding the nonoverlapping panel returns. These nonoverlapping returns are a resource that the SOI Division's major clients are reluctant to exclude from their cross-sectional estimates. To create cross-sectional weights for the entire sample requires a method of dealing with the fact that the nonoverlapping panel returns are not representative of the strata in which they happen to fall. For the most part, all of them are movers from strata with higher income levels. To enable utilization of the entire sample in cross-sectional estimation, we have developed a theoretical approach to weighting the combined sample and have designed procedures to obtain the necessary population and sample data to support the estimation of these weights.

4.1 Theory

To develop the combined panel and cross-section weights, we will utilize an approach that was proposed to us by Roderick Little and Donald Rubin. This approach recognizes that a return in the combined sample in any year may have been selected on the basis of the current primary or secondary filer's 1987 stratum rather than the current year stratum of the return itself. This suggests poststratifying on the combination of 1987 and current year stratum to develop suitable weights.

To implement this weighting methodology requires information on the 1987 and current year stratum assignments for both the sample and the current year population. Moreover, because changes in the composition of tax filing units can affect sample selection, filing status must be taken into consideration as well.

4.2 Proposed Implementation

Because the SOI sample selection already includes determining the sample stratum for every return in the population, the production of a microdata file of sample selection information for the entire population in a given tax year can be achieved as a byproduct of sample selection. The 1987 and current year files can then be linked to provide a data base for generating the required population counts.

To produce the population counts required for weighting the combined sample, we will cross-tabulate the 1987 stratum by the current year stratum for all filing units in the current year population (see Figure 1). Possible changes in filing unit composition carry the following implications. A single taxpayer in the current year who filed in 1987 may have done so as either a single taxpayer or as part of a joint filing unit (to which another current year taxpayer may be matched). Consequently, the number of relevant 1987 classes for persons filing as single taxpayers in the current year is twice the number of 1987 sample strata (plus one additional status for persons who did not file in 1987). Couples who file jointly in the current year may have filed jointly or separately in 1987. If they filed separately, their returns may have been assigned to different strata. Figure 2 illustrates the impact of filing status for a sample design with only two strata. Even in this most simple case, there are seven alternative statuses for the 1987 returns of a couple filing jointly in the current year. Moreover, additional distinctions might be drawn between primary and secondary taxpayers on joint returns, between dependent and nondependent filers, and between single marital status versus married filing separately.

With 39 SOI strata currently (the number will increase significantly after the revision of the stratification), the total number of selection classes implied by the combination of SOI stratum with filing status could exceed 100 in the current year and several hundred in 1987, the base year. The combination of selection classes for the two years implies thousands of possible weight classes. Many of the classes will prove to be empty, but the remaining classes will still be too numerous to serve as individual weighting cells. Consequently, a strategy for collapsing the cells of the cross-strata table will be developed from an analysis of sample data. Cells with small counts may still pose a problem, however. We plan to explore the use of raking to smooth these counts and thereby reduce the variances of the estimated weights.

Because of the uncertainties surrounding the implementation of this approach, the SOI Division will continue to generate its annual aggregate statistics from the cross-sectional sample alone. The advantages of maintaining a single data series may argue for the current method of producing aggregate statistics even after the viability of a combined weighting procedure has been thoroughly established. The principal value of joint weighting may lie in its enhancement of the sample for micro level analysis.

ACKNOWLEDGMENTS

This research has been supported by the SOI Division of the IRS, and for this support the authors are particularly grateful to Fritz Scheuren. The authors wish to acknowledge the other members of the Individual SOI Redesign Planning Team as well as Roderick Little and Donald Rubin for their contributions to the redesign of the individual sample and the proposed methodology for developing combined sample weights. Finally, the authors would like to thank Lauren Haworth of Mathematica Policy Research, Inc. for her preparation of the figures.

TABLE 1

PROJECTIONS OF STATISTICS OF INCOME SAMPLE SIZE THROUGH THE FIRST FIVE YEARS OF THE REDESIGN

Sample Component	1987	1988	1989	1990	1991
(1) Cross-section Sample Returns					
Nondependent returns	114,700	89,100	89,100	89,100	89,100
Dependent returns	5,300	4,100	4,100	4,100	4,100
(2) Cross-section Family Returns					
MFS spouse returns	0	1,410	1,410	1,410	1,410
Dependent returns	0	7,904	7,904	7,904	7,904
(3) Base Year Panel Member Returns					
Nondependent returns	89,755	91,903	96,351	100,755	105,100
Dependent returns	0	7,889	7,837	7,785	7,732
(4) Associated Panel Family Member Returns*					
MFS spouse returns (nonpanel spouses)	0	265	386	433	434
Dependent returns	0	137	324	519	727
(5) Overlap between the Panel and Cross-section Samples					
Nondependent returns	89,755	63,357	50,686	41,055	36,950
Dependent returns	0	5,533	4,426	3,585	3,227
SOI Cross-section Sample	120,000	93,200	93,200	93,200	93,200
Additional Nondependent Returns	0	30,221	47,461	61,543	69,994
Additional Dependent Returns	0	10,397	11,639	12,623	13,136
TOTAL SOI SAMPLE SIZE	120,000	133,818	152,300	167,366	176,330

* These are returns filed by persons who became members of panel families after the base year and who, therefore, are not panel members themselves.

FIGURE 1
ILLUSTRATION OF BASIC WEIGHTING CELLS FOR THE COMBINED SAMPLE

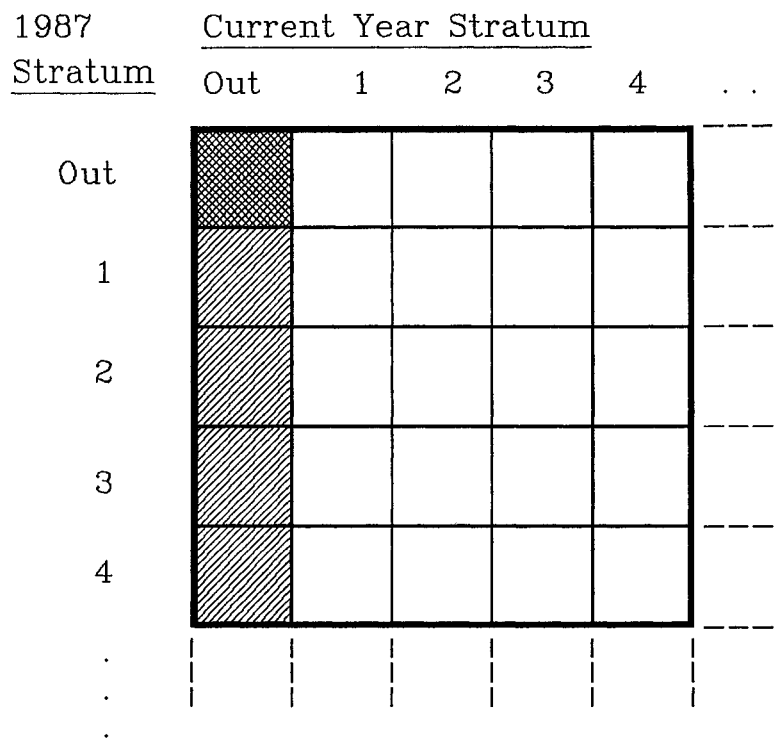

FIGURE 2
IMPACT OF FILING STATUS IN A TWO STRATA DESIGN