OVERLAPPING MEMBERSHIP IN ANNUAL SAMPLES OF INDIVIDUAL TAX RETURNS

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

KEYWORDS: Sample selection, Income change, Mortality

1. INTRODUCTION

Substantial overlap in the membership of repeated cross-sectional samples is desirable because it improves the precision of estimates of change between periods and provides an opportunity for panel-based estimation. Some of the major sources of economic and social indicators utilize rotation groups to build a fixed level of overlap into their sample designs. For example, the Current Population Survey (CPS) employs a design which provides a 75 percent overlap between sample housing units in consecutive months and a 50 percent overlap between sample housing units in two calendar months one year apart.

The annual Statistics of Income (SOI) sample of individual tax returns, an administrative record sample that is the prime resource for income and tax statistics in the United States, incorporates a selection mechanism that yields considerable overlap between consecutive years and even across several years. The methods employed to select the SOI sample differ in a number of respects from those utilized with the major household surveys. The SOI sample does not include rotation groups; nor has it included, until recently, a panel in the usual sense of the term. Yet the annual overlap exceeds that of the CPS.

This paper examines the degree of overlap in SOI sample membership over the 1984-1986 period and seeks to attribute changes in sample membership to alternative dynamic factors in the tax-filing population. An understanding of the relative importance of different factors contributing to the level of overlap is relevant to a number of impending decisions regarding the redesign of the SOI sample--particularly those relating to the integration of the panel and cross-sectional components of the sample. Our findings shed light on longitudinal behavior in one particular context, which may be relevant to other settings, and they provide some insight into problems associated with the sampling of administrative records.

2. THE STATISTICS OF INCOME SAMPLE DESIGN

Familiarity with key elements of the SOI sample design is important to understanding the nature of the year-to-year overlap in sample membership. These key elements include the stratification, the method of selecting sample units within strata, and the specification of sampling rates.

2.1 Stratification of the SOI Sample

Each tax return processed by the IRS during a given calendar year is assigned to an SOI stratum and then subjected to SOI selection. In 1984 and 1985 there were 33 strata, built around nine income classes and three return types, with additional strata added to serve specific needs. The number of strata increased slightly between 1985 and 1988 with the introduction of finer divisions among some of the specialized strata and the addition of two new strata for returns using the abbreviated schedules. The sampling rates utilized in selecting the SOI sample in 1984 ranged from about .02 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.

2.2 Selection of the Sample

Within each stratum, sample selection is based on the first listed or primary taxpayer's social security number (SSN), which is used in two ways. First, returns with specific sets of final four digits in the taxpayer's SSN are selected into a special subsample, the Continuous Work History Sample (CWHS). Returns with any one sequence of four digits represent a one in 9,999 (the sequence 0000 is not used in assigning SSNs) or .01 percent random sample of the entire filing population, and number roughly 10,000 members.

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. Once selected, a particular SSN will continue to be selected so long as it remains in the primary position and 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.

Prior to the 1988 tax year, individual tax returns were sampled at five "levels," which were supplemented on occasion to draw additional returns for special studies. Level one is a representative national sample including approximately 83,000 returns. Levels two and three are supplementary samples, designed to increase statistical precision in specific areas. Level two includes between 30,000 and 40,000 returns selected to improve the national estimates and increase the sample base for a tax model produced for odd-numbered tax years. Level three, which includes more than 200,000 returns, expands the sample for the production of state level estimates. Levels two and three are drawn by increasing the sampling rates within the strata used to select the level one sample. CWHS subsamples are included in all three sampling levels.

Levels four and five are selected in a different manner. Level four returns are selected by applying the selection rules of levels one to three to the secondary SSN (on joint returns). Level four will thus capture taxpayers who formerly filed as primary taxpayers, providing that they have not dropped into strata for which the level three sampling rates are lower than the rates at which they were originally selected. Level five returns have primary SSNs that were selected into the SOI sample (at any level) in a previous year, beginning with 1982, but which have not yet been selected in the current year. The level five sample grows over time as new SSNs are selected into the lower level samples. Together levels four and five included over 300,000 returns in TY 1985.

In odd-numbered tax years, both the level one and level two samples undergo SOI editing and are used in the production of the national level statistics which the SOI Division disseminates. In even-numbered years only level one returns are edited and tabulated. The overlap in the level one sample is the focus of this paper.

2.3 Specification of Sampling Rates

Given the stratum-specific sample size targets, the sampling rates are determined from projections of the population of returns by stratum. In some respects, selection into the SOI sample resembles eligibility for an entitlement program. Once the rules for selection are established, every return meeting these criteria is "eligible" and is selected. Consequently, the sample is vulnerable to errors in the projections. Over- or

underestimates of population size by stratum may produce a sample size that is larger or smaller than was intended. The sample is also vulnerable to classification errors. For example, if the taxpayer's reported cents were to be keyed by IRS as dollars, then the return would be assigned to a high income stratum and sampled at the corresponding rate.

Despite these problems, there are important operational reasons for selecting and editing returns as they are processed, rather than waiting until final population sizes are known.

3. LONGITUDINAL FEATURES OF THE DESIGN

The method of selecting the SOI sample creates the potential for very high overlap between the samples in consecutive years and even for samples several years removed. IRS has estimated the typical year to year overlap at between 60 and 70 percent. Our supposition about the overlap for longer periods of time has been that the rate of decay diminishes fairly rapidly as the filing units with the most volatile economic circumstances become a smaller part of the balance.

Part of the rationale for the SOI sample selection methods stems from the significant movement into and out of the filing population. "Births" or "rebirths" of taxpayers must be taken into account in the design. The incidence of such phenomena is much greater than the housing unit growth which annual household surveys must take into account.

A more significant factor is the substantial mobility of the population among sample strata. Because of the widely varying sampling rates, the sample would drift substantially from its initial composition in just a few years, necessitating supplementation to maintain desired sample sizes by stratum. The sample would have to grow significantly in order to maintain panel membership and statistical precision at the same time. The SOI Division has had to address these problems in the redesign of the sample, which will include a large panel component (Czajka and Walker, 1989). Evidence on movement among strata is presented by Schirm and Czajka (1990).

Another aspect of the rationale is operational. To select a designated sample of individuals each year would require passing a large file of SSNs against each tax return posted to IRS's "individual master file." The returns on the master file are not sorted by SSN at the point of selection, so the entire file of sample SSNs would have to be passed for each return processed through SOI selection. In fact, this methodology will be employed in selecting returns filed by members of the large panel subsample, which is being incorporated into the design as of TY 1988 (Czajka and Walker, 1989). This provides an occasion to reassess the method of selection.

4. FACTORS AFFECTING SAMPLE OVERLAP

Theoretically, 100 percent of the level one sample in one year could appear in the level one sample the next year. In practice, however, this does not occur. The reasons are varied, and they apply to the full sample as well as the level one sample.

4.1 Exclusion from the Level One Sample

There are a number of reasons why a taxpayer whose return is selected into the level one sample in one year might fail to appear in the level one sample the next year. The principal reasons include nonfiling, a change in income sufficiently large to affect stratum assignment, a change in filing status, an error in the recording of an SSN or income component in either year, and a change in the sampling rate for the taxpayer's stratum. In addition, if a taxpayer filed multiple returns the first year, the taxpayer may appear in the level one sample the next year, but one or more of the returns may not have a "match". We describe these factors in detail below.

Nonfiling. A taxpayer may fail to file a return in the next calendar year for any of three principal reasons:

- the taxpayer has died
- the taxpayer is not required to file a return in that year
- the taxpayer has chosen to defer filing

Mortality would account for about one percent attrition annually if the SOI sample is typical of the population as a whole. The frequency of prior year returns in the SOI sample suggests that deferred filing may account for somewhat greater attrition than mortality. Changes in the need to file appear to be the major cause of nonfiling, since nonmatches occur with much greater frequency among low income taxpayers than among higher income taxpayers.

Stratum change. Since the probability of selection varies widely by stratum, any change which subjects the taxpayer to a lower sampling rate may result in that taxpayer's return not being reselected into the level one sample. The probability that a taxpayer making a particular change in stratum assignment will be retained in the sample is given by the ratio of the sampling rates in the two years--specifically, the year two rate divided by the year one rate. A ratio of one or better implies certain selection in the second year. A ratio below 1.0 implies that some of the taxpayers making this transition will not be selected into the level one sample in the second year. For example, a ratio of 0.5 implies that half the taxpayers experiencing this particular change in stratum will not be selected in the second year.

A one-stratum decline carries dramatically different implications, depending on the original stratum. For example, there are two strata where a taxpayer dropping into the next lower stratum faces less than a 14 percent chance of being selected the next year. In other strata, the corresponding probability is as high as 80 percent.

Filing status change. A change in filing status that moves a taxpayer's SSN from the primary position to the secondary position on the return will almost always result in that taxpayer's falling out of the level one sample, even when the taxpayer ends up in a stratum with a higher sampling rate (as may occur when a second person's income is added through marriage or a shift from separate to joint filing). In this case the spouse's SSN now provides the random number that governs selection, with the effect that the original sample member's probability of reappearing in the sample is lowered from certainty to whatever the sampling rate may be for the joint return's stratum. Except in the highest income strata the sampling rates are very small. For example, if a person in the lowest income stratum marries and then files as the secondary taxpayer on a joint return in the next higher stratum, the probability that this return will be selected into the level one sample is only .03 percent.

Obviously only one-half of the persons who marry and file joint returns can appear in the primary position on the tax form, suggesting that, for the population as a whole, transitions from primary to secondary filer could be as numerous as marriages. Several factors may reduce the relative number of transitions, however. Not all persons who marry filed tax returns in the year preceding the marriage. It seems plausible that those who did file would be more likely to use the primary position on the joint return and thus retain their primary filer status. If there is a direct relationship between a partner's income and use of the primary filing status, the highly skewed income distribution of the SOI sample will tilt the balance even further toward newlyweds retaining their primary filing status. Use of the married filing separately status will add to the proportion of

newly married taxpayers who remain primary filers. Nevertheless, we would still anticipate that a substantial proportion of the unmarried sample members who marry will in so doing move out of the level one sample.

Joint filers greatly outnumber single filers in both the population and the SOI sample, and since there is no requirement that partners enter their SSNs in a particular order or be consistent from one year to the next, changes in SSN order by joint filers could produce substantial turnover in the level one sample. The SOI sample is vulnerable to the whims of taxpayers in this regard.

Recording error. An error in the primary SSN recorded by IRS will alter a taxpayer's selection probability--generally lowering it substantially. However, errors are rare since only valid primary SSNs are accepted onto the individual master file.

Errors in the recording of income components can affect selection as well. In particular, a taxpayer who is incorrectly assigned to a high income stratum and selected one year is unlikely to be selected again when assigned to the correct stratum, given the relative magnitudes of sampling rates. Similarly, a taxpayer who is assigned incorrectly to a lower stratum is unlikely to be selected in that year. We have reason to believe that errors overstating income are more common than errors understating income, but both types of errors can depress the year-to-year overlap in the level one sample.

Sampling rate change. Small revisions to the sampling rates or stratum boundaries occur every year. These revisions can affect the selection of taxpayers who are on the margin (i.e., have SSNs with transforms that lie close to the ceiling for their respective strata), even though the taxpayers' characteristics may remain unchanged. Assuming no other changes, a 10 percent reduction in the sampling rate for a stratum will result in 10 percent of the previous year's sample members in that stratum being dropped from the level one sample.

4.2 Exclusion from the Extended Sample

In most cases, a taxpayer dropping out of the level one sample will get selected into one of the remaining four levels of the extended sample. However, if a taxpayer drops out of the level one sample for one of the following reasons, that taxpayer will also fail to appear in the extended sample:

- nonfiling
- a switch from primary to secondary filer, coupled with a change in stratum assignment to one with a substantially lower sampling rate
- an erroneous primary SSN

This last situation will not always result in the taxpayer's being dropped from the sample, but the taxpayer would not be identifiable by an exact match on SSNs.

In addition, if we define overlap in terms of consecutive filing periods, returns for prior filing periods may not have matches in the next year's full sample unless they represent a persistent deferred filing pattern. Matches to prior year returns may be present in the current year's sample or (for returns more than two years old) may have been filed in earlier years.

5. DATA

The analysis described in this paper uses SOI data for three tax years: 1984, 1985 and 1986. According to the SOI scheme, 1984 and 1986 were lean years, with only the level one sample being edited, whereas in 1985 the level two supplement was edited as well. In order to maintain a consistent definition of sample inclusiveness, we need to eliminate the fluctuation created by the in-again, out-again status of the level two supplement. Consequently, in 1985 we define sample membership in terms of just level one status.

The level one sample in 1985 hit the target size with a count of 83,188. The 1984 sample was designed to hit a higher target, and the sample for that year turned out to be 94,385. In 1986 the sample size exceeded the 83,000 figure by more than 6,000. To eliminate the effects of changing sample size upon the estimated overlap we developed a set of sampling rates for 1984 giving us 83,314 returns. These rates were derived from various preliminary and final sampling rates used for the SOI sample around that time. We then applied the SOI selection methodology to identify those tax returns that would have been selected if these alternative rates had been in place instead of the actual level one rates for that year. This 1984 subsample includes barely 100 more returns than the actual 1985 level one sample. We intend to make a similar adjustment to the 1986 sample--basically converting excess level one returns to level two or higher status for the purposes of our analysis. The estimates that we report in this paper utilize the full level one sample for 1986, so they overstate the overlap between 1984 and 1986 relative to the way in which we prefer to define overlap.

Working from the 1984 sample we constructed a panel data base, using the full SOI samples from 1985 and 1986. Records were linked across years on the basis of SSNs and filing period; two returns with the same SSNs in 1984 and 1985 are considered to match if they have consecutive annual filing periods. We were able to match 91 percent of the reduced 1984 sample to returns in 1985, and 89 percent to returns in 1986. The matched records for those who dropped out of the level one sample give us a basis for determining the reasons for noncontinuation in the sample. They also enable us to draw some inferences about the 1984 sample members for whom no 1985 and/or 1986 returns were found in the full sample.

6. EMPIRICAL RESULTS

After presenting our findings with respect to the overlap between the 1984 sample and the 1985 and 1986 samples, we proceed to examine the roles of alternative factors in accounting for the nonoverlap that we observe.

6.1 Overlap with the 1985 and 1986 Samples

Table 1 reports the numbers and percentages of returns in the redefined 1984, or "base year," sample that remained in the level one sample for the next year and the year after that. Among the 83,314 returns in the base year sample, 67.2 percent were matched to returns in the 1985 level one sample, and 61.1 percent were matched to returns in the 1986 level one sample. In all, 73.3 percent of the 1984 sample could be matched to returns in at least one of the next two SOI level one samples while 55.0 percent could be matched to returns in both years. The overlap between 1984 and 1985 is consistent with estimates generated under less controlled conditions. A direct estimate of the two-year overlap has no precedent, but the result confirms our expectations of a significantly reduced rate of decline in overlap after the first year.

To demonstrate the impact of changes in sampling rates we divided the 1984 sample into returns that (1) would or (2) would not have been selected if the (lower) 1985 selection rates had been used in that year. Altogether 74,531 returns fall into the first group while 8,783 fall into the second. The difference between the first group and the 83,188 returns in the actual 1985 level one sample implies that with fixed stratum boundaries and sampling rates (specifically the 1985 configuration) the sample would have grown by nearly 12 percent between 1984 and 1985. This increase would be attributable entirely to growth in the filing population and to upward movement in income.

Not surprisingly there is a substantial difference between the two 1984 subsamples in their overlap with the 1985 and 1986 level one samples. For the 1984 returns that would have been selected under the 1985 sampling rates we were able to match 73.9 percent to returns in the 1985 sample and 65.5 percent to returns in the 1986 sample. For the second 1984 subsample we could match only 10.3 percent to returns in the 1985 sample and 23.9 percent to returns in the 1986 sample. The higher match rate for 1986 than 1985 presumably reflects movement from the 1984 and 1985 strata.

6.2 Accounting for Nonoverlap

We have seen that 32.8 percent of the 1984 base year sample could not be matched to returns in the 1985 level one sample, and 38.9 percent could not be matched to the 1986 sample. How do we account for this nonoverlap? With the matched file, utilizing returns from both the level one and extended samples, we can assess the relative importance of the alternative factors enumerated above.

Changes in sampling rates. A crude estimate of the impact of the change in sampling rates upon the 1984/1985 sample overlap can be obtained from the match rate differential for the two 1984 subsamples distinguished in Table 1. If 73.9 rather than only 10.3 percent of the 8,783 base year returns had been matched to 1985 returns, the total number of matched returns would have been increased by 5,586, and the overall match rate would have been increased by 6.7 percentage points. By this measure, the changes in sampling rates account for more than one-fifth of the nonoverlap.

Table 1--Distribution of 1984 Base Year Sample by Presence in 1985 and 1986 Level One SOI Samples

Presence in 1985 and 1986 Samples	All Returns	Returns Inside 1985 Selection Rates	Returns Outside 1985 Selection Rates
Total 1984 sample size	83,314	74,531	8,783
	Number of returns		
In 1985 level 1 sample	55,998	55,093	905
In 1986 level 1 sample	50,889	48,788	2,101
In both level 1 samples	45,813	45,235	578
In 1985 sample only	10,185	9,858	327
In 1986 sample only	5,076	3,553	1,523
Not in either sample	22,240	15,885	6,355
	Percent of total		
In 1985 level 1 sample	67.2%	73.9%	10.3%
In 1986 level 1 sample	61.1%	65.5%	23.9%
In both level 1 samples	55.0%	60.7%	6.6%
In 1985 sample only	12.2%	13.2%	3.7%
In 1986 sample only	6.1%	4.8%	17.3%
Not in either sample	26.7%	21.3%	72.4%

One weakness of this crude procedure is that the 8,753 base year returns in question have an inherently different distribution by sampling stratum than the balance of the base year sample. Since the year-to-year retention probabilities vary by stratum, as we discuss below, 73.9 percent may over- or understate the expected match rate for the small subsample.

Prior year filing. Prior year returns may reflect a routine filing pattern, with taxpayers consistently filing so late that their returns end up in the next processing year. IRS research has shown that prior year returns often have foreign income, which could contribute to such a filing pattern. We would expect to find year-to-year matches for these returns for as long as the pattern continues. However, prior year returns may also reflect extraordinary circumstances not repeated. The previous year's return may be filed with the current return, for example. Here we would not find a match for the prior year return the next year because the next consecutive return is in the same sample.

There are 2,664 returns with filing periods prior to 1984 in the base year sample. Table 2 reports that only 10.0 percent of these returns were matched to the 1985 level one sample, and only 3.9 percent to the 1986 level one sample. By the same crude measure that we employed in the preceding section, we estimate that prior year returns lower the overall match rate by about 2.0 percentage points and thus account for about 6 percent of the nonoverlap.

Table 2--Distribution of Prior Year Returns in 1984 Sample by Presence in 1985 and 1986 Level One SOI Samples

Presence in 1985 and 1986 Samples	All Returns	Returns Inside 1985 Selection Rates	Returns Outside 1985 Selection Rates
Total prior year returns	2,664	2,452	212
	Percent of total		
In 1985 level 1 sample	10.0%	10.6%	2.8%
In 1986 level 1 sample	3.9%	4.1%	1.4%
Next year return is in 1984 sample	50.9%	50.2%	58.5%
1983 returns	1,985	1,816	169
	Percent of total		
In 1985 level 1 sample	11.3%	12.1%	3.0%
In 1986 level 1 sample	4.7%	5.0%	1.8%
Next year return is in 1984 sample	53.6%	52.9%	60.9%
1982 and earlier returns	679	636	43
	Percent of total		
In 1985 level 1 sample	6.0%	6.3%	2.3%
In 1986 level 1 sample	1.6%	1.7%	0.0%
Next year return is in 1984 sample	42.9%	42.5%	48.8%

If an SOI sample member filed more than one return during a particular year, all of these returns may be picked up in the SOI sample. Consequently, the match for a prior year return may appear in the same sample. Indeed, for 50.9 percent of the prior year returns in the base year sample we found returns for the next filing period in the base year sample as well, thus explaining the paucity of matches in the 1985 file.

Base year returns with 1983 filing periods are of particular interest because one year is the most likely lag if there is any regularity to prior year filing. In the lower panels of Table 2 we separate the base year returns with 1983 filing periods from those with earlier filing periods. The match rate for 1983 returns, 11.3 percent, is nearly double the 6.0 percent match rate for earlier returns, but it is still very low. This suggests that persistent late filers may account for as little as five percent of all prior year returns.

Changes in SSN position. To investigate both the frequency and the impact of transitions between primary and secondary taxpayer, as well as other changes in filing status, we tabulated the 1985 sample selection status for all combinations of 1984 and 1985 filing status. We excluded from the base year sample all prior year returns and all returns that would not have been selected in 1984 if the 1985 sampling rates had applied, thereby eliminating more than 11,000 returns for which the overlap is very low. These exclusions leave a base year subsample of 72,079 returns, for which 76.1 percent have matches in the 1985 level one sample, 17.8 percent have higher level matches, and the remainder (6.1 percent) have no matches.

Our results are presented in Table 3. The returns are divided into three 1984 or origin filing statuses (single, married filing jointly, and married filing separately), each of which is subdivided into several 1985 or destination statuses. The 1985 statuses are defined in terms of the position of the SSN that was primary in 1984. Thus we distinguish primary and secondary taxpayers on 1985 joint returns. We also report as destination statuses nonmatches, and matches based on the 1984 secondary SSN, which are nonmatches to the 1984 primary taxpayer.

Of the 17,485 single taxpayers in the base year subsample, only 372 or 2.13 percent married and filed in the secondary position on a joint return. Predictably, virtually all of these 372 taxpayers--97.6 percent--missed selection into the 1985 level one sample. But their impact on the overall level of overlap is small: they account for only 2.84 percent of the base year sample members who moved from level one to a higher sample level.

Reversals of SSNs by married partners were even less significant in their impact on sample overlap between 1984 and 1985. Of the 53,882 married couples filing jointly in 1984, only 72 or 0.13 percent appear to have reversed the placement of their SSNs on their 1984 and 1985 returns. Another 19 primary taxpayers appear to have filed in the secondary position with a new spouse. While 93.1 percent of the 72 couples and 94.7 percent of the 19 couples moved out of the level one sample, these two groups account for only 0.66 percent of all base year sample members who moved from level one to a higher sample level. Changes in SSN position among married persons who filed separately in 1984 account for only another 0.21 percent of the total moves from level one to higher sample levels. Most of this remarkable consistency in how partners record their SSNs may be attributable to IRS's mailing labels and to professional tax preparers, whose services are employed by a substantial proportion of taxpayers. Clearly, the current sample design's reliance upon consistent ordering of SSNs on joint returns has not given rise to any serious problems.

Changes in SSN position account for only 3.71 percent of the moves from level one to higher sample levels for this large subsample of the 1984 base year sample. As a share of all nonoverlap, they probably amount to less than two percent, implying that they reduce the total sample overlap by less than one percentage point.

Delayed filing and nonfiling. Most of the nonmatches can be attributed to three causes: delayed filing, nonfiling (primarily because the taxpayer has no reason to file), and mortality.

Most late returns are processed by IRS the following year, so we can assess the approximate magnitude of late filing by searching the 1986 sample for the returns that we did not find in the 1985 sample. Of the 6.1 percent of returns that were "missing" from the 1985 sample, nearly one quarter (1.5 percent of the base year sample) did appear in the 1986 sample.

With the data available to us we could not distinguish nonfiling and mortality as sources of the remaining unmatched returns, except where taxpayers filed again in 1986. However, we would expect each cause to produce somewhat different patterns by sampling stratum. Mortality rates should rise with income, as the higher income strata have higher mean ages. By contrast, nonfiling should be far more common at the low end of the income distribution than at higher income levels.

We examined by sampling stratum the 4.7 percent of sample members for whom returns could not be found in the extended sample in 1985, after removing those whose 1985 returns we found in the 1986 sample. Among nonbusiness, nonfarm returns, unmatched returns peaked at 9.1 percent in the

Table 3--1985 Sample Selection for Base Year Subample, by 1984 and 1985 Filing Status

1984 Filing Status by 1985 Filing Status	Total Returns	Percent Level 2-5 in 1985	Share of All Returns by 1984 Filing Status	Share of All Level 2-5 Returns
Total in subsample	72,079	17.8		100.00
Single filers, 1984	17,485	12.1	100.00	16.54
Single	14,707	11.5	84.11	13.24
Married				
Joint primary	688	6.0	3.93	0.32
Filing separately	105	17.1	0.60	0.14
Joint secondary	372	97.6	2.13	2.84
Not matched	1,613		9.23	
Joint filers, 1984	53,882	19.6	100.00	82.66
Joint primary				
Same spouse	49,641	20.7	92.13	80.12
Different spouse	647	13.9	1.20	0.70
Filing separately				
Same spouse	134	13.4	0.25	0.14
Different spouse	12	8.3	0.02	0.01
Joint secondary				
Same spouse	72	93.1	0.13	0.52
Different spouse	19	94.7	0.04	0.14
Single	584	14.4	1.08	0.66
Spouse matched only	84	56.0	0.16	0.37
Not matched	2,689		4.99	
Separate filers, 1984	712	14.3	100.00	0.80
Filing separately				
Same spouse	350	16.9	49.16	0.46
Different spouse	5	20.0	0.70	0.01
Joint primary				
Same spouse	55	5.5	7.72	0.02
Different spouse	8	12.5	1.12	0.01
Joint secondary				
Same spouse	26	92.3	3.65	0.19
Different spouse	2	100.0	0.28	0.02
Single	114	5.3	16.01	0.05
Spouse matched only	7	85.7	0.98	0.05
Not matched	145		20.36	

lowest income stratum. From there it declined sharply to 1.4 percent in the next higher stratum, then rose gradually to 6.6 percent in the highest income stratum. We observed a similar pattern among business returns. The general pattern is consistent with our expectations about mortality and nonfiling.

Eliminating those taxpayers who reappeared in the 1986 sample leaves a residual group that should reflect mortality and extended nonfiling. We attributed these residual nonmatches to mortality except in the lowest income strata, where we felt that extended ineligibility was likely, and in certain specialized strata, where the residual nonmatches were too numerous to be due solely to mortality. In these exceptional strata we substituted the mortality rates estimated for nearby strata. These strong assumptions almost certainly yield an over-estimate of mortality, but without additional data we had no empirical basis for an alternative decomposition.

Summary. On the basis of the findings presented above we have constructed a complete disposition of the 1985 status of the base year sample. These results are presented in Table 4. To do so, we have made assumptions about certain sources of nonoverlap in the subpopulations excluded from Table 3. Basically, we have assumed that whatever nonoverlap was not attributable to prior year filing and changes in the sampling rates was allocated among the other sources in the same proportion as in the subpopulation on which Table 3 is based. Since Table 3 encompasses 72,000 out of the 83,000 base year returns, this assumption would have to be very wide of the mark to have much impact upon our estimates.

Table 4--Estimated Allocation of Total Nonoverlap, 1984-85

Source of Nonoverlap	Proportion of Total 1984 Sample	Proportion of Total Nonoverlap
Total	32.8	100.0
Movement to lower stratum	17.4	53.0
Change in sampling rates	6.7	20.6
Prior year filing	1.9	5.9
Nonfiling	2.9	8.8
Late filing	1.3	4.1
Mortality	1.8	5.5
(Re)marrying and filing as secondary SSN	0.4	1.3
SSN transposition, joint filers	0.1	0.2
Error in primary SSN	0.2	0.5

These results show the dominant role of movement to lower strata and changes in the sampling rates, which together account for 73.6 percent of the returns for which 1985 matches were not found, and which lower the total overlap by 24.1 percentage points. Nonfiling, prior year filing, mortality, and late filing form a second tier, accounting for between 4.1 and 8.8 percent of the nonoverlap. Marriages and remarriages which result in the taxpayer filing in the secondary position lead the bottom tier and account for just 1.3 percent of the nonoverlap. The final two sources, the reversal of SSNs by joint filers, and error in the primary SSN together account for less than one percent of the total nonoverlap.

7. CONCLUSIONS

Several conclusions emerge from this research.

While selecting sample units strictly on the primary SSN puts the edited sample at risk of high turnover among married couples (three-quarters of the SOI sample) and among subpopulations with high rates of marriage, the actual impact is small. Transpositions of SSNs among joint filing units are extremely rare. Sample losses due to new marriages are minimal.

Prior year returns made up 3.2 percent of the base year sample and had few matches to subsequent filing years, thereby lowering the overlap rate by a couple of percentage points. This is not a particularly large impact, but prior year returns do create conceptual problems for panel uses of the SOI data. Prior year returns are included in the SOI sample to represent returns that will be filed (or processed) in a later year. This strategy is rooted in earlier IRS research demonstrating that prior year returns differ from current year returns in important ways. Our research demonstrates that for more than half of the prior year returns there are current year returns for these same taxpayers in the edited SOI sample; in other words, the later year returns that these prior year returns are supposed to represent are in fact already represented. This creates complications when we attempt to use the SOI sample for panel purposes (specifically, what do we do about linking the prior year returns to the next year's sample?), and it will pose further complications for the joint weighting of panel and cross-sectional returns under the new sample design (see Czajka and Walker, 1989). However, the data provided by these returns are a valuable resource that can be used to investigate better ways to handle prior year returns.

Movement between strata accounts for most of the loss of sample members between years. Changing the stratum boundaries can eliminate movement to the extent that taxpayers retain their relative positions with respect to the income stratifier, although the fact that income changes have to be projected is a significant complication. The extent to which taxpayers do maintain their relative positions on the stratifier is a function of how that stratifier is designed. The new SOI sample design features a new stratifier whose intent was in part to provide more stability across years. Schirm and Czajka (1990) address these and other issues related to the impact of sample stratification on year-to-year overlap.

Finally, an issue that remains to be addressed is what implications the nonrandomness of the overlap may have for the statistical precision of estimates made from SOI data. The benign and probably correct view is that the improvements to precision resulting from the overlap are not as great as they would be if the overlapping sample were more representative--i.e., not so skewed away from filing units experiencing reductions in income. With the present overlap structure, estimates of positive change would tend to be more precise than estimates of negative change. It would be useful to be able to quantify this relationship, however, and to provide more specific guidance to data users for whom the estimation of change between years may be particularly important.

ACKNOWLEDGMENTS

This research was performed under contract to the SOI Division of the IRS. The authors are grateful to Fritz Scheuren and members of the Individual SOI Redesign Team for important contributions and helpful suggestions. We would also like to thank Bob Cohen and Gene McKay of Mathematica Policy Research, Inc. for their substantial programming support.

REFERENCES

Czajka, John L. and Walker, Bonnye (1989), "Combining Panel and Cross-Sectional Selection in an Annual Sample of Tax Returns," in Proceedings of the Section on Survey Research Methods, American Statistical Association.

Schirm, Allen L. and Czajka, John L. (1990), "Intertemporal Stability in Total Income and Overlap in Annual Samples of Tax Returns," in Proceedings of the Section on Survey Research Methods, American Statistical Association.