CROSS-SECTIONAL WEIGHTING OF COMBINED PANEL AND CROSS-SECTIONAL OBSERVATIONS

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

KEYWORDS: Income; Sample selection; Stratification

1. INTRODUCTION

For several years the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) has been engaged in the development and implementation of a major redesign of its annual sample of individual income tax returns (Czajka and Walker, 1989; Hostetter et al., 1990). One feature of the new design is a panel sample comparable in size to the annual cross-sectional sample, which includes more than 90,000 returns, typically. The base year panel was selected from the 1987 SOI cross-sectional sample, which was drawn from tax returns processed in 1988, representing primarily (but not exclusively) 1987 filing periods. Returns filed by panel members have been selected along with the cross-sectional sample in each subsequent year (the 1990 SOI sample is being selected and processed currently) and will continue to be selected for several more years.

The design of the panel sample, including its relationship to the cross-sectional sample, has been described by Czajka and Walker (1989). A key feature of this design is the substantial overlap that will exist between the cross-sectional and panel samples during the early years of the panel. The overlap is critical to IRS's ability to support such a large panel.[1] The overlap will diminish over time, however, causing the combined sample to grow in size. For the near term the SOI Division will continue to base its published income statistics and tax model files on just the cross-sectional portion of the combined sample.

Restricting cross-sectional estimation to those returns that were selected into the cross-sectional sample in any given year implies the exclusion of an increasingly larger number of nonoverlapping panel returns. These returns represent a resource that the major users of the data are reluctant to discard.[2] Creating cross-sectional weights for the combined 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 the nonoverlapping panel returns are movers from strata with higher income levels (Czajka and Walker, 1989). Combining the cross-sectional and nonoverlapping panel returns without properly adjusting for these differences would result in biased estimates of cross-sectional characteristics of the tax filing population.

To address these problems, we have developed a methodology for calculating cross-sectional weights for the combined sample. This paper describes the theory and its initial application to the development of cross-sectional weights for the 1988 combined sample. Section 2 provides an overview of SOI sample selection. Section 3 discusses design-based weighting for the combined sample, and Section 4 discusses an alternative approach based on poststratification. Section 5 presents empirical results from our initial application of the methodology discussed in Sections 3 and 4, and Section 6 summarizes our principal findings and conclusions.

2. STATISTICS OF INCOME SAMPLE SELECTION

To fully understand both the problem of cross-sectional weighting of a combined sample and our proposed solution, one must be familiar with both the design of the SOI sample and the procedures for selecting returns--particularly the role of the social security number (SSN).

Each tax return processed by the IRS during a given calendar ("processing") year is assigned to an SOI stratum and then subjected to SOI sample selection. For the 1987 sample there were 39 strata with sampling rates ranging from about .02 percent to 100 percent.[3]

Within each stratum, sample selection is based on the first listed (primary) taxpayer's SSN, which is used for selection in two ways (Czajka and Schirm, 1990). 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. In recent years the SOI sample has included one or two such groups.

For returns not selected into this CWHS subsample, selection is based upon an 11-digit transformation of the SSN (Harte, 1986). Truncation of the transformed value yields a five-digit pseudo-random number that 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 a higher sampling rate. A taxpayer whose income falls sufficiently will drop into a stratum with a reduced probability of selection.

3. DESIGN-BASED WEIGHTING

The basic principle underlying the proposed methodology for weighting the combined sample for cross-sectional estimation, whether by the design-based method discussed here or by the method of poststratification outlined in the next section, is that a return selected into the combined sample in any year may have been selected on the basis of either the current stratum of the return (cross-sectional selection) or the *1987* stratum of the current primary or secondary SSN (panel selection).[4] The implication is that knowledge of both the current and 1987 stratum membership of all SSNs included in the combined sample is required to calculate suitable weights. How to use this information, particularly in light of the complex relationships that may exist between tax filing units over time, is the question that we have had to answer in developing the weighting methodology. Most of our discussion focuses on the construction of weights for individual returns, or filing units, but we conclude this section with a discussion of family unit weights.[5]

3.1 Weights for Individual Returns

The combined sample weighting scheme that we employed utilizes theoretical selection probabilities derived from the panel and cross-sectional sample designs (Little, 1990). Consider the weighting of the 1988 combined sample. For a given return in the 1988 SOI universe let $S_{88} = 1$ if the return was selected into the cross-section sample for that year, and let $S_{88} = 0$ otherwise. Likewise, let $S_{87} = 1$ if an *associated* return was present in the *1987* SOI universe and was selected into the panel, and let $S_{87} = 0$ if such a return either did not exist or, if it did exist, was not selected into the panel. The probability that a return was selected into the 1988 combined sample is given by:

$$(1) \qquad p(C=1) = p(S_{87}=1 \text{ or } S_{88}=1)$$

The design-based theoretical weight is then given by:

$$(2) \qquad w = \{p(C=1)\}^{-1}$$

or the inverse of the probability of selection into the combined sample.

Critical to the implementation of this weighting scheme is the definition of an *associated* 1987 return. For a given return in the 1988 combined sample an associated 1987 return is any return which was categorically eligible for selection into the

panel and which shares an SSN (primary or secondary) with the 1988 return.[6] A categorically eligible return is one that was included in the SOI universe and whose primary filer was not identified (on the return) as a dependent of another taxpayer.[7] A complication in applying this weighting scheme arises from the fact that any one 1988 return might have several associated 1987 returns. For example, a single taxpayer in the 1988 sample may have filed multiple returns for different tax years in the preceding year; all of these returns are associated with the 1988 return. An even more complex but possibly more common situation would involve two persons who married in 1988, with one partner having ended a previous marriage in that year as well. For 1988 the couple might file a joint return, whereas for 1987 one partner filed as single while the other partner filed as married but filing separately. If this previously married partner's SSN also appeared on the former spouse's separate return for 1987, the number of 1987 returns associated with the one 1988 return would be three.[8,9] The major issues in implementing the proposed weighting scheme revolve around how we define $p(S_{87}=1 \text{ or } S_{88}=1)$ in cases such as these.[10]

Let us consider first the simplest cases.

Let π_{88} be the 1988 cross-sectional sampling rate applicable to a particular return in the combined sample. Let π_{87} be the applicable 1987 sampling rate used to select the panel.[11]

If a combined sample member's 1987 return was not categorically eligible for panel selection, we set $\pi_{87} = 0$. Then, $p(C=1)$, the combined sample selection probability, is simply π_{88}. This result obtains for the following reason. If $p(S_{87}=1)$ and $p(S_{88}=1)$ are independent, which they clearly are in this case, then

$$(3) \qquad p(S_{87}=1 \text{ or } S_{88}=1) = p(S_{87}=1) + p(S_{88}=1) - p(S_{87}=1 \text{ and } S_{88}=1) .$$

If $\pi_{87}=0$ then $p(S_{87}=1)=0$ as well, and we have:

$$p(S_{87}=1 \text{ or } S_{88}=1) = p(S_{88}=1) = \pi_{88} .$$

This is the simplest situation that we may observe.

Likewise, if a 1988 return has no associated 1987 return, then π_{87} equals zero, and the combined sample selection probability for that return is simply π_{88}.

If the 1988 return has one associated 1987 return, the correct expression for the combined sample selection probability depends on whether the selection probabilities of the 1987 and 1988 returns are independent. We regard the probabilities as independent if the two returns have different primary SSNs. We do so because sample selection depends on a transformation of the primary SSN, and the transforms of two different SSNs, even those of persons married to each other, are believed to be unrelated. If two returns have the same primary SSN, however, their transforms are identical, and their selection probabilities overlap entirely, meaning that the smaller of the two probabilities is subsumed under the larger probability and has no additional impact on selection.

The implications are as follows. For a 1988 return and an associated 1987 return with the same primary SSN, the combined sample selection probability is simply the larger of π_{88} and π_{87}. For a 1988 return and an associated 1987 return with different primary SSNs, independence of the two selection probabilities implies that the combined sample selection probability is given by the sum of the π_{88} and π_{87}, less their product.[12] These results are displayed in Table 1.

When the number of associated 1987 returns is two or greater, there may occur both independent and nonindependent pairs of selection probabilities. Table 1 lists all three possibilities for two associated 1987 returns: (1) the 1988 and the two 1987 primary SSNs are identical; (2) the 1988 and one of the 1987 primary SSNs (or any two of the three) are identical; (3) no two primary SSNs among the three are identical. Note that we use $\pi_{87,1}$ and $\pi_{87,2}$ to differentiate the selection probabilities of two associated 1987 returns.

For situations involving more than two independent associated returns we apply a general algorithm to obtain the combined sample selection probability. The number of independent selection probabilities that must be taken into consideration in calculating the combined sample weight for a 1988 return is equivalent to the number of unique primary SSNs on all of the eligible returns (nondependent 1987 returns plus the 1988 return being weighted) on which the 1988 primary and secondary SSNs appear. For each of I unique primary SSNs, let π_i represent the maximum selection probability with which that primary SSN appears among all of the eligible returns. The combined sample selection probability, then, is given by

$$(4) \qquad p(C=1) = 1 - [(1-\pi_1)(1-\pi_2) \ldots (1-\pi_I)]$$

where each expression in parentheses thus describes the probability of *nonselection* for a unique primary SSN.

A final observation concerns the relevance of other 1988 returns to the weighting of any one return. While the combined sample selection probability of an individual 1988 return is affected, at least potentially, by all appearances of its one or two SSNs on returns in the 1987 SOI universe, the selection probability does not depend in any way on any other 1988 return. Thus two 1988 sample returns with the same primary SSN, whether this occurrence is attributable to error or to a taxpayer submitting returns for two filing periods, are weighted without reference to each other. The situation is different for family weights, as we explain in the next section, but even there only for married persons filing separately. While a separately filing spouse's 1988 return is irrelevant to a taxpayer's individual probability of selection into the 1988 combined sample, the spouse's return does make an independent contribution to the *couple's* 1988 selection probability, as we explain below.

3.2 Weights for Family Units

While the SOI sample continues to be a sample of filing units (represented by individual tax returns), returns selected on this basis are being supplemented by the identification and collection of the returns of dependents and separately filing spouses of all nondependent sample members (Czajka and Walker, 1989). Family unit weights distinct from filing unit weights will be constructed, the principal differences being that: (1) dependents will not get family weights even if they were selected into the cross-sectional sample, and (2) the family weights of couples filing separately will reflect their dual exposure to selection. As with the individual filing unit weights, family unit weights will be constructed for both the cross-sectional and combined samples.

For the cross-sectional sample, family weights are assigned as follows. First, all dependent returns regardless of how they were selected are assigned family weights of zero because families are not being constructed around dependent sample members. Second, for a nondependent return with any filing status but married filing separately the combined sample family unit weight is identical to the filing unit weight. In many cases the tax family coincides with a single filing unit. If there are dependent filing units within the tax family, they do not affect the selection probability of the unit, and, as already mentioned, they receive family weights of zero. However, these dependent returns *will* be assigned family identification numbers so that they may be linked to other members of their tax families for family level analysis.

The third element of family weighting is that the cross-sectional family weight for a couple filing separately is derived as the sum of the 1988 selection probabilities of the two partners' returns, less their product. This weighting reflects the partners' independent contributions to the selection of their family unit. Note, however, that another 1988 return carrying either partner's SSN (with an earlier filing period or due to the erroneous recording of some other taxpayer's SSN) makes no contribution to the couple's selection probability. For example, if one partner has a second return in the 1988 sample from an earlier filing period, with a filing status of single and a *higher* selection probability, that return could be selected without either of the couple's separate returns being selected. This earlier return does not constitute part of a family unit with the

first two returns, and we would not define a family unit to include all three returns. Instead, we would define two separate family units. Thus there are never more than two 1988 returns that are relevant to the selection and thus the family weighting of a couple. This holds for combined as well as cross-sectional family weighting.

For separately filing couples only one partner's return will receive the family unit weight. The other partner's return will be assigned a family weight of zero, but as with dependents, a common family identification number will enable the two returns to be linked for family level analysis. If one return was selected into the cross-sectional sample and the other was not, the first return will receive the nonzero weight. Otherwise, the nonzero weight will be assigned to the return with the lower primary SSN.[13]

As in the cross-sectional sample, family weights for returns in the combined sample are identical to their filing unit weights, calculated in the manner described in the preceding section, except for dependents (who receive no family unit weights) and couples filing separately. Table 2 summarizes the calculation of design-based combined sample weights for the returns of married persons filing separately. Briefly, if there is no associated 1987 return, the combined sample family weight is the sum of the 1988 selection probabilities of the two partners' returns, $\pi_{88,1}$ and $\pi_{88,2}$, less their product. This is identical to the cross-sectional situation. If there is one associated 1987 return, the combined selection probability for the family unit is given by one of two expressions, depending on whether or not the 1987 return has the same primary SSN as one of the 1988 returns. If a couple changes from joint filing to separate filing between 1987 and 1988, then the 1987 return *will* share a primary SSN with one of the 1988 returns. Finally, if there are two associated 1987 returns, with each one matching one of the 1988 primary SSNs, the combined sample family weight is a function of the larger of the two selection probabilities for each primary SSN. This situation will occur when a couple files separate returns in both years. If one of the 1987 returns does not share a primary SSN with either 1988 return, then there are three independent selection probabilities to be taken into account in deriving the combined sample family weight. The situation is analogous to that presented when there is only one associated 1987 return but it does not share a primary SSN with either 1988 return, except that in this case one of the three probabilities (for the primary SSN that appears on two returns) is the larger of a 1987 and 1988 selection probability.

4. POSTSTRATIFICATION

In developing its annual cross-sectional weights, the SOI Division poststratifies on the design itself, using population and sample counts by sample stratum, with some corrections, to calculate the final weights. There are two ways that we can modify the design-based weighting with poststratification to take advantage of the availability of population aggregates. One is to adjust the design-based weights so that they reproduce the 1988 population counts used to weight the cross-sectional sample. Another is to define poststrata corresponding to all uniquely occurring design-based weights and calculate sample and population counts for these. The population counts would be based on return data linked across years.

We could elaborate on this second approach by developing a finer poststratification than that implied by the design-based weights. While potentially quite cumbersome, such an approach could improve the final estimates by assigning different weights to taxpayers who are making a particular transition in different directions. For example, the design-based method would assign the same or nearly the same weight to a taxpayer making a transition from very low to very high income as to a taxpayer making the reverse transition.[14] While the theoretical weights for such taxpayers may indeed be identical or nearly so, the infrequency of such transitions (and the attendant small sample counts) implies high variability between the theoretical and realized sampling rates. Poststratifying on stratum transitions would improve the precision of combined sample estimates of volatile income items.

This alternative approach is more cumbersome because it implies a cross-tabulation with as many dimensions as the number of different returns whose selection probabilities are relevant to any return being weighted. In the simplest case, where we need consider only one return in each year, we require a two-dimensional table, with each dimension having categories equal to the number of cross-sectional strata (in other words, a 39 by 39 table). For 1988 returns with two associated 1987 returns, we must add a third dimension, which multiplies the number of potential cells by 39. Obviously, many of the cells will have no sample observations or very few, so some collapsing of cells will be required, but effective use of the additional information contained in such a large tabulation implies that the method of collapsing must be carefully designed.

Fortunately, the appearance of any one SSN on multiple 1987 returns with more than two unique primary SSNs is exceedingly rare. Out of 229,592 primary and secondary SSNs in the 1988 combined sample, only 42 such cases were identified in a search of the entire 1987 return population. Only two of the SSNs appeared with more than three unique primary SSNs.

There is another set of circumstances under which weights developed by poststratification might have different (and more correct) expectations than the design-based weights, at least as specified earlier. Our formulation of the design-based weights assumes that the selection probabilities of two returns with different primary SSNs are independent. This assumption rests on the belief that the transforms of SSNs of married persons are unrelated to each other, even though the SSNs themselves may be correlated. Any similarities in partners' SSNs should be limited to the first five digits, which presumably have no effect on the value of the transform.[15] If the transforms *are* in fact correlated, then the design-based estimates of selection probabilities will tend to overstate the true selection probabilities of 1988 returns that are associated with two or more unique primary SSNs (because the product $\pi_1\pi_2$ will understate the probability of both partners being selected), and the estimated weights for these returns will be biased downward.

We can test this critical assumption empirically by generating SSN transforms for married couples and calculating their correlation. We intend to carry out this test as part of our continuing research and, if necessary, modify our formulation of the design-based selection probabilities.

In our initial development of weights for the 1988 combined sample, we have limited our use of poststratification to the adjustment of the design-based weights, as described at the beginning of this section. Future plans call for an evaluation of the merits of poststratifying on transitions between the 1987 and 1988 design strata.

5. EMPIRICAL RESULTS

Our initial test of combined sample weighting was limited to individual filing units. We will calculate family unit weights as part of our continuing research.

We developed preliminary combined sample weights for individual filing units using the methodology described in Section 3.1. Then, using the same poststrata by which the SOI cross-sectional sample is weighted, we adjusted these preliminary weights to reproduce the SOI population totals.[16] Weights of 1.0 were not adjusted, as these indicate returns selected with certainty (based upon either their 1988 stratum membership or the stratum membership of their associated 1987 returns). Except for some of the strata with few sample returns, the adjustments were quite small. Based on the preliminary weights, the combined sample estimate of the total population of returns was within .1% of the true population count. By contrast the population estimate produced by weighting the cross-sectional sample returns by the inverses of their selection probabilities differs from the true population count by .3%.

Table 3 displays combined sample estimates and deviations from the corresponding cross-sectional sample estimates for total returns by filing status. Differences by filing status are of interest because returns with different statuses may be

differentially susceptible to error in panel sample selection and combined sample weighting--particularly with the design-based methodology.

We do find differences by filing status. Single returns are underestimated by somewhat less than .2% while joint returns are overestimated by .6%. Head of household returns are underestimated by 2.1% and widow/er returns by 3.2%. The returns of married persons filing separately (without claiming a spouse exemption) are overestimated by 2.8% while the returns of those who do claim a spouse exemption are overestimated by .3%.

Any error for the statuses widow/er and married filing separately with a spouse exemption is dominated by sampling error, since the combined sample contains fewer than 200 returns between these two statuses. Nevertheless, the findings for both these categories are consistent with an overall pattern: return statuses with one filer are underestimated while those with two filers are overestimated.

This pattern is what we might expect as the result of errors in the SSNs recorded in the data base from which the SOI sample is drawn. For a panel return with a single SSN, an error in that SSN will probably result in the return not being selected (the exceptions being very high income returns and CWHS returns--as long as the error is not in the final four digits). While other returns may be added erroneously through errors that replicate panel SSNs, we would not expect this to happen sufficiently to compensate for the lost returns. For a panel return with *two* SSNs, an error in one SSN will rarely result in that return being lost, as the return can be identified by the other SSN. Moreover, selection on both SSNs implies that we are more likely to pick up erroneous returns, as there are two opportunities for error per return. Furthermore, limited evidence suggests that error rates on secondary SSNs appear to be about five times higher than error rates on primary SSNs (Czajka and Schirm, 1990). In short, it is much more difficult for panel returns with two SSNs rather than one SSN to miss sample selection because of an erroneous SSN while at the same time two-SSN returns have a much greater chance than one-SSN returns of being selected into the combined sample erroneously. Both forces work in the same direction.

The implication is that we may have a number of nonpanel returns--particularly joint and married filing separately returns--in the panel sample while we are missing some panel returns of single taxpayers who actually did file for 1988. However, we can determine the full extent of this problem, and make appropriate corrections, only through a lengthy process of computer-assisted manual review, which is now underway.[17]

To measure the adequacy of the combined sample weighting scheme, even with these deficiencies in the panel sample, we calculated a number of income and tax aggregates for both the cross-sectional and combined samples, using the appropriate weights for each. The generally small discrepancies between these estimates, which are displayed in Table 4, indicate that the combined sample weighting procedures were successful. The combined sample estimate of adjusted gross income (AGI) lies within .05% of the cross-sectional estimate. A number of other combined sample estimates are about equally close to their respective cross-sectional estimates: salaries and wages, net capital gain or loss (as well as the net gain alone), Schedule E net income, and farm net profit. For eight additional items we find the combined sample estimate to be within .25% of the cross-sectional sample estimate, and another eight are within .50%. The seven items for which the combined and cross-sectional sample estimates differ by more than 1% include many of the smallest aggregates, for which sampling error is likely to be a significant factor affecting the comparison. However, the largest discrepancy occurs on an item (long-term capital loss) for which the aggregate, while small, lies close to the median among the 32 items reported in the table.

Coefficients of variation for 18 of these 32 items for the 1988 cross-sectional sample are reported in Schirm and Czajka (1991) and reproduced in the last column of Table 4. Comparing the difference between the two sample estimates to the coefficient of variation for one of the sample estimates does *not* tell us if the difference is "statistically significant," but it does give us a standard against which we can describe the sample differences as small or large.[18] For all but two of the 18 items--long-term capital losses and the net capital loss--the percentage difference between the combined sample and cross-sectional sample estimates is smaller than the coefficient of variation of the cross-sectional sample estimate, and generally substantially so. For example, the difference of .05% on AGI is only one-third the size of the coefficient of variation of that variable, as is the .32% difference on interest received. For net capital gain the difference of .05% compares to a coefficient of variation of 3.05%. Thus the combined sample estimates are indeed quite close to the cross-sectional estimates.

The 11.75% difference on long-term capital losses is one of the two exceptions, being more than twice the size of the 4.70% coefficient of variation, and the .41% difference on net capital loss is about 50% larger than the .28% coefficient of variation of that item. We are inclined to investigate the differences on these and some of the other items where the two sample estimates have large discrepancies relative to the cross-sectional sample coefficients of variation, because there is a seeming inconsistency here. If the combined sample weighting methodology is correct, then differences between the two estimates should be due almost entirely to sampling error plus the nonsampling error that affects both samples; a discrepancy much larger than the cross-sectional coefficient of variation is difficult to explain.

6. CONCLUSION

This paper has described the development and application of a procedure for weighting a combined sample of panel and cross-sectional observations in order to produce an enhanced sample that can be used for cross-sectional analysis. The methodology that we have tested relies on a formulation of the theoretical probability of inclusion in the combined sample, based on the selection probabilities for the current year and for the base year of the panel. Our results provide encouraging evidence that the weighting procedure works quite well but that sample selection errors with respect to panel returns may be nonnegligible. We plan to re-estimate our weights following extensive review of the panel sample.

An alternative to the design-based weighting procedure tested here would rely more heavily on poststratification. We need to look at the merits of poststratifying on stratum transitions--particularly with respect to improving the estimates of volatile income items, whose fluctuations account for large changes in stratum assignment. However, the operational problems in developing suitable population estimates of stratum transitions are not small. Linking the 1987 and current year populations, or at least very large samples, is a sizeable undertaking in and of itself. If erroneous recording of SSNs proves to be a serious problem, false matches between records in the population files will tend to overstate rare transitions--perhaps sufficiently to negate the potential gains from poststratifying. The feasibility of editing the population data to eliminate these false matches may determine the viability of poststratifying on stratum transitions. Nevertheless, this alternative approach should indeed be studied further.

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 are indebted to Bob Cohen of Mathematica Policy Research, Inc. for the substantial programming efforts that made this work possible. We would also like to thank Roderick J. A. Little and Donald B. Rubin of Datametrics Research, Inc., for their significant contributions to the overall design of the weighting procedures and their valuable suggestions along the way. Any errors are entirely our own.

NOTES

[1] Overlapping returns do not add to the total sample size and therefore do not increase the cost of processing the SOI sample.

[2] These major users include the Office of Tax Analysis (OTA) in the Department of the Treasury, the Bureau of Economic Analysis (BEA) in the Department of Commerce, and the Joint Committee on Taxation of the United States Congress.

[3] Unless there is explicit mention of a tax accounting period, the reference year corresponds to the SOI universe for which a return is eligible, which is a function of the year in which the return was processed. More specifically, the reference year is the preceding calendar year. Thus when we refer to a 1988 return we include potentially any return processed during the 1989 calendar year, which is the expected processing year for returns with 1988 accounting periods. Most of the returns processed in 1989 will indeed have 1988 accounting periods, but some of the returns filed and processed during the year will be late returns with tax accounting periods ending in 1987 or earlier. These prior year returns typically represent a few percent of the returns processed during a given calendar year.

[4] Either or both SSNs on a joint or married filing separately (MFS) return in 1988 may have appeared on one or more 1987 returns theoretically eligible for selection into the panel.

[5] A "tax family" consists of all persons associated by marriage or tax dependency and may be represented in a given year by one or more tax returns, each corresponding to a filing unit (Czajka and Walker, 1989).

[6] There is no requirement that the common SSN appear in the same position on the two returns.

[7] A 1988 panel return on which no panel member was selected as a nondependent will not receive a combined sample weight. Persons selected into the panel as dependents would have been selected from the returns of the persons who claimed them, and these "parent" returns would determine the relevant 1987 selection probabilities. While we would be able to identify the parent returns of panel members, we could not do so for nonpanel returns and therefore could not properly weight them. Dependents in the combined sample will be represented almost exclusively by cross-sectional sample returns, which in most cases will be weighted on the basis of their 1988 selection probabilities alone.

[8] A taxpayer using the filing status "married filing separately" is asked to list the spouse's SSN on the return. Thus if two partners file separately, each partner's SSN may appear on two returns for that year.

[9] There would be only two associated returns if the previously married person had filed a joint return for 1987.

[10] Errors in reported or transcribed SSNs may create additional associations which, while incorrect, must still be taken into account because they affect the 1988 selection probability of any return on which these SSNs appear.

[11] The 1987 cross-sectional sample was larger than the panel target size; panel sampling rates were specified to obtain a sample of about 89,000 nondependent returns from the cross-sectional sample, implying panel sampling rates that were less than or equal to the corresponding cross-sectional selection rates.

[12] This result is obtained from equation (3) as follows. If $p(S_{87}=1)$ and $p(S_{88}=1)$ are independent, then the probability of selection in *both* years, $p(S_{87}=1 \text{ and } S_{88}=1)$, is equivalent to the product of the two annual selection probabilities. Hence we have $p(S_{87}=1 \text{ or } S_{88}=1) = \pi_{87} + \pi_{88} - \pi_{87}\pi_{88}$.

[13] While consistent treatment is required, this choice of the lower primary SSN was arbitrary.

[14] Note that transitions involving the 100 percent strata are of no concern, as all returns making these transitions will be represented with certainty.

[15] The first three digits of the SSN contain a geographic code, and the next two are related to the year of issuance (but differently for different geographic codes). Spouses who lived within the same geographic area at the time they received their SSNs may have the same or similar values in the first three positions and potentially the next two digits as well.

[16] The SOI poststrata are the sample design strata with one additional class for returns which turn out to have been selected with certainty only because of error (for example, cents recorded as dollars). For all returns but those assigned to this special poststratum, plus a handful of other returns, the poststratum is identical to the stratum assigned at selection. The SOI cross-sectional sample weights are calculated by dividing the population count in each selection stratum (adjusted to compensate for any sample returns that have been reassigned) by the corresponding sample count.

[17] These results suggest that we should review, in particular, all 1988 returns with secondary SSNs that are panel members and primary SSNs that are not. We should also examine all occurrences of duplicate SSNs--particularly secondary SSNs. Duplicate occurrences in the same position on the return are readily identified by sorting the file on the field in question and searching for consecutive identical numbers.

[18] The standard error of the difference between the two estimates should be much smaller than the standard error of the cross-sectional estimate. We have not yet devised a suitable method of calculating the standard error of the difference, which is affected by the large overlap between the two samples and by the differential weights assigned to returns in the two samples.

REFERENCES

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

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.* Washington, DC: 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.

Hostetter, Susan, Czajka, John L., Schirm, Allen L. and O'Conor, Karen (1990). "Choosing the Appropriate Income Classifier for Economic Tax Modeling." *1990 Proceedings of the Section on Survey Research Methods.* Washington, DC: American Statistical Association.

Little, Roderick J.A. (1990). Personal communication, October 18, 1990.

Schirm, Allen L. and Czajka, John L. (1991). "Alternative Designs for a Cross-sectional Sample of Individual Tax Returns: The Old and the New." *1991 Proceedings of the Section on Survey Research Methods.*

Table 1. Design-Based Combined Sample Selection Probabilities for 1988 Individual Returns (Filing Units) by Number and Relationship of Associated 1987 Returns

Number of 1987 Returns and Relationship to 1988 Return	$p(C_{88} = 1)$
No associated 1987 return	π_{88}
One associated 1987 return	
(1) $PSSN_{88} = PSSN_{87}$	$\text{Max}(\pi_{88},\pi_{87})$
(2) $PSSN_{88} \neq PSSN_{87}$ *and* ($PSSN_{88} = SSSN_{87}$ or $SSSN_{88} = PSSN_{87}$ or $SSSN_{88} = SSSN_{87}$)	$\pi_{88} + \pi_{87} - \pi_{88}\pi_{87}$
Two associated 1987 returns	
(1) $PSSN_{88} = PSSN_{87,1} = PSSN_{87,2}$	$\text{Max}(\pi_{88},\pi_{87,1},\pi_{87,2})$
(2) $PSSN_{88} = PSSN_{87,1} = SSSN_{87,2}$ *or* [$PSSN_{88}$=$PSSN_{87,1}$ and $SSSN_{88}$=(P or S)$SSN_{87,2}$]	$\text{Max}(\pi_{88},\pi_{87,1}) + \pi_{87,2} - \text{Max}(\pi_{88},\pi_{87,1})\times\pi_{87,2}$
(3) [$PSSN_{88}$=$SSSN_{87,1}$ and $SSSN_{88}$=(P or S)$SSN_{87,2}$] *or* $SSSN_{88}$=[(P or S)$SSN_{87,1}$ and (P or S)$SSN_{87,2}$]	$\pi_{88} + (\pi_{87,1} + \pi_{87,2} - \pi_{87,1}\pi_{87,2}) - \pi_{88}\times(\pi_{87,1} + \pi_{87,2} - \pi_{87,1}\pi_{87,2})$

NOTE: The final expression for two associated 1987 returns can be rewritten in an alternative, equivalent form: $1 - [(1-\pi_{88})(1-\pi_{87,1})(1-\pi_{87,2})]$.

Table 2--Design-Based Combined Sample Selection Probabilities for Married Persons Filing Separately in 1988 by Number and Relationship of Associated 1987 Returns

Number of 1987 Returns and Relationship to 1988 Return	$p(C_{88} = 1)$
No associated 1987 return	$\pi_{88,1} + \pi_{88,2} - \pi_{88,1}\pi_{88,2}$
One associated 1987 return	
$PSSN_{88,1} = PSSN_{87}$	$\text{Max}(\pi_{88,1},\pi_{87}) + \pi_{88,2} - \text{Max}(\pi_{88,1},\pi_{87}) \times \pi_{88,2}$
$PSSN_{88,1} = SSSN_{87}$	$1 - [(1-\pi_{87})(1-\pi_{88,1})(1-\pi_{88,2})]$
Two associated 1987 returns	
$PSSN_{88,1}$=$PSSN_{87,1}$ *and* $PSSN_{88,2}$=$PSSN_{87,2}$	$\text{Max}(\pi_{87,1},\pi_{88,1}) + \text{Max}(\pi_{87,2},\pi_{88,2}) - [\text{Max}(\pi_{87,1},\pi_{88,1}) \times \text{Max}(\pi_{87,2},\pi_{88,2})]$

Table 3--Combined Sample Estimates of Total Returns by Filing Status

Filing Status	Combined Sample Estimate (1,000s)	Percentage Deviation from Cross-sectional Sample Estimate	Combined Sample Size
One filer			
Single	48,542	-0.18	30,021
Head of household	11,066	-2.12	6,843
Widow/er	91	-3.19	111
Two filers			
Married filing joint	48,456	0.61	92,821
Married filing separately			
Without spouse exemption	1,881	2.76	4,071
With spouse exemption	51	0.33	67

Table 4--Error for Combined Sample Estimates of 1988 Income Aggregates

Income Item	Percentage Deviation from Cross-sectional Sample Estimate	Coefficient of Variation of Cross-sectional Estimate
Adjusted gross income or deficit	-0.05	0.15
Income	-0.04	
Deficit	0.54	
Salaries and wages	-0.04	0.23
Interest received	0.32	0.98
Dividends	-0.39	1.42
Pensions and annuities in AGI	-0.21	1.44
Short-term capital gain	1.64	2.89
Short-term capital loss	5.45	7.55
Long-term capital gain	0.22	0.96
Long-term capital loss	11.75	4.70
Business net profit or loss	-0.28	1.42
Profit	-0.28	
Loss	-0.21	
Net capital gain or loss	0.04	
Gain	0.06	3.05
Loss	0.41	0.28
Supplemental gain or loss	-8.83	
Gain	-0.59	4.10
Loss	2.85	6.26
Schedule E net income or loss	-0.41	
Income	0.01	
Loss	0.38	
Farm net profit or loss	1.25	
Profit	0.03	4.66
Loss	0.16	3.35
Total itemized deductions	0.13	0.50
Total tax liability	-0.17	0.26