Longitudinal Estimates and Permanent Random Numbers in Administrative Records Studies

Paul B. McMahon, Internal Revenue Service

For the past two decades, the Internal Revenue Service has used a version of a "permanent random number" in the selection of samples for the various Statistics of Income (SOI) studies. The value of this procedure in the administrative records area lies in the ease of use, as well as the statistical virtues. These virtues are undiminished by the burden of repeated selection on the companies included in these studies, for they are never contacted. Only the data on the returns they must file are used in any case.

We examine the usefulness of this procedure, for a medium-sized study, on partnerships where the design and sampling rates have been unchanged for several years. First, however, we present some background on the studies, design, and environment. We will look into some trends in the data, then examine the variance on some year-to-year comparisons.

Background

The earliest studies in the Statistics of Income series predated computer processing and so used a manual sequential sampling procedure. With the introduction of computer processing in the late Sixties, it became possible to select the sample using ending digits of the Employer Identification Number (EIN). Reliance on an administrative records processing system has limitations, including the need to operate on its schedule.

In practice, this has meant that the sample selection process is integrated into the weekly processing. This constraint has meant that the population size is unknown at the time that the sampling scheme is placed in practice, so only the sampling rates might be preset. Our clients need detailed records for their analysis, and the administrative data on the Service's computer files are not as complete as they desire. Thus, the Statistics of Income programs must rely on samples of those return filings and supplement the data from those administrative files with additional extracted information. Fortunately, as a sampling frame, the Internal Revenue Service's Master File Systems have a reasonable number of potential stratifiers and identifiers that are relatively stable.

The structure of EIN assignment led to a serious constraint. The lead two digits were assigned based on the Internal Revenue Service District Office that served the area in which the company (or other entity) was headquartered. For some early years, when an organization claimed tax-exempt status, it was assigned a "9" in the third digit. This was later discontinued, but we still see the effect in the overall distribution.

The fourth and fifth digits were often zeroes, due to the nonuniform distribution of firms and organizations across the various districts. This leaves only the last four available for use in sample selection.

The distribution of the last four digits was not (and still is not) uniform, with significant clustering effects on the final digit in particular. This limited the differentiation in the sampling rates across the strata, since the smallest viable sampling rate was approximately two in a thousand. As the population grew, the amount of the studies' resources demanded by the expanding class of very large firms forced reductions in the selection probabilities for records in the smallest size category. Moreover, there was a certain amount of clustering of the EIN's in some classes of organizations. In the case for Fiduciaries, for example, a bank might obtain a block of sequential account numbers for their trust department. The differences between the entities within that block were expected to be minor, so selections of sequential organizations would be undesirable.

To get around these limitations, the Individual Income Tax Returns Studies experimented with sequential sampling. The weakness of this strategy arose from the need to integrate the sample's selection with the complex weekly batch processing of the administrative systems across ten sites. Controlling this operation proved difficult, expensive, and incomplete. Another solution was needed.

These problems, the limitation on the smallest sampling rate and the replication of the selected returns across reruns of the weekly sampling, were resolved by Benjamin Tepping. The method he described used the EIN, prime numbers, and modular arithmetic to create a "Transformed Taxpayer Identification Number" (TTIN). He noted that the values of C and N had to be large to create effective randomness for sample selection.

$$TTIN = [(EIN)*C] \bmod N .$$

Among this procedure's favorable qualities were straightforward computer programming and the existence of an inverse; that is, given the TTIN, one can compute the EIN that was used in its creation (Harte, 1983).

This TTIN has 11 digits, but only the last four are used in sample selection for the Business Master File Sampling Operations. It is these last four digits that we consider as the permanent random number. The programming advantages arise from the simple selection test of whether that random number (last four digits of the transform divided by 10,000) is less than a prescribed sampling rate. Since the number is generated from the EIN, in the event of a rerun during the weekly sample selection, any record previously selected would be retained, unless the sampling rate or strata boundaries were amended. Clearly, this effect also continues across years, since a business will file using the same EIN time and time again, and, therefore, if it remains the same size (and the design is unchanged), the firm will be retained in the sample across those years.

The real questions about this procedure are:

1. *How closely do the achieved sampling rates match the prescribed?*
2. *What does the retention rate look like?*
3. *What is the impact on the estimates of year-to-year change?*

■ Partnership Sample Design

To answer these questions, we use the SOI Partnership Studies for Tax Years 1993 through 1996. These studies focus on those businesses that have more than one owner, yet are not incorporated. The Partnership Studies do not have the largest sample of those produced by the Statistics of Income Division, nor the smallest, starting at 30,000 and growing to almost 40,000 by the end of this period. For these years, the design and sampling rates were constant, which gives us a good opportunity to investigate the qualities of this sample selection process.

The design employs 73 strata, divided along industry groupings, assets size classes, and a measure of operational size. This later stratifier is a composite forced on us by the way the tax code views different types of income. We used the available information to approximate the net income and receipts measures more commonly used, but we cannot recreate those items at the time of sample selection.

As Table 1 on the next page shows, about one-third of the strata are reserved for the Real Estate Partnerships. This single industry dominates the population, containing about one-third of all businesses. If we proportionately allocated the sample, we would have much less reliable estimates for the less populous industrial divisions, so about half the proportional sample is assigned to those strata. Conversely, we increased the allocation to the smaller divisions to improve those estimates.

Previous reports on the effectiveness (McMahon, 1995) of the sample design demonstrated that the current version improved the estimates of the industry divisions while maintaining the level of reliability of the major national estimates. While we alluded to some year-to-year changes in that study, we could not address those issues at that time. With the revised design came higher sampling probabilities for records in the strata for the smallest firms.

On the surface, this would not seem to give rise to any questions from the clients or public, but recall that the selection mechanism tends to retain firms in the sample. This can mean that a small company that was selected for the 1992 study (the previous sample design) in the stratum with the least probability of selection would be selected under the latest design (for Tax Year 1993) as well. Since the weights depend on the prob-

Table 1: Tax Years 1993-1996 Partnerships, Strata Definitions, and Sampling Rates

Assets $100,000,000 or more 1.00
Assets less than $100,000,000
and Receipts/Income $25,000,000 or more 1.00

Real Estate Operators

Assets ($)	Under 50,000	50,000 under 100,000	100,000 under 250,000	250,000 under 500,000	500,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 under 25,000,000
	Absolute Value of Receipts/Income ($)						
Under 250,000	0.0018	0.003	0.009	{ ...		0.030 ... }	^
250,000 under 750,000	0.0020	0.0035	0.006	{ ...		0.018 ... }	|
750,000 under 2,500,000	{ ... 0.0040 ... }		0.0065	0.008	{ ... 0.025 ... }		0.300
2,500,000 under 5,000,000	{ ...	0.010 ... }		0.015	0.013	0.030	
5,000,000 under 25,000,000	{ ...	0.020 ... }		0.020	0.040	0.050	|.
25,000,000 under 100,000,000	{ ...			0.300	...	}	0.130

Farms, Trades, Finance, and Services

Assets ($)	Under 40,000	40,000 under 100,000	100,000 under 250,000	250,000 under 1,000,000	1,000,000 under 2,500,000	2,500,000 under 5,000,000	5,000,00 under 25,000,000
Under 250,000	0.0055	0.0060	0.009	0.017	{ ...	0.065 ... }	^
250,000 under 750,000	0.0055	0.0090	0.015	0.020	{ ...	0.070 ... }	|
750,000 under 2,500,000	{ ... 0.01 ... }		0.017	0.030	0.060	0.090	0.400
2,500,000 under 5,000,000	{ ... 0.045 ... }		0.050	0.040	{ ...	0.10 ... }	
5,000,000 under 10,000,000	{ ...	0.055 ... }		0.070	0.085	0.120	
10,000,000 under 25,000,000	{ ...	0.090 ... }		{ ... 0.150 ... }		0.23	|.
25,000,000 under 100,000,000	{ ...			0.35	...	}	1.00

Mining, Construction, Manufacturing, and Transportation

Assets ($)	Under 40,000	40,000 under 100,000	100,000 under 250,000	250,000 under 500,000	500,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 under 25,000,000
Under 250,000	0.003	0.008	0.0085	0.015	{ ...	0.006 ... }	^
250,000 under 1,000,000	{ ... 0.030 ... }		0.060	{ ... 0.040 ... }		0.090	| 0.50
1,000,000 under 5,000,000	{ ... 0.070 ... }		0.120	{ ...	0.050 ... }	0.140	
5,000,000 under 25,000,000	{ ...	0.30 ... }		{ ... 0.30 ... }		0.230	|.
25,000,000 under 100,000,000	{ ...		0.40 ...			}	1.00

ability of selection, that small firm will have its weight decrease (and for the years in question, it might be more than a third). Indeed, for some smaller industries, many of the same firms were the basis for the small domain estimate in both study years. Thus, an apparently significant decrease in the estimated number of firms could simply be due to the design change.

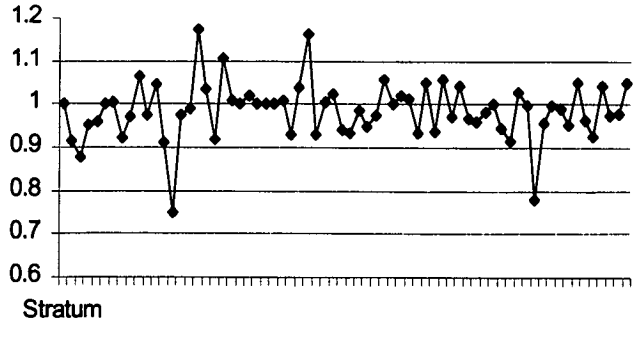
Of course, the decreased probability of selection could also result in a new record being included and result in a relatively steep rise for some other estimate, but the author has seldom had to field calls about such growth situations – decreases seem more readily apparent.

■ Sampling Rates

With a Bernoulli sampling design, the sample size is a random variable, with the probability of selection being set before sampling begins. In our case, we must develop the design at least two years before the last of the sample is chosen. That is, we do not have the sampling frame available until after the sample is selected. Thus, in the design, we use population estimates projected at least two years into the future. This gives rise to the difference between the actual proportion of the population selected for the sample and that which we might plan.

But how much does the actual rate differ from that which we set a priori? After all, this directly affects the project planning. Chart 1 shows the relative difference between the actual and planned sampling rates for the

Chart 1: SOI 1993 Partnership Sampling Rates
Actual / Expected



1993 Tax Year Study. As you see, the difference is usually less than ten percent. There are, however, four strata where the differences exceed this range. When we extend our consideration across several years, three cases retreat back into the usual ± 10 percent range.

The one stubborn case, shown on the graph in the lower left corner, should have been sampled at a 1.5-percent probability. The observed rate was approximately 1.2 percent in each year. We thought, at first, that this might have been the result of a programming error, but a careful review of the program code showed that the proper procedure was applied. This meant that the expected sample size of about 100 firms in this stratum was short by 25. Yet the impact on the estimates is ignorable because this class is only one of the 23 used in selecting the real estate operators (with an overall sample size of at least 5,000 for that single industry).

■ Sample Retention

The third reason for using a permanent random number scheme (after improved rate selection and operational simplicity) was to improve the retention of sample units from one time period to the next. We first constructed a comparison between the result of a match from the 1993 study's file to the 1994 file and what we would expect if independent selection were used. The matching routine used the Employer Identification Number, and since all of the selected records posted to the Business Master File, we can be reasonably certain that there are very few false matches on that criterion. However, we did not restrict the match to the study years' accounting periods, so some small number of extraneous records might have been introduced.

We calculated the expected sample retention by applying the sampling rates to the observed sample. This could cause a small understatement of the expected retention for those cases where the sample drops from a higher probability stratum to a lower one (by a factor that depends on the difference in the two strata probabilities). However, this method does have the quality of accounting in part for the migration of firms among the classes. The results are shown in Table 2, below.

Table 2: Sample Retention Efficiency
Tax Years 1993 to 1994

	Matching	Expected
All firms	22,986	7,700
Non-certainty	17,539	2,300

The use of the Transform Taxpayer Identification Number yields a threefold overall increase in the retained sample. When we exclude the large number of high asset or income firms that are selected for the sample with certainty, though, the improvement is a quite significant sevenfold increase over an independent selection.

Now, the population is always undergoing changes, especially births and deaths. These factors increase the value of the permanent random number procedure over a simple panel study because the change that they represent is better captured by repeated surveys. Just as clearly, though, they affect the size of the retained sample over the years, even when the sampling rates are stable. For studies of partnerships, though, births and deaths are a particular hazard because the nature of this business structure is quite suitable to ad hoc operations, such as the floatation of stocks or small construction jobs.

Since the reporting deadline follows the closure of the tax year, the sample is selected in the period after that year is done. Allowance has to be made, of course, for filing extensions and IRS's processing, which leads to the sample being drawn for an entire year. Table 3 shows that the sample size increased for each subsequent study, with a considerable jump for Tax Year 1996.

Table 3: Sample Retention Profile
Tax Years 1993 through 1996

Selection	Tax Year			
Year	1993	1994	1995	1996
1994	28,941			
1995	22,980	30,630		
1996	20,691	24,567	33,824	
1997	18,604	21,808	27,290	39,957

The mild growth in the other years arose from the increase in the number of large firms, but the 1996 growth came from a sudden increase in the population as a whole.

The retention of sample units over the four years in this review has a second year falloff of about 20 percent. The 1993 and 1994 projects had a further drop of about 9 percent in the third year. This pattern confirms the effect of the ad hoc operations on the partnership population, while further illustrating the value of the permanent random number selection procedure.

■ Effect on Estimates

The impact of this procedure on the reliability of the estimates has two aspects: the first on nonsampling errors and, of course, the second on the variability. The nonsampling errors are reduced by the increased availability of information from prior years for use in the identification and resolution of data abstraction faults. Since these Statistics of Income studies use the tax forms as the survey instrument, we must take the data our sponsors require from wherever the administrative design puts them on the various forms.

As was noted in a previous paper (McMahon, 1996), the remoteness of a datum (hidden among text on a back page, perhaps) has a strong effect on whether the clerk abstracting the information notices it. By comparing matched records, one can identify those reports that are likely to contain these obscurities.

But most of these items are not used in the published tables, so their impact on the general user is ignorable. On the other hand, all users want to know something about the distribution across industries. Here, the ability to cross-check with prior years could reduce errors by providing abstraction clerks with the codes used in previous years.

Table 4 shows that the vast majority of the retained entities received an industry code that was either the same as the previous year or in an adjacent industry. There were, of course, some miscodings in the earlier year, including a handful of records for which the in-

Table 4: Retained Sample's Industry Migration From 1995 to 1996
(Percent)

	Agriculture	Mining	Construction	Manufacturing	Transportation	Trade	Finance	Real Estate	Services
Agriculture	98.0	0.1	0.1	0.3	0.0	0.6	0.7	0.0	0.3
Mining	0.0	98.4	0.1	0.4	0.1	0.3	0.4	0.0	0.3
Construction	0.1	0.1	97.9	0.4	0.2	0.2	0.2	0.5	0.4
Manufacturing	0.5	0.3	0.2	96.2	0.9	1.3	0.2	0.1	0.4
Transportation	0.0	0.2	0.1	0.3	97.4	0.4	0.3	0.2	1.0
Trade	0.2	0.0	0.1	0.7	0.1	97.5	0.2	0.1	1.1
Finance	0.1	0.0	0.1	0.0	0.0	0.0	98.5	0.9	0.4
Real Estate	0.0	0.0	0.0	0.0	0.0	0.1	1.3	98.4	0.2
Services	0.1	0.0	0.1	0.2	0.5	0.7	0.6	0.3	97.5

dustry was not reported or discernable, but most of the changes were the result of firms changing their operations. A builder, for example, might temporarily rent out equipment, switching from construction to services.

Variance of Longitudinal Estimates

The example described above also illustrates a problem in estimating the variances across the studies, for such a change would also result in a strata migration. Roughly two-thirds of the retained sample remained in the same sampling class from one year to the next, and most of those that changed were in adjacent strata. However, of the total sample used in making the estimate of, say, asset growth, the proportion of selected firms remaining in strata declines to less than 40 percent (after allowing for births and deaths).

We do not at this time have population counts for births, deaths, or continued operations for any strata, let alone information on migrations among the sampling classes. Thus, we cannot post-stratify the samples to simplify the estimation. This situation will be remedied soon, but, in the meantime, we wish to estimate the effect the retention of firms in the sample has on the estimates.

The variance of the difference between an estimate for, say, 1995 and 1996 includes the variance for each of the two annual estimates, as well as the covariance, as shown below.

$$Var(\hat{Y}_2 - \hat{Y}_1) = Var(\hat{Y}_1) + Var(\hat{Y}_2) - 2Cov(y_1, y_2)$$

The estimates we publish in the *SOI Bulletin* each fall (e.g., Wheeler, 1994) are conditioned on the sample chosen. This means that we need to estimate a conditional variance for the year-to-year growth. This is straightforward for the variances of the individual years, but, lacking the population data, the form of the conditional covariance is not clear.

Our initial attempt to estimate the covariance used the higher weight from the two years. Since the selections are chosen using the permanent random number, the probability of selection in two studies is the smaller of the two. That is:

$$p_{(12)} = p_{2|1} p_1$$

If the firm were selected for the same strata in both years (or one with a higher sampling probability), then the probability that it would be selected in the second year, given selection in the first, $p_{2|1}$, is certainty. Hence the joint probability is the original stratum's times 1. If the firm drops into a lower probability stratum, then the conditional probability of selection in the second year, given selection in the first, is $p_{2|1} = (p_2 / p_1)^* p_1$, once again the smaller selection probability.

Since we were exploring the problem, we began by

slightly modifying the basic p estimator to reflect the observed joint selection probabilities:

$$Cov(y_1, y_2) = \frac{\sum (y_{1i} y_{2i} / p_{(12)i})}{(\hat{N} - 1)} - \frac{(\sum y_{1i} / \pi_i)(\sum y_{2i} / p_{(12)i})}{\hat{N}(\hat{N} - 1)}$$

This clearly was not the correct form, and our first estimates showed it. For example, the relative sampling error for the change in Total Assets, from 1995 to 1996, in the industry division "Wholesale and Retail Trade," computed with this covariance, was 5.1653 percent. If we had assumed the covariance were zero, that number would have been 5.1656 percent: an ignorable difference. In this case, though, the roughly $20 billion growth between 1995 and 1996 was entirely attributable to the new firms.

Yet this form of the covariance does indicate what we can do to improve year-to-year comparisons. First, we need to replace the estimated populations with counts from a sampling frame. Second, we can expect a real improvement by separating the births, deaths, and ad hoc operations from the firms that are continuing concerns. And lastly, we need to continue research in this area.

■ Further Research

To these ends, we are constructing longitudinal sampling frames for nearly all of the Statistics of Income studies. This will take considerable time, for there currently is no source that can exactly replicate the population we actually subjected to sampling over the past few years. Such a data base will, of course, be of significant use in the next round of sample redesign.

We are also funding research into efficient methods of estimating this variance. These results will affect the sample allocation in the near term and the strata design soon thereafter. Based on the strata migration patterns, it now appears that we currently have too many sampling classes, which may interfere with the post-stratification suggested above.

■ References

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

McMahon, P. (1996), "Non-Sampling Errors in Data Abstraction from Administrative Records," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, P. (1995), "Statistics of Income Partnership Studies: Evaluation of the Expanded Sampling Plan," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, P.; Collins, R.; and O'Conor, K. (1990), "Revising the Statistics of Income Partnership Sampling Plan," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Wheeler, T. (1994), "Partnership Returns, 1992," *Statistics of Income Bulletin, Fall 1994*, Internal Revenue Service.