REVISING THE STATISTICS OF INCOME PARTNERSHIP SAMPLING PLAN

Paul McMahon, Karen O'Conor and Richard Collins, Internal Revenue Service
P. McMahon, Internal Revenue Service R:S:P, P.O. Box 2608, Washington, D.C. 20013-2608

Key Words: Sample Design, Administrative Records

INTRODUCTION

This paper is a report on the first stages of building a revision to the Internal Revenue Service's (IRS) Statistics of Income Partnership sample design. First, there is a review of some background for the studies, including both historical and environmental information. This is followed by a description of the current design, including an assessment of its strengths and opportunities for enhancements. The third section deals with some features of the revised sampling plan. Finally, the paper closes with a discussion of research planned for the near future.

BACKGROUND

As a form of business organization, "Partnerships" have been around for a very long time. At present, a partnership must have at least two owners, but they may be corporations, individuals, fiduciaries, estates or other partnerships. While the Internal Revenue Service does not collect income taxes from partnerships (each partner is taxed on his or her own share of the income), an information return is required reporting income and partner interests. It is from these reports that IRS' Statistics of Income program has gathered data on partnerships for annual publication (for example, see Middough, 1990).

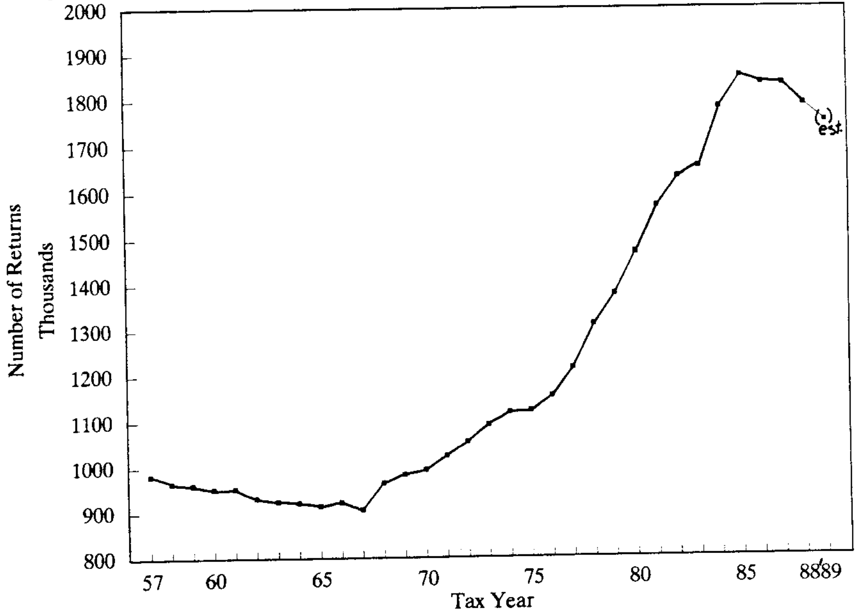
The population of partnerships has changed considerably since 1957, when the first annual sample design was employed for this series, and this shows in the number of returns filed each year. As seen in Figure A, the number of returns filed between 1957 and 1968 slowly eroded from about one million companies to just over 900,000. Between 1968 and 1976 the population began to grow at a rate of about 2 percent a year. This coincided with IRS's introduction of centralized filing and computer processing of returns in 1968; in fact it may be more indicative of the better control and lowered reporting requirements made possible by mechanization.

The growth rate more than doubled, to 5 percent, in the decade 1977 to 1986, the population reaching a high of 1,850,000 returns. This rapid increase was fueled by the appearance of tax shelters, which subsequently became one of the major targets of the 1986 Tax Reform Act (see Nelson, 1989 and Petska, 1990). Since then, the population has decreased by about 60,000 returns, and our projection for the current processing (during 1990) indicates a further decrease of about 37,000 for tax year 1989, to about 1,750,000 returns.

Over the years the sample has been through a number of changes as well, some in response to changes in the population, some due to budget fluctuations, but most brought on by the opportunities and constraints of the administrative environment.

The environment is IRS' tax processing system. This system encompasses the abstraction of selected data from the various tax forms to computer records, the verification of that data for internal consistency and the matching of the

Figure A.--Trend of Partnership Population: 1957-1989



record to an account on a master file of entities (thus verifying the entity data). Once this portion of the process is complete, the records are subjected to sampling for the studies. Because the sampling operation is dependent on the data available on the computer record, the choice of stratification variables is limited. Thus, any change in the computer records' content can cause a modification of the sample design.

It should also be noted that the Partnership program is only one of several statistical studies going on simultaneously in the Statistics of Income Division. The Business Master File system that we use to select the sample contains a large variety of return types: Corporations, Estates, and Exempt Organizations, for example, as well as tax deposit forms and a number of records used internally for tax administration. Since we select samples of a number of these forms, it is most efficient from a systems standpoint to combine the sampling programs into a single operation. Thus, we must manage the amount of change and its severity in a given year, for otherwise this complex process would fail in operation (and putting things straight is usually expensive, time consuming and very difficult).

CURRENT DESIGN

The Partnership program sample design now in place is over a decade old and rooted in the environment of that time. The original sample size for the design was 40,000 returns, with about 3,000 allotted for the five classes reserved for extra large partnerships. (See Figure B.) The stratifying variables were defined as follows:

- Income/Loss was the amount of money the company made or lost (the Bottom Line);
- Total Assets was (and still is) a measure of the financial size and includes such items as land, buildings and investments;
- Receipts was the total revenues of the company (before most costs were

Figure B.--Statistics of Income 1988 Partnerships Design and Population Counts

Total Assets $25,000,000 or more	8,664
Total Assets less than $25,000,000 and Income/Loss $5,000,000 or more	1,987
Total Assets less than $25,000,000, Income/Loss $2,500,000 under $5,000,000 and Receipts $5,000,000 or more	1,428
Total Assets $5,000,000 under $25,000,000, Income/Loss $2,500,000 under $5,000,000 and Receipts less than $5,000,000	1,399
Total Assets under $5,000,000, Income/Loss $2,500,000 under $5,000,000 and Receipts less than $5,000,000	1,029

Income/Loss under $2,500,000:

Absolute Value of Total Assets ($)	Absolute Value of Receipts ($)			
	Under 250,000	250,000 under 1,000,000	1,000,000 under 5,000,000	5,000,000 or more
		Real Estate Companies		
0 or not reported	105,659	3,367	562	48
1 under 250,000	209,678	2,334	364	36
250,000 under 1,000,000	126,858	12,859	540	29
1,000,000 under 5,000,000	42,415	42,329	6,452	88
5,000,000 under 25,000,000	5,295	6,712	10,237	622
		Non-Real Estate Companies		
0 or not reported	440,535	19,260	2,827	236
1 under 100,000	382,286	49,077	5,456	180
100,000 under 1,000,000	165,870	42,083	16,850	899
1,000,000 under 5,000,000	34,346	10,250	8,536	2,156
5,000,000 under 25,000,000	8,367	1,984	3,441	1,873

subtracted).

IRS' administrative rules were the first to cause changes to the design -- even before it went operational. An amendment to the regulations permitted some partnerships not to report asset data, which gave rise to the categories "Assets Zero or Not Reported."

All in all, there are currently 45 sampling strata. The set of strata that are reserved for Real Estate Operators traces to earlier designs. One third of the Partnership population is in that industry, but our users (primarily, the Bureau of Economic Analysis) are interested in the various industry divisions. Thus, to maintain the quality of estimates for the non-Real Estate Operators industries in an era of reduced sample sizes, only half as much sample is allocated to that industry as its proportion would seem to dictate.

The strata boundaries are also rooted in the past, both methodologically and due to constraints of the computer systems. The current certainty strata, for example, are set so low that more than half the sample size is allocated to those strata. This is eroding the distributional aspects of the sample, for while the overall sample size is stable, the growth of these certainty strata requires more and more resources to be diverted from other strata. In order to compensate for this tendency, the boundaries of the current strata must be raised from the current $25 million for Total Assets to $75 million; and for Net Income and Receipts, from $5 million to $10 million, so as to reduce the proportion of the sample allocated to these strata. This growth of large unbounded strata raises an issue that crosses many Statistics of Income (SOI) projects. In fact, many SOI staff members are currently involved in the search for longer term solutions than periodic redesigns. (See for example: Hinkins, 1988 or 1990; Jones, 1984; Hostetter, 1990; or Mulrow, 1990.)

Nevertheless, while the distributional coverage of the sample has eroded, the coefficients of variation (CV) for the various estimates at the national, all industries level have maintained quite reasonable values. Total Assets, for example, has a CV (the standard error of an estimate divided by that estimate) of less than 2 percent; Receipts, under 1 percent; and Inventories (from the Cost of Goods Sold Schedule), of only 3.5 percent.

Our users are most interested in the industry divisions' data. Figure C shows the CV's for selected estimates across the divisions. The CV's for Manufacturing and Transportation are all below 5 percent; at the other extreme, however, Construction and Agriculture have some values exceeding 25 percent. These estimates do not take advantage of the post-stratification we already use (because we want to compare these CV's to those generated by the candidate replacement designs), but the improvement is, of course, dependent on the accuracy of the population data available. Unfortunately, the dependability of industry codes for smaller industries is questionable, so both separate strata and post-strata may not hold the entire answer.

We have already mentioned that Real Estate Operators are one third of the population, far more than any industry division (except Finance, of which they are part), but at the divisional level there are three others that stand out as quite large. (See Figure D.)

Finance (including Real Estate Operators) dominates the landscape. It is a high ground that greatly influences both the estimates and their coefficients of variation at the national level. Even without the inclusion of Real Estate, though, Finance would be one of the three largest divisions. As shown in the chart, there are two secondary peaks alongside Finance: Trade and Services. These three areas contain over 80 percent of the Partnership population. (Agriculture contains another 7 percent.) The chart also presents two other features of the population -- Final Returns (filed by companies going out of existence) and companies not required to report on their assets. Final Returns are important to the development of longitudinal estimators, while the other group

Figure C.--Coefficients of Variation For Selected Variables by Industry Division

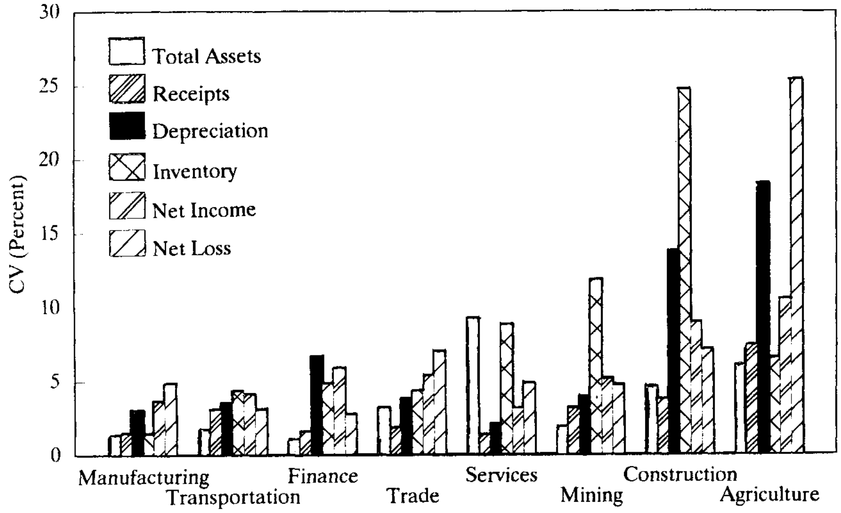
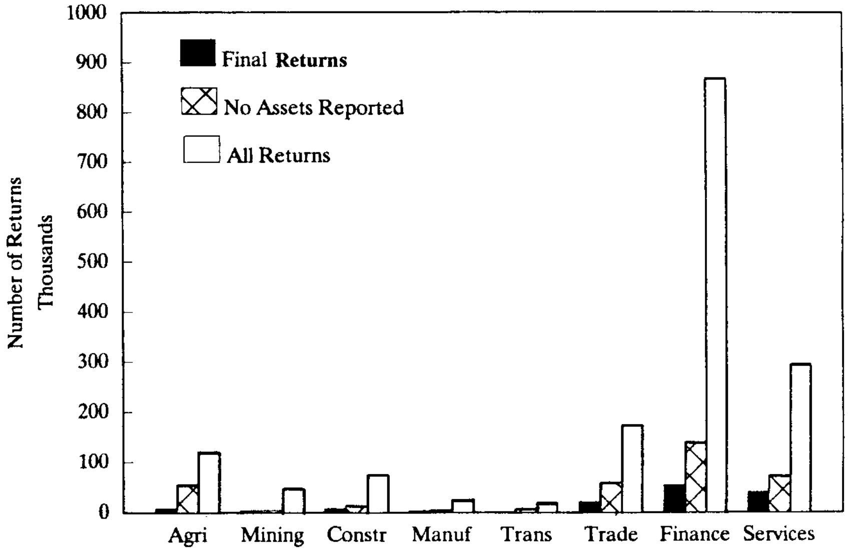


Figure D.--Distribution of Partnership Population by Industry



is, as the next section indicates, one focus of the redesign effort.

FEATURES OF THE REVISED DESIGN

Our primary users are particularly interested in income by industry and changes in income patterns over time. Because we are moving from static measures of the economy (as presented in tables) to modeling and distributional representations of the economy as a whole, we need two measures of the size of a company: one based on current activity and the other stable over time.

Total Assets is the most stable longitudinal measure available on the Partnership return. As the Industry chart (Figure D, above) shows, however, a significant number of companies are not required to report Total Assets nor any Balance Sheet information (such as cash, land, buildings, mortgages and equity). We plan to eliminate this problem by predicting a value for Total Assets. (Since this would be an intermediate step in determining a Partnership return's stratum, a reasonably close estimate will do.) Fortunately, it is relatively straightforward to identify records where Total Assets should be predicted, distinguishing between the not reported and the "true zero" asset reports. For example, a Final Return (from a company that has ceased doing business) has zero Assets, and these can be readily identified. Further, a balance sheet exemption code can be used to identify which companies do not need to report. Thus, we can focus our attention on records for which Assets are nonzero but not reported.

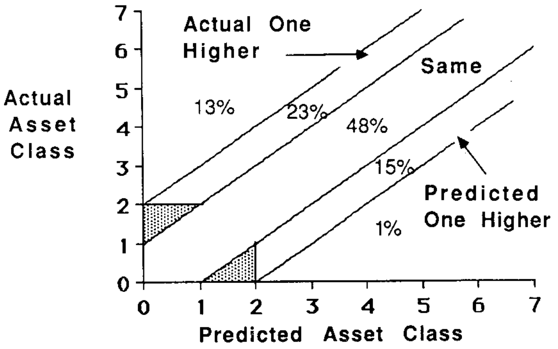
To predict the assets strata, we turned to regression formulae. We had hoped that four equations would be sufficient, one each for Trade, Finance, Services and a catchall, to minimize the increase in complexity for the computer selection programs. The initial intercept models we inspected had a major drawback in that the values for the Y-intercepts were so large that they forced many records into strata with high selection probabilities (or even into the certainty strata) on that value alone. As a result, we constrained the regressions through the origin, so that small and inactive companies would remain in sparsely sampled strata.

The variables used to predict Total Assets included Depreciation, Inventory (from the Cost of Goods Sold Schedule), Receipts, Portfolio Income items and Number of Partners. (Other items considered but rejected were Salary and Wages, Payments to Partners and Total Income.) The initial set of regression equations were used along with the actual asset values from reporting companies to determine strata boundaries. These data showed that in many cases the variable "Number of Partners" would act in place of the rejected Y-intercept to propel returns into the largest strata. This led to a requirement for at least seven equations: one each for Wholesale and Retail Trade, two for Finance, and two more for Services, along with the All Others formula.

Almost half of the companies that qualified as exempt from the assets reporting requirement still provided the data. This allows us to evaluate the affect of the projection scheme by comparing the sampling class generated by the regression equations to that arising from the actual, reported amount of Total Assets. As Figure E demonstrates, about half of the returns are placed in the same stratum under both methods, with most of the balance in adjacent strata. The two shaded triangles represent the majority (60 percent) of the mis-predicted cases. They are in the smallest two classes for both reported and predicted values. The affect of the non-agreement is, therefore, small compared to the population as a whole.

Reducing the number of asset classes to six increases the width of the classes sufficiently to increase the "same class" agreement to about 58 percent. A decrease in the number of asset

Figure E.--Error Pattern For Seven Asset Classes



classes may increase the variance of the estimates, but the computations to assess the impact are not complete. Also, this analysis does not take into account the second stratifier, the current activity measure.

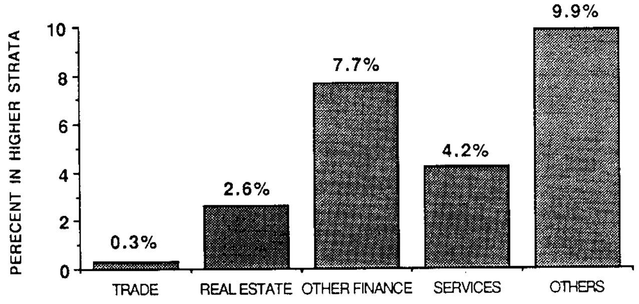
The sample design now in use employs two current activity measures: Receipts and Income/Loss, although the latter is used only to identify the largest cases. While this approach has been effective, we may be able to improve on it.

We should note that both of these are composite stratifiers, the sum of several items each. This arose as a result of tax reform changes dictated by the 1986 Tax Reform Act, which separated income into "active" and "passive" varieties. Since our users are interested in economic analysis, this distinction was not very helpful, so we combined the items. For the most part, this resulted in a reconstruction of the original Receipts and Income/Loss as conceived a decade ago.

For the future, we propose using a single set of strata boundaries for the current activity measure and using the larger of Receipts or Income/Loss, plus various portfolio items, such as Royalties. This approach will cause about 5 percent of the returns now stratified on Receipts, alone, to be placed in a higher stratum (which would have a higher probability of selection). As Figure F illustrates, though, this effect is not the same in all industries.

For the most part, the effect of using the new current activity measure does not appear to be a very large change, and most of the records fall into adjacent strata. Indeed, under 1.5 percent climbed more than one stratum, so the impact will be a marginal improvement. Why, then, should we even do it? At the present, it is unlikely that the redesigned selection process can be operational before 1993. The cost of implementing these preliminary changes is modest, however, and they do permit us to more easily construct a bridge design for use in 1992. We have decided, therefore, to make the changes to improve our design in the short term.

Figure F. -- Proportion of Returns in Higher Strata Using Current Activity Measure



The bridge design (see Figure G on the next page) for processing during 1992 will retain the 45 strata of the design now in use, so that only one of the dozens of computer programs is affected. We will use seven asset classes and six income categories, and retain the Real Estate Operators separate strata. The bridge design will also incorporate the seven asset size prediction formulae, although we expect further improvements are forthcoming.

FURTHER RESEARCH

We have developed a number of alternative stratification plans, and to choose among them for the final design we need to develop estimates of the variances for assorted major variables under each plan. Only a few of these calculations have been completed. We are also concerned about the longitudinal stability of the regression equations. Here the problem arises that the change in the tax law could give us different answers using data from before the Tax Reform Act of 1986. Thus, to test our regressions we need to wait for the 1989 data, and that sample is still being selected.

We also need to complete the bridge design described above. It has become obvious that we will not be able to complete the design in time for the 1992 processing year (for that deadline is only a couple of months away), but we can make use of what we have learned much sooner. Indeed, we have already used the information to set in place a sample reduction plan that will maintain the accuracy of the national grand total estimates while avoiding a cost overrun.

The imputation of Total Assets for those not reporting that amount brings up the question of imputing other estimates from the balance sheet. In fact, for a number of variables -- such as land, buildings, accumulated depreciation, and owners equity -- different and perhaps multiple methods may need to be employed.

A longer term goal is to develop an explicit longitudinal estimator. The method used to select the sample makes it likely that the same companies' returns are selected year after year (see Harte, 1986). However, we do not make use of this information in estimating the year-to-year growth. Perhaps, by the time we are ready to do this, we will have output from the new design in operation, and an examination of its affects will be in order. Once that is done it will, no doubt, be time to begin work on the next design for the Partnership study.

Figure G.--A Tentative Outline for the Bridge Year Design (with Population Estimates)

Total Assets $75 Million or more	2,000
Total Assets less than $75 Million and CAM $10 Million or more	5,000
Total Assets $15 Million under $75 Million and CAM less than $1.25 Million	1,600
Total Assets $15 Million under $75 Million and CAM $1.25 Million under $10 Million	4,500
Total Assets under $15 Million, CAM $1.25 Million under $10 Million	8,000

Absolute Value of Total Assets ($)	Current Activity Measure (CAM)				Absolute Value of Total Assets ($)	Current Activity Measure (CAM)			
	Under 40,000	40,000 under 150.000	150,000 under 350.000	350,000 under 1.250.000		Under 40,000	40,000 under 150.000	150,000 under 350.000	350,000 under 1.250.000
	Real Estate Operators					Non Real Estate Operators			
Under 100,000	182,000	18,000	9,300	2,100	Under 35,000	377,000	127,000	42,000	21,000
100,000 under 350,000	67,000	78,000	9,200	1,000	35,000 under 150,000	78,000	78,000	41,000	30,000
350,000 under 1,000,000	21,000	43,000	28,000	8,400	150,000 under 600,000	46,000	39,000	27,000	30,000
1,000,000 under 4,500,000	5,300	14,300	30,000	37,000	600,000 under 3,500,000	18,000	18,000	13,000	18,000
4,500,000 under 15,000,000	670	900	1,700	10,900	3,500,000 under 15,000,000	3,000	1,600	2,300	8,000

ACKNOWLEDGMENTS

The authors would like to thank Wendy Alvey and Beth Kilss for their assistance in preparing the paper and its presentation. Clem Brittain and Bettye Jamerson were very helpful in converting drafts of charts into their final polish. We are also grateful to Stephen Libster for coming to our aid every time the computer system left us confused.

REFERENCES

Harte, J. M. (1986), Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn., 603-608.

Hinkins, S., Jones, H. and Scheuren, F. (1988), Design Modification for the SOI Corporate Sample: Balancing for Multiple Objectives, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn., 216-221.

Hinkins, S., Mulrow, J. and Collins, R. (1990), Design and Use of an Imbedded Panel in the SOI Corporate Sample, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn.

Hostetter, S., Czajka, J. L., Schirm, A. L. and O'Conor, K. (1990), Choosing the Appropriate Income Classifier for Economic Tax Modeling, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn.

Jones, H. W., and McMahon, P. (1984), Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn., 437-442.

Middough, J. (1990), Partnership Returns, 1987, Statistics of Income Bulletin, Internal Revenue Service, Vol. 9, No. 3, pp 5-30.

Mulrow, J. and Woodburn, R. L. (1990) An Investigation of Stratification Errors, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn.

Nelson, S. and Petska, T. (1989), Partnerships, Passive Losses and Tax Reform, Proc. of the Sect. on Bus. and Econ. Stat., Amer. Stat. Assn.

Petska, T. and Nelson, S. (1990), Partnerships and Tax Shelters: An Analysis of the Impact of the 1986 Tax Reform, Proc. of the Sect. on Survey Res. Meth., Amer. Stat. Assn.