Proxies in Administrative Records Surveys

Paul B. McMahon, Internal Revenue Service, P.O. Box 2608, Wash., D.C. 20013

Key Words: Sampling Frame, Administrative Records, Proxy

Introduction

The administrative records we are concerned with are the financial records filed with the Internal Revenue Service (IRS). These documents are either electronically filed or must have data abstracted from them for processing through the administrative systems of the Service. The question that naturally arises, then, is why a sample survey is needed at all. Moreover, since the filing is mandatory and enforced (with real penalties for noncompliance), the need for the use of proxies might not be obvious.

We will address these issues first, starting with a brief description of the IRS's processing and the needs of our sponsors, then examining the impact of the proxies on the three largest and longest running annual surveys in the Statistics of Income series. These studies are Corporation Income Tax Returns, Individual Income Tax Returns, and Partnership Returns of Income.

Background

When tax documents are received, the IRS extracts selected information from them, both for posting to the accounts on the various Master Files, and for verifying the amounts within and across records dealing with the same transactions. Interest income, for example, is a component of net income (on which the tax is based), and so will be used in checking that calculation, and is reported by both the receiver and payer.

Abstracting all the information on all the various forms is a prohibitively expensive proposition. Thus, the Service abstracts only those amounts that directly show revenue, indicate a likelihood that an examination will yield significant changes in revenue, or are separately funded.

The extent of the data abstraction depends on the type of record [McMahon, 1999]. Individual Income Tax Returns have large amounts of information placed in electronic media, while major corporations on the other hand have only a relative handful of items extracted. In 2001 about 3,300 of the top 10,000 underwent an examination [*Internal Revenue Service Data Book, 2001*], though. It appears, then, that the reason for selecting those firms was based on some external criteria, and thus, beyond the direct revenue items, additional information was simply not useful.

Our sponsors, Treasury's Office of Tax Analysis and Congress's Joint Committee on Taxation, need data that allow them to evaluate the operation of the current tax law and estimate the effects of proposed revisions. For these uses, the electronic data used in the administrative operations are simply not enough. Thus, an extensive data abstraction and editing process is needed to collect such detailed information.

The incomplete nature of the computer records, while not directly supporting the needs of Treasury and Congress, do form a rather effective sampling frame that permits us to quickly locate the original records filed by the taxpayers. There is also sufficient information on those computer records to permit the statisticians in the Statistics of Income Division to devise complicated stratification plans that isolate important subpopulations and minimize the variances of the estimates.

Our sponsors also need the data as early as possible, so that they can respond to inquiries from those proposing changes to the law on the effects of their modifications. This pressure for the most recent data is intense enough that the Statistics of Income Division has a policy of providing preliminary data based on the data in hand at some date. These preliminary estimates are biased, since they are based on a cutoff sample, particularly underestimating losses [McMahon, 1994]. This underestimation is not unexpected since it is due to the cutoff dates being close to the date that the initial filing extensions expire. However, even the complete sample doesn't fully cover the population for the target period of interest.

The reason is that extensions for longer periods are granted in certain cases. The additional delay is warranted because, sometimes, there are unresolved issues. In other cases, there might be lost records or other extenuating circumstances. The IRS recognizes that taxpayer records destroyed in a flood or other natural disaster will take an extended time to reconstruct. A recent example of such a blanket extension was the 6 month waiver granted to the areas affected during the events of September 11, 2001.

Hence, to adjust for delayed records that will be filed after our cutoff, we use proxies in these studies.

The proxies have two distinct types: records with values derived from prior studies that are updated using publicly available information, and records for recent prior years that are filed during the selection period. The first type of proxy is most often present in the Corporations Studies, particularly where very large firms are concerned. We will not be addressing the effects of this group of proxies because the small number would quickly lead to disclosure problems.

The second set of proxies assumes that records arriving late from a previous year are much like the records for the current target period that will arrive after our cutoff date. The proxies are included in the population and subjected to sampling as if they were in

the target fiscal periods. This standard practice for the Statistics of Income Corporation, Individual, and Partnership Studies has been used for at least the past 3 decades.

For the purpose of this paper, we will use the "Study Year" definition from the Corporations Studies. Corporations may choose any month to end their fiscal periods (with certain State-Law-based exceptions). Thus, to provide a consistent comparison across years, the definition has a target year running from July of the first year through June of the next. For example, a Tax Year 1999 study is focused on firms with fiscal years ending in July 1999 through June 2000.

We will use the results from three project areas, Corporations, Individuals, and Partnerships, to see what the effect is overall. We will "correct" certain key estimates by removing the proxies, then replacing their values with those from target period records included in later programs. We begin with an overview of each of these studies and the particular deadlines that affect them.

Corporations

Corporations are entities created by the States. Usually, the firms choose this sort of organizational framework to limit the liability its owners might face. Businesses in certain lines of endeavor, such as insurance or banking, though, are required by the States to use the corporate form of organization, while others such as accountancies have been restricted from using corporate status. The States also place certain requirements on the fiscal periods of some industries, such as requiring insurers to use December to end their fiscal year. For most companies, though, it is left to their own best judgment.

Despite this relative freedom on accounting periods, almost 80 percent elect to use the calendar year. The filing instructions require that the report on the completed year (or other tax period) be sent to the IRS within 2½ months. This places the bulk of the filing in the subsequent 12 months. However, about 6.5 percent have ending dates in the first 3 months of the target period, and another 4 percent in the following June.

This dispersion across the year has meant that the sampling period for the Statistics of Income Corporation Studies needed to begin shortly after the close of the first of these fiscal periods, and last through the filing period, with extensions, for the firms with the latest fiscal period, for a total of nearly 21 months. Actually, though, the period is a few months longer to allow for processing through the systems.

The number of proxies will vary over the years, not necessarily following the pattern of the overall growth of the population. As Table 1 shows, though, the number in the Tax Year 1998 Study was reasonably close to the number of target fiscal period records that appeared in the following year. The number of proxies is also quite small, compared to the population, just slightly more than half a percent. There were about the same number of proxies in the sample, 926 records, as were in the delayed class, 954 (which are actually the proxies included in the Tax Year 1999 Study).

Table 1: Tax Year 1998 Corporation Proxies

	Estimated Number of Records	Average Net Income (or Deficit)	Average Total Assets
Proxies	31,148	29,900	6,723,000
Delayed Records	29,030	-58,300	2,625,000
Target, Timely	4,818,738	172,400	7,707,000

While the numbers of delayed and proxy records are close, though, key observations on them show that they do have large differences. Yet while the difference for the average amount of net income between the proxies and the records they replace (Delayed Records) is large, the difference between the delayed records and the regular filers (Target, Timely) is greater still. This is really not surprising, given the legal environment, because firms that are unlikely to owe additional tax are more readily granted further extensions. Since firms with losses rarely have income tax liabilities outstanding, they are predominant in the delayed filing population.

On the other hand, the proxies are more like the regular Tax Year 1998 filers than the delayed records are for Total Assets. This may be associated with administrative operations arising from the IRS reorganization.

As we have already noted, though, the proxies form only a small proportion of the sample and estimated population. Are the effects of the proxies of any note in the estimates produced?

Table 2: Effect of Tax Year 1998 Corporation Proxies

	Estimated Number of Records	Net Income or Deficit (Millions)	Total Assets (Billions)
Including Proxies	4,850,000	831,700	37,300
Including Delays	4,848,000	829,100	37,200
No Proxies or Delayed Records	4,819,000	830,800	37,100

In Table 2, we see that the various estimates have very nearly the same values. Less than 0.05 percent separates the estimates of the total population that either include the proxies or the delayed records, and with a sample size nearing 130,000, this is not an important difference. The differences for Net Income and Total Assets are each under a third of one percent,

yet here we may have significance. Computing the variance of the estimate for the adjusted population is not straightforward, and outside of the resource limit for this review. However, the sample includes all records with more than $10,000,000 of Total Assets, or more than $2,500,000 in absolute value of Net Income (Deficit). It is reasonable, then, to conclude that even though the difference is small, it is significant.

Individuals

Natural persons, as the laws tend to phrase it, may only have noncalendar tax periods with the consent of the Internal Revenue Service. Not surprisingly, such an occasion is quite rare. But this does not mean that there are no prior-year records in the Statistics of Income Individual Income Tax Studies, for filing extensions are automatically granted for 6 months, with further delays allowed if the cause is reasonable. These additional delays are not often required, as evidenced by the approximately 97.8 percent of the records processed in a calendar year that are reports for the subject tax year.

That is, our proxies account for only about 2.2 percent of the estimated population and 2.5 percent of the sample. There are about 4,400 proxies and 172,000 core filers in the sample of Individual Income Tax Returns. (The proxies from earlier years were omitted from the tables below, about 1,000 records in the sample and an estimated population of about 950,000.)

The Individuals Studies have an imbedded panel. Records included for this reason are, therefore, retained no matter what tax year the filing covers. Very large records, since they tend to be rich in rare types of data, are of high interest to our sponsors so they are also retained without regard to the tax year. Ordinary records, though, are only included if they are from the most recent 3 years.

Since we are using a recent study year, 1999, as the basis for this review, records that are delayed in filing for more than an additional year are not yet available. Therefore, we will examine only the contribution of the nearly 3,400 proxies from the most recent year, in this case, Tax Year 1998 Returns in the 1999 Study. Since those records are about two-thirds of the estimated population (and more than 75 percent of the proxies in the sample), we capture most of the effect.

We exclude the other proxies from this analysis. This means that the data we cite here are not the same as those presented in the publicly available tabulations. The Tax Year 1999 filers who were included in that year's study are the Core Filers in the tables below. The Delayed Filers are the Tax Year 1999 records that were included in the 2000 Study. This allows us to directly compare at least a part of the effect of the proxies directly, in the context of a corrected estimate for 1999.

The estimates are based on stratified samples of tax returns subjected to sampling at various rates. Records containing rare or large amounts were classified into strata where the probability of selection is 100 percent, while relatively simple records reporting small sums of money went to strata with probabilities as low as 0.05 percent.

The overall sampling fraction is about 0.14 percent, while the effective rate for the proxies and delayed records is about 0.18 percent, which reflects the greater complexity of the later filers' records. This is also reflected by the coverage of Adjusted Gross Income, where the core filers in the sample reported almost 6.5 percent of the estimated total, compared to the higher proportions that were reported by the proxies, 7.3 percent, and the delayed filers, at almost 7.5 percent.

But these data present an incomplete view of the impact of proxies, as the sample was selected with a large variety of probabilities. Thus, we now turn to estimated population characteristics.

Table 3: Tax Year 1999 Individual Proxies' Estimated Averages for Key Variables

	Estimated Number of Records	Average Adj. Gross Income	Average Tax Liability
Proxies	1,877,000	35,100	5,500
Delayed Filers	2,007,000	38,500	5,800
Core Filers	124,008,000	46,500	6,900

There are nearly 7 percent more delayed filers than proxies in the population, as shown in Table 3, and those delayed filers had, on average, a 10-percent higher Adjusted Gross Income (AGI) and an associated 6-percent greater tax liability than the proxies. Still, those average AGI and tax figures are closer than the delayed records are to the average of the core filers, which are about another 20 percent higher yet.

Table 4: Tax Year 1999 Individual Proxies, Effect on Overall Estimates

	Estimated Number of Records	Adjusted Gross Inc. (Billions)	Tax Liability (Billions)
Core & Proxies	125,882,000	5,833	870.0
Core & Delayed	126,012,000	5,844	871.4
Core Filers	124,008,000	5,767	859.7

With Corporation records, the use of proxies clearly improved the estimates, and here in Table 4, we see that there is significant improvement as well. However, where the Corporation amounts tended to be

overstated by the inclusion of the proxies, the Individual study is still marginally understated, both in the estimated population and on key variables.

The understatement is not constant across subpopulations, as we see in Table 5. The difference for the records that are reporting income from a business or profession (Schedule C), or from farms (Schedule F), is very small indeed. It appears that nearly all of the difference arises from non-business, non-farm sources.

Table 5: Tax Year 1999 Individual Proxies and Attached Schedules

	Estimated Number of Records	Adjusted Gross Inc. (Billions)	Tax Liability (Billions)
Non-Business & Non-Farm:			
Core & Proxies	107,000,000	4,654	677.9
Core & Delayed	107,126,000	4,663	679.3
Schedule C Attached:			
Core & Proxies	16,824,000	1,072	173.8
Core & Delayed	16,809,000	1,073	173.8
Schedule F Attached:			
Core & Proxies	2,058,000	106	18.3
Core & Delayed	2,076,000	107	18.5

Partnerships

The organizations of interest for this series of studies are active businesses that have more than one owner. These firms are not incorporated under the applicable State laws, either by their own election or because the State prohibits it for their line of business. Beyond that, however, there is a wide variation in the nature of these companies, with some having publicly traded interests, some with limited liabilities, and others where all the liabilities are common to the owners.

The Statistics of Income Partnership Studies select a stratified sample of about 35,000 records annually from a population that is currently growing at a rate of about 5 percent annually, reaching nearly 2,200,000 reports for the 2000 Study. There are over 70 strata, based on the amount of total assets, net income, net receipts and industry. We employ a permanent random number selection mechanism [Harte, 1986], as all the Statistics of Income studies do, along with a small panel of firms with very rare circumstances. This last condition can mean that multiple records from the same firm for different accounting periods can end up in a single study, as the records from earlier years are considered proxies. However, multiple records from very large businesses are removed, with only the report for the most recent year retained.

For the 1998 Partnership Study, about 96.3 percent of the reports in the sample were for the target accounting periods, and most of the rest, over 3 percent, were from the prior year's fiscal periods. When adjusted for the variety of sampling probabilities, which range from under 0.1 percent to certainty, the population estimates show that nearly 97.3 percent have target accounting periods.

Table 6: Proxies for Various Tax Years' Partnership Studies

Study Year	Sample Proxies	Estimated Population	Total Assets (Billions)
1998	1,423	50,100	150.6
1999	1,468	35,800	185.4
2000	1,283	36,200	247.7

Although the number of proxies in the sample seems to have dropped off significantly in the 2000 Study, this figure is actually in line with the 2 previous years, given that the overall sample size was reduced from 42,000 in 1999 to 35,000.

We had intended to use the 1998 Study as the basis for this review, but the proxies in that study used the old Standard Industrial Classification based codes. This would compromise any comparisons we attempt using the current North American Industry Classification System (NAICS), because there would be another source of error, namely in the assignment of the NAICS Codes by the data abstraction clerks. In fact, a review of these data suggests that there were problems with that conversion. Thus, we decided to use the 1999 Partnership Study as the basis, and restrict the review to those proxies with accounting periods from July 1998 through June 1999 filed during Calendar Year 2000. The delayed filers, then, are the records with periods of July 1999 through June 2000 that were filed in 2001.

Table 7: Tax Year 1999 Partnership Proxies' Estimated Averages for Selected Variables

	Estimated Number of Companies	Average Total Assets	Average Business Receipts
Proxies	28,800	6,210,000	3,900,000
Delayed Filers	28,100	8,830,000	5,130,000
Core Filers	1,902,000	3,060,000	1,070,000

There is very close agreement between the number of firms estimated from the proxy records and those delayed into the following study, amounting to only about 2.5 percent. However, total assets and receipts are understated by a quarter. Still, as in the

Individual Studies, the proxy averages are closer than the averages of the core filers. The populations here are small, however, so we expect to see little effect on the overall estimates.

Table 8: Partnership Proxies' Effect
On Selected Estimates

	Proxies And Core Filers	Delayed And Core Filers	1999 Coefficient Of Variation
Partnerships	1,931,000	1,931,000	0.31%
Partners	15,333,000	15,286,000	5.01%
Total Assets	5,995,350	6,056,138	0.23%
Receipts	2,141,655	2,173,411	0.20%
Net Income	348,129	344,275	0.50%
Net Deficit	119,564	119,004	1.61%

(Amounts in millions)

Because we have merged samples from various studies in the estimates for Table 8, computation of the standard errors is problematical. There are, for example, a number of strata that contain single observations with the target fiscal periods. Had we actually extended the sampling period, though, most of those problems would vanish. Under that situation, the variances would not have been too different from those calculated for the full 1999 Study. Hence, we may use those figures as a reasonable guide.

As one would expect from Table 7, the number of firms is the same, after rounding to thousands. What might come as a surprise is the larger coefficient of variation for the number of partners. Historically, this estimate has had much larger relative errors than monetary variables have because it is not closely related to any of the stratification items. From that perspective, then, the difference between the proxy-influenced estimate and the corrected (with the delayed filers) figure is not important.

The differences for total assets and receipts, however, are another story. Here, the relative difference between the estimates for assets is 1.0 percent, and, for receipts, it is 1.5 percent, or more than four times the size of the related coefficients of variation. The relative difference for net income is also above one percent, but that is only slightly more than twice the relative error.

It might be that the data above are the result of economic effects, and, in particular, the large increases observed in the valuation of securities in 1998 and 1999. If so, an examination of the industry distribution might confirm this hypothesis. Table 9 presents this information at the industry division level, based on the NAICS codes reported. (Note, please, that these estimates in Table 9 do not sum to the totals in Table 8, in part due to rounding, but also because we excluded records for which an industry could not be determined.)

A clear majority of the firms electing partnership status are in the Finance Division. The difference between the number of delayed filings and those used as proxies is not of any importance. However, about two-thirds of the difference for total assets appears in that industry division. The effect is even more pronounced for net income, but muted for receipts. This fits the assumption that asset growth during 1999 could explain the underestimation due to the use of proxies.

Yet the estimates for net deficit show very little effect. This is in line with previous research on preliminary estimation [McMahon, 1994], showing that firms with large losses tend to predominate the population of late filers.

Table 9: Adjusted Partnership Estimates by NAICS Industry Division for Selected Items
And the Effect of Proxy Use

	Partnerships		Total Assets		Receipts		Net Income		Net Deficit	
Industry Division	Core & Delayed Filers	Delayed Minus Proxies	Core & Delayed Filers	Delayed Minus Proxies	Core & Delayed Filers	Delayed Minus Proxies	Core & Delayed Filers	Delayed Minus Proxies	Core & Delayed Filers	Delayed Minus Proxies
Raw Materials	145,800	300	222,775	700	123,682	1,611	18,780	-102	9,663	169
Goods Prod.	163,600	0	380,692	771	451,269	-983	33,800	-971	12,253	58
Distribution	164,200	2,300	186,625	-1,589	429,264	9,202	16,657	284	7,807	-56
Information	20,600	600	264,464	7,265	133,632	2,785	20,755	647	26,786	-249
Finance et al.	1,074,200	-2,400	4,514,565	41,344	583,282	13,646	172,044	-4,251	42,808	-235
Prof. Services	166,100	-800	276,861	13,392	244,342	3,754	58,497	270	9,250	46
Education, etc.	46,300	400	47,866	75	73,631	1,007	11,686	257	2,817	-30
Leisure, etc.	96,200	-600	52,673	-1,258	120,803	737	10,585	66	7,002	-279
Other Services	51,500	-300	9,154	209	12,970	153	1,421	2	549	16

(Note: the numbers of partnerships are rounded to hundreds, and the monetary values are in millions of dollars.)

Conclusion

The use of proxies in these administrative records studies often results in underestimation of many parameters, but the lack of reliable information on the number or distribution of those records that will be delayed in filing beyond the studies' completion deadlines leaves little alternative, at least for the foreseeable future. Those firms and individuals who file late tend to have particular characteristics, especially in showing losses from economic activities. Since the taxes that arise from those situations are not significant, the administrative operations do not tend to require stringent filing deadlines.

There are other administrative effects that impact on the presence and characteristics of the late filing populations, some due to extraneous events, such as floods, and others to changes in the Internal Revenue Service's structure.

This review was not able to take a longer timeline into account due to the recent industry classification changes and to data availability issues. We hope that this issue will be addressed in a future paper.

Acknowledgements

This review would not have been possible without the assistance of Richard Collins and Valerie Puckett, who provided the data from Corporations and Individual Income Tax Returns Studies.

Notes and References

Internal Revenue Service, *Data Book 2001*, Publication 55B, Washington DC, 2001. Table 10, on page 15, shows that 10,300 Corporation Income Tax Returns were filed, reporting $250,000,000 or more in total assets during Calendar Year 2000. Of these returns, 3,305 were subjected to an examination.

McMahon, Paul (1999), "Administrative Records, Regulations, and Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, Paul (1994), "Statistics of Income Partnership Studies: Evaluation of Preliminary Estimates," *Proceedings of the Section on Survey Research Methods*, 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," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.