Changing Industry Code Systems: The Impact on the Statistics of Income Partnership Studies

Paul B. McMahon, Internal Revenue Service

The Statistics of Income series of studies are primarily designed for use by the United States Congress and Treasury Department in evaluating tax law provisions, both current and proposed. The partnership area studies have been conducted annually for almost 50 years now, with increased prominence during periods when tax shelters were popular. Over that time, a number of different sample designs have been employed, with the current strata outline dating back about a decade.

This design is now under review for four reasons. First, it is prudent to re-examine a sampling plan periodically, and this study's is past due. Second, the population is no longer contracting, as it was when the current design was implemented, and the length of the distribution's tail has also grown. Third, regulatory changes have forced us to make ad hoc changes to the existing design, and the impact is not well studied. One regulatory change that seriously affects the existing design is the switch from industry codes based on the Standard Industrial Classification Manual to codes based on the North American Industry Classification System. This shift also changes the analysis, which is the fourth reason.

But first, some background on the administrative environment and the nature of the population. We will follow the path from the firm's classification through the processing environment and onto the sampling frame, with comments about the impact of the industry coding change. We will then examine the effect on the current design and the revision being planned.

◆ Background

The Statistics of Income studies use the tax forms filed with the Internal Revenue Service as questionnaires on economic issues. The use of administrative records for such purposes has many limitations, but the mandatory nature of the filing does constrain the nonresponse aspect. For partnerships, the filing of a return does not, under normal circumstances, arise from taxes due, but from the requirement that firms report the earnings of their owners, much the same as corporations report dividend income.

Partnerships are unincorporated businesses with more than one owner. Firms with this organizational structure have a form of business that falls between Individuals (acting as sole proprietorships) and highly structured Corporations. The definition for these firms is in the hands of the States, for they decide what a corporation is; they also have a hand in determining what a partnership is. For example, there are a small number of businesses that have the interests in them traded on the open market like stocks, yet State laws call them Publicly Traded Partnerships. On the other hand, some joint operating agreements appear on the surface to be partnerships, but under the laws are not. In addition, States forbid certain types of professional operations from becoming corporations, like law firms, forcing the confederations to become partnerships.

These examples might give the impression that the owners of the companies must all be individuals. Such is not the case. Corporations, tax-exempt organizations, individuals, and other partnerships may all be owners in any combination. The only constraint is that there must be two owners. Partnerships are not subject to an income tax directly; instead the income, tax liabilities, and credits are passed through to the owners.

These organizations are required to report their earnings and the distribution of that income and so on amongst the partners annually. The report includes a small space for a word, or perhaps two, describing the sort of business and product, and another box for the "Business code number." This business code was, before 1998, based on the Standard Industrial Classification (SIC) codes. The IRS's business codes were a consolidated subset of the SIC list. There were different tables of codes for different types of organizations, as well. Since Corporations, as an example, tended to have many manufacturing firms, there were far more codes

for that division than were listed in the table for Partnerships, where that sort of business was scarce.

With the introduction of the North American Industry Classification System (NAICS), the Corporation and Partnership filers used the same list of industry codes. In part, this was because the Service was able to persuade the panel constructing the NAICS Codes that there were some sorts of businesses that had to have distinct classifications. Two such were for Regulated Investment Corporations (Open End Investment Funds, 525910) and Real Estate Investment Trusts (525930).

However, that list of codes was far too extensive for inclusion in the instruction booklet the filers are asked to follow, so the Service combined individual categories to fit their needs. The construction of the Service's codes followed the consequences of the tax and legal environment, rather than the economic concerns that influenced the development of the NAICS Codes. These legal concerns also led to a redefining of certain codes in the Insurance area as well.

◆ Administrative Environment

That list of codes is used by taxpayers in filling out their returns. Those records, whether on paper or, increasingly, electronic media, are then transmitted to the IRS. When the Service receives these records, it then converts the data to its own form of electronic record for tax administration purposes.

One might ask why not simply tabulate that data file. Certainly, with an entire population of less than two million firms, it would not be too great a burden on today's computers to produce tallies for any question at will. And, since the primary sponsors are permitted by law to view the individual records, a simple copy of the file would seem to suffice.

But the initial "Transaction Record" has faults, both in content and quality. First, beyond the Employer Identification Number, Tax Period, and industry, there are only 51 monetary amounts and 15 indicator fields available. We are ignoring the processing codes and other internal fields here, of course, as they have no bearing on our studies. Even so, not all of the germane fields are considered useable (because some are rarely applicable and, thus, overlooked). So, we are left with 36 monetary and 5 indicator variables.

The quality issue also applies to the entity information at this stage of the process. Discrepancies between the Employer Identification Number and other information have not yet been identified, let alone researched and corrected. Among these bits of information untested at this point is the industry code.

The familiar problems endemic to self-reporting of codes are present, but there are other sources for apparent errors to arise as well. Table 1 shows the results for returns filed during 1999 and as of this writing for 2000. There is considerable error to these counts because some records get counted more than once (as they cycle through the error correction process). The total population processed through the sampling operation in 1999 was 1,972,000, meaning that the data are overstated by 1.6 percent.

Table 1: Partnership Returns Industry Coding

	1999		Through May 2000	
	Count	Percent	Count	Percent
Valid NAICS	1,619,000	80.8%	624,000	86.7%
Invalid NAICS	152,000	7.6	44,000	6.1
Unknown NAICS	98,000	4.9	24,000	3.3
Valid SIC	122,000	6.1	26,000	3.5
Invalid SIC	12,000	0.6	3,000	0.4
Unknown SIC	1,000	0.1	--	0.0
Total Returns	2,004,000	--	719,000	--

The categories in the table above are somewhat misleading. A "Valid NAICS" Code for this purpose was one that the IRS defined in its publications. "Unknown NAICS" would be records that did not have any industry code reported, but had a Tax Year of 1998 or later. Similarly, "Unknown SIC" means that no code was reported for records of Tax Years before 1998. The two "Invalid" lines contain a mix of those records that have SIC or NAICS codes that are not on the IRS's list, along with reporting and keying errors.

The trend clearly shows that the NAICS-based coding is improving. Last year, only slightly more than 80 percent were valid (improving by only 1/2 percent between May and yearend), while, this year, the figure for May (2000) is closer to 87 percent. About half this increase arises from a diminishment in the number of SIC-based codes that appeared, but the rest arises from improved processing and reporting.

The reporting is a particular factor, because, while the filing deadline for nearly all partnerships is early in the year, it is quite straightforward to get an extension of 6 months, and not too difficult to get an additional 3 months beyond that. There are also some respondents who do not provide the industry information, and others who simply copy whatever they reported in that part of the form the previous year. Hence, even though the basic quality of the NAICS-based codes is encouraging, the processing had to accommodate both the NAICS- and SIC-based codes. The strategy IRS used was to insert two leading zeroes before the old SIC-based codes.

The administrative processing has one more step that needs to be addressed: Posting. The transaction records that gave rise to the statistics in Table 1 are matched to the IRS's Business Master File records. This posting process affirms the entity information in the transaction, updates the Master File, and allows certain permanent information from the Master File to be appended to the transaction record. The piece that interests us is that the last reported SIC-based code, posted prior to January 1999, is among the additional data.

This file from the posting process, with the enhanced transaction records, forms our sample frame.

◆ Current Design

This sampling frame, then, has both SIC-based and NAICS-based industry codes, along with various monetary variables. But just as the industry coding is not quite what it seems, so it is with the other fields. Total Assets, for example, need not be reported for firms below a certain size if they meet some other, fairly relaxed, conditions. A similar situation exists for business receipts and net income. Here, in response to tax law definitions, the various sources of revenue are labeled "active" or "passive," then held to different procedures.

We will not go into the specifics of the current monetary classes, nor the statistical properties of this design, as that information was published elsewhere. The outline of the strata and the improvements in the coefficients of variation were described in McMahon (1995). The impact of the data abstraction process's quality was explored in McMahon (1996), and the use of permanent random numbers in sample selection in McMahon (1998).

In general outline, though, the immediate predecessor had four main sections. The first section contained a pair of strata for firms with assets of at least $100 million, or income or receipts of at least $25 million. The rest of the population was divided into the three remaining sections based on Industry. The set of strata reserved for Real Estate Operators (except developers) and Lessors of Buildings (SIC Code 6511) has been a staple of the Partnership design since the mid-Seventies. This predecessor design split the remaining population into a third section for Mining, Construction, Manufacturing, and Transportation companies (SIC Codes 1000 through 4999), and the fourth for Agriculture, Trade, Finance, Services, and companies without industry data.

The strata within these design sections depend on Total Assets, Income, and Receipts. This classification has been in use for most of the decade, with only minor adjustments. We recently added an additional pair of strata at the upper end of the Asset and Income distributions to control the growth of the certainty class, for example. We have also had to set aside special classes for records identified as Publicly Traded Partnerships or for having been filed electronically.

But the key change was the conversion to the new six-digit industry code.

The instructions for abstracting the new industry code made provision for records reporting from previous years' industry classification. When one of the SIC-based codes was encountered during processing in 1999 (Tax Year 1998, mainly), two lead zeroes were inserted to distinguish them from the NAICS-based IRS industry codes. We used this industry code for the stratification, even though there was the old SIC code for continuous businesses.

The decision on how to handle the migration from one coding scheme to the other had to be made in the fall of 1997, long before any data on the actual pattern of filing amongst the industries could be known. We also did not want to confound any analysis by instituting design changes beyond a minimum. Hence, we redefined the old categories in terms of the new-NAICS based codes, as shown in Table 2, below.

We also prepared for the advent of a large influx of records that had no industry information present. In this case, we deferred to the economists involved in the project. The parsing of the economic "receipts" into active and passive sources, while normally unhelpful, came to the fore. Where an amount of rent was present and the industry was missing or clearly wrong (in the ranges of less than 000100, 009000 through 110999, or greater than 820000), we declared those records to be in the Real Estate Operators strata.

Table 2: Industry Groups Used in the Tax Year 1998 Sample Design

	Principal Business Activity Codes	
Industry/Division	Standard Industrial Classification	North American Industry Classification System
Real Estate Operators	6511	531110 and 531120
Mining, Construction, Manufacturing, and Transportation	1000 through 4999	200000 through 350000, and 480000 through 519999
Farms, Trades, Finance, and Services	All other codes	

Table 3: Partnerships by SIC Industry Division

Industry Division (SIC)	Estimated Tax Year 1997 Population	Tax Year 1998 Population
Real Estate Operators	592,000	581,900
Finance	382,300	364,500
Services	311,000	320,100
Trade	173,000	152,400
Agriculture	127,100	118,600
Construction	72,100	67,300
Manufacturing	40,000	32,900
Transportation	30,900	25,500
Mining	28,000	23,700
Unknown		287,000

But why use the industry in the sample design at all? Since Real Estate Operators comprise about a third of the entire population (see Table 3, above), about 12,000 of the target 35,000 sample selections would have been in this one industry (about 2 percent of the estimates). By creating separate strata for them, we maintain the accuracy of that industry's estimates while halving the sample in that domain. We then use the roughly 6,000 records to reinforce estimates elsewhere. It is clear that Divisions like Transportation and Manufacturing are overshadowed and, thus, needed additional sample (beyond proportional allocation) to permit the level of analysis desired.

The data in Table 3 show another factor as well as the industry domination. The data for 1998 are derived from the historical SIC-based industry from the Master File, yet there is a segment of the population without this information. There can be only one reason: these are the records of new businesses and, thus, do not have prior-year data. Thus, we cannot rely on the historical industry code for stratification.

◆ Designing with NAICS

The goals, in so far as the industry perspective is concerned, are first to identify any NAICS industry that would, if proportionately sampled, be allocated more sample units than are needed for analysis. Second, identify which, if any, are at risk of receiving too few observations. The third goal is to minimize the effect of records with unknown industry classification on the sample design.

Table 4 presents the migration observed in the frame using the records subjected to sampling during 1999 for the Tax Year 1998 Study. Since this is the first year of the new coding system, it is unsurprising that there would be a number of firms that received erroneous classification. This is, we believe, one source of the smallest frequencies reported below.

The "Total" column shows where the largest industries are. The Finance Division, under the new system, has the lion's share of the population, with more than

Table 4: Sampling Frame Transition From SIC to NAICS

NAICS Division	SIC Division										
	Total	Unknown	Agri-culture	Mining	Con-struction	Manu-facturing	Trans-portation	Trade	Finance	Real Estate	Services
Total	1,973,800	287,000	118,600	23,700	67,300	32,900	25,500	152,400	364,500	581,900	320,100
Raw Materials	134,300	9,300	91,200	19,200	500	2,600	1,300	1,300	4,400	1,600	2,800
Goods Production	129,500	22,500	1,700	200	51,400	17,100	400	3,900	23,000	3,200	6,000
Distribution	140,200	25,300	1,100	100	600	3,100	11,100	84,800	900	600	12,500
Information	17,800	3,600	*	*	100	1,700	3,800	600	300	200	7,500
Finance, etc.	936,600	100,500	5,100	400	2,500	400	1,700	3,800	275,800	507,900	38,400
Prof. Services	133,600	22,400	4,100	100	800	1,100	1,600	2,600	3,800	2,000	95,200
Education, etc.	36,700	5,000	*	*	*	*	100	200	300	300	30,800
Leisure, etc.	77,300	11,300	400	*	100	700	200	29,000	1,100	1,000	33,500
Other Services	56,500	9,200	1,500	*	900	500	400	2,100	1,200	400	40,300
Unknown	311,200	77,900	13,400	3,600	10,300	5,700	4,800	24,200	53,700	64,600	53,200

(Note: Rounded in hundreds, with an asterisk in cells where the count was less than 50)

half arising out of the Real Estate Operators (SIC 6511) class. However, the non-Real Estate portion of Finance still contains about twice the population of the other divisions. So, there may be other candidates as well. In fact, as Table 5 shows, there are four main components to the Finance Division.

The two largest industries of Table 5 arise almost exclusively from the SIC Real Estate Operators category. This means that they share the same sorts of financial and business profile and are compatible for stratification. The "Other Activities Related to Real Estate" firms are less closely tied to that profile, with only about 40 percent of their populations having been previously identified as Real Estate Operators. Still, since they do share the same industry sector, there are reasonable gains to be had by including them in the special strata.

This does not appear to hold for "Other Financial Investment Activities." Less than 4 percent of that population were formerly identified as Real Estate and, as the Asset column shows, they clearly have a distinct distribution in that regard. Moreover, this collection of investment groups tends to have significant reported amounts of short- and long-term capital gains. These are areas with a history of large variability across the years, so it seems wise to leave them to a proportional representation, yielding about 1,600 sample observations.

On the other side of the coin, a proportionate share of the sample would lead to only about 300 records for the 4 published Information Division industries and a bit over 600 records for the Education, Health, and Social Assistance Division's 10 published categories. Such small sample sizes for the individual industries would support only the most cursory analysis. However, combining these NAICS Sectors means generating strata that are not homogeneous. Of course, the solution for this is clear: Post-stratification. In this case, the bias would be ignorable (effectively zero) because the post-stratification population data are collected during the sampling process.

But just as clearly, there is a problem in the category "Unknown." As we saw in Table 1, this population should be declining, but we will still need to make provision for this sizable group of records in the design. The real problem is that the population of records with unknown industry classification (at the time of sample selection) may contain about the same proportion of Lessors of Buildings as the population with known industries. The ad hoc plan used in the current sampling program, which simply uses the presence of rental income, tends to identify non-real estate operators too often.

We propose to substitute an "80-percent rule," where a firm will be categorized (for sampling purposes only) as a Lessor of Buildings if the proportion of its receipts that are real estate rents exceeds 80 percent of the total. As Table 6 shows, this would have correctly identified over 78 percent of the known Lessors and about 35 percent of the industry "Other Activities Related to Real Estate." Only the Agriculture Sector would have a mi-

Table 5: Largest Finance Division Industries

NAICS Industry	Number of Firms	Assets (Millions)
Other Financial Investment Activities	113,500	1,029,000
Lessors of Residential Buildings and Dwellings	285,300	459,000
Lessors of Non-Residential Buildings	237,000	596,000
Other Activities Related to Real Estate	116,700	253,000
All Other Finance Division	184,100	660,000

Table 6: Real Estate Rents as a Proportion of Total Receipts

	No Rents Reported	Under 70 Percent	Between 70 and 80 Percent	Between 80 and 90 Percent	At Least 90 Percent
Lessors of Buildings	15.6%	4.1%	2.0%	37.8%	40.5%
Other Real Estate	59.0%	4.6%	1.4%	14.6%	20.3%
Agriculture	83.8%	4.0%	1.0%	5.0%	6.3%
Other Valid NAICS	96.5%	1.3%	0.1%	0.9%	1.1%

(Percentage of firms within each industry with a given proportion)

nor negative impact, misidentifying about 7,000 farms as belonging to the Real Estate areas, and this is not a significant concern to our users. Those companies that fail the 80-percent rule would be placed in one of the strata for sectors Trade, Raw Materials, and so on.

◆ Conclusion

The conversion of the sampling strata from the SIC-based codes to the NAICS scheme was not entirely successful. The initial quality of the reported codes has improved, but the basic incompatibility of the two systems means that a completely new stratification plan is needed for the Partnerships Studies.

The replacement sample design will have five main categories. The largest firms and those with peculiar conditions will still have strata set aside for them, and the three industry groupings will be retained, although with updated particulars. What was once Real Estate Operators will now hold their successor industries, but the sparse industry group is entirely redefined. At this time, the full details of the revision are not known, for analysis of the data sets has only just begun.

◆ References

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, pp. 650-655.

McMahon, P. (1996), "Non-Sampling Errors in Data Abstraction From Administrative Records," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 184-189.

McMahon, P. (1995), "Longitudinal Estimates and Permanent Random Numbers in Administrative Records Studies," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 709-714.

■