Assessing Industry Codes on the IRS Business Master File

Paul B. McMahon, Internal Revenue Service

An early process in the development of any business survey is the construction of a sampling frame, and a list of establishments is usually the preferred frame. The most favored sources for such a frame are records systems with lots of auxiliary information, which permit stratification, probability proportional to size sampling, calibration estimation, and other options. The Internal Revenue Service's Business Master File System is one such source.

The records on that system are not available to any who would survey this population, but the laws do provide that certain agencies do have access. Limited data are available to the Census Bureau, for example. However, the Service's Master File Systems are designed with accounting and administration in mind, not survey sampling. Thus, there are a number of conventions that, if not understood, could degrade the usefulness of records from that system.

These issues were addressed in past papers, most recently in the areas of processing conventions (McMahon, 1999), delayed filing effects (McMahon, 2002), and regulatory exemptions (McMahon, 2003). Another issue is the quality of the data on that system when the information is not directly connected to matters of tax collection, but is of considerable interest for a sampling frame. One such variable is the industry code.

We examine this code using records processed during Calendar Year 2003 both because it is the latest full year available and because it shows the effects of the latest revisions to the North American Industry Classification System (NAICS). Since Corporation data for Tax Year 2002 are not available as of this writing, we confined this review to businesses organized as partnerships.

Sources of the Data

The records that the Service provides for use in sampling frames arise from the filing of tax forms. In this particular case, we are concerned with the annual records filed on Form 1065, *Partnership Return on Income*. The entities providing these forms are businesses that have two or more owners and are not incorporated, though there are a small number of exceptions.

The exceptions involve some legal forms of business permitted by some States, like "Publicly Traded Partnerships" and "Limited Liability Companies." The existence of these variations on the partnership theme arises from the power of the States under the constitution, which means that the Federal Government must deal with the consequences, in this case by having these hybrid organizations file the partnership form.

That form has four pages, although attachment pages, such as Schedule K-1, *Partner's Shares of Income, Credits, Deductions, Etc.* (one for each partner), and depreciation forms are usually present as well. The associated instructions for the basic form are 34 pages in length, including the mailing instructions and industry classification rules. Contrast this with the 42 pages devoted to the short title list in the 2002 manual for NAICS. In the full classification system, there are 1,179 separate industries, which are far too many to expect the taxpayer to search through [1] and would cost too much to mail to each requestor. As a result, the Service reduced this list to 427 six-digit industry codes that list in just three pages of the instructions.

The industry codes used by the Service differ only by combining industries into more general categories. That is, the Service did not create any special group from a subset of one of the NAICS codes. Moreover, with the exception of the sole proprietorships, the Service uses the same codes across the various types of businesses.

Businesses, however, do change their focus from time to time, and this might result in a change of industry. For example, a company might build residences, rent models, and sell completed units. Depending on the circumstances, then, it could be in one of three industries. The IRS instructions set the rule that the code to be assigned depends on the activity that provides the greatest share of a firm's total receipts.

Total receipts, however, appear nowhere on the tax form. Instead, a detailed computation is required that requires 17 amounts from three schedules, which in turn reference still other forms and schedules [2].

Taken together, the long list of codes and the complicated process of deciding the industry, as well as the taxpayer's time, make it very likely that the code used in a previous year will simply be copied onto the current version of the tax form. This is a process quite like that used by the various Individual Income Tax softwares, which, while consistent over the years, may not reflect the current status. This situation may well explain why roughly 4,000 partnership returns were received during 2003 with industry codes that were based on the obsolete Standard Industrial Classification (SIC) codes (see Table 1, below).

Although only a small proportion of the partnership returns are filed electronically, in order to use the data effectively in a sampling frame, the data must be accessible in that format. This means that the paper returns must be transcribed, at least in part. In practice, and as we have noted elsewhere, only a relatively small number of items are abstracted, but the industry code is one of them.

Sometimes, the respondent's handwriting is illegible, or they have provided clearly incorrect

values. Those cases are directed to a reviewer for correction, though that may result in assigning a code "999000" for "unknown." This may occur more frequently during periods where large numbers of records must be processed, but we have not examined this possibility.

For administrative reasons, the electronically-filed returns are automatically edited to include the same data items as those abstracted from the paper returns. The resulting records are known as "Transaction Records," following the usage in accounting practice.

Table 1: Tax Year 2003 Partnerships:
Transaction Records Validity

	Number	Proportion
Valid NAICS	2,297,000	95.9%
Valid SIC	3,700	0.2
Invalid NAICS	95,000	4.0
Invalid SIC	600	--

(*Proportions do not add to 100% due to rounding*)

The validity code on which Table 1 depends is the result of a simple test of whether a given industry code entry is on a list, and does not mean that the code is appropriate for the firm in question. Ascertaining the verity of a code for any particular record would require a separate source of that information.

Fortunately, there are other sources for an industry code available on the sampling frame. Once a partnership transaction record is complete and passed a series of perfunctory tests, it is ready for a process called "Posting." This process involves matching a transaction to a Business Master File Account based on the Employer Identification Number and selected other data, updating that account, and transferring some information to the transaction. We are interested here in the "Entity" part of the data, which includes such items as the name and address for contacting the firm, and an industry code. (We will, henceforth, refer to this code as the "Entity NAICS" code to distinguish it from the code on the Return Transaction.)

Table 2. Tax Year 2003 Partnerships:
Entity Industry Sources

	Number	Proportion
NAICS-Based Codes		
Transaction	2,157,000	90.0%
Social Security	219,000	9.1
Exam	4,900	0.2
Other	30	--
SIC-Based Codes		
Transaction	6,000	0.3
Social Security	600	--
Code Not Available	8,800	0.4

The information from the Social Security Administration is introduced at the time a firm receives an Employer Identification Number. Part of the processing of an application at Social Security involves assignment of a NAICS code, which is then passed to the Service along with other data needed to initiate an account.

Revisions to industry codes can arise as part of those administrative actions where agents contact the businesses, and these are grouped under the title "Exam" in Table 2. The other sources are really too small to detail, though they can include information about exempt organizations (since there are no constraints on the nature of an owner of a partnership).

The nearly 9,000 records with an industry code "Not Available" might be those with NAICS codes not on the Service's list. We tested this hypothesis by matching a copy of the 2002 version of these codes to those records. There were no matched records. A manual review of a handful suggests that data from an adjacent area of the return had been erroneously entered as the industry.

While most of the Entity NAICS entries arise from returns, via transactions, the codes are not necessarily from the current tax year. Almost 3 percent of such transactions had either invalid transaction NAICS codes or some SIC-based entry. We know these data must be from another source due to the rules on updating the Master File Accounts.

Those rules for updating the industry on the Master File accounts start with permitting only valid codes to be considered. Next, NAICS-based codes have higher priority than the SIC-based versions. And then, the source matters too: data from Exempt Organizations, over Social Security, over IRS's Examination, over the return transaction, over the occasional information from Collections, in that order. Finally, the posting program selects the code that has the greater specificity if all other factors are equal. (This routine applies to all records that are posted to the Business Master File, not just partnership records.)

In short, the process favors new over old, for greater source reliability (at least in the opinion of those designing the system), and for greater detail over lesser.

Given the strong reliance on information from the tax returns, we would expect significant agreement between the Entity NAICS and the transaction's code. Overall agreement, however, may hide real problems in some sectors.

For the balance of this review, we will confine our attention to the sectors, based on the first two digits of the NAICS Code. In part, this is due to space constraints for this article; but mostly, it is due to concerns about disclosure and the distribution of the Statistics of Income Partnership sample.

Table 3: Tax Year 2002 Partnerships Sector-Level Agreement Between Industry Codes

2002 North American Industry Code System (NAICS) Title	NAICS Sector	Records With NAICS	Entity NAICS from Transaction		Entity and Transaction Sectors Agree	
			Number	Percent	Number	Percent
Agriculture, Forestry, Fishing, and Hunting	11	125,763	119,463	95.0%	123,276	98.0%
Mining	21	26,046	23,700	91.0%	25,530	98.0%
Utilities	22	2,528	2,213	87.5%	2,326	92.0%
Construction	23	133,448	106,613	79.9%	123,180	92.3%
Manufacturing	31-33	40,263	35,101	87.2%	37,427	93.0%
Wholesale Trade	42	35,776	28,013	78.3%	31,310	87.5%
Retail Trade	44-45	124,100	107,755	86.8%	115,394	93.0%
Transportation and Warehousing	48-49	27,922	25,082	89.8%	26,234	94.0%
Information	51	25,585	20,458	80.0%	23,112	90.3%
Finance and Insurance	52	281,027	225,095	80.1%	266,524	94.8%
Real Estate and Rental and Leasing	53	1,008,948	976,126	96.7%	986,818	97.8%
Professional, Scientific, and Technical Services	54	157,084	138,160	88.0%	148,020	94.2%
Management of Companies and Enterprises	55	18,353	15,889	86.6%	15,866	86.4%
Administrative and Support and Waste Management and Remediation Services	56	37,691	26,842	71.2%	30,331	80.5%
Educational Services	61	6,141	4,158	67.7%	5,027	81.9%
Health Care and Social Assistance	62	47,350	40,861	86.3%	45,154	95.4%
Arts, Entertainment, and Recreation	71	33,951	27,696	81.6%	31,598	93.1%
Accommodation and Food Services	72	73,359	67,112	91.5%	70,769	96.5%
Other Services (except Public Administration)	81	70,881	62,192	87.7%	68,148	96.1%
Public Administration	92	48	32	66.7%	30	62.5%
Unknown	99	104,499	104,494	100.0%	103,981	99.5%
Total		2,380,763	2,157,055	90.6%	2,280,055	95.8%

Analysis of the Frame

The data in Table 3 are from the sampling frame (not a sample), using the Entity NAICS as the source for the sector, and with records excluded where the industry code is based on the Standard Industrial Classification (SIC) or is invalid. The rate of agreement between the two industry codes is almost 96 percent, which is not too surprising given the source for most of the codes. Over 90 percent of the codes arise from a Return Transaction, though some will be from prior-year records instead of the current tax year. The agreement rate for those records with the industry code arising from the transaction is, unsurprisingly, over 99.9 percent.

The agreement rate for records where the Entity NAICS did not arise from the transaction was 67.4 percent.

Sixteen of the 21 categories shown in Table 3 have agreement rates greater than 90 percent, with 7 higher than 95 percent. Most of the other groups have rates in the 80-to-90-percent range, and these sectors are among those with the fewest firms. Indeed, the smallest, Public Administration, has the lowest rate of agreement between the two NAICS codes.

This sector, though, would seem to be out of scope for a business survey. It may be that these organizations are charities forming some sorts of joint operations; we cannot tell from the data available, which are too sparse to begin with.

The other "sector" that is out of place is the group of "Unknown" firms. Since these comprise about 4.4 percent of the population, larger than most sectors, the characteristics of this group are of immediate interest. Three main variables are of particular interest: Net Income or Loss, Total Assets, and Total Receipts, because they indicate the size and activity of a firm.

Table 4: Tax Year 2002 Partnerships--Distributions of Firms by Selected Variables

	All	Valid NAICS		Unknown Industry	
Net Income/Loss		Number	Percent	Number	Percent
-1,000,000 or More	24,094	24,044	1.1%	50	0.0%
-250,000 Under -1,000,000	54,924	54,792	2.4%	132	0.1%
-1 Under -250,000	828,178	821,171	36.1%	7,007	6.7%
0 or Not Reported	173,815	85,554	3.8%	88,261	84.5%
1 Under 250,000	1,141,527	1,132,816	49.8%	8,711	8.3%
250,000 Under 1,000,000	112,347	112,086	4.9%	261	0.2%
1,000,000 or More	45,878	45,801	2.0%	77	0.1%
Total	2,380,763	2,276,264		104,499	
Total Assets					
0 or Not Reported	679,896	582,588	25.6%	97,308	93.1%
1 Under 250,000	792,447	787,636	34.6%	4,811	4.6%
250,000 Under 1,000,000	437,614	436,231	19.2%	1,383	1.3%
1,000,000 Under 25,000,000	439,259	438,307	19.3%	952	0.9%
25,000,000 or More	31,547	31,502	1.4%	45	0.0%
Total	2,380,763	2,276,264		104,499	
Total Receipts					
0 or Not Reported	373,559	283,159	12.4%	90,400	86.5%
1 Under 250,000	1,450,103	1,437,916	63.2%	12,187	11.7%
250,000 Under 1,000,000	347,008	345,586	15.2%	1,422	1.4%
1,000,000 Under 25,000,000	198,720	198,248	8.7%	472	0.5%
25,000,000 or More	11,373	11,355	0.5%	18	0.0%
Total	2,380,763	2,276,264		104,499	

The data in Table 4 depend on the transaction records, and, thus, the monetary variables do have some limitations. For example, some items that would belong in an economic definition of Total Receipts or Net Income/Loss are not available from those records. Still, the main contributing items are present, such as gross receipts and net rent from real estate.

The firms that have an unknown industry have a disproportionate number showing no net income or loss among the items available on the frame. Not only do nearly 85 percent show zero for that amount, but that group provides more than half of the firms without net income or loss during 2002. Even when we exclude those with a zero for that amount, the distribution of net income or loss drops off much more rapidly, at roughly thrice the pace, than for firms with reported industries.

The picture for Total Assets is less clear, but this is due in large part to a regulation that permits firms with less than $250,000 in total receipts and less than $600,000 in total assets to withhold that information from their filings. The dropoff is not as steep as it is for Net Income, but the effect is still there.

This pattern of concentration at zero with attenuated tails of the distributions continues for Total Receipts. Actually, all but a few hundred of the records that reported no net income or loss also had zeros for amounts of total assets and total receipts.

This raises the question of what industry these firms actually belong in. Remembering that the instructions for filing asks the respondent to use total receipts as the basis, if that amount is in fact zero, then should not the response be "unknown?"

These firms may be characterized as inactive, with the filings being in response to the form the Service mailed. In fact, using the Statistics of Income Partnership Study, we estimate that there are about 137,000 such firms, nearly 27,000 more than the frame counts. The difference is likely due to the variations between the tax law definitions and those based on economic concepts used for the SOI study.

Table 5: Tax Year 2002 Partnerships--Sample Estimates of Industry Distribution

2002 NAICS Title	Sector	Entity NAICS	Edited NAICS	Entity & Sample Agree	Error Rate
Agriculture, Forestry, Fishing, and Hunting	11	117,048	117,667	110,941	5.2%
Mining	21	28,095	29,549	27,896	0.7%
Utilities	22	2,331	2,507	2,019	13.4%
Construction	23	126,423	134,114	115,173	8.9%
Manufacturing	31-33	36,787	38,364	33,185	9.8%
Wholesale Trade	42	37,240	37,800	30,470	18.2%
Retail Trade	44-45	118,595	122,013	109,400	7.8%
Transportation and Warehousing	48-49	26,573	26,007	23,569	11.3%
Information	51	23,613	28,580	21,334	9.7%
Finance and Insurance	52	256,820	263,024	248,520	3.2%
Real Estate and Rental and Leasing	53	985,603	999,786	966,940	1.9%
Professional, Scientific, and Technical Services	54	155,372	145,612	133,832	13.9%
Management of Companies and Enterprises	55	17,896	18,773	15,450	13.7%
Administrative and Support and Waste Management and Remediation Services	56	37,794	44,405	30,337	4.1%
Educational Services	61	5,569	6,269	4,575	17.9%
Health Care and Social Assistance	62	46,321	47,468	44,411	4.1%
Arts, Entertainment, and Recreation	71	39,227	42,691	35,859	8.6%
Accommodation and Food Services	72	73,881	77,698	71,099	3.8%
Other Services (except Public Administration)	81	67,177	57,121	49,332	26.6%
Unknown or SIC-Based Code	Unknown	39,804	2,724	2,053	94.8%
Total	All	2,242,169	2,242,169	2,074,342	7.5%

Partnership Sample

Thus far, the discussion has focused on the data from the administrative systems only. If we assume that agreement between the Transaction Record and the Entity NAICS implies validity, then we see that the proportion of partnership records with "valid" industry sectors is about 95.8 percent. Removing those records where the industry is "unknown" only drops this figure to 95.6 percent.

These conclusions rest, however, on a simple list matching, not on inspection of source records. Fortunately, the Statistics of Income Partnership Study for Tax Year 2002 included a significant effort to verify the NAICS codes (though without contacting the respondents). This effort included researching publicly available published and Internet data.

Of the 34,800 records selected for this sample, 33,600 were considered "in scope" and received the extra attention. In the end, only 17 records could not be assigned a NAICS code. The corresponding estimated population for the "unknown industry" is about 2,700, or slightly over 0.1 percent. The coding used the Service's version of NAICS, not the full set of codes.

Note that matching the full NAICS list's 6-digit codes against those assigned to the sample results in about 16,400 records, almost half, being identified as having invalid codes. That is, if the full population were treated as the sample was, about a third (761,000) would not have valid codes under the naïve assumption.

The sample was drawn from the frame, described in the previous section, as the records were filed during 2003. Strata were defined by size of total assets, net income (or loss) or receipts, industry, and select other characteristics of special importance to our sponsors.

We included industry in the design because division level estimates were deemed important. With the real estate leasing businesses comprising over a third of all partnerships, a proportionate distribution of the sample over all the groups would have left several sparsely sampled. Hence, we reduced the sample in real estate and increased the sample for

other industry divisions, and particularly those with few firms. This resulted in a sample with sufficient records at the sector level to assess the accuracy of the NAICS codes, at that level of aggregation, on partnership transaction records.

We compare, in Table 5, the estimated distribution across industry (for active partnerships) using the Entity NAICS codes, and the codes assigned during the data abstraction. The frequencies are quite similar. Most of the estimates using the validated codes are a bit higher than those based on the Entity NAICS, with the greatest proportionate differences in the less populous sectors.

Some difference is expected, of course, because there was a recoding of most of the nearly 40,000 records without a NAICS code. There was also a large movement from "Other Services," which may be what the respondents decided to use when they could not easily find an answer.

However, the similarity of the distributions masks a greater disagreement between the two sets of codes. The overall accuracy drops to 92.5 percent from over 95 percent, but even this needs to be qualified. "Real Estate Rental and Leasing," which contains almost 45 percent of the population, has an error rate of only 1.9 percent. This low error rate is undoubtedly due to the ease that the original coding clerks for the transaction records have in determining an industry: these returns all have Form 8825, *Rental Real Estate Income and Expenses of a Partnership or an S Corporation*, attached.

On the other hand, we should also consider that the category "Other Services" is the equivalent of "miscellaneous." That list of codes is rather long, at three pages; so, having a large number of records from that category being reassigned is to be expected.

Removing those sectors from consideration reduces the overall agreement to only slightly more than 89 percent. "Educational Services" has a small sample, and only a dozen or so were reassigned to other sectors. "Wholesale Trade," however, presents quite a puzzle, with over 100 records reclassified, and only about a third into "Retail Trade" where we might expect them.

Conclusion

A major reason for this review was to ascertain whether the industry codes on the IRS's Business Master File system for partnerships is sufficiently reliable for stratification purposes. With respect to real estate firms, the quality is quite sufficient, at least for the Entity NAICS. The picture is less clear with respect to those sectors with small populations, where, in some cases, the proportion reclassified is modest, while, in others, the error rates are quite high.

We cannot, of course, generalize to other types of administrative records maintained on the Business Master File, such as Corporation Income Tax Returns, though we note that they appear to have a similar situation with respect to having clearly invalid codes. That investigation will have to be the subject of another paper.

Nor can we attribute the error to any source. The nature of the data before us does not allow us to distinguish between errors by the respondent or the reviewer, except, of course, where the form contains an old SIC-based industry code. This is, however, only a small piece of the non-NAICS coded records.

The sample was too small for more detailed analysis, but it is certain that the finer the coding, the more relative error we can expect. It is also clear that the methods employed to refine the sample cannot be used on the entire population with any hope of success.

Notes and Cites

[1] North American Industry Classification System, United States (2002), Executive Office of the President, Office of Management and Budget, Introduction, page 16.

[2] Total receipts is the sum of:

Form 1065, pg .1: Gross Receipts, Ordinary Income From Other Partnerships, Net Farm Profit, Net Gain or Loss From the Sale of Business Property, and Other Income;

Schedule K: Non Real Estate Rents, Interest Income, Ordinary Dividends, Royalty Income, Short Term Capital Gains, Long Term Capital Gains (Taxed at the 28 Percent Rate), Other Portfolio Income, Income Under Section 1231, and Other Income;

Form 8825: Gross Real Estate Rents, Net Gain or Loss From the Sale of Business Property, and Income From Other Real Estate Partnerships.

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

McMahon, Paul (2002), "Proxies in Administrative Records Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

McMahon, Paul (2003), "Regulatory Exemptions and Item Nonresponse," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.