Section 3 Description of the Sample and Limitations of the Data

This section describes the 2001 Corporate sample design, sample selection, data capture, data cleaning, and data completion. The techniques used to produce estimates and an assessment of the data limitations, including sampling and non-sampling errors, are also discussed.

Background

From Tax Year 1916 through Tax Year 1950, data were extracted for the Statistics of Income (SOI) program from each corporate return filed. Stratified probability sampling was introduced for Tax Year 1951. Since that time, the sample size has generally decreased while the population has increased. For example, for Tax Year 1951 the sample comprised 41.5 percent of the entire population, or 285,000 of the 687,000 total returns filed. In comparison, for 2001, the sample proportion was about 2.6 percent of the total population of over 5.5 million.

For 1951, stratification was by size of total assets and industry. From 1952 through 1967, the stratification was by a measure of size only. The size was measured by volume of business (1953-1958) or total assets (1952 and 1959-1967). Since 1968, returns have been stratified by both total assets and, for Form 1120, 1120-A and 1120S returns, a measure of income [1].

Target Population

The target population consists of all returns of active corporations organized for profit that are required to file one of the 1120 forms that are part of the SOI study.

Survey Population

The survey population includes the returns that filed one of the 1120 forms selected for the SOI study and posted to the IRS Business Master File (BMF). Amended returns and returns for which the tax liabilities changed because of a tax audit are excluded. Figure C gives the actual number of corporate returns by form type that were subject to sampling during Tax Years 1998 through 2001. These population counts differ from all the estimated population counts in this publication because they include out-of-scope returns.

Bertrand Überall, Richard Collins, and Kim Henry were responsible for the sample design and estimation of the SOI 2001 Corporation Program under the direction of Yahia Ahmed, Chief, Mathematical Statistics Section, Statistical Computing Branch.

Figure C.--Population Counts by Corporate Form Type, Tax Years 1998-2001

	Tax Year			
Form Type	1998	1999	2000	2001
1120	2,190,409	2,165,338	2,146,170	2,142,542
1120-A	259,696	244,339	235,459	231,622
1120S	2,716,507	2,866,963	3,008,022	3,147,642
1120-L	1,572	1,522	1,465	1,479
1120-PC	3,352	3,437	3,593	3,930
1120-RIC	10,044	10,449	11,157	11,479
1120-REIT	969	1,079	1,114	1,057
1120-F	22,157	22,270	22,385	23,912
Total	5,204,706	5,315,397	5,429,365	5,563,663

Sample Design

The current sample design is a stratified probability sample, with stratification by form type, and either size of total assets alone, or both size of total assets and a measure of income. Forms 1120 and 1120-A are stratified by size of total assets and size of "proceeds." Size of "proceeds", which is used as the measure of income, is defined to be the larger of the absolute value of net income (or deficit) or the absolute value of "cash flow," which is the sum of net income, several depreciation amounts, and depletion. Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, and 1120-REIT are each stratified by size of total assets only. Form 1120S is stratified by size of total assets and size of ordinary income.

The design process began with projected population totals that were derived from those used to estimate IRS administrative workloads and adjusted based on previous years' population distributions. Using projected population totals by sample strata, an optimal allocation, based on stratum standard errors and cost estimates, was carried out to assign a sample size to each stratum such that the overall targeted sample size was approximately 137,000. A Bernoulli sample was selected independently from each stratum with sampling rates ranging from 0.25 percent to 100 percent. Figure D on the following page shows the stratum boundaries, the sampling rates, and the frame population and sample counts from the BMF for each form type. This table also shows the population and sample counts after adjustments for missing returns, outliers, and weight trimming. The total realized sample for Tax Year 2001, including inactive corporations and non-eligible returns, is 147,093 returns.

Figure D.--Corporation Returns: Number Filed, Number in Sample, and Sampling Rates, by Selection Class

Sample class number	Description of sample selection classes		Sampling rates (%)	Number of returns			
				BMF counts		After adjustments**	
	Size of total assets	Size of proceeds*		Population	Sample	Population	Sample
	All Returns, Total			**5,563,663**	**147,093**	**5,563,781**	**146,479**
	Form 1120 w/ Form 5735 attached, Total			**358**	**358**	**358**	**352**†
1	Under $100,000,000		100.00	286	286	286	282
2	$100,000,000 - $250,000,000		100.00	36	36	36	35
3	$250,000,000 or more		100.00	36	36	36	35
	Form 1120 (no Form 5735 attached), 1120-A, Total ***			**2,364,422**	**73,789**	**2,364,531**	**73,347**
4	Under $50,000	Under $25,000	0.40	884,902	3,580	884,902	3,568
5	$50,000 - $100,000	$25,000 - $50,000	0.40	311,960	1,221	311,960	1,221
6	$100,000 - $250,000	$50,000 - $100,000	0.50	408,684	1,992	408,684	1,984
7	$250,000 - $500,000	$100,000 - $250,000	1.00	278,273	2,712	278,273	2,707
8	$500,000 - $1,000,000	$250,000 - $500,000	1.60	193,431	3,086	193,431	3,079
9	$1,000,000 - $2,500,000	$500,000 - $1,000,000	4.00	147,742	5,772	147,742	5,763
10	$2,500,000 - $5,000,000	$1,000,000 - $1,500,000	5.60	57,731	3,208	57,731	3,198
11	$5,000,000 - $10,000,000	$1,500,000 - $2,500,000	10.00	32,727	3,246	32,725	3,237
12	$10,000,000 - $25,000,000	$2,500,000 - $5,000,000	100.00	23,238	23,238	23,250	23,163
13	$25,000,000 - $50,000,000	$5,000,000 - $10,000,000	100.00	11,259	11,259	11,258	10,914
14	$50,000,000 - $100,000,000	$10,000,000 - $15,000,000	100.00	6,891	6,891	6,891	6,856
15	$100,000,000 - $250,000,000	$15,000,000 or more	100.00	5,024	5,024	5,024	4,997
16	$250,000,000 - $500,000,000		100.00	1,253	1,253	1,247	1,247
17	$500,000,000 or more		100.00	1,307	1,307	1,413	1,413
	Form 1120S, Total ***			**3,146,661**	**45,651**	**3,146,663**	**45,578**
18	Under $50,000	Under $25,000	0.25	1,226,355	3,034	1,226,354	3,027
19	$50,000 - $100,000	$25,000 - $50,000	0.25	506,944	1,243	506,944	1,242
20	$100,000 - $250,000	$50,000 - $100,000	0.25	567,751	1,435	567,751	1,433
21	$250,000 - $500,000	$100,000 - $250,000	0.40	373,615	1,544	373,615	1,541
22	$500,000 - $1,000,000	$250,000 - $500,000	0.76	206,053	1,556	206,052	1,553
23	$1,000,000 - $2,500,000	$500,000 - $1,000,000	2.10	146,080	3,022	146,078	3,019
24	$2,500,000 - $5,000,000	$1,000,000 - $1,500,000	3.30	57,293	1,908	57,292	1,905
25	$5,000,000 - $10,000,000	$1,500,000 - $2,500,000	6.40	32,774	2,113	32,772	2,109
26	$10,000,000 - $25,000,000	$2,500,000 - $5,000,000	100.00	19,224	19,224	19,227	19,189
27	$25,000,000 - $50,000,000	$5,000,000 - $10,000,000	100.00	6,175	6,175	6,172	6,160
28	$50,000,000 - $100,000,000	$10,000,000 - $15,000,000	100.00	2,375	2,375	2,374	2,373
29	$100,000,000 - $250,000,000	$15,000,000 or more	100.00	1,302	1,302	1,295	1,290
30	$250,000,000 or more		100.00	720	720	737	737
	Form 1120-L, Total			**1,239**	**673**	**1,247**	**676**
31	Under $10,000,000		43.00	971	405	955	385
32	$10,000,000 - $50,000,000		100.00	146	146	151	150
33	$50,000,000 - $250,000,000		100.00	61	61	63	63
34	$250,000,000 or more		100.00	61	61	78	78
	Form 1120-F, Total			**23,850**	**3,823**	**23,858**	**3,818**
35	Under $10,000,000		13.00	23,034	3,007	23,031	2,993
36	$10,000,000 - $50,000,000		100.00	444	444	446	445
37	$50,000,000 - $250,000,000		100.00	191	191	191	190
38	$250,000,000 or more		100.00	181	181	190	190
	Form 1120-PC, Total			**3,651**	**1,068**	**3,652**	**1,061**
39	Under $2,500,000		10.00	2,211	199	2,206	191
40	$2,500,000 - $10,000,000		25.00	769	198	769	197
41	$10,000,000 - $50,000,000		100.00	536	536	541	538
42	$50,00,000 - $250,000,000		100.00	127	127	126	125
43	$250,000,000 or more		100.00	8	8	10	10
	Form 1120-REIT, Total			**829**	**700**	**831**	**695**
44	Under $10,000,000		25.00	178	49	181	47
45	$10,000,000 - $50,000,000		100.00	188	188	188	187
46	$50,000,000 - $250,000,000		100.00	248	248	248	247
47	$250,000,000 or more		100.00	215	215	214	214
	Form 1120-RIC, Total			**10,989**	**9,367**	**10,992**	**9,359**
48	Under $10,000,000		15.00	1,887	265	1,875	251
49	$10,000,000 - $50,000,000		100.00	2,409	2,409	2,416	2,413
50	$50,000,000 - $100,000,000		100.00	1,459	1,459	1,459	1,458
51	$100,000,000 - $250,000,000		100.00	2,009	2,009	2,008	2,003
52	$250,000,000 - $500,000,000		100.00	1,342	1,342	1,343	1,343
53	$500,000,000 or more		100.00	1,883	1,883	1,891	1,891
54	**Special Studies**		100.00	**11,664**	**11,664**	**11,649**	**11,593**†

* Proceeds is defined as the larger of absolute value of net income (deficit) or absolute value of cash flow (net income + depreciation + depletion).
** Includes adjustments for missing returns, outliers, and weight trimming.
*** Returns were classified according to either size of total assets or size of proceeds, whichever corresponded to the higher sample class. Example: A Form 1120 return with total assets of $750,000 and proceeds of $75,000 is in sample class 8 (based on total assets), rather than in sample class 6 (based on proceeds).
† The adjusted sample count is lower than the adjusted population count due to returns unavailable for processing.

Sample Selection

Corporation income tax returns are filed at the Cincinnati, Ogden, and Philadelphia IRS Submission Processing Centers. All corporate returns are processed initially to determine tax liability. All tax data are transmitted and updated on a weekly basis to the IRS Business Master File (BMF) system located in Martinsburg, West Virginia. These returns are said to "post" to the BMF. This BMF database serves as the SOI sampling frame. The SOI sample is also selected on a weekly basis.

Sample selections for Tax Year 2001 occurred over the period of July 2001 through June 2003. A 24-month sampling period is needed for two reasons. First, approximately 16.9 percent of all corporations had noncalendar year accounting periods. In order to take these filings into consideration, the 2001 statistics represent all corporations filing returns with accounting periods ending during the period from July 2001 to June 2002. Also, many corporations, including some of the largest, request six-month filing extensions. The combination of noncalendar year filing and filing extensions means that the last Tax Year 2001 returns that the IRS received (those with accounting periods ending in June 2002, which must therefore be filed by October 2002) could be timely filed as late as March 2003, if the six-month extension of the October 2002 due date is taken into account. Normal administrative processing time lags required that the sampling process remain open for the 2001 study until June 30, 2003. However, a few very large returns for Tax Year 2001 were added to the sample as late as November 2003.

Each tax return posted to the BMF and belonging to the survey population (as defined above) is assigned to a stratum and then subjected to sampling. Each filing corporation has a unique Employer Identification Number (EIN). An integer function of the EIN, called the Transformed Taxpayer Identification Number (TTIN), is computed. The number formed by the last four digits of the TTIN is a pseudo-random number. A return for which this pseudo-random number is less than the sampling rate multiplied by 10,000 is selected in the sample.

The algorithm for generating the TTIN does not change from year to year. Consequently, any corporation selected into the sample in a given year will be selected again the next year, providing that the corporation files a return using the same EIN in the two years and that it falls into a stratum with the same or higher sampling rate. If the corporation falls into a stratum with a lower rate, the probability of selection will be the ratio of the second year sampling rate to the first year sampling rate. If the corporation files with a new EIN, the probability of being selected will be independent of the prior year selection [2].

Data Capture

Data processing for SOI begins with information already extracted for administrative purposes; over 100 items are available from the BMF system, and are checked and corrected as necessary. Some 1,300 additional data items are extracted from the tax returns during SOI processing. The SOI data capture process can take as little time as fifteen minutes for a small, single entity corporation filing on Form 1120-A, or up to a week for a large consolidated corporation filing several hundred attachments and schedules with the return. The process is further complicated by several factors:

- The 1,300 separate data items that may be extracted from any given tax return often require totals to be constructed from various other items on other parts of the return.
- Each 1120 form type has a different layout with different types of schedules and attachments, making data extraction less than uniform for the various form types.
- There is no legal requirement that a corporation meet its tax return filing requirements by filling in, line by line, the entire U.S. tax return form. Therefore, many corporate taxpayers report many of their financial details in schedules of their own design.
- There is no single accepted method of corporate accounting used throughout the country, but rather several accepted accounting "guidelines," many of which are unique to geographic locations. SOI staff attempt to standardize these differences during data abstraction and editing.
- Different companies may report the same data item, such as other current liabilities, on different lines of the tax form. Again, SOI staff attempt to standardize these differences.

In order to help overcome these complexities and differences due to taxpayer reporting, SOI staff prepare detailed instructions for the SOI editing unit at the IRS Submission Processing Centers each tax year. For Tax Year 2001, these instructions consisted of more than 900 pages covering standard and straightforward procedures and instructions for exceptions that might be encountered.

Data Cleaning

Statistical processing of the corporate returns is performed in an online computer environment and the data from returns are entered directly into the SOI corporation database. In this context, the term "editing" refers to the combined interactive processes of data extraction, consistency testing, and error resolution. There are over 900 of these tests, which look for such inconsistencies as:

- Impossible conditions, such as incorrect tax data for a particular form type;
- Internal inconsistencies, such as items not adding to totals;
- Questionable values, such as a bank with an unusually large amount reported for cost of goods sold and/or operations; and
- Improper sample class codes, such as when a return has $100 million in total assets, but was selected as though it had $1 million because the last two digits of the total assets were keyed in as cents.

Data Completion

In addition to the tests mentioned above, missing data problems must be addressed and returns that are to be excluded from the tabulations must be identified. The data completion process focuses on these issues.

If the missing data items are from the balance sheet, then imputation procedures are used. If data for a whole return are missing because the return is unavailable to SOI during the data capture process, imputation procedures are also used in certain cases.

A ratio-based imputation procedure is used to estimate missing balance sheet items for all 1120 forms except those with less than 12-month accounting periods. The ratios are determined using the most recent data available, either the corporation's 2000 return (if the corporation filed a return in 2000) or the 1999 aggregate data for the corporation's minor industrial group, which are the most recent aggregate data available at the time the editing for Tax Year 2001 begins. If the reported items in the balance sheet do not balance (i.e., the sum of asset items does not equal the sum of liability and shareholders' equity items), then missing items are imputed. If the total assets amount is among the missing items, this item is imputed first based on the ratio of total assets to business receipts (or total receipts) from either the corporation's 2000 return, or the 1999 aggregate data for the corporation's minor industry. The other missing asset and liability items are then imputed based on the ratios so that the total of all asset items and the total of all liability items are both equal to the total assets amount, whether this amount was reported or imputed. A detailed description of the balance sheet imputation process is given in reference [3]. The following chart shows the number of sampled returns that had balance sheet items imputed, as well as the percentages they represent of the total sample sizes, for Tax Years 1998 through 2001.

Returns with imputations	Tax Year			
	1998	1999	2000	2001
Number of imputed returns	70	68	38	41
Percent imputed	0.05	0.05	0.03	0.03

For Tax Year 2001, none of the 41 imputed returns had imputed total assets.

Data for unavailable critical corporations are imputed in various ways, depending on what information is available at the time the SOI database is produced. Critical corporations include corporations with total assets greater than or equal to 5 percent of the total assets for their minor industrial group, and corporations for which total assets are over a specified limit, which is dependent on form type or minor industry. For critical corporations selected for the sample but unavailable for statistical processing, taxpayer-surveyed data are used. There are six such returns in the Tax Year 2001 data. For critical corporations not selected for the sample, if the current tax return is not found in any of the IRS Submission Processing Centers and no other current tax data are available, data from the previous year's return are used with adjustments for tax law changes. There are 21 prior year returns in the Tax Year 2001 data.

Another part of the data cleaning process is identifying sampled returns that are not eligible for the sample. The BMF system used for sample selection can include duplicate tax returns and other out-of-scope returns, such as returns of nonprofit corporations, returns having neither current income nor deductions, prior-year tax returns, amended or tentative returns, returns of nonresident foreign corporations having no effectively connected income with a trade or business within the United States, fraudulent returns, and returns of corporations exempt from taxation.

Figure E on page 11 displays the number of inactive sampled returns that were excluded from tabulations, as well as the percentages they

represent of the total sample sizes, in Tax Years 1998 through 2001.

Figure E.--Number of Inactive Sampled Returns for Tax Years 1998-2001

	Tax Year			
Type of inactive return	1998	1999	2000	2001
No Income or Deductions	1,460	1,450	1,615	1,668
Duplicate*	799	770	1,044	1,421
Other**	3,645	3,725	3,684	4,294
Total	5,904	5,945	6,343	7,383
Percent of sample	4.29	4.22	4.38	5.02

* Duplicate returns are those that appear more than once in the sample.
** Includes prior-year returns.

Estimates of the number of active corporations by form type for Tax Years 1998 through 2001 are provided in Figure F below.

Figure F.--Estimated Number of Active Returns for Tax Years 1998-2001

	Tax Year			
Form Type	1998	1999	2000	2001
1120	2,021,929	1,990,782	1,970,777	1,936,066
1120-A	211,801	191,769	186,177	185,114
1120S	2,588,088	2,725,775	2,860,478	2,986,486
1120-L	1,620	1,551	1,520	1,474
1120-PC	3,624	3,739	3,732	3,949
1120-RIC	9,897	10,318	10,991	11,318
1120-REIT	932	1,071	1,099	1,031
1120-F*	10,996	10,898	10,498	10,154
Total	4,848,888	4,935,904	5,045,274	5,135,591

* Foreign Insurance Companies file on Forms 1120-L and 1120-PC, but are counted in Form 1120-F Tables 10 and 11.
Note: Detail may not add to total due to rounding.

Estimation

The estimates of the total number of corporations and associated money amounts produced in this report are based on weighted sample data. Either a one-step process or a two-step process was used to determine the weights, depending on the return's form type.

Under the one-step process, the weights are assigned as the reciprocal of the achieved sampling rate, adjusted for missing returns, outliers, and weight trimming. These weights, referred to as the "national weights", are used to produce the aggregated total frequencies and money amounts published in this report for Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, 1120-REIT and Form 1120 with Form 5735 attached, as well as for Form 1120 and 1120S returns that were sampled with certainty.

The two-step process was used to improve the industry estimates for Form 1120-A, and Form 1120 and 1120S returns that are not self-representing. The first stage is the one-step process described above, which provides an initial weight for the return. The second stage involves post-stratification by industry and sample selection class. The industry classification used for post-stratification weighting is a two-digit code based on the North American Industrial Classification System (NAICS). In the course of this two-dimensional post-stratification, certain cells may have small sample sizes. To handle this problem, a bounded raking ratio estimation approach is applied in order to determine the final weight [4]. Restrictions are placed on the raking process to produce final weights that fall within the range /(2/3) x original weight to /(3/2) x original weight. These final weights are used to produce the aggregated frequency and money amount estimates that are published in this report for these forms.

Data Limitations and Measures of Variability

Several extensive quality review processes are used to improve data quality, beginning at the sample selection stage with weekly monitoring of the sample to ensure that the proper number of returns is being selected. They continue through the data collection, data cleaning, and data completion procedures with consistency testing. Part of the review process includes extensive comparisons between the 2001 data and the 2000 data. A great amount of effort is made at every stage of processing to ensure data integrity.

Sampling Error

Since the corporation estimates are based on a sample, they may differ from the population aggregates that would have been obtained if a complete census of all income tax returns had been taken. The particular sample used to produce the results in this report is one of a large number of possible samples that could have been selected under the same sample design. Estimates derived from one of the possible samples could differ from those derived from other samples and from the population aggregates. The deviation of a sample estimate from the average of all possible similarly selected samples is called the sampling error. The standard error (SE) is a measure of the average magnitude of the sampling errors over all possible samples.

The standard error is the most commonly used measure of the sampling error and can be estimated from the sample. The estimated standard error is usually expressed as a percentage of the value being estimated. This is called the estimated coefficient of variation (CV) of the estimate, and it can be used to assess the reliability of an estimate.

The estimated coefficient of variation of an estimate is calculated by dividing the estimated standard error by the estimate. Estimated coefficients of variation by industrial groupings for the estimated number of returns, as well as for selected money amount estimates, are shown in Table 1 beginning on page 29. For the estimated number of returns by asset size and sector, estimated coefficients of variation are given in Figure G on page 13. The corresponding estimates can be found in Table 4.

The estimated coefficient of variation, CV(X), can be used to construct confidence intervals for the estimate X. The estimated standard error, which is required for the confidence interval, must first be calculated. For example, the estimated number of companies in the manufacturing sector with net income and the corresponding estimated coefficient of variation can be found in Table 1 and used to calculate the estimated standard error:

$$SE(X) = X \bullet CV(X)$$
$$= 147{,}291 \times 3.27/100$$
$$= 4{,}816$$

Assume that a 95-percent confidence interval for the estimated number of returns in manufacturing is desired. The 95-percent confidence interval is constructed as follows:

$$X \pm 2 \bullet SE(X) = 147{,}291 \pm (2 \times 4{,}816)$$
$$= 147{,}291 \pm 9{,}632$$

Thus, the interval estimate is 137,659 returns to 156,923 returns. This means that if all possible samples were selected under essentially the same general conditions and using the same sample design, and if an estimate and its estimated standard error were calculated from each sample, then approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average estimate derived from all possible samples. Thus, for a particular sample, it can be said with 95-percent confidence that the average of all possible samples is included in the constructed interval. This average of the estimates derived from all possible samples would be equal to or near the value obtained from a census.

Nonsampling Error

In addition to sampling error, nonsampling error can also affect the estimates. Nonsampling errors can be classified into two groups: random errors, whose effects may cancel out, and systematic errors, whose effects tend to remain somewhat fixed and result in bias.

Nonsampling errors can be categorized as coverage errors, nonresponse errors, processing errors, or response errors. These errors can be the result of the inability to obtain information about all returns in the sample, differing interpretations of tax concepts or instructions by the taxpayer, inability of a corporation to provide accurate information at the time of filing (data are collected before auditing), inability to obtain all tax schedules and attachments, errors in recording or coding the data, errors in collecting or cleaning the data, errors made in estimating for missing data, and failure to represent all population units.

Coverage Errors: Coverage errors in the SOI Corporation data can result from the difference between the time frame for sampling and the actual time needed for filing and processing the returns. As stated previously, many of the largest corporations receive extensions to their filing periods and, as a result, may file their returns after sample selection has ended for that tax year. However, any of the largest returns found are added into the file until the final file is produced.

Coverage problems within industrial groupings in the SOI Corporation study result from the way consolidated returns may be filed. The Internal Revenue Code permits a parent corporation to file a single return, which includes the combined financial data of the parent and all its subsidiaries. These data are not separated into the different industries but are entered only into the industry with the largest receipts. Thus, there is undercoverage of financial data within certain industries and overcoverage in others. Coverage problems within industrial groupings present a limitation on any analysis done with the sample results.

Nonresponse Errors: Unit nonresponse occurs when a sampled return is unavailable for SOI processing. For example, other areas of the IRS may have the return at the time it is needed for statistical processing. These returns are termed "unavailable returns." In 2001, there were 326 unavailable returns in the corporation study, which constituted about 0.22 percent of the total sample size. The number of unavailable returns and the percentages of the total sample size for Tax Years 1998 through 2001 are shown in the following chart.

Unavailable returns	Tax Year			
	1998	1999	2000	2001
Number of unavailable returns	154	228	412	326
Percent unavailable	0.11	0.16	0.28	0.22

Figure G.--Coefficients of Variation (CVs) for Number of Returns, by Asset Size and Sector, for Tax Year 2001

Sector	All asset sizes	Size of total assets: Zero assets	$1 under $ 500,000	$500,000 under $1,000,000	$1,000,000 under $5,000,000
	(1)	(2)	(3)	(4)	(5)
All industries[1]	**0.19**	**2.87**	**0.34**	**0.81**	**0.35**
Agriculture, forestry, fishing, and hunting	2.59	20.56	3.58	4.33	2.43
Mining	6.89	39.25	9.78	15.19	7.61
Utilities	16.05	83.44	20.61	44.89	18.22
Construction	1.09	8.80	1.50	2.94	1.30
Manufacturing	2.24	12.96	3.61	4.21	1.65
Wholesale and retail trade	1.03	7.37	1.40	2.11	0.95
Transportation and warehousing	3.08	15.69	3.91	6.80	4.32
Information	4.19	18.12	5.18	12.46	5.60
Finance and insurance	2.50	13.16	3.43	6.32	3.23
Real estate and rental and leasing	1.24	8.32	1.75	2.31	1.50
Professional, scientific, and technical services	1.18	7.25	1.43	5.18	3.04
Management of companies (holding companies)	5.97	18.51	9.76	15.38	7.24
Administrative and support and waste management and remediation services	2.91	14.38	3.33	9.97	5.85
Educational services	7.38	31.67	8.30	29.93	14.50
Health care and social assistance	1.52	12.48	1.79	8.30	5.45
Arts, entertainment, and recreation	4.06	16.95	5.09	10.47	5.24
Accommodation and food services	1.71	15.15	2.16	6.05	2.75
Other services	2.18	13.78	2.44	6.65	5.22

Sector	Size of total assets – continued: $5,000,000 under $10,000,000	$10,000,000 under $25,000,000	$25,000,000 under $50,000,000	$50,000,000 under $100,000,000	$100,000,000 under $250,000,000
	(6)	(7)	(8)	(9)	(10)
All Industries[1]	**1.01**	**0.03**	**0.07**	**0.05**	**0.05**
Agriculture, forestry, fishing, and hunting	5.88	0.24	0.95	0.68	1.45
Mining	8.29	0.31	0.90	0.66	0.86
Utilities	21.82	0.78	2.07	1.21	1.18
Construction	3.62	0.09	0.35	0.26	0.59
Manufacturing	2.13	0.08	0.23	0.17	0.23
Wholesale and retail trade	1.45	0.06	0.24	0.18	0.32
Transportation and warehousing	6.76	0.21	0.79	0.58	0.76
Information	6.03	0.23	0.55	0.37	0.51
Finance and insurance	3.97	0.10	0.20	0.09	0.09
Real estate and rental and leasing	2.70	0.10	0.43	0.32	0.61
Professional, scientific, and technical services	4.50	0.20	0.52	0.37	0.50
Management of companies (holding companies)	8.45	0.18	0.41	0.16	0.16
Administrative and support and waste management and remediation services	11.28	0.30	0.93	0.67	1.01
Educational services	25.55	0.86	2.98	1.76	2.55
Health care and social assistance	8.81	0.33	1.09	0.71	1.10
Arts, entertainment, and recreation	9.18	0.32	1.21	0.84	1.16
Accommodation and food services	6.51	0.23	0.87	0.65	0.96
Other services	46.14	0.39	1.79	1.16	1.79

[1] Includes returns not allocable by sector.

Note: Returns with assets of $250,000,000 or more are self-representing.

Processing Errors: Errors in recording, coding, or processing the data can cause a return to be sampled in the wrong sampling class. This type of error is called a mis-stratification error. One example of how a return might be mis-stratified is the following: a corporation files a return with total assets of $100,000,023 and net income of $5,000. A processing error causes the last two digits of the total assets to be keyed in as cents, so that the return is classified according to total assets of $1,000,000.23 and net income of $5,000.00. The return would be mis-stratified according to the incorrect value of the total assets stratifier. To adjust for mis-stratification errors, only returns selected in a non-certainty stratum which really belonged in a certainty stratum were moved to this stratum.

Response errors: Response errors are due to data being captured before audit. Some purely arithmetical errors made by the taxpayer are corrected during the data capture and cleaning processes. Because of time constraints, adjustments to a return during audit are not incorporated into the SOI file.

References

[1] Jones, H. W., and McMahon, P. B. (1984), "Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present," *1984 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 437-442.

[2] Harte, J. M. (1986), "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *1986 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 603-608.

[3] Überall, B. (1995), "Imputation of Balance Sheets for the 1992 SOI Corporate Program," *1995 Proceedings of the Section on Survey Research Methods,* American Statistical Association, pp. 275-280.

[4] Oh, H. L. and Scheuren, F. J. (1987), "Modified Raking Ratio Estimation," *Survey Methodology*, Statistics Canada, Vol. 13, No. 2, pp. 209-219.