Section 3 Description of the Sample and Limitations of the Data

This section describes the 2000 Corporate sample design, including the methods used in the selection of returns, data capture, data cleaning, and data completion. The techniques used to produce estimates and an assessment of the data limitations, including measures of sampling variability, 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 size of the samples 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 2000, the sample proportion was about 2.7 percent of the total population of over 5.4 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 a measure of income, the definition of which depends on the return's form type [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 1997 through 2000. These population counts will differ from all the estimated population counts in this publication because they include out-of-scope returns.

Bertrand Überall, Richard Collins, and Kim Saint André were responsible for the sample design and estimation of the SOI 2000 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 1997-2000

	Tax Year			
Form Type	1997	1998	1999	2000
1120	2,219,131	2,190,409	2,165,338	2,146,170
1120-A	272,858	259,696	244,339	235,459
1120S	2,574,150	2,716,507	2,866,963	3,008,022
1120-L	1,613	1,572	1,522	1,465
1120-PC	3,228	3,352	3,437	3,593
1120-RIC	9,420	10,044	10,449	11,157
1120-REIT	674	969	1,079	1,114
1120-F	21,780	22,157	22,270	22,385
Total	5,102,854	5,204,706	5,315,397	5,429,365

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, as the measure of income, size of ordinary income.

The design process began with projected population totals derived from those used to estimate IRS administrative workloads and are adjusted based on previous years' population distributions. Using projected population totals by sampling strata, an optimal allocation, based on variance and cost estimates, was carried out to assign sampling rates such that the overall targeted sample size is approximately 136,000. A Bernoulli sample is selected independently from each stratum with rates ranging from 0.25 percent to 100 percent. Figure D on the following page shows the stratum boundaries, sampling rates, and population and sample counts for each form type. This table also shows the adjusted population and sample counts after reclassification of returns that were mis-stratified due to errors in the stratifying variables. (See sub-section on Processing Errors, page 14, for further information on the handling of mis-stratified 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			**5,429,365**	**145,517**	**5,429,473**	**144,917**
	Form 1120 w/ Form 5735 attached, Total			**294**	**294**	**294**	**289**
1	Under $100,000,000		100.00	228	228	228	224
2	$100,000,000 - $250,000,000		100.00	32	32	32	32
3	$250,000,000 or more		100.00	34	34	34	33
	Form 1120 (no Form 5735 attached), 1120-A, Total			**2,381,335**	**83,440**	**2,381,417**	**83,083**
4	Under $50,000	Under $25,000	0.40	873,388	3,557	865,841	3,556
5	$50,000 - $100,000	$25,000 - $50,000	0.40	316,820	1,303	319,445	1,343
6	$100,000 - $250,000	$50,000 - $100,000	0.50	415,284	2,040	418,927	2,288
7	$250,000 - $500,000	$100,000 - $250,000	1.00	281,485	2,914	284,179	3,074
8	$500,000 - $1,000,000	$250,000 - $500,000	1.60	195,530	3,235	196,465	3,351
9	$1,000,000 - $2,500,000	$500,000 - $1,000,000	4.00	149,393	5,934	149,048	6,348
10	$2,500,000 - $5,000,000	$1,000,000 - $1,500,000	5.60	58,356	3,349	58,055	3,492
11	$5,000,000 - $10,000,000	$1,500,000 - $2,500,000	10.00	33,368	3,397	32,919	3,429
12	$10,000,000 - $25,000,000	$2,500,000 - $5,000,000	100.00	23,742	23,742	23,184	23,061
13	$25,000,000 - $50,000,000	$5,000,000 - $10,000,000	100.00	11,551	11,551	11,281	11,198
14	$50,000,000 - $100,000,000	$10,000,000 - $15,000,000	100.00	7,231	7,231	7,100	7,067
15	$100,000,000 - $250,000,000	$15,000,000 or more	100.00	7,942	7,942	7,858	7,811
16	$250,000,000 - $500,000,000		100.00	2,786	2,786	2,751	2,729
17	$500,000,000 or more		100.00	4,459	4,459	4,364	4,336
	Form 1120S, Total			**3,008,022**	**45,415**	**3,008,024**	**45,224**
18	Under $50,000	Under $25,000	0.25	1,180,440	2,940	1,177,227	3,031
19	$50,000 - $100,000	$25,000 - $50,000	0.25	482,572	1,152	484,538	1,257
20	$100,000 - $250,000	$50,000 - $100,000	0.26	538,857	1,401	540,188	1,607
21	$250,000 - $500,000	$100,000 - $250,000	0.41	354,332	1,543	355,401	1,675
22	$500,000 - $1,000,000	$250,000 - $500,000	0.80	196,277	1,597	196,198	1,648
23	$1,000,000 - $2,500,000	$500,000 - $1,000,000	2.20	139,406	3,059	139,676	3,248
24	$2,500,000 - $5,000,000	$1,000,000 - $1,500,000	3.50	54,785	1,973	54,606	2,052
25	$5,000,000 - $10,000,000	$1,500,000 - $2,500,000	7.00	31,894	2,291	31,620	2,289
26	$10,000,000 - $25,000,000	$2,500,000 - $5,000,000	100.00	19,077	19,077	18,761	18,674
27	$25,000,000 - $50,000,000	$5,000,000 - $10,000,000	100.00	5,874	5,874	5,708	5,674
28	$50,000,000 - $100,000,000	$10,000,000 - $15,000,000	100.00	2,285	2,285	2,221	2,202
29	$100,000,000 - $250,000,000	$15,000,000 or more	100.00	1,605	1,605	1,505	1,494
30	$250,000,000 or more		100.00	618	618	375	373
	Form 1120-L, Total			**1,465**	**883**	**1,477**	**878**
31	Under $10,000,000		43.00	1,016	434	1,006	414
32	$10,000,000 - $50,000,000		100.00	160	160	163	159
33	$50,000,000 - $250,000,000		100.00	97	97	97	96
34	$250,000,000 or more		100.00	192	192	211	209
	Form 1120-F, Total			**22,385**	**3,648**	**22,387**	**3,642**
35	Under $10,000,000		13.00	21,513	2,776	21,530	2,791
36	$10,000,000 - $50,000,000		100.00	482	482	472	468
37	$50,000,000 - $250,000,000		100.00	185	185	181	180
38	$250,000,000 or more		100.00	205	205	204	203
	Form 1120-PC, Total			**3,593**	**1,187**	**3,601**	**1,170**
39	Under $2,500,000		10.00	2,111	227	2,054	210
40	$2,500,000 - $10,000,000		25.00	713	191	764	196
41	$10,000,000 - $50,000,000		100.00	444	444	448	436
42	$50,00,000 - $250,000,000		100.00	210	210	213	211
43	$250,000,000 or more		100.00	115	115	122	117
	Form 1120-REIT, Total			**1,114**	**994**	**1,114**	**986**
44	Under $10,000,000		50.00	257	137	256	132
45	$10,000,000 - $50,000,000		100.00	183	183	184	184
46	$50,000,000 - $250,000,000		100.00	312	312	314	312
47	$250,000,000 or more		100.00	362	362	360	358
	Form 1120-RIC, Total			**11,157**	**9,656**	**11,159**	**9,645**
48	Under $10,000,000		15.00	1,762	261	1,756	252
49	$10,000,000 - $50,000,000		100.00	2,361	2,361	2,361	2,358
50	$50,000,000 - $100,000,000		100.00	1,436	1,436	1,434	1,433
51	$100,000,000 - $250,000,000		100.00	1,978	1,978	1,986	1,985
52	$250,000,000 - $500,000,000		100.00	1,300	1,300	1,301	1,300
53	$500,000,000 or more		100.00	2,320	2,320	2,321	2,317

* Proceeds is defined as the larger of absolute value of net income (deficit) or absolute value of cash flow (net income + depreciation + depletion).
** These adjustments include restratification (see subsection on Processing Errors, page 14).
*** Includes added returns not posted to the BMF during the two-year IRS processing period.
**** Does not include missing returns, but does include added returns not posted to the BMF during the two-year IRS processing period.
Note: 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 total realized sample for Tax Year 2000, including inactive corporations and rejected returns, is 144,917 returns.

Sample Selection

Corporation income tax returns are filed at the Cincinnati, Ogden, and Philadelphia Submission Processing Centers. (Prior to January 2002, for part of the 2000 Corporate Program, returns were also filed at seven additional centers located throughout the country). All corporate returns are processed initially to determine tax liability and are then made available for other programs including SOI. All tax data are transmitted and updated on a weekly basis to the IRS Business Master File (BMF) system located in Martinsburg, West Virginia. This system serves as the point of selection for the sample, which was also selected on a weekly basis.

Sample selections for Tax Year 2000 occurred over the period of July 2000 through June 2002. A 24-month sampling period is needed for two reasons. First, approximately 17.9 percent of all corporations have noncalendar year accounting periods. In order to take these filings into consideration, the 2000 statistics represent all corporations filing returns with accounting periods ending during the period from July 2000 to June 2001. Also, many corporations, including some of the largest, request 6-month filing extensions. The combination of noncalendar year filing and filing extensions means that the last returns due to be received by IRS for the Tax Year 2000 (those with accounting periods ending in June 2001, which must therefore be filed by October 2001) could be timely filed as late as March 2002, if the 6-month extension of the October 2001 due date is taken into account. Normal administrative processing time lags required that the sampling process remain open for the 2000 study until June 30, 2002. However, a few very large returns for Tax Year 2000 were added to the sample as late as November 2002.

Each corporation is assigned a permanent and unique Employer Identification Number (EIN). The EIN is used as the basis for random selection. A pseudo-random number (PRN) is generated using the EIN as the seed. The last four digits of the PRN, called the transformed taxpayer identification number (TTIN), are compared to the sampling rates; a corporation for which the value of its TTIN is below 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 is the ratio of the second year to the first year selection probabilities. If the corporation files with a new EIN, the probability of being selected will be independent of the prior year selection. Due to the fact that corporations typically maintain the same EINs, this use of the EIN as the basis for sample selection results in many of the same corporations selected into the sample from year to year. This also results in a reduction of the sample variance for estimates of year-to-year change [2].

Data Capture

Data processing for SOI begins with information already extracted for administrative purposes; over 100 items are available from the BMF system. Some 1,300 additional items are extracted from the tax returns during SOI processing. The administrative data are checked and corrected as necessary. 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 as long as 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 2000, these instructions consisted of more than 800 pages covering standard and straightforward procedures and instructions for exceptions and nonstandard situations that might be encountered.

Data Cleaning

Statistical processing of the corporate returns was performed in an online computer environment and the data from returns were entered directly into the 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 $10,000 in total assets, but was selected as though it had $1 million.

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 1999 return (if the corporation filed a return in 1999) or the 1998 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 2000 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 1999 return, or the 1998 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]. Figure E below shows the number of sampled returns that had balance sheet items imputed for Tax Years 1997 through 2000. This figure also shows the number and percentage of unavailable returns, as discussed in the subsection on Nonresponse Errors, page 14, as well as the number of mis-stratified returns, as discussed in the subsection on Processing Errors, also on page 14.

Figure E.--Number of Returns with Imputed Balance Sheets, Number of Unavailable Returns, and Number of Mis-stratified Returns for Tax Years 1997-2000

Tax Year	1997	1998	1999	2000
Imputed returns	61	70	68	38
Unavailable returns	38	154	228	412
% of returns unavailable	0.04	0.11	0.16	0.28
Mis-stratified returns	1,495	3,059	2,482	2,790

For Tax Year 2000, none of the 38 imputed returns have 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 the minor industrial group in which they are classified, and corporations for which total assets are over a specified limit, which is dependent on the form type or the minor industry. For critical corporations selected for the sample but unavailable for statistical processing, taxpayer-surveyed data are used. For the critical corporations not selected for the sample, if the current tax return is not found in any of the IRS service 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 33 prior year returns in the Tax Year 2000 data.

Another part of the data cleaning process is identifying sampled returns that are not used in the tabulation. The BMF system used for sample selection can include duplicate tax returns and other out-of-scope returns, such as returns for nonprofit

corporations and prior-year tax returns. These include the following types of inactive returns:

- Returns having neither current income nor deductions;
- Duplicate returns;
- Amended returns not associated with the original returns, as well as tentative returns not associated with the revised returns;
- Corporations exempt under Section 931 of the IRC;
- Corporations exempt under Section 1247 of the IRC;
- Corporations exempt under Section 883 of the IRC;
- "Cost corporation" returns exempt under Revenue Ruling 52-542;
- Corporations exempt under Code section 501(c)(15);
- Nonresident foreign corporations having no income effectively connected with a trade or business within the United States;
- U.S. Virgin Island corporations exempt under Code section 934;
- Political organizations filing under Code section 527;
- General stock ownership corporations exempt from tax;
- Homeowners' associations exempt under Code section 528;
- Information returns reporting no tax because of tax treaty or convention according to Code section 894;
- Most prior-year returns with total assets under $250 million filed on tax forms for years prior to 1999 and with accounting periods ending before July 2000;
- Returns filed on a form type which is not included in the SOI sample;
- Fraudulent returns;
- Returns of businesses incorporated in a tax-exempt U.S. Possession.

Figure F below displays the number of inactive sampled returns that were excluded from tabulations and the percentages they represent of the total sample sizes in Tax Years 1997 through 2000.

Figure F.--Number of Inactive Sampled Returns for Tax Years 1997-2000

	Tax Year			
Type of inactive return	**1997**	**1998**	**1999**	**2000**
No Income or Deductions	1,321	1,460	1,450	1,615
Duplicate*	665	799	770	1,044
Other**	2,654	3,645	3,725	3,684
Total	4,640	5,904	5,945	6,343
% of Sample	4.72	4.29	4.22	4.38

* 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 1997 through 2000 are provided in Figure G below.

Figure G.--Estimated Number of Active Returns for Tax Years 1997-2000

	Tax Year			
Form Type	**1997**	**1998**	**1999**	**2000**
1120	2,009,866	2,021,929	1,990,782	1,970,777
1120-A	221,940	211,801	191,769	186,177
1120S	2,452,254	2,588,088	2,725,775	2,860,478
1120-L	1,685	1,620	1,551	1,520
1120-PC	3,595	3,624	3,739	3,732
1120-RIC	9,098	9,897	10,318	10,991
1120-REIT	666	932	1,071	1,099
1120-F*	10,977	10,996	10,898	10,498
Total	4,710,083	4,848,888	4,935,904	5,045,274

* 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 results. 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. These 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.

The two-step process was used to improve the industry estimates. The first stage is the one-step process described above and provides an initial

weight for the return. The second stage involves post-stratification by industry. During post-stratification, certain cells have small sample sizes. To handle this problem, a raking ratio estimation approach is applied during post-stratification in order to determine the final weights for Form 1120, 1120-A, and 1120S records that are not self-representing [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 frequencies and money amounts that are published in this report for these Forms.

Beginning with Tax Year 1998, the industry classification used for post-stratification weighting is a two-digit code based on the North American Industrial Classification System (NAICS). This new classification system replaces the Standard Industrial Classification (SIC) system used in years prior to 1998.

Data Limitations and Measures of Variability

Several extensive quality review processes were used to improve the quality of the data. The review processes began at the sample selection stage with weekly monitoring of the sample to ensure that the proper number of returns was being selected. They continued through the data collection, data cleaning, and data completion procedures with consistency testing. Part of the review process included extensive comparisons between the 2000 data and the 1999 data. A great amount of effort was 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 figures 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. Sometimes, for convenience, the standard error is expressed as a percentage of the value being estimated. This is called the coefficient of variation (CV) of the estimate, and it can be used to assess the reliability of an estimate.

The coefficient of variation of an estimate is calculated by dividing the standard error by the estimate. 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, coefficients of variation are given in Figure H on the following page. The corresponding estimates can be found in Table 4.

The coefficient of variation, CV(X), can be used to construct confidence intervals of the estimate X. The 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 coefficient of variation can be found in Table 1 and used to calculate the standard error:

$$SE(X) = X \bullet CV(X)$$
$$= 168{,}580 \times 2.86/100$$
$$= 4{,}821$$

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

$$X \pm 2 \bullet SE(X) = 168{,}580 \pm (2 \times 4{,}821)$$
$$= 168{,}580 \pm 9{,}642$$

Thus, the interval estimate is 158,938 returns to 178,222 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 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.

Figure H.--Coefficients of Variation (CVs) for Number of Returns, by Asset Size and Sector, for Tax Year 2000

Sector	All asset sizes	Size of total assets: Zero assets	$1 under $100,000	$100,000 under $250,000	$250,000 under $500,000	$500,000 under $1,000,000
	(1)	(2)	(3)	(4)	(5)	(6)
All industries[1]	**0.19**	**2.88**	**0.55**	**1.04**	**0.99**	**0.68**
Agriculture, forestry, fishing, and hunting	2.61	20.58	6.35	6.72	4.84	3.65
Mining	7.17	33.12	12.89	27.16	18.53	14.87
Utilities	15.66	85.52	27.87	43.21	40.18	37.76
Construction	0.99	9.24	2.00	3.52	3.35	2.52
Manufacturing	2.06	14.52	5.20	6.15	4.80	3.69
Wholesale and retail trade	0.91	7.57	2.07	2.37	2.37	1.80
Transportation and warehousing	2.58	17.40	4.24	8.29	7.75	6.04
Information	3.92	16.98	6.13	13.16	12.37	10.50
Finance and insurance	2.38	12.72	4.26	8.43	7.47	6.27
Real estate and rental and leasing	1.17	8.19	2.66	3.49	2.91	2.07
Professional, scientific, and technical services	1.15	7.46	1.69	4.65	5.30	5.27
Management of companies (holding companies)	5.87	18.67	12.91	17.74	17.80	14.73
Administrative and support and waste management and remediation services	2.89	13.82	4.00	8.06	10.26	10.10
Educational services	7.48	31.67	9.29	27.55	29.61	25.09
Health care and social assistance	1.40	12.67	2.32	4.50	7.11	8.85
Arts, entertainment, and recreation	3.99	20.19	5.70	13.37	15.11	11.69
Accommodation and food services	1.57	12.50	3.05	4.64	5.47	5.03
Other services	2.03	11.74	3.05	5.39	5.84	5.76

Sector	Size of total assets—continued: $1,000,000 under $5,000,000	$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
	(7)	(8)	(9)	(10)	(11)	(12)
All Industries[1]	**0.33**	**0.57**	**0.04**	**0.05**	**0.04**	**0.05**
Agriculture, forestry, fishing, and hunting	2.14	5.06	0.34	0.65	0.77	1.50
Mining	6.21	9.18	0.38	0.55	0.62	0.72
Utilities	18.05	22.14	0.85	1.20	1.44	1.10
Construction	1.14	2.00	0.13	0.27	0.27	0.62
Manufacturing	1.45	1.85	0.09	0.14	0.16	0.21
Wholesale and retail trade	0.82	1.22	0.07	0.15	0.19	0.30
Transportation and warehousing	3.28	6.93	0.26	0.50	0.62	0.61
Information	7.36	4.99	0.21	0.32	0.39	0.43
Finance and insurance	2.82	3.72	0.15	0.15	0.10	0.09
Real estate and rental and leasing	1.32	2.60	0.45	0.27	0.36	0.60
Professional, scientific, and technical services	2.80	4.42	0.18	0.30	0.38	0.49
Management of companies (holding companies)	6.32	8.80	0.24	0.25	0.18	0.18
Administrative and support and waste management and remediation services	5.70	9.93	0.37	0.59	0.83	0.78
Educational services	15.30	24.66	0.97	1.57	1.93	2.55
Health care and social assistance	5.21	9.00	0.36	0.61	0.71	0.94
Arts, entertainment, and recreation	5.48	9.78	0.43	0.73	0.79	1.35
Accommodation and food services	2.18	5.05	0.31	0.55	0.67	0.67
Other services	4.84	12.44	0.48	0.93	0.99	1.64

[1]Includes returns not allocable by sector.

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

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 above, 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 2000, there were 412 unavailable returns in the corporation study, which constituted about 0.28 percent of the total sample size. The number of unavailable returns and their percentages of total sample sizes for Tax Years 1997 through 2000 are shown in Figure E, page 10.

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 $10,000.23 and net income of $5,000.00. A processing error causes the cents to be keyed in as dollars so that the return is classified according to total assets of $1,000,023 and net income of $5,000. The return would be mis-stratified according to the incorrect value of the total assets stratifier. The number of mis-stratified returns for Tax Years 1997 through 2000 is given in Figure E, page 10.

Mis-stratified returns in the sample were reclassified into their proper sampling classes after complete data capture. The population of returns that needed to be reclassified was estimated from the sample and the stratum population sizes were adjusted accordingly [5]. Population and sample totals were minimally affected by reclassification, and an analysis of the sample results tended to confirm that mis-stratified returns occurred randomly. Steps are being taken by staff in both the Submission Processing Centers and the SOI Division to minimize the number of mis-stratified returns.

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.

[5] Mulrow, J. M. and Jones, H. W. (1989), "Sampling Administrative Records: Detection and Correction of Stratification Error," *Statistics of Income and Related Administrative Record Research: 1988-1989*, Internal Revenue, December 1990, pp. 139-144.