Section 3

Description of the Sample and Limitations of the Data

This section describes the sample design, sample selection, data capture, data cleaning, and data completion processes for the Statistics of Income (SOI) 2011 Corporation statistics program. It also presents the techniques used to produce estimates of the total number of corporations and associated variables as well as an assessment of the data limitations, including sampling and nonsampling errors.

Background

From Tax Years (TY) 1916 through 1950, the Statistics of Income (SOI) program for Corporations extracted data from each corporate return filed. Beginning with TY 1951, however, SOI introduced stratified probability sampling. Since that time, the sample size has generally decreased while the corporate tax return population has increased. For example, for 1951, the sample accounted for 41.5 percent of the entire population, or 285,000 of the 687,000 total returns filed. For 2011, the sample represented about 1.76 percent of the total population of just over 6.26 million returns. This population count differs from the estimated population count cited elsewhere in this publication because the sampling frame includes out-of-scope and duplicate returns.

For 1951, SOI stratified the sample by size of total assets and industry. However, from 1952 through 1967, SOI stratified the sample by a measure of size only. The size was measured by either business volume (1953–1958) or total assets (1952 and 1959–1967). Since 1968, SOI has stratified returns by both total assets and, for Forms 1120 and 1120S, 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.

Bertrand Überall, Richard Collins, and Elliot Mountjoy were responsible for the sample design and estimation of the SOI 2011 Corporation Program under the direction of Tamara Rib, Chief, Mathematical Statistics Section, Corporation Statistics Branch.

Survey Population

The survey population includes corporate tax returns filed with one of the 1120 forms selected for the study and posted to the IRS Business Master File (BMF). Excluded are amended returns and returns for which the tax liabilities changed because of a tax audit. Figure E gives the number of corporate returns by form type that were subject to sampling during Tax Years 2008 through 2011.

Figure E. Population Counts by Corporate Form Type, Tax Years 2008–2011

Form type	Tax year			
	2008	2009	2010	2011
1120	2,001,930	1,927,971	1,867,941	1,835,482
1120S	4,293,544	4,332,077	4,336,365	4,367,077
1120-L	891	825	748	700
1120-PC	7,828	8,104	8,572	9,237
1120-RIC	13,221	13,106	13,385	14,193
1120-REIT	1,679	1,672	1,798	1,928
1120-F	30,620	30,295	32,414	35,149
Total	**6,349,713**	**6,314,050**	**6,261,223**	**6,263,766**

Sample Design

The current design is a probability sample stratified by form type, and either by size of total assets alone or both size of total assets and a measure of income. Form 1120 returns are stratified by size of total assets and size of "proceeds," which is the measure of income for this form. Size of proceeds is defined as 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. Form 1120S is stratified by size of total assets and size of ordinary income. All other 1120 forms (1120-F, 1120-L, 1120-PC, 1120-RIC, and 1120-REIT) are stratified by size of total assets only.

SOI began the design process with projected population totals derived from IRS administrative workload estimates, adjusted according to the distribution by population strata from several previous survey years. Using projected population totals by sample strata, SOI carried out an optimal allocation based on stratum standard errors to assign sample sizes to each stratum such that the overall targeted sample size was

approximately 110,000 returns. Mathematical statisticians selected a Bernoulli sample independently from each stratum, with sampling rates ranging from 0.25 percent to 100 percent. Figure F shows the stratum boundaries, sampling rates, frame population, and sample counts from the BMF for each type of 1120. This table also shows the population and sample counts after adjustments for missing returns, outliers, and weight trimming. The total realized sample for 2011, including inactive and noneligible corporations, is 110,447 returns.

Sample Selection

The IRS Cincinnati and Ogden Submission Processing Centers initially process all corporate returns to determine tax liability before transmitting the data weekly to the IRS Business Master File (BMF). These returns are said to "post" to the BMF, which serves as the SOI sampling frame. SOI also selects the sample on a weekly basis.

Sample selection for TY 2011 occurred over the 24-month period, July 2011 through June 2013. SOI requires a 24-month sampling period for two reasons. First, approximately 10 percent of all corporations use noncalendar-year accounting periods. To capture these returns, the 2011 statistics include all corporations filing returns with accounting periods ending between July 2011 and June 2012. Second, many corporations, including some of the largest corporations, request 6-month filing extensions. This combination of noncalendar-year accounting periods and filing extensions means that the last TY 2011 returns the IRS received had accounting periods ending in June 2012, and therefore, had to be filed by October 2012. However, taking into account the 6-month extension, these returns could have been filed as late as March 2013 and still be considered timely. To account for the normal processing time, the sample selection process remained open for the 2011 study until the end of June 2013. However, SOI added a few very large returns for TY 2011 to the sample as late as August 2013.

Each tax return in the survey population is assigned to a stratum and subject 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. Therefore, corporations selected for the sample in any given year may be selected the following year, providing the corporation files a return using the same EIN and 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 selection will be independent from the prior-year selection [2].

Data Capture

Data processing for SOI begins with information already extracted for IRS administrative purposes; over 100 items available from the BMF system are checked and corrected as necessary. SOI extracts some 1,630 additional data items from the corporate tax returns during processing. This data-capture process can take as little as 15 minutes for a small, single-entity corporation filing Form 1120, or up to several weeks for a large, consolidated corporation filing several hundred attachments and schedules with the return. The process is further complicated by several factors:

- Over 1,630 separate data items may be extracted from any given tax return. This often requires constructing totals from various other items elsewhere on 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 forms.
- There is no legal requirement for a corporation to meet its tax return filing requirements by filling in, line by line, the entire U.S. tax return form. Therefore, many corporate taxpayers report financial details using schedules of their own design or using commercial tax-preparation software packages.
- There is no single accepted method of corporate tax accounting in the United States, but rather, several accepted "guidelines," which can vary by geographic location. 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. SOI staff also attempt to standardize these differences.

To help staff overcome these complexities and differences in taxpayer reporting, for each tax year, SOI prepares detailed instructions for the editing units at the IRS Submission Processing Centers. For TY 2011, these instructions consisted of almost 1,000 pages, covering standard and straightforward procedures and instructions for addressing data exceptions.

Data Cleaning

SOI staff enter data directly into the database from the corporate tax returns selected for the sample. In this context, the term "editing" refers to the combined interactive processes of data extraction, consistency testing, and error resolution. SOI runs over 860 tests to check for inconsistencies, such as:

- Impossible conditions, such as incorrect tax data for a particular form type;
- Internal inconsistencies, such as items not adding to totals;

Figure F. 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			
	Size of total assets	Size of proceeds*		BMF counts		After adjustments**	
				Population	Sample	Population	Sample
	All Returns, Total			**6,263,766**	**110,447**	**6,263,768**	**110,130**
	Form 1120, Total ***			**1,830,099**	**50,567**	**1,830,099**	**50,442**
1	Under $50,000	Under $25,000	0.40	782,102	3,149	782,102	3,096
2	$50,000–$100,000	$25,000–$50,000	0.40	198,633	807	198,633	806
3	$100,000–$250,000	$50,000–$100,000	0.40	260,854	1,044	260,854	1,043
4	$250,000–$500,000	$100,000–$250,000	1.09	195,566	2,106	195,566	2,105
5	$500,000–$1,000,000	$250,000–$500,000	1.81	145,753	2,651	145,753	2,646
6	$1,000,000–$2,500,000	$500,000–$1,000,000	3.48	117,973	4,098	117,973	4,090
7	$2,500,000–$5,000,000	$1,000,000–$1,500,000	5.94	48,229	2,838	48,229	2,835
8	$5,000,000–$10,000,000	$1,500,000–$2,500,000	10.55	29,484	3,163	29,484	3,161
9	$10,000,000–$25,000,000	$2,500,000–$5,000,000	27.00	21,457	5,784	21,457	5,777
10	$25,000,000–$50,000,000	$5,000,000–$10,000,000	50.00	10,324	5,203	10,324	5,195
11	$50,000,000–$100,000,000	$10,000,000–$15,000,000	100.00	6,216	6,216	6,215	6,205
12	$100,000,000–$250,000,000	$15,000,000 or more	100.00	6,852	6,852	6,851	6,838
13	$250,000,000–$500,000,000		100.00	2,798	2,798	2,799	2,797
14	$500,000,000 or more		100.00	3,858	3,858	3,859	3,848
	Form 1120S, Total ***			**4,365,896**	**33,655**	**4,365,896**	**33,495**
15	Under $50,000	Under $25,000	0.25	1,705,448	4,229	1,705,448	4,196
16	$50,000–$100,000	$25,000–$50,000	0.25	649,766	1,609	649,766	1,598
17	$100,000–$250,000	$50,000–$100,000	0.25	750,809	1,920	750,809	1,902
18	$250,000–$500,000	$100,000–$250,000	0.31	545,142	1,674	545,142	1,662
19	$500,000–$1,000,000	$250,000–$500,000	0.56	312,180	1,744	312,180	1,737
20	$1,000,000–$2,500,000	$500,000–$1,000,000	0.99	218,344	2,140	218,344	2,128
21	$2,500,000–$5,000,000	$1,000,000–$1,500,000	1.56	84,256	1,278	84,256	1,276
22	$5,000,000–$10,000,000	$1,500,000–$2,500,000	2.52	50,174	1,233	50,174	1,227
23	$10,000,000–$25,000,000	$2,500,000–$5,000,000	20.00	31,062	6,170	31,062	6,140
24	$25,000,000–$50,000,000	$5,000,000–$10,000,000	30.00	10,040	2,983	10,040	2,972
25	$50,000,000–$100,000,000	$10,000,000–$15,000,000	100.00	4,274	4,274	4,274	4,262
26	$100,000,000–$250,000,000	$15,000,000 or more	100.00	3,202	3,202	3,202	3,197
27	$250,000,000 or more		100.00	1,199	1,199	1,199	1,198
	Form 1120-L, Total			**545**	**312**	**545**	**310**
28	Under $10,000,000		43.00	370	137	370	136
29	$10,000,000–$50,000,000		100.00	87	87	87	87
30	$50,000,000–$250,000,000		100.00	43	43	42	41
31	$250,000,000 or more		100.00	45	45	46	46
	Form 1120-F, Total			**35,044**	**5,396**	**35,046**	**5,381**
32	Under $10,000,000		13.00	32,946	4,256	32,947	4,245
33	$10,000,000–$50,000,000		13.00	1,100	142	1,100	140
34	$50,000,000–$250,000,000		100.00	566	566	565	564
35	$250,000,000 or more		100.00	432	432	434	432
	Form 1120-PC, Total			**8,811**	**1,897**	**8,811**	**1,896**
36	Under $2,500,000		10.00	6,311	605	6,311	604
37	$2,500,000–$10,000,000		25.00	1,599	391	1,599	391
38	$10,000,000–$50,000,000		100.00	720	720	720	720
39	$50,000,000–$250,000,000		100.00	175	175	175	175
40	$250,000,000 or more		100.00	6	6	6	6
	Form 1120-REIT, Total			**1,911**	**1,590**	**1,911**	**1,590**
41	Under $10,000,000		25.00	428	107	424	103
42	$10,000,000–$50,000,000		100.00	420	420	420	420
43	$50,000,000–$250,000,000		100.00	518	518	517	517
44	$250,000,000 or more		100.00	545	545	550	550
	Form 1120-RIC, Total			**14,181**	**9,751**	**14,181**	**9,749**
45	Under $10,000,000		15.00	2,906	438	2,903	435
46	$10,000,000–$50,000,000		30.00	2,746	784	2,745	783
47	$50,000,000–$100,000,000		100.00	1,275	1,275	1,274	1,274
48	$100,000,000–$250,000,000		100.00	2,000	2,000	1,997	1,997
49	$250,000,000–$500,000,000		100.00	1,539	1,539	1,540	1,540
50	$500,000,000 or more		100.00	3,715	3,715	3,722	3,720
51	**Special Studies (All Form Types)**		**100.00**	**7,279**	**7,279**	**7,279**	**7,267†**

* 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, undercoverage, 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 5 (based on total assets), rather than in sample class 3 (based on proceeds).
†The adjusted sample count is lower than the adjusted population count due to returns unavailable for processing.

- 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, SOI addresses missing data items and identifies returns to be excluded from the tabulations. The data completion process focuses on these issues.

SOI uses a ratio-based imputation procedure to estimate missing balance sheet items for certain returns included in the sample. It uses the most recent data available to determine the imputation ratios. These data are either: 1) the corporation's TY 2010 return, if it filed a return for the previous year and the balance sheet was not already imputed for that year, or 2) the TY 2009 aggregate data for the corporation's minor industrial group, which were the most recent aggregate data available when editing for TY 2011 began.

SOI imputes the missing items when the balance sheet items do not balance (i.e., the sum of asset items does not equal the sum of liability and shareholders' equity items). If the amount of total assets is among the missing items, then it is imputed first based on the ratio of total assets to business receipts (or total receipts) from either the corporation's TY 2010 return, or the TY 2009 aggregate data for the corporation's minor industry. Then, SOI imputes the additional missing items based on ratios so that both the total of all asset items and the total of all liability items are equal to the total assets amount. Reference [3] provides a description of the balance sheet imputation process.

Figure G shows the number of sampled returns that had balance sheet items imputed, as well as the percentages of the total sample sizes they represent for Tax Years 2008 through 2011. For TY 2011, the total assets from returns having imputed total assets represent only a negligible fraction of the total estimated assets for all active returns in the sample.

Figure G. Number of Imputed Returns for Tax Years 2008–2011

Returns with imputations	Tax year			
	2008	2009	2010	2011
Number of imputed returns	52	63	42	47
Percent imputed	0.05	0.06	0.04	0.04

SOI uses various methods to impute data for some certainty returns unavailable for editing, depending on the information available at the time the return needs to be completed for the sample. These corporations are identified from the previous year's sample using a combination of assets and receipts. Additional corporations may be identified to ensure industry coverage. SOI uses data filed electronically for those corporate returns selected for the sample, but unavailable for statistical processing. For TY 2011, there were 43 returns that met these criteria. For some returns not selected for the sample, if the current tax return was not located and no other current tax data were available, then SOI used data from the previous year's return, with adjustments for tax law changes, if needed. There is only a negligible number of returns derived from prior-year returns in the Tax Year 2011 data.

The data cleaning process also includes identifying returns not eligible for the sample as the BMF may have duplicate and other out-of-scope returns. These returns include those filed by nonprofit corporations, returns having neither current income nor deductions, and prior-year tax returns. Additionally, amended or tentative returns, nonresident foreign corporations having no effectively connected income with a trade or business located within the United States, fraudulent returns, and returns filed by tax-exempt corporations are not eligible for the sample. Figure H displays the number of inactive sampled returns excluded from the tabulations, as well as the percentages of the total sample size they represent for 2008 through 2011.

Figure H. Number of Inactive Sampled Returns for Tax Years 2008–2011

Type of inactive return	Tax year			
	2008	2009	2010	2011
No Income or deductions	1,480	1,360	1,608	1,959
Other*	5,367	5,145	4,686	4,236
Total	6,847	6,505	6,294	6,195
Percent of sample	6.09	5.95	5.80	5.60

*Includes duplicate returns (returns that appear more than once in the sample) and prior-year returns.

Figure I provides estimates of the number of active corporations by form type for 2008 through 2011. For Forms 1120-L and 1120-PC, these estimates may differ from the population counts in Figure E due to changes made during the data capture and data cleaning processes.

Figure I. Estimated Number of Active Returns for Tax Years 2008–2011

Form type	Tax year			
	2008	2009	2010	2011
1120	1,762,483	1,694,869	1,649,285	1,624,888
1120S	4,049,943	4,094,562	4,127,554	4,158,572
1120-L	945	866	796	752
1120-PC	7,670	7,890	8,244	8,822
1120-RIC	13,140	13,043	13,256	14,120
1120-REIT	1,660	1,635	1,766	1,894
1120-F*	11,379	11,680	12,824	14,077
Total	**5,847,221**	**5,824,545**	**5,813,725**	**5,823,126**

*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

SOI bases the estimates of the total number of corporations and associated variables produced in this report on weighted sample data using either a one-step or two-step process, depending on the form type filed. Under the one-step process, SOI assigns a weight for the return, which is the reciprocal of the realized sampling rate, adjusted for unavailable returns, outliers, weight trimming, and any other necessary adjustments. SOI uses these weights, referred to as the "national weights," to produce the estimates published in this report for Forms 1120-F, 1120-L, 1120-PC, 1120-RIC, and 1120-REIT, as well as Form 1120 and 1120S returns that were sampled with certainty.

The two-step process is used to improve the estimates by industry for returns filed on either Form 1120 or 1120S that are not selected in self-representing strata. The first stage of the two-step process is to assign an initial weight for the return as described above. The second stage involves post-stratification by industry and sample selection class. SOI uses a bounded raking ratio estimation approach to determine the final weights because certain post-stratification cells may have small sample sizes [4]. 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

SOI uses several extensive quality review processes to improve data quality. This starts at the sample selection stage with weekly monitoring to ensure the proper number of returns is selected, especially in the certainty strata. These processes continue through the data collection, data cleaning, and data completion procedures with consistency testing. Part of the review process includes extensive comparisons between the sample year (2011) and prior-year (2010) data. SOI designed each processing stage to ensure data integrity.

Sampling Error

Since the TY 2011 estimates are based on a sample, they may differ from population aggregates resulting from a complete census of all corporate income tax returns. The TY 2011 sample is one of many possible samples that could have been selected under the same sample design. Estimates derived from one possible sample could differ from those derived from another and also 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), a measure of the average magnitude of the sampling errors over all possible samples, can be estimated from the realized 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 smaller the CV, the more reliable the estimate is deemed to be.

SOI calculates the estimated coefficient of variation of an estimate by dividing the estimated standard error by the estimate itself and taking the absolute value of this ratio. Table 1 shows the estimated coefficients of variation by industrial groupings for the estimated number of returns, as well as selected money amounts. Figure J shows estimated coefficients of variation for the number of returns, by asset size and sector. Table 4 provides the corresponding estimates.

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:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= 146{,}580 \text{ x } 3.46/100 \\ &= 5{,}072 \end{aligned}$$

A 95-percent confidence interval for the estimated number of returns in manufacturing is constructed as follows:

$$\begin{aligned} X \pm 2 \bullet SE(X) &= 146{,}580 \pm (2 \text{ x } 5{,}072) \\ &= 146{,}580 \pm 10{,}144 \end{aligned}$$

The interval estimate is 136,436 returns to 156,724 returns. This means that if all possible samples were selected under the same general conditions and 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 include coverage errors, nonresponse errors, processing errors, or response errors. The inability to obtain information for all sampled returns, differing interpretations of tax concepts or taxpayer instructions, inability to provide accurate information at the time of filing (data are collected before auditing), and inability to obtain all tax schedules and attachments may cause these errors. These errors may also be caused by data recording or coding errors, data collecting or cleaning errors, estimation errors, and failure to represent all population units.

Figure J. Coefficients of Variation (CVs) for Number of Returns, by Asset Size and Sector, for Tax Year 2011

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]	**0.18**	**1.59**	**0.47**	**1.10**	**0.59**
Agriculture, forestry, fishing, and hunting	2.67	12.50	4.44	4.25	3.24
Mining	7.02	21.60	10.43	19.55	9.75
Utilities	17.21	78.87	23.62	35.56	23.32
Construction	1.00	4.69	1.69	3.82	2.18
Manufacturing	2.56	9.80	4.35	6.02	2.87
Wholesale and retail trade	1.01	4.64	1.57	2.55	1.43
Transportation and warehousing	2.46	7.52	4.07	8.82	4.85
Information	4.10	11.16	5.77	14.45	8.54
Finance and insurance	2.31	8.56	3.52	8.05	4.51
Real estate and rental and leasing	1.16	4.90	1.97	2.96	1.83
Professional, scientific, and technical services	1.09	4.03	1.59	5.68	3.69
Management of companies (holding companies)	5.79	13.36	12.30	13.83	8.64
Administrative and support and waste management and remediation services	2.80	7.09	3.71	10.04	7.24
Educational services	7.06	14.25	9.01	31.22	19.08
Health care and social assistance	1.26	6.99	1.73	6.34	6.05
Arts, entertainment, and recreation	3.81	10.71	5.13	14.74	10.87
Accommodation and food services	1.55	8.04	2.22	6.87	4.66
Other services	2.01	6.48	2.68	6.70	5.98

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]	1.00	0.46	0.59	0.05	0.08
Agriculture, forestry, fishing, and hunting	9.36	4.44	5.79	0.86	0.00
Mining	10.45	4.17	4.49	0.43	0.63
Utilities	28.62	12.50	10.61	0.00	0.00
Construction	3.70	1.84	2.53	0.19	0.38
Manufacturing	3.71	1.31	1.52	0.11	0.07
Wholesale and retail trade	2.48	0.94	1.38	0.12	0.10
Transportation and warehousing	10.08	3.95	4.51	0.47	0.58
Information	10.07	3.84	7.27	0.29	0.38
Finance and insurance	5.39	1.80	1.85	0.08	0.20
Real estate and rental and leasing	3.92	1.77	2.12	0.47	0.68
Professional, scientific, and technical services	5.45	2.48	2.88	0.18	0.25
Management of companies (holding companies)	9.33	3.83	2.93	0.16	0.17
Administrative and support and waste management and remediation services	12.09	6.01	6.48	0.57	0.00
Educational services	26.45	12.57	12.73	0.00	0.00
Health care and social assistance	13.14	5.30	5.62	0.64	0.00
Arts, entertainment, and recreation	16.96	7.16	7.39	1.12	0.00
Accommodation and food services	10.70	3.55	5.96	0.82	0.00
Other services	15.70	7.52	9.24	0.00	0.00

[1] Includes returns not allocable by sector.
NOTE: Returns with assets of $250,000,000 or more are self-representing and thus are not subject to sampling error.

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. Since many of the largest corporations receive extensions to their filing periods, they 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 into the industry with the largest receipts. Thus, there is undercoverage of financial data within certain industries and overcoverage in others. Coverage problems within industries present a limitation on any analysis of the sample results.

Nonresponse Errors: There are two types of nonresponse errors: unit and item. 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 2011, there were 278 such unavailable returns in the corporation study, which constituted about 0.25 percent of the total sample. Figure K shows the number of unavailable returns and the percentage of the total sample size for Tax Years 2008 through 2011.

Figure K. Number of Unavailable Returns for Tax Years 2008–2011

Unavailable returns	Tax year			
	2008	2009	2010	2011
Number of unavailable returns	293	141	150	278
Percent unavailable	0.26	0.13	0.14	0.25

Item nonresponse occurs when certain items are unavailable for a return selected for SOI processing, even if the return itself is available. An example of item nonresponse would be items missing from the balance sheet, even though other items have been reported.

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 misstratification error. One example of how a return might be misstratified 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 misstratified according to the incorrect value of the total assets stratifier. To adjust for misstratification errors, only returns selected in a noncertainty stratum which really belonged in a certainty stratum were moved to this certainty 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, SOI does not incorporate adjustments to a return during audit into the 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.