EVALUATION OF THE IRS CORPORATE SOI SAMPLE

Robert P. Clickner, Glenn J. Galfond, and Lawrence A. Thibodeau, Applied Management Sciences, Inc.

This is a progress report on the current effort by the Statistics of Income (SOI) Division of the Internal Revenue Service to redesign the corporate SOI sample. To focus and direct the redesign effort, an evaluation of the current design and method of estimation has been conducted by Applied Management Sciences under contract to the Internal Revenue Service.

This evaluation was carried out in three phases:

(1) Identify the principal users of the data and their data needs.

(2) Empirically assess the quality of the SOI sample data relative to the users' needs.

(3) Construct alternative designs and assess them relative to the current design.

This report concentrates on the third phase and only briefly summarizes Phases 1 and 2 to set the stage for and to motivate Phase 3. Applied Management Sciences has issued separate reports on each of these three phases. [1], [2], [3]

All empirical analyses discussed in this paper were conducted on the 1980 corporate SOI data file. Also, certain special subpopulations of corporate returns were excluded from the scope of the research because of their special characteristics, for example, mutual insurance companies and Domestic International Sales Corporation returns. These exclusions reduced the population size from 2,867,219 to 2,689,245 returns and cut the sample size from 85,593 to 74,383 returns.

PHASE 1

The many users of corporate SOI data have varying needs; some are transient and some ongoing. However, three agencies of the federal government extensively use the SOI data on an ongoing basis and are therefore extremely important to Internal Revenue Service. These users are:

- Office of Tax Analysis, Department of the Treasury
- Joint Committee on Taxation, The Congress.
- Bureau of Economic Analysis, Department of Commerce

Data requirements of the tax policy analysts in the Office of Tax Analysis and the Joint Committee on Taxation are quite complex. These agencies use the SOI data to estimate the revenue impacts of proposed tax law changes. Due to interactions among the different deductions/credits, thresholds, and other complexities, these revenue estimations require data on individual corporate returns. Therefore, the tax analysts need:

- Maximum possible item detail and coverage of subdomains defined by the presence or absence of certain deductions/credits because it is difficult to predict which parts of the tax code will be considered for revision in any particular year.
- Good coverage of small-sized and medium-sized returns (as well as large-sized) within industrial categories to produce distributional cross-section analyses.

The Bureau of Economic Analysis has comparatively simple data needs. It uses the corporate SOI data for its national income and product accounts program, plant and equipment expenditures survey, and other economic estimation surveys. For these, the Bureau needs the estimated totals and counts published annually in the Corporation Source Book. There are 97 variables totaled in the 1980 Corporation Source Book detailed by industrial classification, by asset size class, and by income class (into two classes with and without net income).[4] Only in limited circumstances does the Bureau need more detailed information.

These two sets of needs are quite different and imply quite different designs. For the Bureau's needs, an optimal design would emphasize the largest returns and might well be similar to the current design. The tax analysts' needs suggest: sampling a larger number of smaller returns; an industrial stratification; and possibly a design targeting certain deductions/credits. The redesign effort thus seeks a better balance between these divergent requirements than exists in the current design.

PHASE 2

Phase 2 was an empirical analysis of the 1980 corporate SOI sample data. A few of the conclusions are stated here to motivate the alternative designs constructed in Phase 3. Details can be found in [2]. The conclusions are:

- Estimates of money amounts totals in the current design are very precise, nationally, and in most industries. (See also [8].)
- Industries are not equal in the SOI sample. The industries with lower asset distributions are less accurately represented in the SOI sample, as measured by the coefficient of variation.
- Despite industry misreporting, letting selection probabilities depend extensively

on industry is feasible and potentially beneficial.

PHASE 3

Two designs, which were moderately different from the current design, were constructed and appraised relative to the current design. They were found to be roughly equivalent to the current design--better by some measures of quality, poorer by others. As a useful analogy, imagine these measures of quality as defining a multidimensional response surface with each design being a reference point on the surface. The conclusion is then that the response surface is flat in the neighborhood of the current design. Its behavior elsewhere is unknown.

The structure of the current design will now be reviewed, followed by outlines of the two alternative designs. It is convenient to reference these designs by number--0, 1 and 2--with Design 0 being the current design.

For purposes of comparison, all three designs were constructed to achieve the same total sample size--74,383--the actual size of 1980 sample, as indicated above.

Since no complete frame exists with all variables, the 1980 SOI file, with its sampling weights, was used as a "pseudo-population." That is, if a return in the sample had a weight of 3.0, say, it was assumed to represent exactly three returns, all having the same data values. This is a cost-efficient procedure that permits the estimation of sample characteristics in expectation. However, it has some disadvantages. Primarily, it biases the results in favor of the current design.

Design 0 (Current)

The current design is detailed in [5] and [8]. It is stratified by total assets and net income or deficit into 10 L-shaped strata, as shown in Figure 1. Industrial classification enters the design only through the existence of a stratum defined exclusively for financial returns. The sample is allocated across the strata using an approximate Neyman allocation to minimize the variance of total assets and net income or deficit. An explanation of Neyman allocation is given by Cochran [6]. The sampling rates increase as one moves up and to the right from the origin in Figure 1.

Design 1

The first alternative design is cross-stratified by industrial classification (with 58 strata) and total assets with 12 strata. The cut points between the total asset strata are defined to coincide with the cutpoints published in the Corporation Source Book.[4] The sample is allocated proportionally across industries. Neyman allocation is again used, this time to minimize the variance of total assets within industries. Figure 2 displays the structure of Design 1.

Design 2

Design 2, like Design 1, begins with a cross-stratification of industrial classification (58 strata) by total assets. However, Design 2 has eight asset strata, and the cut-points between strata are determined using the Dalenius-Hodges rule to find the cut-points that minimize the variance of total assets. Cochran discusses the Dalenius-Hodges rule.[6]

To address the tax analysts' need for coverage of subdomains defined by deductions/credits, the four lower-asset strata were each split into two strata. A return was put into one stratum if all three of the variables--Investment Credit, Net Ordinary Gain or Loss, and Dividends Received --were absent on the return. It was put into the other stratum if one or more of these three were present on the return. It was not felt necessary to split the strata containing higher-asset returns; both the sampling rates and incidence rates of the deductions/credits are large enough to ensure sufficient numbers of returns in the sample with the deductions/credits. Figure 3 displays the structure of the Design 2 strata. These three deductions/credits were selected for several reasons: they are perennially important to the tax analysts; their incidence rates are small enough that random sampling will not ensure their presence in the sample in sufficient numbers; and their incidence rates are not so small that forcing them into the sample would greatly skew the sample. (See [3] for details.)

Impact on Sample Sizes

The three designs distribute the sample of 74,383 quite differently across the various industries. Figure 4 shows the distributions, with the 58 industries grouped into 9 classes. Since Design 1 employs proportional allocation across industries, its distribution is the same as the population's. Some industries--Mining, for example--are treated similarly by all three designs. Others are treated quite differently. Finance is less emphasized in Design 1 than in Designs 0 or 2. In contrast, Services are emphasized more in Design 1 than in Design 0 or 2.

The sample sizes also are different across corporate sizes, as measured by total assets and shown in Table 1. All three designs over-sample the larger returns: 85 percent of the population have assets under $500,000, but no design allocates more than 32 percent of the sample of these returns. Design 1, with proportional allocation, samples the largest number of small returns; Design 2, with several small strata for the small returns, samples the smallest number of small returns.

Precision of Estimators

To assess the precision of the estimated money amount totals--that is, to assess the designs vis-a-vis the Bureau of Economic Analysis' needs--we computed the coefficients of variation (CVs) for selected money amounts considered important by the Bureau within each of the 58 industrial categories. For each variable, the 58 CVs were plotted into a box-plot. An explanation of box-plots can be found in Tukey [7]. Box-plots for three of the money amounts are displayed below. Others may be found in [3].

Figure 5 shows the box-plots for Total Assets. Three box-plots are shown, one for each design. Each box-plot is based on 58 CVs. Figure 5 shows that Designs 1 and 2 estimate Total Assets for most industries better than Design 0. Design 1 allows a few industries to be very poorly estimated, as indicated by the outliers. These industries tend to:

FIGURE 1: DESIGN 0 (CURRENT) STRUCTURE

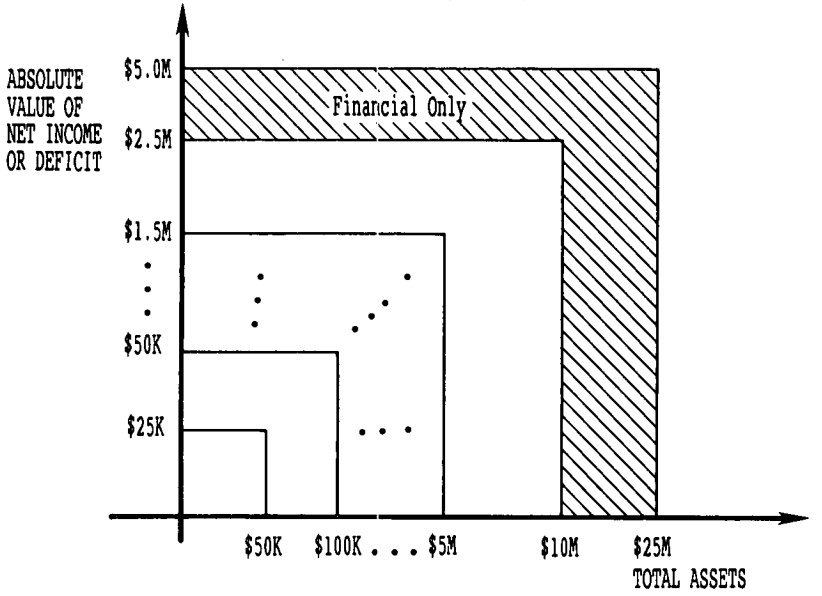


FIGURE 2: ALTERNATIVE DESIGN 1 STRUCTURE

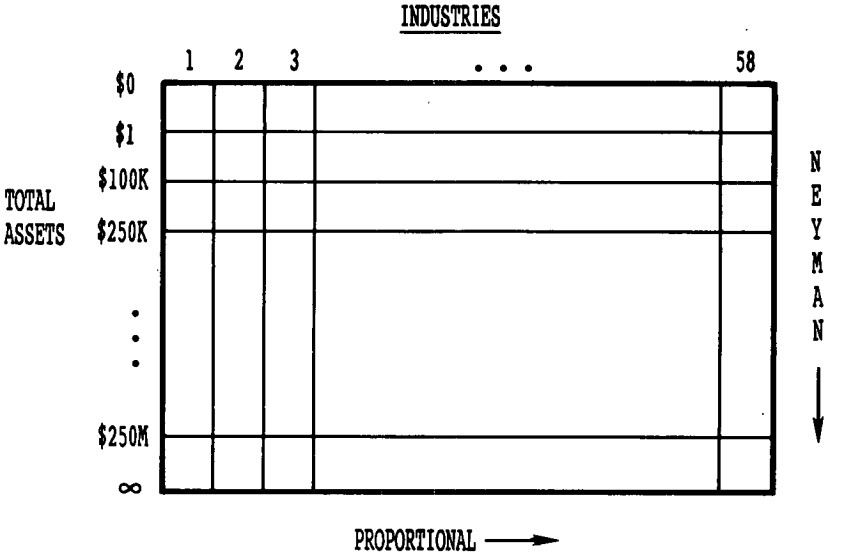
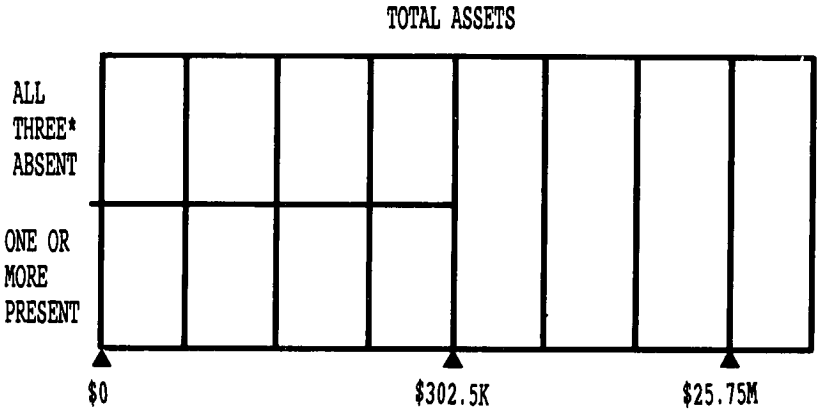
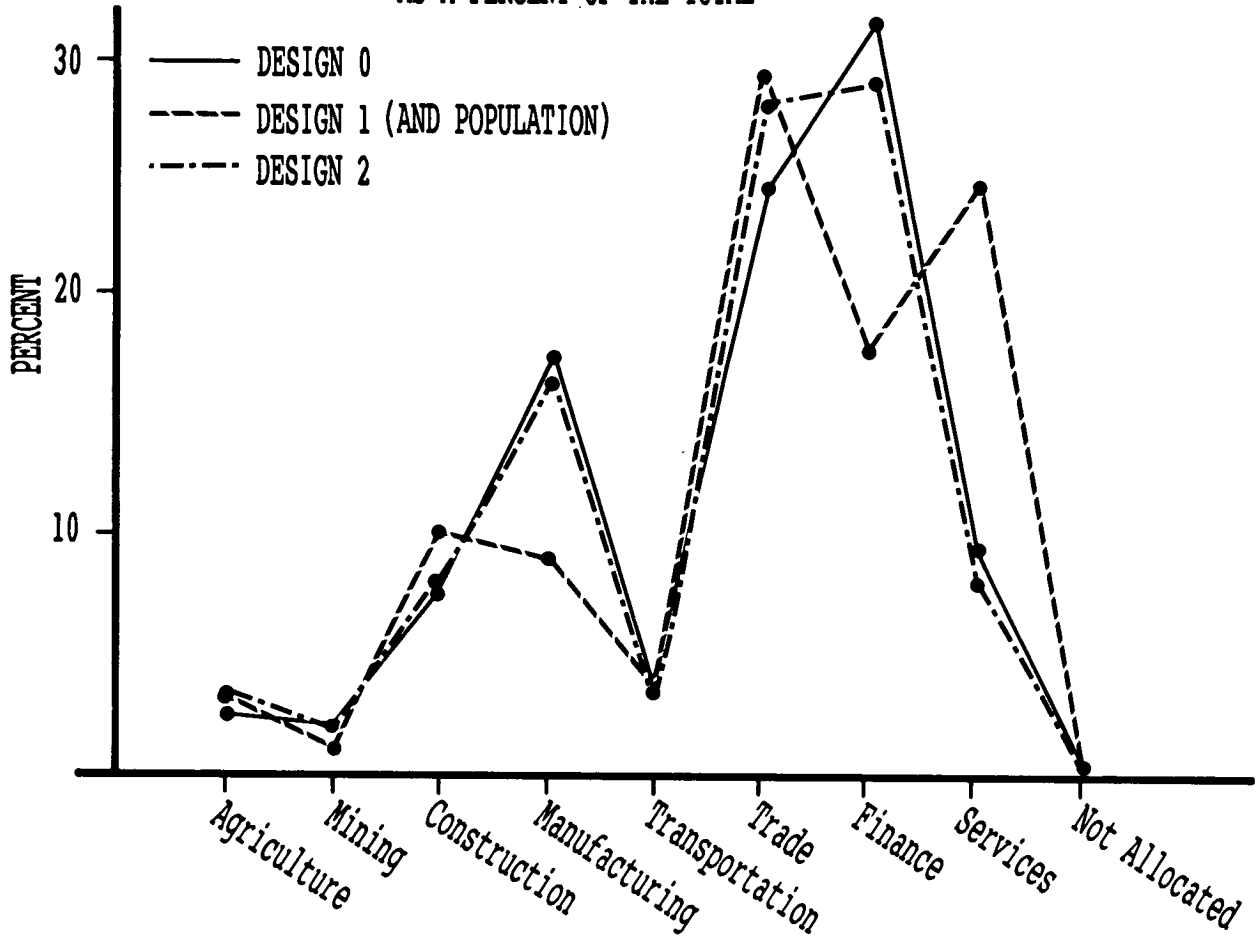


FIGURE 3: DETAIL OF ALTERNATIVE DESIGN 2 STRUCTURE



*INVESTMENT CREDIT, NET ORDINARY GAIN/LOSS AND DIVIDENDS RECEIVED

FIGURE 4: POPULATION AND SAMPLE SIZES BY INDUSTRY AS A PERCENT OF THE TOTAL*



*The total is 74,383 returns in the sample, and 2,689,245 returns in the population.

TABLE 1: POPULATION AND SAMPLE SIZES BY ASSET SIZE AS PERCENTS OF THE TOTALS

SIZE OF TOTAL ASSETS	POPULATION SIZE (Percent of 2,689,245)	SAMPLE SIZE (Percent of 74,383)		
		DESIGN 0	DESIGN 1	DESIGN 2*
Under $500,000	84.9	22.6	31.8	8.7
$500,000 to 25,000,000	14.4	53.8	56.6	68.0
$25,000,000 And Up	0.7	23.6	11.6	23.3

*The Design 2 breakpoints are actually $542,500 and $25,750,000.

FIGURE 5: BOX-PLOTS OF CVs OF TOTAL ASSETS

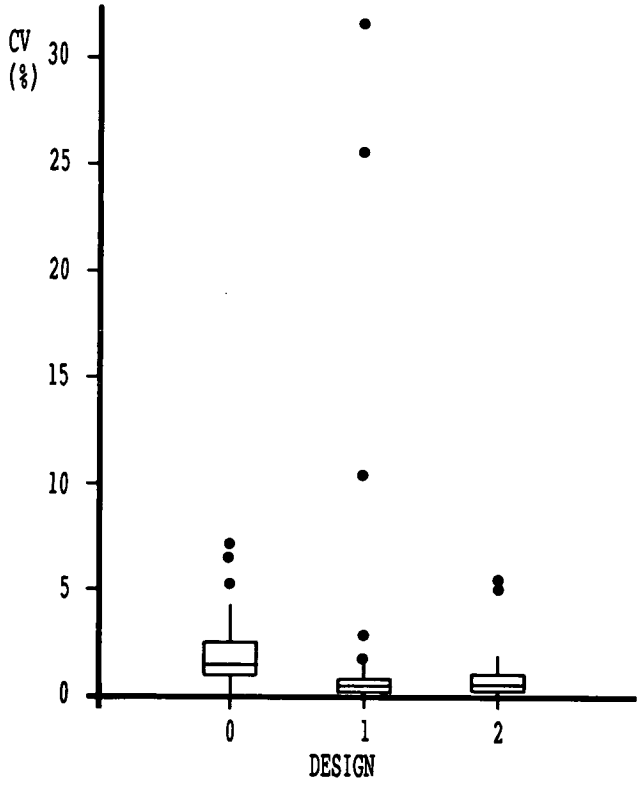


FIGURE 6: BOX-PLOTS OF CVs OF INVENTORIES

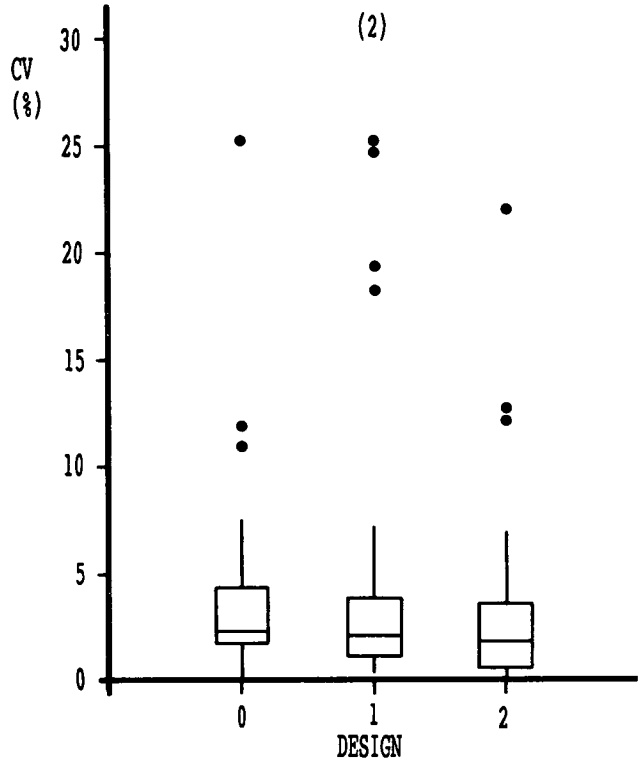
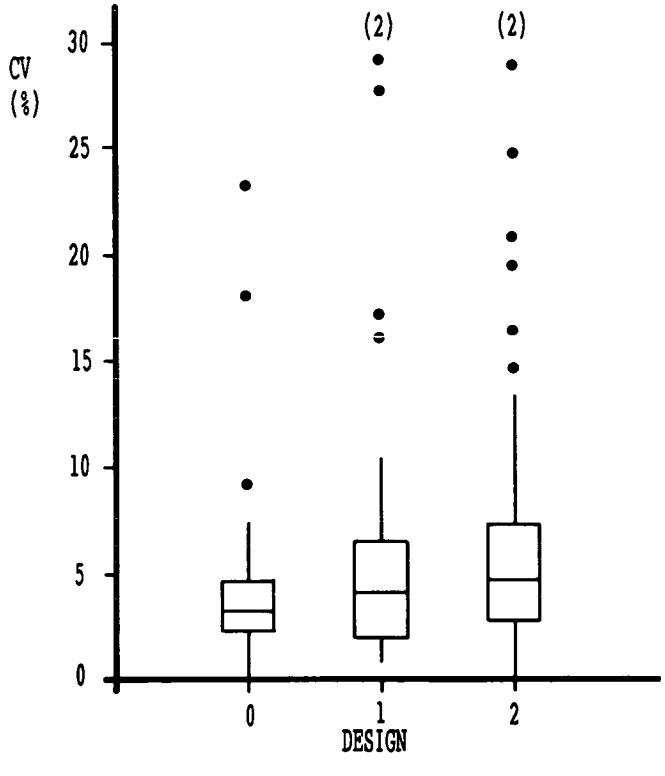


FIGURE 7: BOX-PLOTS OF CVs OF INCOME TAX AFTER CREDITS



(2) = 2 CVs above 30%

have small sample sizes; be in the finance group, e.g., banks or insurance companies; or be ill-defined industrial classes, i.e., "not-allocable."

Figures 6 and 7 give similar displays for two other variables but make different statements about the relative quality of the three designs. Figure 6 shows the results for Inventories. The three designs are essentially tied, except for the few industries again poorly estimated by Design 1. Finally, Figure 7 shows the results for Income Tax After Credits. This variable is well-estimated by the current design but relatively poorly estimated by the two alternative designs.

CONCLUSIONS

As stated previously, the two alternative designs are roughly equivalent to the current design--better by some measures, poorer by others. Since neither design substantially improves on the current design, we would not recommend that either replace the current design.

There are two conclusions that can be drawn from all three phases of this evaluation research. Both conclusions are supported by the data in hand at this time, albeit not conclusively established. The first is that the current design is very good and that it will be difficult to construct a substantially improved design. The second is one cannot hope to realize substantial improvements in quality by moderate changes in design structure. Substantial changes that may or may not be feasible in the current operating environment would be required.

ACKNOWLEDGMENTS

The authors are indebted to Michael Cohen, Huseyin Goskel, Marshall Hellmann, Winnie McMahon, and numerous other members of Applied Management Sciences' staff for statistical assistance, computing support, typing, and many other contributions to this research. We would also like to thank Fritz Scheuren, Daniel Rosa, Homer Jones, Paul McMahon, and many other members of the SOI Division's staff for their support on this project.

REFERENCES

[1] Applied Management Sciences, Inc. (1984a), Revised Final Report-- Identification of Corporate SOI Objectives, Key Variables, and Design Criteria (Unpublished Report), Silver Spring, Maryland.

[2] Applied Management Sciences, Inc. (1984b), Revised Final Report--Review of Current Corporate SOI Design and Method of Estimation (Unpublished Report), Silver Spring, Maryland.

[3] Applied Management Sciences, Inc. (1984c), Final Report--Evaluation of Corporate SOI Sample Design and Method of Estimation (Unpublished Report), Silver Spring, Maryland.

[4] Internal Revenue Service (1983), Source Book: Statistics of Income: Active Corporation Income Tax Returns: July 1980 - June 1981, Washington, D.C.

[5] Internal Revenue Service (1983), Statistics of Income - 1980: Corporation Income Tax Returns, U.S. Government Printing Office, Washington, D.C.

[6] Cochran, W.G. (1977), Sampling Techniques, Third Edition, John Wiley and Sons, New York.

[7] Tukey, J.W. (1977), Exploratory Data Analysis, Addison-Wesley, Reading, Massachusetts.

[8] Jones, H., and McMahon, P. (1984), "Sampling Corporation Income Tax Returns For Statistics of Income, 1951 to Present" Proceedings of the Section on Survey Research Methods, American Statistical Association