

# Turning Administrative Systems Into Information Systems

by Fritz Scheuren and Tom Petska

**S**tatistical work conducted within large administrative structures can have a somewhat different character than that carried out in exclusively statistical settings. Experience as the statistical component of a large tax collection organization, the Internal Revenue Service (IRS), certainly bears this out. Although the relationship between statistical and administrative activities has its complications, the Statistics of Income (SOI) function can (and does) act as a focal point for a broader and better use of administrative systems for statistical purposes.

As an introduction to the celebration of 80 years of Statistics of Income data (1913-1992), this article presents an overview of the current SOI tax and economic statistics program [1]. In addition to providing a glimpse of the operations and statistical projects, the article describes how the program is attempting to modernize in order to better meet the needs of its many customers within the Treasury Department, certain other Federal Government agencies, Congress and the public at large.

## Background

Histories of statistics, depending on the author, have the profession beginning at least as a descriptive activity in China around 2 A.D. (with its first census). In those early days, there was no separation between the statistical and administrative undertakings of government; indeed, the term "statistics" comes from the word "state" and, in its original usage, meant simply numerical "facts about the state." Modern statistical inference probably began around 1750, although the concept of probability emerged around 1660 [2]. SOI has ties to these foundations of the statistics profession and proudly sees itself directly linked to these traditions.

SOI statistical work actually began about 80 years ago with ratification of the sixteenth amendment to the U.S. Constitution (in 1913) and, later that year, the enactment

of the first modern income tax law [3]. The Revenue Act of 1916 required the annual publication of statistics. Despite many revisions to the tax law, the original goal of that Act, to collect statistical information on taxes, continues today. Specifically, the current Internal Revenue Code (which is based on the Tax Reform Act of 1986) states that the Secretary of Treasury will —

**Tax law requires the annual publication of statistics.**

"....prepare and publish, not less than annually, *statistics* reasonably available with respect to the operations of the internal revenue laws, including classifications of taxpayers and *of income*, the amounts claimed or allowed as deductions, exemptions, and credits..." (italics added) [4].

For reasons now obscure, the words italicized above were joined together to give the IRS statistical operation its name—the "Statistics of Income (SOI) program."

The IRS uses a self-assessment system for the collection of most Federal taxes. Under this system, taxpayers, whether individuals or businesses, report their financial affairs and calculate their tax liabilities, which are then subject to audit examination. The main basic data source for SOI is the tax returns and related documents. While this source is quite different from that of Government statistical agencies that are survey-oriented, the SOI program still has essentially the same overall mission—to collect and process data so that they become meaningful information and to disseminate this information to its customers and users; hence, the title of this article is "Turning Administrative Systems into Information Systems."

The costs of administering the Federal income tax system are substantial; the annual budget of the Internal Revenue Service for the current (1993) fiscal year is \$7.4 billion. The SOI program presently requires an annual budget of about \$28.5 million (almost 0.4 percent of the IRS total) to accomplish its statutory responsibilities. If revenues from reimbursable studies are also considered, the SOI figure amounts to nearly \$31 million. Thus, despite its key role in converting administrative data into statistical information, the SOI program represents a very small portion of IRS resources. Moreover, it often has a relatively low priority in the overall IRS mission.

## From Data to Information

SOI core statistical systems have much in common with those of other statistical organizations [5]. This section describes, in some detail, how these statistical activities are applied in SOI programs. Statistical sampling is a

*Fritz Scheuren is Director, Statistics of Income Division, and Tom Petska is Chief, Coordination and Publications Staff. This article was originally published in the Journal of Official Statistics, Volume 9, Number 1, June 1993, and has been updated and expanded with information primarily from recently-published articles in Business Economics (Part 1 was published under the title "The Statistics of Income Program of the Internal Revenue Service," April 1992; the second part under the title "Information from Tax Returns: The Statistics of Income Program of the Internal Revenue Service: Part 2," July 1992.) and the Proceedings of the Eighty-Fifth Annual Conference on Taxation, National Tax Association, Salt Lake City, Utah, 1993, under the title "Turning Administrative Systems into Information Systems: The Statistics of Income Program of the IRS."*

## Turning Administrative Systems Into Information Systems

major tool in study designs, and computers are a ubiquitous element in this environment. Data collection is a highly structured and disciplined process. Sample estimates are usually obtained by randomization-based weighting of selected cases; public-use files are made available after the data for individual taxpayers are "sanitized" to meet confidentiality concerns. Aggregate data are compiled as tables which are frequently published, and research on methods, often driven by operating concerns, is conducted in ongoing attempts at improvement.

### SOI Sample Design and Selection

Tax returns are filed and administratively processed in one of ten IRS regional locations, called "service centers." Once processed, the data from each of these centers are compiled into a computerized "Master File" system which is the administrative backbone of the Agency. SOI operations begin by sampling returns from the Master File system; the Master File offers a sampling frame that enables efficient and sophisticated sample designs to be used.

Generally, statistics compiled for the SOI studies are based on stratified probability samples of returns. As the returns are processed in the Master File system, they are assigned sampling strata based on criteria such as income (or other measures of economic size), industry and presence of supplemental forms or schedules. Each taxpayer, whether an individual or business, has a unique number—the social security number for individuals and the employer identification number for businesses. These unique taxpayer identification numbers (TIN's) are used as the seed for a pseudo-random number (a mathematical transform of the TIN) which, along with the sampling strata, determine whether a given return is to be selected for the sample.

The algorithm for generating the TIN transform stays the same from year to year. Consequently, any return is selected into an SOI sample provided that it falls into a stratum with the same or higher probability of selection [6]. If it falls into a stratum with a lower probability of selection, the likelihood of selection will correspond to the ratio of the second year's to the first year's selection probabilities [7,8].

Of over 200 million tax returns processed each year for

administrative purposes, only about half a million are sampled for statistical analysis. However, since sampling rates generally increase with increases in the size of financial amounts (for example, income or assets), the returns in the SOI samples are, on average, disproportionately larger and more complex than those in the administrative (population) files. Thus, in comparison to IRS administrative processing, which captures 100 percent of the tax returns but only limited item content, SOI programs collectively represent a small overall volume — however, with a proportionately higher fraction of complex returns and much more item content.

### Data Capture Techniques

After sampling, the machine readable information available from the Master File system is substantially augmented with additional data items captured from the (still largely paper) tax returns themselves. (Some of the "captured" data are also revised in order to improve data consistency.) Statistical abstracting can take as little as a few minutes for a simple return to as long as several days for a large multinational corporate return.

Until a few years ago, the basic SOI information processing was conducted in a "batch-mode," involving several operational units at all ten service centers. Within each center, data were manually abstracted from the returns, key-entered and error-corrected by different functional units. A fragmented system such as this denied "ownership" and accountability and was not conducive to maintaining high levels of quality.

To improve the quality and efficiency of SOI field processing, a network of minicomputers was built solely for statistical processing. This new system uses on-line transaction processing so that all data capture operations are completed in a single operation. In addition to reducing handling costs and removing overlapping responsibilities, accountability and ownership have been improved because one person is now responsible for assuring the validity of all data processing for any sample case. Although implemented with different hardware and software, this data processing system resembles Blaise, developed by the Netherlands Central Bureau of Statistics [9].

Another processing improvement was to reduce the number of field sites to just five of the ten service centers. Programming is done mainly by staffs of computer specialists primarily at two "hub" sites. The system is connected via a Treasury Department telecommunications network, which electronically links the geographically-dispersed operations so that data can be efficiently transferred between locations. This capability enables

**Of over 200 million tax returns processed each year, only about half a million are sampled for statistical analysis.**

## Turning Administrative Systems Into Information Systems

"experts," wherever located, to better monitor processing and to accelerate efforts to attain still higher levels of quality.

### Data Cleaning and Completion

Due to substantial penalties for misreporting, the detailed income and expenditure data on tax returns are generally regarded as more reliable than similar survey data. Even so, SOI employees go to great lengths to protect against nonsampling errors, such as those due to taxpayer or data entry error. Extensive on-line tests for consistency and reliability are made based on the structure of the tax law and the improbability of various data combinations. Subsamples of work are independently reprocessed and compared as a further check.

Missing data problems arise, albeit infrequently (under 1 percent of the cases). Most of the missing data problems involve tax return schedules in support of totals. To handle these, missing items can sometimes be obtained through telephone or written follow-ups. More often, though, the missing data are obtained through imputation. For example, an estimate can be made using: other information on a return (or in an accompanying schedule); prior-year data for the same taxpayer; or data from a "similar" return for the same year. The multiple imputation techniques have proven highly successful, and their use is increasingly applied in SOI work [10,11].

### Weighting and Estimation

On the whole, the SOI approach to making statistical summaries, using design-based inferences for the calculation of estimates and their standard errors, is quite straightforward. In the applications, the probability with which a return is selected for an SOI sample depends on the sampling rate prescribed for the stratum in which it is classified. Weights are computed by dividing the population count of returns filed for a given stratum by the count of sample returns for that same stratum. In some studies, it is possible to improve the estimates by employing *post-strata*, based on supplemental criteria or refinements of those used in the original stratification. Weights are then computed for these post-strata using additional population counts—oftentimes with fairly computer-intensive methods, such as *raking ratio estimation* [12].

*Model-assisted estimates* and *bootstrapping techniques* have been explored for selected SOI programs, but their deployment remains infrequent [13-15]. A combination of *randomization weighting* and model-assisted techniques is now used to make preliminary estimates prior to the completion of sampling. In one application, because the

cases obtained late in the sampling period are not random (they tend to be more complex), *propensity score weighting* has been tried [16].

### Published Tables and User Analyses

Extensive tabulations containing aggregated data have always been produced as part of the SOI program. While many of these continue to be primarily for Government analysts and policymakers, there is also a large (paper) publication effort in the quarterly *Statistics of Income Bulletin* and other annual and periodic SOI reports [17-20]. Electronic media products are increasingly available on magnetic tape, floppy disk, CD ROM and in a computer bulletin board format [21].

Microsimulation modeling "experiments" have become the *modus operandi* for policy analysis, and this is the case for SOI data users as well [22,23]. Recently, the National Academy of Sciences made a number of important recommendations for improvements to microsimulation modeling, and SOI has begun to rethink its work as a result [24]. Prior to the National Academy report, a major redesign effort for microdata products had already begun, and further efforts are now being planned, especially on improving the documentation of the data collection process and of the data themselves, for each project [25].

### Programs and Customers

An organizational strength of SOI is its close relationships with the Treasury Department's Office of Tax Analysis (OTA) and the Congressional Joint Committee on

Taxation (JCT), key decision-makers on tax policy within the Federal executive and legislative branches, respectively. This closeness assures a high degree of relevance for SOI work. SOI data are the principal source of information for revenue estimation and analyzing the functioning of the tax system. SOI data are also used extensively to measure and analyze the economy in the national economic accounts produced by Commerce Department's Bureau of Economic Analysis (BEA). Other users of SOI data

cover a broad spectrum of researchers, tax practitioners and the public at large (see Figure A).

**Due to penalties for misreporting, detailed data on tax returns are generally regarded as more reliable than similar survey data.**

## Turning Administrative Systems Into Information Systems

For most of its 80-year history, the main emphasis of the SOI program has been individual and corporation income tax information. However, growth has occurred in both the nature and number of studies undertaken. For 1980, the SOI program consisted of 26 projects; since then, the number of studies has more than doubled to nearly 60, comprising a wide range of business statistical programs, international studies and other special studies. This two-fold growth in programs was accompanied by a parallel four-fold increase in the amount of data extracted from the various tax and information returns, all at virtually no increase in inflation-adjusted costs.

Although the major product of SOI's information mission has historically been the Statistics of Income series of publications, the mission has shifted from one where the principal output was a "product" (e.g., statistical tables) to one where the output is increasingly information "services." These services involve providing OTA and JCT with files of tax return data and offering assistance on how to interpret the information correctly.

### Individual SOI Program

Detailed income and tax statistics from individual income tax returns have been published in annual reports by the IRS beginning with Tax Year 1916. The content of the program is largely determined by OTA and JCT for use in tax policy research and for revenue estimation. The needs of other researchers for individual income tax data are addressed on a cost-reimbursable basis.

Figure A

### Statistics of Income User Inquiries by Type of Request, 1992

Inquirer	Telephone request	Written request
<b>Total</b> .....	<b>100.0%</b>	<b>100.0%</b>
<b>Business</b> .....	<b>32.7</b>	<b>37.3</b>
Consultant/researcher.....	19.4	17.5
Accounting firm.....	3.4	8.0
Law firm.....	3.0	5.9
Other private business.....	6.9	5.9
<b>Government</b> .....	<b>37.7</b>	<b>18.2</b>
Internal Revenue Service.....	16.1	1.5
Other Federal Government....	8.5	1.3
Congressional.....	4.9	1.7
State or local Government.....	7.7	13.5
Public library.....	0.5	0.2
<b>Other</b> .....	<b>29.6</b>	<b>44.4</b>
Association.....	6.0	10.9
College or university.....	5.4	15.2
Private citizen.....	11.5	13.5
Student.....	1.2	0.8
Foreign.....	0.5	1.7
Media.....	5.0	2.1
Other.....	--	0.2

NOTE: Detail may not add to totals because of rounding.

While the individual program has historically been based on an annual cross-sectional sample of individual tax returns, a major redesign of the program has been undertaken [26]. From detailed discussions with OTA, it became apparent that the individual program needed to be refocused in three aspects:

**Currently, the SOI program consists of nearly 60 projects.**

- since the annual cross-sectional samples were not conducive to multi-year economic modeling for such events as sales of capital assets, the redesigned sample includes a large continuing "panel" of (the same) individuals embedded within the annual cross-sectional samples;
- because "family economic units" (reflecting households rather than individuals) are a more desirable focus of tax analysis, social security numbers of dependents, now reported on individual income tax returns, are used to obtain dependents' returns, which are then combined with the parents' returns to form "tax families;" and
- sampling stratifiers and selection rates have been restructured to enhance the samples of returns with greater policy interest, such as those with very high or low incomes or those of the aged.

SOI individual taxation information is published annually in a complete report; in addition, annual articles are published in the *Statistics of Income Bulletin* presenting preliminary individual income and tax data, individual tax rates and tax shares, and on high-income tax returns [27-28].

Other studies closely related to individual taxation issues are the sales of capital assets (SOCA) panel studies and "information returns" studies [29]. The SOCA studies have been conducted periodically by creating a panel file of returns to track capital asset transactions by type of transaction.

The Tax Year 1989 sample of individual income tax returns has been linked, on a record-by-record basis, to information returns filed by employers and financial institutions (for example, Forms W-2 for wages paid and Forms 1099 for interest and dividends paid) in order to create a more complete data base for analysis. As an offshoot of sampling "family economic units," information documents of individuals claimed as dependents on returns in this sample will be added. Finally, a representative selection of information documents of individuals not covered by individual income tax returns (i.e., non-filers:

## Turning Administrative Systems Into Information Systems

and non-dependents) will be added. Using the number of personal exemptions claimed as the indicator, the resulting data base is expected to cover nearly 98 percent of the U.S. population [30].

### **SOI Business Programs**

Although businesses can be legally organized in a variety of ways, most business activity is conducted by corporations, partnerships or sole proprietorships. These three annual SOI programs are, thus, often referred to as the SOI business studies.

**Corporations.**— Like those for individuals, SOI corporate income tax data have been published annually, beginning with tax returns for 1913. These data are the only publicly-available source of financial information on all corporations, since other sources include only large or publicly-held corporations or those in certain industries. In addition to its use in tax policy simulation and revenue estimation at OTA and JCT, the corporation program provides the basic source data used for estimating corporate profits for BEA's national economic accounts.

The corporation program is rich in item content; complete income statement, balance sheet, and tax computation information have been mainstays of the program almost since its inception. As with the individual income tax studies, this program is being restructured to better meet the needs of OTA and JCT. Increased longitudinality (using panels of taxpayers) is being incorporated into future studies, and enhancements are underway to compile corporate "families" by linking parent and subsidiary entities of corporations, whether or not consolidated returns are filed. Through the financial support of BEA, the delivery of preliminary corporation statistics has been revived beginning in 1992. SOI corporation data are published annually in the *Statistics of Income Bulletin*, in a complete report, and in a detailed Source Book which is also available on magnetic tape [31,32].

**Partnerships.**— The annual SOI partnership program is vital to the national accounts since it is the only source of data on these businesses.

Partnerships are required to file annual information returns including an income statement, balance sheet and schedules showing the shares of income or losses and other items distributed or allocated to partners. Partners are required to report the distri-

butions or allocations from partnerships on their own income tax returns.

For many years, partnerships commanded only modest interest because they were not taxed directly and, thus, had no direct effect on Federal revenue; however, the proliferation of partnerships in tax shelters has substantially increased interest. For example, curbing deductions of partnership losses by individual partners, through passive loss limitations, was a key provision in the 1986 Tax Reform Act (TRA) [33]. Partnership data are annually published in the *Statistics of Income Bulletin* and have also been examined in other articles focusing on the impact of TRA [34-36].

**Sole Proprietorships.**— Information about nonfarm proprietorship business activities is reported on Schedule C of the individual income tax return. Profits from these activities are combined with income from other sources in order to compute individual "adjusted gross income." Data on proprietorships provide the other half of information on the unincorporated business sector for the national accounts. Here, again, the tax return is the only annual source of financial information about these businesses. Proprietorship information is also published annually in the *Statistics of Income Bulletin* [37].

### **SOI International Studies**

"International" studies are conducted biennially or periodically in two broadly-defined areas: foreign investment and activity abroad by U.S. "persons" and investment and activity in the United States by foreign "persons."

Foreign investment and activity abroad by U.S. persons includes the following studies: corporation and individual foreign tax credit, Controlled Foreign Corporations of U.S. corporations, Interest-Charge Domestic International Sales Corporations and Foreign Sales Corporations, U.S. Possessions Corporations, international boycott participation, individual income earned abroad and foreign trusts. Treasury uses many of these studies for mandated reports to Congress.

Investment and activity in the United States by foreign persons includes the following studies: foreign-owned U.S. corporations, foreign corporations with income derived from a U.S. business, nonresident alien income and tax withheld, nonresident alien estates, U.S. partnership income of foreign partners and sales of U.S. real property interests by foreign "persons."

Much of the recent analyses of these data is published in the *Statistics of Income Bulletin*. In addition, they have been compiled into a compendium of studies of interna-

**SOI data are the principal source of information for revenue estimating and analyzing the functioning of the tax system.**

## Turning Administrative Systems Into Information Systems

tional income and taxes, available for purchase upon request [38].

### Other SOI Studies

Annual, biennial or periodic statistical programs are also conducted on tax-exempt (i.e., nonprofit) organizations, certain tax-exempt obligations, estates and personal wealth, and certain excise taxes.

**Tax-Exempt Organizations.**— Annual studies are conducted of information returns filed by private foundations, nonprofit charitable and other organizations exempt under Internal Revenue Code section 501(c), exempt organizations with taxable “unrelated business income,” and tax-exempt private activity bonds.

In recent years, there has been a major focus by Congress and the business sector on the movement of exempt organizations into commercial activities and the resulting impact of this movement on for-profit businesses. Beginning with Reporting Year 1993, the statistical sampling of returns filed by nonprofit charitable and other organizations exempt under Internal Revenue Code subsections 501(c)(3) through 501(c)(9) will be linked with the sampling of returns filed to report business income from activities considered unrelated to the organization’s exempt purposes. This will ensure that information on both tax-exempt income and taxable income is available for the nonprofit charitable and other exempt organizations studied.

Studies of tax-exempt organizations are periodically released in historical compendiums that consolidate data on these organizations, as well as include articles that were previously published in professional journals or in the *Statistics of Income Bulletin*. The first volume in the series includes information for Reporting Years 1974-1987; the second, for Reporting Years 1986-1992 [39].

**Estate Tax and Personal Wealth.**— Estate tax studies are now conducted annually based on a specific year of death. Periodically, studies are undertaken to estimate the wealth of top (living) wealthholders by combining applicable mortality rates with estate data [40]. Many of these analyses of estates and wealth have been published in the *Statistics of Income Bulletin* and a historical compendium on estate and personal wealth is now in preparation [41,42].

A long-term research project is underway examining intergenerational transfers of wealth through inheritance based on estate tax filings from 1916 to the present [43]. SOI has also had a partnership role with the Federal Reserve Board in periodically mounting the Survey of Consumer Finances, a complex series of household

interviews designed to estimate personal wealth [44].

**Excise Taxes.**— These studies include (or have included) returns on the quarterly crude oil windfall profit tax and environmental excise taxes on certain hazardous substances (i.e., the so-called “Superfund taxes”) [45]. Additional studies are being planned to provide a more comprehensive and timely picture of excise taxes [46].

### Shaping the Future

Some areas where the SOI program is shaping, or being shaped by, the future include: the continuing revolution ongoing in statistics, computing, and allied professions;

changes in management styles and organizational practices (driven largely by Japanese successes in quality and productivity); and, last, attempts to address continuing deficiencies in meeting growing customer needs and demands. Each of these “opportunity challenges” is looked at briefly, followed by some concluding remarks.

**Many Government statistical agencies have not kept up with the explosive growth of statistical theory and methods.**

### Changes In Statistical Practice

Historically, SOI, like most Government statistical agencies, has retained a much stronger “enumerative” or descriptive focus than an “analytic” or “inferential” one [47]. For SOI to continue this enumerative focus is a major impediment to developing an improved structure of information collection and analysis. Without a doubt, the “science” side of SOI work must be given greater emphasis [48]. In addition, a more analytic focus will not only allow SOI to continue to attract and retain outstanding employees, but will also bring it closer to its customers [49,50]. Recent years have seen encouraging trends, with much work now being done jointly with the major customers. The coming improvements in microsimulation modeling, partly driven by new longitudinal samples, should see this trend continue.

Many Government statistical agencies, including SOI, have not kept up with the explosive growth of statistical theory and methods. This is ironic because, in some cases, important methodological developments are being made within the Government [51]. Many such agencies, again including SOI, are trying to change this. Most SOI statistical methods are close to best practice; however, the range of tools employed, while growing, remains fairly narrow.

An example of the way SOI’s methods will continue to

## Turning Administrative Systems Into Information Systems

grow and modernize is the partnership role SOI has taken with the Federal Reserve Board in the Survey of Consumer Finances and with the Bureau of Labor Statistics in the use of "cognitive" methods as a way of improving the tax forms [52]. Indeed, there has been an increasing institutional acceptance within IRS of this improved approach because it both reduces taxpayer burden and elicits more accurate responses.

### Restructuring Management Practices

Long-term initiatives are underway to improve "quality" management in SOI. During the 1980's, conventional quality control techniques for detecting errors were gradually replaced by quality improvement techniques designed primarily to prevent errors. Japanese practices have been studied, and the ideas of both Deming and Juran are being implemented [53].

Early piecemeal improvement attempts are now giving way to a more integrated approach. SOI is now deploying its second annual total quality organization (TQO) plan, in which the five key strategies are [54]:

1. *Customer Focus.* — Provide greater access to SOI data in a more timely and flexible manner. Develop and market new products and services designed to increase benefits to customers.
2. *Employee Focus.* — Make SOI a more desirable, fulfilling and productive place for people to work. Build communication systems that facilitate a freer exchange of information within SOI and with customers and suppliers.
3. *Lean Production.* — Create data processing systems that are "best in class." For example, develop the capability to accept changes throughout project life cycles. Reduce the amount of rework or corrections needed at each processing stage of a project. Help suppliers develop and maintain a steady workflow and a stable workforce [55].
4. *Measurement Systems.* — Integrate and improve existing quality and resource measurement systems to aid project teams in achieving self-management.
5. *Planning Processes.* — Develop an increasingly systematic planning process to improve the focus of SOI quality initiatives.

As a result of these initiatives, the relationships between SOI and its customers and suppliers are changing. Substantial productivity and quality gains already achieved are due to many factors, including methodological enhancements and computing improvements, but

perhaps most of all to the still early efforts to embrace modern management techniques.

SOI is proud of what has been achieved so far. On the other hand, much more remains to be done if new ways of doing business are to fully succeed. The key to the eventual outcome will be the extent to which SOI staff is drawn into the process and the degree to which teamwork structures—*emphasizing reciprocal responsibility*—replace traditional hierarchies.

### Informational Deficiencies

Despite recent strides, SOI systems continue to have chronic weaknesses. Currently, SOI is focused on several initiatives to address deficiencies in the SOI systems to better meet customers' needs. Five examples of these are described, followed by a sketch of some of the plans proposed to address these shortcomings. These five are the needs for: improved data consistency, greater program timeliness, better tracking of demographic changes, improved data access and preservation of historical information.

**Data Consistency.**— Problems of data consistency are of two general types: statistical and conceptual. Despite extensive validity testing, inconsistent or erroneous data still escape undetected in some SOI data files in spite of "heroic" efforts. Sometimes such errors are the result of "creative reporting" by taxpayers that, when appropriately weighted to produce national estimates, result in large discrepancies. For whatever reason, efforts continue to rid these out of the systems. Any error that affects any estimate by a "significant" amount (where "significant" is defined by the customer or user) is unacceptable and efforts must continue to drive the error rate down toward zero.

The ever-changing concepts of the economic, financial and tax variables, as reflected on tax and information returns, create a multitude of difficulties in trying to examine tax data over time. Unlike other statistical agencies where the questionnaire content can be customized to meet a statistical need, SOI's "questionnaires" are tax returns, which are developed specifically for tax compliance and other administrative purposes rather than for statistical studies. Not only is there limited flexibility in changing tax return content, but the information provided

**Timeliness of SOI studies has become a primary focus for improvement and one in which some successes have been achieved.**

## Turning Administrative Systems Into Information Systems

on returns is often complicated, not uniformly reported, and conceptually inconsistent over time, thereby impeding multi-year economic analysis.

Improving the conceptual clarity and year-to-year consistency of tax and information return content is a problem that has no easy solution. Since the tax laws have been frequently revised by the Congress, preserving year-to-year consistency can be challenging. For some data items, such as business depreciation deductions, no attempt is made to make adjustments to the basic source data provided on tax and information returns as long as they are consistent with current provisions of the tax code. In other cases, such as when the overall profits of partnerships were redefined by a tax form change, efforts have been made to re-create a consistent series over time.

As noted above, the contribution that cognitive psychologists have been making to the restructuring of survey instruments is a new and exciting breakthrough [56]. SOI has supported this approach as a user of the Bureau of Labor Statistics' Collection Procedures Research Laboratory since its inception in 1989 and plans to send IRS tax forms as a means of improving tax administration and, as a by-product, the quality of the SOI data.

**Timeliness.**— The fact that users never have enough current information from tax returns is a key weakness of the SOI program. Timeliness of SOI studies has become a primary focus for improvement and one in which some successes have been achieved. All of the major SOI studies have a sampling period that extends for one year (or more) beyond the close of the applicable accounting period to ensure the inclusion of returns for which there are filing extensions. To illustrate, most Tax Year 1992 individual taxpayers filed by April 15, 1993, but significant numbers (about 6 million) sought and were granted extensions until August 15, 1993. For taxpayers living abroad, further extensions are common. Thus, to insure inclusion of returns filed late in that year, the sample for Tax Year 1992 is kept open throughout Calendar Year 1993 [57].

Significant efforts are being made throughout SOI to complete statistical processing within a minimum time after the close of the return sampling period. The *Statistics of Income Bulletin* periodically summarizes the release dates of SOI information [58]. What is not

**Finding ways to obtain wider public access, while protecting taxpayers' confidential information, is an SOI goal.**

apparent in these summaries is the extent to which delivery dates have been improving. In the individual SOI program, for example, early economic estimates are available 2 months after the end of the 1992 filing period, more detailed preliminary data before the close of the filing year and final information is available 6 months later. Both of these represent a significant acceleration over earlier schedules.

In addition, internal discussions are underway to provide even more current information to our principal customers on a preliminary basis by providing access to files in process and by early closure of the sample periods, as has been done for the individual SOI program for many years now. For example, an early cut-off has just been reinstated in the corporation program allowing delivery of preliminary data 9 months ahead of the final data.

**Tracking Demographic Changes.**— The redesign of the individual program at the request of OTA has underscored the need to continually improve longitudinality in SOI studies. Such "unusual" transactions as capital gains realizations can only be understood by means of a panel data base. As a result of this redesign, additional efforts are needed to ensure completeness, as well as inter-temporal consistency.

A similar need for greater longitudinality applies to business sector studies, where tracking business legal entity changes has become a difficult and costly aspect of file processing. For example, assuring the inclusion of all large corporations in the annual corporation program is a very labor-intensive effort. Often mergers, acquisitions, liquidations, reorganizations and "spin-offs" are determined to be the source of what had initially appeared to be missing entities. Recent tax reform legislation has heightened the occurrence of changes of legal form, such as switching from a corporation to a limited partnership. Unfortunately, an automated system that is also readily accessible and which will facilitate tracking such changes for effective tax administration does not yet exist.

**Greater Public Access.**— Finding ways to obtain wider public access, while protecting taxpayers' confidential information, is an SOI goal. Tax returns are protected by law from public scrutiny, and strict procedures govern the handling of returns and computer tape files containing such information. Even after specific identifiers (e.g., name, address and social security or employer identification number) are removed, the remaining tax return data are usually still confidential.

OTA and JCT, SOI's primary customers, are authorized to receive detailed tax return (microdata) files, so com-



## Turning Administrative Systems Into Information Systems

puter tape files of tax return information are regularly provided. Most other users of SOI data can only have access to summary tabulations. However, public-use microdata files of individual tax data have been produced regularly since 1960. For these other users, these files are the only source of "sanitized" tax return data available for simulations.

Making more tax microdata publicly available to researchers outside of Government has been studied in SOI for some time. A research project to study the feasibility of public-use business files was undertaken in the early 1980's [59]. Various measures were considered to make the data available while protecting taxpayer confidentiality, including purging names and other unique identifiers, rounding data items to make it more difficult to identify businesses, and averaging the financial data of "similar" taxpayers. While this work demonstrated the feasibility of creating public-use data for unincorporated businesses, the cost of producing such data would be prohibitively expensive.

**Preservation of Historical Information.**— Although current efforts are focused on better meeting current and future customer needs, SOI has become the "keeper" of an abundance of tax information documents in a variety of media. Much of this information, though cumbersome to use, is irreplaceable. However, as new technologies become more available, the cost of moving this information into more user-friendly formats will drop considerably. A very difficult decision must be addressed as to how much current resources should be diverted from present work to safeguard this historical information.

### Concluding Comments

This article has discussed what the Statistics of Income program is and what it is trying to become. In comparison to some other Government statistical agencies, SOI is small, and because the SOI mission is highly focused, most of what it does is not widely known. The SOI organization and its programs have strong traditions that provide a sense of continuity and confidence—unfortunately, sometimes at the price of being overly conservative in the face of a changing environment.

As part of efforts to meet future goals, SOI has participated in and contributed, in a modest way, toward many of the worldwide "paradigm shifts" now sweeping statistics and statistical organizations. It has benefitted especially from the revolution in computing systems, albeit belatedly; the quality revolution is also one where SOI started late but where some important strides have been made. Applying newly invented or improved concepts and tools to old problems has been energizing;

indeed, the excitement has not only led SOI to tackle new problems, it has provided the impetus for SOI to "reinvent itself." The belief is that only by reinventing itself will SOI be able to successfully address the present and especially the future needs of all customers. Toward this goal, comments and suggestions are sought. An invitation is extended to those having similar challenges to join in a common effort.

For more information on SOI programs and services, contact the SOI Statistical Information Services office on (202) 874-0410 (FAX number (202) 874-0922), or write the Director, Statistics of Income Division R:S, P.O. Box 2608, Washington, DC 20013-2608. A computerized bulletin board, now in operation, enables SOI data users to keep in touch better with current data releases and other program events; it can be accessed on (202) 874-9574.

### Notes and References

- [1] See, also, *Statistics of Income Bulletin—75th Anniversary, 1913-1988*, Fall 1988, Volume 8, Number 2.
- [2] Fienberg, Stephen E., "A Brief History of Statistics in Three and One-Half Chapters: A Review Essay," *Statistical Science*, Volume 7, Number 2, pp. 208-225, 1992.
- [3] Actually, SOI-related statistical work at IRS relating to corporations began with 1909. Data on the excise tax on corporate net income for 1909-1913 are included in *Statistics of Income* for 1916.
- [4] See section 6108, "Statistical Publications and Studies," *Internal Revenue Code of 1986*.
- [5] See Fellegi, Ivan, "Some Mathematical Statistical Research Problems in Statistics Canada," paper presented at the Annual Meeting of the Statistical Society of Canada, Quebec, Quebec, 1987; and Scheuren, Fritz, "Statistical Research Problems in Government," paper prepared for the Annual Meeting of the Statistical Society of Canada, Ottawa, Ontario, May 1989.
- [6] See Tepping, Benjamin and Hansen, Morris, "Results of a Study to Improve Sampling Efficiency of Statistics of Corporation Income," unpublished technical report, Westat, Inc., 1974.
- [7] See Sunter, Alan, B., "Implicit Longitudinal Sampling from Administrative Files: A Useful Technique," *Journal of Official Statistics*, Volume 2, Number 2, pp. 161-168, 1986.

## Turning Administrative Systems Into Information Systems

- [8] Harte, James M., "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *1986 Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 603-608, 1987.
- [9] Pierzchala, Mark, "A Review of Three Editing and Imputation Systems," *1990 Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 111-120, 1991.
- [10] Rubin, Donald B., *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, 1987.
- [11] Hinkins, Susan and Scheuren, Fritz, "Hot Deck Imputation Procedures Applied to a Double Sampling Design," *Survey Methodology*, Volume 12, Number 12, pp. 181-196, 1986.
- [12] Oh, H. Lock and Scheuren, Fritz, "Modified Raking Ratio Estimation," *Survey Methodology*, Volume 13, Number 2, pp. 209-219, 1987.
- [13] See, for example, Sarndal, Carl-Erik; Swensson, Bengt; and Wretman, Jan, *Model Assisted Survey Sampling*, Springer-Verlag, 1992.
- [14] Mulrow, Jeri; Oh, H. Lock; and Collins, Richard, "An Evaluation of Bounded Raking Ratio Estimation in the Statistics of Income Corporate Programs," *1991 Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 597-601, 1992.
- [15] See, also, Wong, William and Ho, Chih-Chin, "Bootstrapping Post-Stratification and Regression Estimates from a Highly Skewed Distribution," *1991 Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 608-613, 1992.
- [16] Czajka, John; Hirabayashi, Sharon M.; Little, Roderick J.A.; and Rubin, Donald B., "Projecting From Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics," *Journal of Business and Economic Statistics*, Volume 10, Number 2, pp. 117-131, 1991.
- [17] *Statistics of Income—1989, Individual Income Tax Returns*, U.S. Department of the Treasury, Internal Revenue Service, 1992.
- [18] *Statistics of Income—1989, Corporation Income Tax Returns*, U.S. Department of the Treasury, Internal Revenue Service, 1992.
- [19] *Source Book of Statistics of Income—1989, Corporation Income Tax Returns*, U.S. Department of the Treasury, Internal Revenue Service, 1992.
- [20] See, also, Alvey, Wendy; Kilss, Beth; and Jamerson, Bettye, editors, *Statistics of Income and Related Administrative Record Research*, periodic reports, U.S. Department of the Treasury, Internal Revenue Service, 1981-1993; also, Kilss, Beth and Alvey, Wendy, editors, *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, Volumes I and II, U.S. Department of the Treasury, Internal Revenue Service, 1984, and *Record Linkage Techniques—1985*, U.S. Department of the Treasury, Internal Revenue Service, 1985.
- [21] See, for example, "General Description Booklet for the 1988 Individual Public Use File," Statistics of Income Division, Internal Revenue Service, 1990.
- [22] See Wolfson, Michael; Gribble, Steven; Bordt, Michael; Murphy, Brian; and Rowe, Geoff, "The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration," *Survey of Current Business*, Volume 69, Number 5, pp. 36-40, U.S. Department of Commerce, Bureau of Economic Analysis, 1990. (A discussion by Fritz Scheuren is included in this same issue.)
- [23] Atrostic, B.K. and Nunns, James R., "Measuring Tax Burden: A Historical Perspective" in Ernst Berndt and Jack Triplett, editors, *50 Years of Economic Measurement*, National Bureau of Economic Research Studies in Income and Wealth, Volume 54, pp. 343-408, University of Chicago Press, 1990.
- [24] See Citro, Constance F. and Hanushek, Erik A., editors, *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling, Volume 1, Review and Recommendations*, National Academy Press, 1991.
- [25] See David, Martin, "Metadata Systems for Statistics of Income Data Products," unpublished working paper, Department of Economics, University of Wisconsin, 1992.
- [26] See Czajka, John and Walker, Bonnye, "Combining Panel and Cross-sectional Selection in an Annual Sample of Tax Returns," *1989 Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 463-468, 1990.

## Turning Administrative Systems Into Information Systems

- [27] See footnote 17.
- [28] See Gross, Edward B., Jr., "Individual Income Tax Returns, Preliminary Data, 1991," *Statistics of Income Bulletin*, Spring 1993, Volume 12, Number 4, 1993; Day, Charles and Strudler, Michael, "Individual Income Tax Rates and Tax Shares, 1988-1989," *Ibid.*; and "High-Income Tax Returns, 1989," *Ibid.*
- [29] Information returns are filed by employers or financial entities to report wages and other monies paid to taxpayers. They tell the taxpayers how much of the amounts they must report on their income tax returns. IRS then uses information returns to cross-check what the taxpayer reports.
- [30] Sailer, Peter; Windheim, Barry; and Yau, Ellen, "Toward a New Census Paradigm: Comparing Coverage of the Population in the Census and on Tax Documents," presentation at the Annual Meeting of the American Statistical Association, Boston, Massachusetts, 1992. For similar presentations exploring alternatives to census-taking, see Jensen, Poul, "The Quality of Administrative Data From a Statistical Point of View: Some Danish Experience and Considerations," *1987 Proceedings of the Statistics Canada Symposium on Statistical Uses of Administrative Data*, pp. 291-300; Leyes, John, "An Administrative Record Paradigm: A Canadian Experience," *Statistical Policy Working Paper 20, 1990 Seminar on Quality of Federal Data*, Federal Committee on Statistical Methodology, U.S. Office of Management and Budget, pp. 66-76, 1991; and van de Stadt, Huib and Vliegen, Mathieu, "An Alternative for the Census? The Case of the Netherlands," *1992 Proceedings of the Annual Research Conference*, Bureau of the Census, U.S. Department of Commerce, pp. 591-596, 1992.
- [31] See, for example, Shumofsky, Nina, "Corporation Income Tax Returns, 1990," *Statistics of Income Bulletin* (this issue).
- [32] See footnotes 18 and 19.
- [33] See, for example, *Explanation of the Tax Reform Act of 1986 for Individuals*, U.S. Department of the Treasury, Internal Revenue Service, 1987.
- [34] See, for example, Shekhter, Elaina, "Partnerships, 1990," *Statistics of Income Bulletin*, Summer 1992, Volume 12, Number 1, 1992.
- [35] Petska, Tom, "Partnerships, Partners, and Tax Shelters after Tax Reform, 1987-1989," *Statistics of Income Bulletin*, Summer 1992, Volume 12, Number 1, 1992.
- [36] Petska, Tom and Nelson, Susan, "Partnerships and Tax Shelters: An Analysis of the Impact of the 1986 Tax Reform," *1990 Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1991.
- [37] See, for example, Ammon, Craig, "Sole Proprietorships, 1990," *Statistics of Income Bulletin*, Fall 1992, Volume 12, Number 2, 1992.
- [38] *Statistics of Income—Compendium of Studies of International Income and Taxes, 1984-1988*, U.S. Department of the Treasury, Internal Revenue Service, 1991.
- [39] *Statistics of Income—Compendium of Studies of Tax-Exempt Organizations, 1974-1987*, U.S. Department of the Treasury, Internal Revenue Service, 1991, and *Statistics of Income—Compendium of Studies of Tax-Exempt Organizations*, Volume 2, 1986-1992 (in preparation).
- [40] Mallet, B., "A Method of Estimating Capital Wealth from the Estate Duty Statistics," *Journal of the Royal Statistical Society*, Volume 71, pp. 65-84, 1908.
- [41] Johnson, Barry W., "Estate Tax Returns, 1989-1991," and Johnson, Barry W. and Schwartz, Marvin, "Personal Wealth, 1989," *Statistics of Income Bulletin*, Spring 1993, Volume 12, Number 4, 1993.
- [42] *Statistics of Income—Compendium on Estates and Personal Wealth* (in preparation).
- [43] McCubbin, Janet, "The Intergenerational Wealth Study: Basic Estate Data, 1916-1945," *Statistics of Income Bulletin*, Spring 1990, Volume 9, Number 4, 1990.
- [44] Kennickell, Arthur and Woodburn, Louise, "Methodological Issues in the Estimation of the Distribution of Household Net Worth: Results from the 1989 Survey of Consumer Finances," *1992 Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1993.
- [45] See Boroshok, Sara P., "Environmental Taxes, 1991," *Statistics of Income Bulletin* (this issue).
- [46] See Davie, Bruce F., "Tax Expenditures in the Federal Excise Tax System," unpublished working

## Turning Administrative Systems Into Information Systems

paper, Statistics of Income Division, Internal Revenue Service, 1993.

- [47] Norwood, Janet L., "The Influence of Statistics on Public Policy," *1989 Proceedings of the Symposium on Statistics in Science, Industry and Public Policy*, pp. 30-42, 1990.
- [48] Triplett, Jack, "The Federal Statistical System's Response to Emerging Data Needs," *Journal of Economic and Social Measurement*, Volume 17, Number 3, pp. 155-177, 1991.
- [49] O'Connor, Karen; Atrostic, B.K.; and Gillette, Robert, "Moving from Descriptive Statistics to Inference," *Proceedings of Symposium 90; Measurement and Improvement of Data Quality*, Statistics Canada, 1991.
- [50] An analytical focus that includes gaining more insights into the data SOI produces from a subject-matter point of view is also needed. This should also help to attract and retain outstanding employees, as well as to bring SOI closer to its customers.
- [51] See Bailar, Barbara A., "Contributions to Statistical Methodology from the U.S. Federal Government," *Survey Methodology*, Volume 16, Number 1, pp. 51-57, 1990.
- [52] See Tanur, Judith M. and Fienberg, Stephen E., "Cognitive Aspects of Surveys: Yesterday, Today, and Tomorrow," *Journal of Official Statistics*, Volume 8, Number 1, pp. 5-17, 1992.
- [53] See Deming, W. Edwards, *Out of the Crisis*, MIT Press, 1986; and Juran, Joseph M., *Juran on Planning for Quality*, Macmillan, 1988.
- [54] See "Total Quality Organization (TQO) Plan: Objectives, Plans and Milestones," unpublished working paper, Statistics of Income Division, Internal Revenue Service, 1992.

[55] See, for example, Womack, James P.; Jones, Daniel T.; and Roos, Daniel, *The Machine That Changed the World*, Harper Perennial, 1990.

[56] See footnote 52.

[57] The effect of filing extensions, in the case of corporations, is more complicated because over half of all corporations file for noncalendar year accounting periods; also, corporations are allowed a second extension of time in which to file (under certain conditions). Because of the incidence of noncalendar year reporting, the definition of the "tax year" adopted for SOI covers accounting periods ending in July of one calendar year through June of the following calendar year. Therefore, after taking filing extensions into account, the last timely filed return for "Tax Year 1992," for example, is not received until March 1994.

[58] See "Public Release of SOI Information," *Statistics of Income Bulletin*, Fall 1992, Volume 12, Number 2, pp. 3-4, 1992.

[59] Spruill, Nancy, "The Confidentiality and Analytic Usefulness of Masked Business Microdata," *1983 Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1984. For related research on this topic, see Dalenius, Tore, *Controlling Invasion of Privacy in Surveys*, Statistics Sweden, 1988; Duncan, George and Lambert, Diane, "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics*, Volume 7, pp. 207-217, 1989; Paass, Gerhard, "Disclosure Risk and Disclosure Avoidance for Microdata," *Ibid.*, Volume 6, pp. 487-500, 1988; and Rubin, Donald B., "Comments on Confidentiality—A Proposal for Satisfying All Confidentiality Constraints Through the Use of Multiply-Imputed Synthetic Micro-Data," unpublished paper, 1991.