Exporting a Statistical System: Towards Establishing a Tax Statistics Function in South Africa

Tom Petska, Internal Revenue Service

The U.S. income tax system has a long history since enactment of the 16th amendment in 1913 and subsequent law requiring the annual publication of statistics on its operations [1]. This responsibility established the Internal Revenue Service's Statistics of Income (SOI) function. Despite many revisions to the Internal Revenue Code, this requirement continues to this day [2].

With the end of the apartheid era and the initiation of open elections in 1994, the Department of Finance of the Republic of South Africa requested assistance though the U.S. Departments of State and Treasury to initiate a tax statistics function which would enable microsimulation modeling for tax policy analysis and revenue estimation. This paper reports on the experience of advising South African (S.A.) government officials in the Department of Finance's Tax Policy Chief Directorate (TPCD) and the South African Revenue Service (SARS) on the technical and resource needs to initiate this function. The paper uses the U.S. SOI system as a potential "model" and applies this model to the unique features of the South African system.

This paper is organized as follows. In the first section, issues concerning principles, practices, and mission of a tax statistics function are presented. Next, resources and organizational placement are discussed, followed by an overview of operational functions. In later sections, benefits to the revenue service are examined; plus concluding thoughts on progress toward developing the system in South Africa are addressed.

◆ Principles, Practices, and Mission

In attempting to build a statistical function where none had existed, it is important to develop a strong foundation. In this regard, the U.S. Committee on National Statistics (CNSTAT) published in 1992 *Principles and Practices for a Federal Statistical Agency*, with the assistance of several U.S. Federal statistical agencies, which provided a concise but well-articulated "blueprint" for statistical agencies [3]. Although it is now 8 years old, its thoughtfulness and thoroughness are substantiated by the fact that, when the U.S. Federal statistical agencies were recently given an opportunity to update the document, few had many changes.

Principles and Practices addressed three necessary ingredients of Federal statistical agencies, those being: relevance to policy issues; credibility among data users; and trust among data providers and subjects. Statistical practices were the most detailed section of the book, which specified the necessary ingredients for statistical agencies, including a clearly defined mission, a strong measure of independence, cooperation with data providers and users, wide dissemination, and caution in conducting non-statistical activities.

The SOI mission is to collect and process tax return data so that they become meaningful information and to disseminate this information to its many customers. The primary customers are the U.S. Department of the Treasury's Office of Tax Analysis (OTA) and its Legislative Branch counterpart, the U.S. Congress's Joint Committee on Taxation (JCT). A third major customer is the Department of Commerce's Bureau of Economic Analysis, an organization responsible for maintaining the U.S. national income and product accounts. SOI has many other customers, including academic researchers, policy analysis "think tanks," Congress, libraries, and the general public.

Although there is an immediate need in South Africa for an SOI-like staff to improve revenue estimation and policy analysis, the mission of "SOI/S.A." should be broad-based and include widespread data dissemination. Statistical publications and electronic data dissemination on the Internet are highly recommended to develop a wide range of users, as well as to provide the citizens of South Africa with information on the functioning of their tax laws. These activities can contribute to a dialogue on the equity of the tax system, which should result in increased critical review as well as improvements in perceptions of balance and fairness.

◆ Resources and Organizational Placement

The annual budget of the SOI program is about $30 million, consisting of nearly 500 staff years plus equipment, training, travel, and overhead. The organizational placement of SOI is in the Internal Revenue Service, whose function is tax administration and collection. When SOI was established (over 80 years ago), it was primarily a clerical operation, with statistical summaries compiled manually. Later, even with the initiation of the IRS Master File System and computerized sampling, statistical abstraction and editing of tax returns designated for SOI samples remained very labor-intensive.

Because of SOI's organizational placement, special care is needed to ensure that the extensive data needs of OTA and JCT are adequately met and that the organizational priorities of SOI are not diverted to other IRS operational priorities. In this regard, senior executives of the Treasury and the Joint Committee on Taxation meet periodically with their counterparts in the Internal Revenue Service to ensure an adequate level of budget support and to establish the priorities for the SOI function.

Concerning the organizational placement in South Africa, the two principal reasons from the U.S. experience are largely absent. First, with the current capabilities of both computerized sample selection and data editing, the dependency on a substantial clerical function can be avoided if the South African New Income Tax System (NITS) can be used as a reliable sampling frame and source of high-quality financial data. Further, the likelihood of establishing a U.S.-size function, numbering in the hundreds of staff years, is unlikely.

With the development of capabilities to sample and edit tax data, the SOI/S.A. function would essentially be the link between the TPCD and SARS. This "link" would be staffed by:

- ❒ *Survey statisticians*, who develop sample designs and monitor the execution of those designs;
- ❒ *Population file programmers*, who would sort and stratify the population files and select and extract samples; and
- ❒ *Economists/tax law specialists*, who would direct logistics, provide subject-matter expertise, and analyze and publish findings.

The TPCD and SARS are in different departments of the South African national government, and current law does not clearly establish sharing tax data by SARS to the TPCD. Therefore, regardless of organizational placement, full cooperation, accompanied by complete sharing of all SARS data, is required.

◆ Operational Functions

Statistical operations require a structured and disciplined approach, consisting of planning and design, sampling and estimation, data abstraction and editing, and dissemination and publication.

Planning and design--Planning consists of communicating with study sponsors and customers to gain content needs and operationalizing this into a workable design. As a planning tool, sample size and item content need to be integrated in a study plan that can be realistically accomplished by available resources.

In many instances, there is a tendency to put little work into planning and to progress quickly to implementation. However, this can lead to problems, as staff can become over-committed and completions of project functions do not converge. As a result, a conscious planning effort is highly recommended to clarify roles and responsibilities, as well as to anticipate potential bottlenecks.

Sampling and estimation--Statistics compiled for SOI studies are generally based on stratified probability samples of tax or information returns. As returns are processed into the IRS Master File systems, they are assigned sampling classes, based on criteria such as size of income or assets (or other measures of economic size), industrial activity, accounting period, or the presence of certain supplemental forms or schedules.

Each taxpayer, whether an individual or a business, has a unique number--the Social Security Number (SSN) for individuals or the Employer Identification Number (EIN) for businesses. These unique taxpayer identification numbers (TIN's) are used as the seed for a pseudo-

random number which, along with the sampling strata, determines whether a given return is to be selected for the SOI sample [4]. The probability of a return being designated for the SOI sample depends on the sampling rate prescribed for its sample class or stratum and may range from a fraction of 1 percent to 100 percent.

The U.S. system has three clear advantages over the S.A. tax system concerning development of a tax statistics operation. These advantages include the following:

- The presence in the U.S. system of unique and (for the most part) unchanging Taxpayer Identification Numbers;
- Coverage through tax returns of high percentages of the study populations; and
- Relatively shorter tax return filing extensions.

It is highly recommended that the S.A. tax system adopt unique and unchanging taxpayer identification numbers for sampling, as well as other operational and research purposes. Unique and unchanging TIN's would facilitate file matching, as well as benefit sampling.

Concerning coverage, certain features of the South African "Pay as You Earn" (PAYE) system complicate the situation. The innovative PAYE system is a process that removes the filing burden for many low income, South African taxpayers with only wage or salary income by allowing employers to adjust tax withholdings to exactly equal tax liabilities. Such taxpayers would thus not have to file tax returns.

To construct a statistical profile of the population of South African individual taxpayers, it was envisioned to integrate PAYE data with data from tax returns to create a statistical "model" of all current (and potential) taxpayers. However, income and tax liabilities *for each individual* are not currently available in the PAYE system--only aggregate tax payments from employers are available. This limitation restricts the ability to construct a statistical profile of *all individuals*. So, until an alternative means is determined, individual income tax analysis will have to be confined to the tax filing population.

In the U.S. SOI system, sampling periods are generally kept open for 8-12 months after closure of the final ending accounting date to ensure the inclusion of late filed returns, which are often atypical. Since the U.S. system uses a calendar-year basis for individual tax returns, the sample period is kept open until December 31.

The situation is further complicated for corporation taxpayers, because many have staggered fiscal year accounting periods. In the U.S. SOI corporation program, corporations are included in Tax Year 1999 if their accounting period ends between July 1, 1999, and June 30, 2000. However, corporations frequently request filing extensions. So, to keep open the sampling period for 12 months after the ending accounting period would require sampling until June 30, 2001.

In the U.S., extensions to file tax returns are granted quite readily for 4 months, but less so for longer durations, and estimated tax payments may be required. In South Africa, 15-month extensions are frequently granted. Since SOI studies are a compilation of information *for one tax period*, such extensions are problematic. For example, for any tax year, such delays could delay file completion for nearly 4 years after much of the financial activity.

From exploratory tabulations of S.A. corporate data, the distribution of multiple tax years filed within a processing year is very evident. How to address this issue is not clear, since this could delay the completion of a file for any tax period for an unacceptably long duration.

Finally, whether or not to sample or use the entire population is based in part on the available computer platform. Recently, we were able to access and tabulate complete population file extracts for corporations, so it is not clear that sampling from population files is a necessity, although resources needed to edit the data enter into this issue. This is addressed in the next section.

Data abstraction and editing--In the U.S. SOI system, data items for sampled cases from the master file system are copied into a minicomputer network, where data content is substantially augmented with additional items

manually extracted from tax returns. Statistical abstraction can take as little as a few minutes for a simple return, to as long as several days for a large corporate return. This editing system uses on-line transaction processing, so that all data capture operations are completed in a single pass. One editor is thus responsible for ensuring the validity of all data processing for a given return.

In order that final statistics are consistent and reliable, SOI economists and subject-matter experts have developed extensive on-line tests and error correction procedures that are applied to each sampled return. These tests and correction procedures are based on the structure of the tax laws, generally accepted accounting principles, and the improbability of various data combinations. Subsamples of returns are independently reprocessed and analyzed for a quality evaluation.

An operational goal is to test every data item and code for reasonableness, as well as its relationship to other items. If any such relationship is not upheld, some form of edit is usually made.

The SOI data editing systems are thus very labor-intensive; few SOI studies have been completed without substantial manual data abstraction and editing. In addition, there is a reluctance to change to less labor-intensive processes, since the tax data are complex, and overall data quality is heavily dependent on the accuracy of the editing process. In general, the need for manual abstraction and editing is dependent on two issues:

- Is the statistical item content key entered from the population files adequate for analytical and revenue estimation purposes?
- Does the level of quality and complexity necessitate extensive manual editing and review?

Concerning item content, it appears that the South African NITS system data will have an acceptable level of data content. Concerning data quality and complexity, the situation is less clear. In the U.S., many relatively low-income individual income returns are quite simple, with limited income types and other taxpayer-reported characteristics. For such cases, an automated or high-level system of outlier review and edit would, in all likelihood, yield reasonable results.

But as complexity increases, this may not be the case. In the U.S. system, large corporation returns, often with multinational financial operations, are extremely complex, and hundreds of hours are spent reviewing and correcting these data, even after initial abstraction and editing.

In the S.A. system, exclusive use of NITS data in place of large-scale manual data abstraction is a reasonable way to begin, at least for individual returns. But even for these data, a series of "structured queries" and consistency tests would need to be developed to detect and correct data relationships that were deemed to be incorrect. Limited samples of tax returns could be selected to improve the knowledge of taxpayer reporting problems, and tax returns for limited subsamples of large or complex cases could be accessed to help in editing. Ideally, subject-matter experts should develop "edit rules" to be deployed in full-scale studies.

Many agencies in the U.S. and elsewhere have begun to develop automated data edit systems, building an "artificial intelligence" knowledge base. Most such examples have started with fairly simple returns, where the editing relationships can be specified to handle most cases.

Since data editing can be resource-intensive, whether or not to sample or to use population files is dependent on the editing methods used. Clearly, if the *quality* of NITS data is acceptable, so that most data editing can be accomplished without acquiring substantial volumes of tax returns, and if an acceptably large computer platform to handle the population files is available, there is really no need to sample.

This is an empirical question, which can only be answered by examination of the NITS data for all types of tax returns. At the Pretoria, S.A. Receiving Center, individual tax return data are key-entered twice, and, if discrepancies are present, they are manually reviewed and resolved. However, at the Ramburg Receiving Center, a center focused exclusively on large company tax

returns, no such double-key system is in place. This was quite surprising, considering the complexity and importance of these returns, and it raises a concern that acceptable data quality for these records may require substantial review and correction.

Dissemination and publication--U.S. SOI information is made publicly available through both printed publications and electronic media. The *Statistics of Income (SOI) Bulletin* is published quarterly, with each issue containing four to eight articles and data releases plus historical tables covering tax collections, taxpayer assistance, and tax return projections [5]. Separate "complete reports" on individual and corporation income tax returns, as well as a corporation source book, are also published annually [6-8].

Periodically, special compendiums of research and analysis, covering such topics as nonprofit organizations, estate taxation, and international business activities, are published. Research articles documenting methodological and analytical issues are also published in a series of annual reports [9].

SOI has expanded information dissemination through its Internet worldwide web site, providing users a quick and easy option for accessing SOI data. At present, 40,000 files are downloaded monthly from this site [10].

An SOI/S.A. should extensively publish data on taxation. In addition to SOI-like financial summaries, statistics on compliance, tax processing, and auditing should also be published, as is now the case in the *IRS Data Book* [11].

◆ Benefits to the Revenue Service

In the U.S., the SOI function is primarily focussed on the tax policy needs of OTA and JCT. However, SOI and its sister agency, the IRS Research Division, both have substantial roles in assistance within the IRS.

The SOI data system is used as an early warning in IRS. In addition to the annual individual taxpayer study, SOI has constructed a small individual sample study that shows weekly reporting trends in the primary filing season (January 1-April 15). This study, the Taxpayer Usage Study (TPUS), uses a separate sample of individual returns and reports on characteristics of the individual taxpayer population, such as use of paid tax return preparers and the reporting of certain forms, schedules, or items, especially those that are new for the year [12].

The "publication expertise" developed in SOI has led to taking over the responsibility for producing the IRS annual *Data Book* [11]. This publication, once known as the *Commissioner's Annual Report*, has extensive tabular information on the processing of tax returns and the compliance and audit processes.

SOI has also developed a small staff (approximately 10) of mathematical statisticians whose role is to provide statistical direction, guidance, and support within IRS. This section serves as resident consultants for non-SOI areas in the IRS on a wide variety of statistical issues. They provide guidance on systems and sample design, statistical analysis and estimation, quality measures, customer satisfaction surveys, and cognitive research. Their projects currently include measuring employee satisfaction, alternative methods of filing, customer service satisfaction programs, and remittance processing strategy studies.

In addition, SOI samples have been used as screening devices in the audit process. SOI has cooperated with the IRS examination function for many years, mainly in the form of providing sample files of domestic and multinational corporations. In recent years, this process was expanded to include small, unincorporated businesses. The SOI samples were closely examined to ascertain the reporting characteristics of different types of business--by size, industry, and profitability. However, special care must be taken not to let SOI-sampled cases be targeted for audits, since these would bias the "representativeness" of the samples.

For many years, on a cycled basis, IRS undertook a Taxpayer Compliance Measurement Program audit study to measure the overall level of compliance and the size of the tax gap. For any return subjected to these audits, the taxpayer would have to justify each entry on his or her tax return. Upon completion of the audits, before

and after audit amounts are compared for each return in the sample, and the sample is weighted to population totals to estimate the tax gap and to help develop criteria for operational audits. This program has been on hold for over 10 years, mainly because the audits were perceived as intrusive. As a result, the reliability of the estimates of the tax gap has diminished.

SARS could benefit from periodic individual statistical studies, and a TPUS-like system could be an early warning on reporting characteristics. And the SOI sample designs could be used as a first approximation for compliance samples. But, as previously noted, the actual sampled cases for the SOI measurement system cannot be selected for audit, unless for those strata where the sample rate is 100 percent. Alternately, the statistical expertise developed in an SOI function could be used to independently design audit samples.

◆ Final Comments

The development of an SOI function would be highly beneficial to the tax policy and revenue estimation functions in South Africa, as well as provide other benefits to SARS. Without such a function, the complicated processes of selecting reliable, stratified random samples from the population files, editing these data, and weighting them to population totals have a high likelihood of failure.

Until an SOI function is developed, revenue estimators in the Department of Finance will have to make *ad hoc* data requests from SARS staff. These requests, which would have to be for tax return population or sample files for individuals, trusts, corporations, and closely-held businesses, would require guidance on how to interpret SARS file record structures and processing idiosyncrasies.

Tax systems are very complex, so construction of a disciplined measurement system can only be accomplished if all important aspects are addressed. SOI's successes are attributable to both a very capable staff as well as to development of mathematically proven and relatively stable systems with regular and continuous incremental improvements. It is clearly an investment in information infrastructure.

In most areas of SOI, first attempts at selecting new, complex samples have met with only marginal success at best. But processes were improved, corrections were made, and the systems became more and more reliable. This was not only accomplished from incremental, technical improvements but also from stable staffing and human capital development.

To launch an SOI-like function in South Africa, the system should start small and build on success. Since the system is truly a bridge between the very different roles of the tax policy analysis and tax compliance processing and collection, it would need an interdisciplinary group of talented and highly trained statisticians, programmers, and economists. Not only should all members be fully cognizant of each other's areas of specialization, but each should also fully understand the intricacies of SARS processes and its impacts on the samples, as well as the uses of the files in microsimulation modeling and analysis.

◆ Acknowledgments

Special thanks to members of the South African Department of Finance's Tax Policy Chief Directorate and the South African Revenue Service, particularly Martin Grote, Iuliana Clayton, and Pumla Bam, for their assistance in this work. Additional thanks to current and former colleagues in the U.S. Department of the Treasury's international advisory program, Bob Klayman, Sam Thompson, Selcuk Caner, Susan Hinkins, and Fritz Scheuren, for their outstanding assistance in this effort, and to Beth Kilss and James Dalton for their review of the manuscript. Any errors that remain are the responsibility of the author.

◆ Notes and References

[1] Portions of this paper have benefited from prior overviews of the SOI function written by Fritz Scheuren and Tom Petska for *Business Economics* (1992); the *Proceedings of the National Tax Association* (1992); and the *Journal of Official Statistics* (1993). The most recent version published is in *Turning Administrative Statistics Into Infor-*

mation Systems (1994) by Tom Petska, and it was recently updated, expanded, and placed on the SOI Internet site by James Hobbs: http://www.irs.gov/tax_stats.

[2] See Internal Revenue Code, Section 6108(a).

[3] *Principles and Practices for a Federal Statistical Agency*, Martin, Margaret E. and Straff, Miron L. (editors), Committee on National Statistics, National Academy Press, 1992.

[4] The algorithm used for generating the taxpayer identification number (TIN) transform generally stays the same from year to year. This longitudinal character of the sample design improves the estimates of change from one year to the next.

[5] *Statistics of Income (SOI) Bulletin*, Publication 1136, Internal Revenue Service.

[6] *Statistics of Income--1997, Individual Income Tax Returns*, Publication 1304, Internal Revenue Service.

[7] *Statistics of Income--1997, Corporation Income Tax Returns*, Publication 16, Internal Revenue Service.

[8] *Source Book of Statistics of Income--1997, Corporation Income Tax Returns*, Publication 16, Internal Revenue Service.

[9] *Statistics of Income: Turning Administrative Systems Into Information Systems, 1999*, Publication 1299, Internal Revenue Service.

[10] The SOI Internet website, Tax_Stats, can be accessed at http://www.irs.gov/tax_stats.

[11] See, for example, *1998 Data Book*, Publication 55B, Internal Revenue Service. The responsibility for this publication was transferred to the SOI Division in 1999.

[12] Sailer, Peter and Parisi, Michael, "Taxpayer Usage Study, 1997," *Statistics of Income (SOI) Bulletin*, Volume 18, Number 1, Summer 1998, Publication 1136, Internal Revenue Service. ■