Statistics of Income

2007 SOI Paper Series

Improving the Quality of U.S. Tax Statistics: Recent Innovations in Editing and Imputation Techniques at the Statistics of Income Division of the U.S. Internal Revenue Service

by Scott M. Hollenbeck, Melissa Ludlum, and Barry W. Johnson

Improving the Quality of U.S. Tax Statistics: Recent Innovations in Editing and Imputation Techniques at the Statistics of Income Division of the U.S. Internal Revenue Service

Scott M. Hollenbeck, Melissa Ludlum, and Barry W. Johnson, Internal Revenue Service

The Statistics of Income (SOI) Division of the United States Internal Revenue Service (IRS) is charged by the U.S. Congress with preparing and publishing statistics on the U.S. tax system. SOI was established in 1916, soon after the adoption of a Federal income tax, and the first SOI report, based on tax returns filed by individuals and corporations, was released in 1918. Early SOI reports were used primarily by the U.S. Treasury Department, the Congress, and the Commerce Department for tax research, estimating revenue, and constructing the National Income and Product Accounts. As SOI programs and products have expanded, users in other Government agencies, academic researchers, the media, and the general public have come to rely on tax data produced by SOI for studying the U.S. economy and evaluating tax policy initiatives (see Wilson, 1988, for a more complete history of the SOI program).

In order to fulfill its directive, SOI has created a structured system for transforming administrative data into statistical files, using its own data collection systems, wholly autonomous of main IRS tax return processing. SOI annually conducts approximately 110 different projects involving data collection from tax returns and information documents. Project content is developed by working closely with data users to ensure both continuity and utility. Teams of SOI economists, computer specialists, statisticians at SOI headquarters in Washington, DC, and specially trained employees located in IRS submissions processing centers in Georgia, Missouri, Ohio, Texas, and Utah work together to extract and perfect information from tax documents in order to create statistically valid data. For most studies, data are extracted from stratified random samples of returns as they are filed.

This paper will provide an overview of SOI data collection systems, focusing on three main programs—studies of individual income tax returns, corporate income tax returns, and information returns filed by tax-exempt charities and private foundations. It will briefly outline the three programs and highlight recent innovations, including the use of digital images as source documents and the integration of electronically filed tax return information with data provided on traditional, paper returns. The paper will also discuss procedures used to impute record-level data for returns that were selected for SOI samples but unavailable for processing, as well as detail the challenges and benefits of automating the statistical processing of certain electronically provided data. Finally, SOI's use of imputation to approximate values for missing data items will be discussed.

▶ SOI Individual, Corporate, and Tax-Exempt Programs

SOI conducts annual studies of returns filed by individuals and corporations to report and pay income taxes, as well as information returns filed by tax-exempt organizations. The SOI individual income tax program includes information reported on Form 1040 and its attachments (see Internal Revenue Service 2006b), while the corporation income tax program includes information from Form 1120 and its attachments (see Internal Revenue Service 2006a). SOI studies of tax-exempt organizations include information captured on Forms 990 and 990-PF filed by charities and private foundations, respectively. These organizations operate for charitable purposes, such as those that are religious, scientific, literary, or educational, and are exempt from Federal income tax, but are required to file information returns annually with the IRS that detail asset holdings, revenue, and expenses (see Arnsberger, 2006; Ludlum and Stanton, 2006). For each of these SOI studies, a stratified random sample of returns is selected based on a variety

of return characteristics, using information captured on the IRS Masterfile during administrative processing.[1]

In producing statistical files from tax return information, SOI employs state-of-the-art computer technology and rigorous data perfection procedures. Custom data collection applications, using Graphical User Interface (GUI) technology, are designed to correct taxpayer errors, reduce nonsampling error, and minimize data collection costs. For most studies, certain core data items extracted from the IRS Masterfile are preloaded for each sampled return. Specially trained workers, known as "editors," then transcribe and code additional information from the returns, schedules, and attachments. They also modify taxpayer reported data as needed in order to ensure that the data conform to SOI customer specifications. Data are automatically validated as they are entered, using computer validity checks to verify coded values and key mathematical relationships. In most cases, editors are required to resolve potential errors identified by these checks before entering additional data. To monitor overall data quality, subsamples of edited returns are subjected to item-by-item quality review. Finally, subject-matter experts carefully review all files for accuracy before releasing them to customers.

▶ Recent Technological Innovations

Advances in computer technology that have transformed almost every aspect of daily life in the U.S. have also had a tremendous impact on both IRS and SOI operations. Paper documents submitted to the IRS can now be displayed and transmitted in a "paperless" environment, and electronic data provided to the IRS are beginning to replace traditional paper tax and information returns. Improved software systems and increased computer-processing capacity have allowed SOI to expand interactive testing of data, providing editors with instant feedback when money amounts or codes are inconsistent with pre-established editing rules. The use of sophisticated editing tools, online dictionaries, and calculators has also greatly expanded. Increased storage capacity has allowed SOI to use data reported for previous tax periods to validate current-year values. The addition of digital dashboards, which provide current statistics on inventory, productivity, and quality, has helped improve the management of many SOI studies. The following section discusses the impact of two important advances, digital imaging and electronic filing, on SOI programs.

SOI Return Imaging and Split-Screen Edit Systems

In 1998, SOI began producing digital images of tax returns and information documents.[2] Gradually, SOI expanded this operation to include all other tax forms and information documents that historically had been microfilmed and stored for research and data correction or validation. SOI currently images more than 30 different IRS forms. For some forms, the entire population is captured digitally, while, for others, only returns selected into SOI samples, or those with select characteristics, are imaged. Depending on the type of return, the images are made available to a wide range of users, including SOI staff, other IRS functions, the U.S. Congress, the U.S. Treasury Department, and, in the case of tax-exempt organizations, the general public. In 2006, SOI imaged over 71.5 million tax and information return pages.

Digital Tagged Image File Format (TIFF) images have provided SOI with several opportunities to improve the quality and efficiency of its operations. One such innovation is the transition from single-view GUI Oracle-based data editing applications to more sophisticated split-screen systems. These systems display an electronic copy of the tax return on one side of a 24-inch, wide-aspect monitor and the GUI data editing application on the other. The return image is displayed at full size, although editors are able to use zoom features to magnify the image. In addition, the image and the editing system are synchronized, meaning that, as data are collected or verified and the editor scrolls or moves to new data entry screens, the application automatically

[1] The IRS transcribes selected data items during initial processing of all tax and information returns. These data are used for administrative purposes, such as verifying tax computations and recording payments. Collectively, these data are referred to as the IRS Masterfile in this paper.

[2] Initially, this work was done in partnership with the Tax-Exempt and Government Entities business unit at IRS and The Urban Institute, a Washington, DC, research organization, to fulfill IRS regulations, which require that information returns filed by nonprofit institutions be made available to the general public.

changes the view of the return presented. Editors are given both online and paper copies of data editing and error correction instruction manuals.

The first SOI project to take full advantage of digital images was the Private Foundations study. SOI analysts and computer specialists developed a system that uses Adobe Acrobat software to present Portable Document Format (PDF) images of the tax return, created from the TIFF files. Editors are able to use all standard features of Adobe Acrobat Reader to view and manipulate the images. While field personnel were initially apprehensive about working with images, rather than paper documents, when surveyed 5 months after the system was introduced, most felt that their work was more enjoyable in the split-screen environment and advocated the adoption of this technology by other SOI projects. Figure 1 shows that, while productivity and data quality diminished somewhat when the split-screen system was introduced, these statistics quickly rebounded to levels comparable with paper processing. After the success of this project, split-screen applications were developed for other SOI studies, including those of public charities, tax-exempt bonds, and, most recently, corporation income taxes. Several other split-screen applications are in various stages of development.

Figure 1: Tax-Exempt Organization Studies Production Statistics

Fiscal year	Document format	Returns per hour	Accuracy rate*
Private foundations			
2004	Paper	5.9	99.9
2005	Images	4.8	99.7
2006	Images	5.7	99.9
Charities			
2004	Paper	4.5	99.8
2005	Paper	4.5	99.7
2006	Images	4.8	99.8

*Accuracy rates are calculated based on data from quality review samples and represent the number of error-free data items divided by the total number of data items collected.

While the use of digital images has not had a substantial impact on the speed or quality with which data are input, use of this technology has had important effects in other ways. Significantly, because imaging allows SOI to process paper returns quickly, SOI's impact on other areas of IRS that work with paper documents has been minimized. The availability of images has also reduced the number of missing returns, which are returns that were selected for an SOI study but were not available to SOI, usually because they were controlled by another IRS function This has reduced the need to impute returns or make sample weight adjustments to reduce sampling bias.

Perhaps the greatest benefit of working with digital images has been the increased availability of documents to a geographically disbursed work force. Economists and statisticians in Washington, DC, and editors in field locations, such as Ogden, UT, are able to view documents simultaneously, greatly simplifying problem resolution and eliminating the need to mail or fax sensitive information between offices. Images are also helpful during the post edit phase of data collection. Once data have been collected for the sample of returns selected for a study, economists and statisticians further test and analyze these data before providing them to customers. In the past, when errors were suspected, paper documents were ordered from the IRS files function and sent to Washington. This time-consuming and costly process effectively limited research to only those documents that appeared to contain the most significant errors. Access to digital images for the entire SOI sample of returns for some studies has allowed analysts to look at many more returns, improving final data quality.

Electronic Filing and the Tax Return Database—SOI Individual Income Tax Study

Improved communication technology and broad dissemination of computer technology in the U.S. have allowed IRS to expand its capacity to receive return information from filers in electronic, rather than paper, formats. In 1986, IRS introduced a pilot electronic filing program, allowing certain taxpayers in three U.S. cities to file their individual income tax returns electronically, via licensed "transmitters," resulting in about 25,000 submissions. In 1992, the IRS achieved another milestone by allowing taxpayers to e-file these

returns from home computers, with more than 125,000 individual income tax filers participating in a pilot conducted in the State of Ohio. IRS annually has expanded the individual income tax "e-file" program to include additional forms and schedules so that approximately 98 percent of all individual income tax filers were eligible to file electronically by 1994, with the eventual goal of enabling electronic filing for all taxpayers. Electronic filing of individual income tax returns grew to more than 73 million in Calendar Year 2006 (see Figure 2). In 2007, electronic filers whose reported Tax Year 2006 adjusted gross income was less than $52,000 were eligible for free electronic filing through selected software vendors. For those with larger incomes, this software could be purchased from commercial vendors or returns could be electronically filed by most paid tax return preparers, usually for a nominal fee.

Figure 2: Electronically Filed Individual Income Tax Returns

(Numbers in millions)

Fiscal year*	2004	2005	2006
Total filing population	131.3	132.8	133.9
Number e-file, population	61.5	68.5	72.8
Percent e-file, population	46.8%	51.6%	54.4%

* Fiscal year runs from October 1 through September 30

Source: Internal Revenue Service Data Book (2004, 2005, 2006) Publication 55B

Electronic filing provides several benefits to taxpayers, including convenience, faster refunds, and accuracy. Taxpayers can file returns using a number of convenient and expedient methods, including from their home computers or via their tax preparers. IRS issues refund checks for electronically filed returns more quickly than for those that are paper-filed. Generally, IRS issues a refund check within 3 weeks of acknowledging an electronically filed return; refund checks for paper returns are issued within 6 weeks of receipt.[3] In addition, electronically filed returns are less likely to contain errors, due to embedded mathematical tests and program logic that automatically provide the proper additional forms and schedules based on information entered into the program by the filer. As a result, electronically filed returns are 99-percent less likely to generate any correspondence with IRS submissions processing personnel.[4]

Electronically filed individual income tax return data that are transmitted to the IRS are currently stored in the Tax Return Database (TRDB). This data set contains all of the data items provided for each return, as opposed to the more limited number of data items retained on the IRS Masterfile. Data derived from the TRDB have become an important component of SOI's annual individual income tax studies. Traditionally, Masterfile data were combined with extensive data extracted manually from source documents by SOI editors, to produce a file containing nearly 2,000 variables. With the introduction of e-file data to the SOI individual income tax program, all data items available from the TRDB for e-filed returns, as well as data items from the Masterfile, were preloaded to the SOI editing system. SOI editors would then validate the data by manually triggering validation tests and then making any necessary corrections. This reduced cost by decreasing the time it took to process the e-filed returns, since the editors did not have to transcribe data for these returns, just validate them. As the IRS has expanded the number of forms and schedules that can be e-filed, SOI data transcription costs have decreased significantly.

Although much of the data from e-filed returns can be easily validated using the consistency tests embedded in the SOI edit system, not all line items can be handled in this manner. One of the added benefits of SOI statistical data files over administrative data collected by the IRS when returns are received, is that during SOI processing, some data items are reassigned, or reallocated, from the way they are originally reported by the taxpayer. For example, in some instances, the computer programs that are used to e-file Form 1040 allow the taxpayer to report several similar items on

[3] Source: IRS Web site, http://www.irs.gov/newsroom/articl/0,,id=108001,00. html.

[4] Source: IRS Web site, http://www.irs.gov/efile/article/0,,id=118450,00.html.

a single line of the tax form, but SOI customers often require separate data fields in order to distinguish among different types of information. "Other income" is a data item that typically must be reallocated to new fields by SOI editors. Taxpayers are allowed to make multiple entries for various "other income" sources on Form 1040, line 21, all of which are stored together on the TRDB, along with brief text descriptions for each separate money amount. In addition to amounts properly reported as other income on line 21, some taxpayers improperly allocate income amounts to this line that should have been reported elsewhere on the return. In both cases, SOI editors will reallocate these amounts to different SOI data fields, based on the descriptions, in order to provide a more accurate picture of a taxpayer's income source. The result for SOI customers is a file of data from electronic and traditional paper sources that is as consistent and as accurate as possible. Recently, SOI has automated much of the processing of e-filed individual income tax data using a combination of validation tests and data correction and imputation procedures described later in this paper.

Modernized Electronic Filing—SOI Businesses and Tax-Exempt Entities Studies

Electronic filing has gradually spread from forms related to the individual income tax to other types of tax and information documents processed by the IRS. In 1987, IRS introduced e-filing for certain business income tax returns. A major milestone for the IRS was the introduction of Modernized Electronic Filing (MeF) in 2004. Unlike the earlier system, which collected only numeric and character data strings and stored the information in traditional databases, MeF, based in Extensible Markup Language (XML), collects both taxpayer data and information tags.[5]

With the advent of MeF, the IRS greatly expanded its capacity for accepting electronically filed return data, including tax returns filed by businesses and corporations, as well as information returns filed by tax-exempt organizations. In 2005, the IRS mandated that certain types of filers submit their tax returns and information

Figure 3: Selected Electronically Filed Returns, by Fiscal Year

Type of tax return	2004		2005		2006	
	Number	Percent of total	Number	Percent of total	Number	Percent of total
Estate and trust income tax	1,328,445	35.6%	1,350,186	36.6%	1,360,876	36.8%
Corporation income tax	12,477	0.5%	51,224	2.1%	136,311	5.6%
Small corporation income tax	35,053	1.0%	149,704	4.1%	389,133	10.2%
Partnership income tax	91,159	3.6%	107,571	4.0%	274,721	9.9%
Tax exempt organizations	465	0.1%	3,228	0.4%	11,115	1.3%

* Fiscal year runs from October 1 through September 30

Source: Internal Revenue Service Data Book (2004, 2005, 2006) Publication 55B

Figure 4: MeFile Returns in SOI Samples, by Tax Year

	2004			2005		
	MeF	Total sample	Percent	MeF	Total sample	Percent
Tax-exempt organizations	223	21,700	1.0%	3,700	27,500	13.5%
Private foundations	10	11,450	0.1%	210	11,500	1.8%
Corporations	2,702	146,269	1.8%	23,000	112,400	20.5%

[5] XML allows developers to set standards for the types of information that should appear in a document, and in what sequence, making it possible to define the content of a document separately from its formatting. This simplifies the task of reusing the content in other applications but also allows for the recreation of the look and feel of a traditional paper document if desired. XML also provides a basic syntax that simplifies the process of sharing information between different kinds of computers and different applications.

documents electronically. For tax years ending on or after December 31, 2006, all large corporations, those with total assets of at least $10 million and that filed at least 250 annual Federal returns (including all excise and employment tax returns, as well as wage and income statements that must be filed for each employee) are required to file their corporate income tax returns (Form 1120) electronically. Organizations are primarily required to provide data in XML format, although some PDF documents are allowed, and organizations may apply for an exemption to the rules if they are able show that the rules impose an undue technological or financial burden. For SOI studies of Form 1120, the relatively large organizations subject to the new requirement constitute a significant portion of the annual sample.

Beginning in 2005, the IRS established a mandatory schedule for electronic filing of Forms 990 and 990-PF by charities and private foundations, similar to that imposed on corporations. For tax years ending on or after December 31, 2006, all public charities with $10 million or more in assets that file at least 250 returns annually, and all private foundations and nonexempt charitable trusts, regardless of asset size, that file 250 or more returns annually are required to file electronically. Again, these relatively large organizations represent significant portions of SOI samples of charities and private foundations. Figure 3 shows the growth in electronic filing for selected entities.

Integration of XML Data

Beginning in 2006, SOI, working with other functional areas in IRS, developed programs to render return images from the MeF data in XML format by converting the data to TIFF images stored on SOI's computer network. These images are then available for SOI processing and are being seamlessly integrated with scanned digital images of traditionally filed returns in SOI split-screen data collection applications. There are subtle differences between the rendered and traditional images, the most significant being that electronically filed returns generally have fewer attached supporting documents, such as balance sheets, appraisals, and income statements, than returns filed on paper. Figure 4 details the number of MeF returns that were included in selected SOI study samples.

For corporate and tax-exempt returns filed electronically, SOI is now working to extract statistical data directly from the XML code, rather than simply rendering images. The extracted data are being stored in Oracle databases and will be made available to SOI projects. Field personnel will access the data using either existing data editing systems or new systems that look and feel similar to the systems used to edit data from paper returns. In both cases, data will be subjected to extensive consistency testing, and editors will have opportunities to make corrections and apply codes and other adjustments needed to make data conform to the analytical requirements of data users. It is anticipated that this innovation will greatly reduce the cost of data collection for the SOI corporate and tax-exempt programs.

▶ Unit and Item Nonresponse in SOI Samples

In addition to improving the data collection processes of SOI programs, recent technological advances have allowed SOI to refine its methodologies for addressing unit and item nonresponse in the individual, corporate, and tax-exempt organization programs. A variety of computer tests, balancing routines, and ratio-based procedures that alter and impute return information are used in these techniques. To assist with the imputation of missing information, analysts use data derived from a variety of sources, including Masterfile information, prior-year data, and electronically filed returns.

Because individuals and organizations that fail to file required tax and information returns are subject to strict penalties and fines, unit nonresponse is not a large problem in most SOI programs. However, the need for timely data to use for budgeting and planning means that a few late filed returns will be missing at the close of an SOI study period, resulting in a sample that "does not fully cover the population for the target period of interest" (McMahon, 2002). To adjust for records that will be filed after the close of a study period, SOI uses proxies. These fall into two groups: 1) records created using values from prior studies that are updated using either survey or publicly available information and 2) records for recent prior years that are filed during the selection period.

Item nonresponse in SOI data files most often arises when filers fail to follow preparation instructions fully or when SOI programs require data that are not directly reported on IRS tax and information returns, such as certain corporation and nonprofit balance sheet items. In a few instances, missing data result when taxpayers neglect to file required supporting schedules and documents timely. SOI frequently uses data from prior-year returns as the basis for imputing these missing values. Finally, for some programs, a small number of timely filed returns are selected for SOI processing, but are unavailable to SOI during a study period, primarily because some other function of the IRS has control of the documents. For these "missing" returns, limited Masterfile data are available, but the detailed data needed to satisfy SOI program requirements are not. For some key returns, imputation can be used to construct a valid record for statistical purposes; for others, sample weight adjustments can be made to ensure that final samples represent key characteristics of all returns in a filing population.

SOI Individual Income Tax Program

In SOI's sample of individual income tax returns, approximately 3.0 percent of returns that will ultimately be filed for a particular tax year study are unavailable to SOI during the period allotted for data collection. For these cases, proxy returns are used as substitutes if the proxies are from 1 of the most recent 3 tax years. McMahon (2002) points out that these proxy returns more accurately represent data for late-filed returns than the core of timely filed returns, because late-filed returns are not randomly distributed among tax filers. Thus, making typical adjustments to design-based sample weights of timely filed returns to represent the unfiled returns would bias the resulting estimates. The use of proxy returns is a better alternative; however, this practice also seems to introduce some bias. Research has shown that proxy returns systematically understate "true" values for late-filed individual income tax returns, especially for those filers who derive substantial income from nonbusiness, nonfarm sources (McMahon, 2002).

For timely filed returns, the use of computer tests and balancing routines has become essential for correcting errors and estimating missing values for certain returns, and are proving especially useful for automating the processing of certain electronically filed returns in the individual income tax program. Known as "forced balancing," this methodology relies on SOI's Post Edit Reconciliation Process (PERP). PERP is an automated system of computer programs originally designed to ensure that data collected from the myriad forms and schedules that can be filed by individuals in fulfillment of their annual income tax reporting requirements were in balance with one another after SOI edit processing had been completed. At its inception, PERP was only used to review data, not to alter them in any way. If forms were not in balance, subject-matter experts in SOI headquarters would manually review them. Any changes to the final data file were initiated through the SOI editing system, and, after all changes had been made, the return would then be re-evaluated using the PERP program.

Use of the PERP system to automatically impute, or force into balance, return information was initially limited to those returns that were considered "missing" after they were selected to be a part of the SOI sample, typically about 250 returns per year. Using the limited data available for these returns from the Masterfile, routines were created to impute the missing details of forms and schedules. These routines were designed to ensure that detailed data summed to available totals for each form and that data carried from one form to another were consistent. Ratio-based adjustments were automatically applied to bring detail into balance with totals that had been proven correct through other tests embedded in the program.

As the number of electronically filed returns increased, SOI experimented with using the PERP system to process relatively simple e-filed returns, bypassing the normal field review. Only returns for which all fields needed for the SOI program were available from the TRDB were initially processed using PERP; returns containing data items which required any sort of reallocation or reclassification continued to be processed through the regular SOI

editing program. Returns eligible for automated processing were identified at the time of sampling. If these returns satisfied all the PERP tests, meaning the return was internally consistent, the return was considered finished and added to the final SOI study file. If an error was detected, the return was either passed through the tests of the regular editing process by an editor, or manually reviewed and adjusted by a National Office analyst, depending on the complexity of the problem.

Tax Year 2004 was the first year that e-filed returns were processed automatically through the PERP program without having first been processed through the regular SOI editing system. In Tax Year 2004, the basic individual income tax program had a total sample size of 200,295 returns, of which 64,670 were e-filed. Of the e-filed returns in the sample, 18,193 returns were processed solely through the PERP program (see Figure 5). For returns that were identified as potentially containing errors, most could be resolved by a simple review. Only 28 returns processed using the PERP system required National Office analysts to make corrections that were so extensive that it was necessary to pass the return data through the regular editing system to make the adjustments. Once corrected and retested using PERP, these records were considered "forced closed" and added to the final data file.

Figure 5: E-filed Returns Processed Through PERP

Tax year	2004	2005
SOI sample	200,295	292,837
E-file portion of SOI sample:		
Number e-file, SOI sample	64,670	114,897
Percent e-file, SOI sample	32.3%	39.2%
Returns processed through PERP	18,193	47,753
Percent processed through PERP	28.1%	41.6%
Returns requiring manual editing	28	8
Returns with change in AGI	14	23
Returns with change in total tax liability	36	191
Approximate editing hours saved	1,400	4,100

Income and tax data from electronically filed returns closed through PERP proved to be quite reliable. Only 14 of the more than 18,000 returns processed required a change in adjusted gross income (AGI) in order to satisfy the PERP tests. Nearly all of the changes to these 14 returns were minimal, most likely due to rounding. There were 36 returns that required a change to total tax liability. As with the AGI corrections, these changes were small and resulted in virtually no overall change to aggregate tax liability reported for the entire sample of returns that were closed through PERP.

Using PERP to close selected e-filed returns as a part of SOI sample processing proved to be cost-effective. The ability to "force close" these returns saved a substantial amount of editing time over that which would have been required for editors to validate the reported data manually. To estimate these cost savings, the average edit time per return for each sample code, computed for e-filed returns processed manually, was multiplied by the number of returns that were closed through PERP for each SOI sample code. For Tax Year 2004, using the PERP system to force close e-filed returns saved approximately 1,400 editing hours, including overhead costs.

After reviewing the success of using the PERP system to process a large number of returns in Tax Year 2004, a number of new tests were added to expand the program for Tax Year 2005. In 2005, the SOI sample size increased to 292,837 returns, including more than 50,000 additional electronically filed returns. Over 40 percent of the e-filed returns in the sample, 47,753 returns could be forced closed through the expanded PERP program. Of these, only 8 returns required a National Office analyst to process the corrections using the regular SOI editing program on them. The rest of the returns either passed all the tests or were reviewed by National Office analysts and accepted as filed.

Like their Tax Year 2004 counterparts, electronically filed Tax Year 2005 returns that were closed through PERP required few changes to the data. Just 23 returns had a change in AGI, mostly small changes due to rounding. Just 191 returns had a change in total tax liability after being force closed. Again, the changes to these returns resulted in no significant change to the aggregate AGI or tax liability of the overall sample of electronically filed returns. Closing nearly 48,000 returns automatically through PERP resulted in saving 4,100 hours of editing time, including overhead costs. By expanding the number of returns that could be forced closed using the PERP program, SOI has greatly reduced the cost of collecting data, freed resources for

other purposes, and enabled the expansion of the Form 1040 sample size at virtually no cost.

Another important tool used to validate, correct, and impute tax return data in SOI samples is the ability to browse SOI data collected for a taxpayer in a prior tax year. Prior-year SOI data are not only useful to analysts during professional review of PERP error logs, but also to editors who are using the regular SOI editing programs. When appropriate, prior-year data can be used to fill in "missing" data for the current year. For example, editors are required to assign an industry code based on information the taxpayer provides on Form 1040, Schedule C, Profit or Loss from Business. If editors are unable to determine the appropriate code, they can immediately access the assigned code and associated business description for the previous year and, if the business description is the same for both years, assign the previous year's code for the current year.

SOI Corporation Income Tax Study

SOI studies of corporation income tax data use several approaches to impute values in their "Advance" and "Final" data files for returns considered essential to the sample. In the corporate sample, large or "supercritical" cases comprise just 0.3 percent of total returns filed, but represent approximately 58.0 percent of total assets reported in any given year. Thus, their absence would distort the statistical estimates, particularly at the industry level (Davitian, 2005). For those supercritical cases for which no tax return is available, an alternate record is built using available data. Alternate records can be constructed using a combination of data from the IRS Masterfile (in those cases where a return was filed but for some reason is not available for SOI processing), data collected from questionnaires sent by SOI directly to these corporations, and imputed values based on prior-year data. When possible, data provided from a questionnaire are used as the basis for an imputed corporate tax record in the SOI Advance Data File; however, nonresponse is often a significant problem. Figure 6 shows the number of supercritical cases for which returns were missing, by study year, along with the number of corporations that responded to SOI questionnaires. It should be noted that some apparent nonrespondents in each year are actually corporations that were not required to file because they had filed as a subsidiary of a parent corporation that year. In most cases, actual return data are available for nearly all supercritical cases by the time the final corporation income tax data file is prepared.

Figure 6: Corporate Income Tax Study, Missing Critical Cases

Tax year	1998	1999	2000	2001	2002
Critical cases missing, Advance Data File	140	161	242	193	170
Cases that responded to SOI Questionaires	76	80	119	89	97
Imputed Cases, Final Data File	5	4	0	27	19

Like most SOI studies, missing data items are rare in the corporation income tax program. For those relatively few cases where a balance sheet entry is missing, a ratio-based imputation procedure is used. The ratios are determined using the most recent data available, either the specific corporation's prior-year return data (if those data were not imputed) or the most current tax year data available for the minor industrial group that includes the corporation. If the total asset and liabilities amounts are reported, details are imputed to equal these key sums. If these items are missing as well, they are first imputed, and then the details are imputed to insure that the detail balances with the imputed totals (see IRS, 2006; Uberall, 1995). Figure 7 shows that few records in SOI corporation data files contain imputed balance sheet amounts.

Figure 7: Corporation Income Tax Returns, Imputed Balance Sheet Items

Tax year	1998	1999	2000	2001	2002
Number of returns containing imputed data	70	68	38	41	33
Percent of returns containing imputed data	0.05	0.05	0.03	0.03	0.02

SOI Tax-Exempt Organization Studies

SOI studies of tax-exempt organizations also occasionally impute missing large-case returns. Data are imputed based on information reported on previous and

subsequent year returns, as well as limited current-year data available on the Masterfile. In some cases, data for the year of interest are reported directly on returns filed for the previous and subsequent years. Current-year data items that cannot be obtained directly from the two returns are imputed by applying ratio-based methods to available data from both alternate years, as well as to information available from the Masterfile. Less typically, if only one alternate-year return is available and no supplemental Masterfile data exist, data from the available previous-or subsequent-year are used as a proxy for current-year values.

In addition, SOI studies of tax-exempt organizations make extensive use of IRS Masterfile and alternate-year return information in validating and correcting data included on final study-year files. Charities and private foundations file information returns to report detailed balance sheet and income statement information annually in order to demonstrate that they are complying with IRS regulations that govern their tax-exempt status. Balance sheet information is reported on annual Forms 990 and 990-PF for a 2-year period, and, for private foundations, assets are valued at both book and fair market values. Private foundations report information on charitable distributions made over a 5-year period on the annual returns that they file. SOI's tax-exempt organization program makes use of these interdependencies by integrating return information for prior years into the data collection systems, regularly using these data to validate and improve information provided by a filer. Information reported on subsequent-year information returns, when available, is also incorporated into adjustments and imputations. Thus, if values for key fields are missing or appear incorrect for a particular tax year, information from a previous year can often be substituted. IRS Masterfile data are also integrated into these systems and can be used to verify or correct a limited number of fields. For example, Masterfile data are used to verify and correct tax-exemption type codes, which are assigned by the IRS when an organization is granted tax-exemption, but are also self-reported by filers on their annual returns.

All studies of tax-exempt organizations use ratio-based procedures to impute missing or incorrectly reported items, incorporating either prior-year data or similar information reported elsewhere on the return. Frequently, filers will provide lump-sum figures for key items, such as expenses, assets, or income, while SOI programs require that such figures be allocated to detailed subcategories. When detailed data for a similar item are reported elsewhere on the return, editors use automated computer routines to impute detailed amounts from the reported lump-sum values. For example, if a private foundation reports its end-of-year fair market value of total assets as a lump-sum value, but detailed data are available for these assets at book value, the system uses book value ratios to impute the fair market value detail lines. In cases where similar data are not reported on the return and a lump-sum value is reported, editors use automated computer routines to impute detailed amounts from reported lump-sum values, based on prior-year data. For example, charities are required to report detailed categorizations of their expenses, annually. If only a lump-sum value is reported, an automated routine, using ratios based on prior-year values, will impute amounts in order to allocate the total among the various detailed categories. In addition, to improve the longitudinal consistency of the annual study files, editors consistently substitute prior-year data for certain current-year values. For example, the system generally imputes the beginning-of-year book value of total assets for tax-exempt organizations based on the end-of-year book value reported on the prior-year return. Figure 8 shows the frequency of imputed balance sheet items from returns filed by charities and private foundations.

▶ The Future

It is anticipated that the nature of SOI field operations and SOI products will change markedly over the next decade as the number of returns and information documents filed electronically increases and data processing technology continues to evolve. SOI anticipates using technological and efficiency gains to provide more information, to provide information more quickly, and to produce and provide these data more efficiently. In a recent draft 10-year plan, many changes to processes and products are outlined. Known collectively as "SOI 2016," this vision of the future assumes that within 10 years, SOI will be collecting data in a nearly paperless environment, using either the popula-

Figure 8: Tax-Exempt Organization Returns, Imputed Balance Sheet Items

Item imputed	Tax Year 2003		Tax Year 2004	
	Number	Percent of sample	Number	Percent of sample
Beginning-of-year assets imputed based on prior year return	24,100	77.6	28,409	86.8
Beginning-of-year assets imputed based on current year return	206	0.7	187	0.6
End-of-year book or fair market value assets imputed based on current year return	229	0.7	194	0.6

tion of data provided electronically by filers or digital images created by SOI or other functions in the IRS. These data will be available in real time–that is, as the IRS receives returns.

A significant future change to SOI processing will be the introduction of Optical Character Recognition, Intelligent Character Recognition, or other similar technologies, which will be used to capture data from paper-filed returns in order to speed the data editing process. SOI is currently experimenting with these technologies and hopes to have a prototype in production by 2008. The introduction of automated editing software, coupled with increased use of electronically provided data, will change the nature of SOI field operations. SOI staff will continue to edit data and resolve data inconsistencies; however, the data transcription burden will be nearly eliminated. In addition, more data testing, error resolution, and coding will be performed in an automated, batch mode, prior to editors accessing the data. As a result, existing editing resources will be available to perform more complicated imputation, correction, and analysis. One area that will almost certainly see an increase is the use of longitudinal editing, the use of prior-year data to identify and correct outliers and anomalies in the data.

SOI products will also change significantly in the next decade. Automated data cleaning routines and decreased transcription costs are already enabling larger sample sizes and the use of population files for some analysis. Expansion of unedited or forced balanced data will allow for sample size increases needed to support small-area estimates. In addition, customers have expressed particular interest in making greater use of panel data, obtaining population data for the creation of ad hoc panels and for researching infrequent data items, and linking data across tax forms. SOI is working to develop routines that will find and fix large value errors in the entire population of individual income returns, currently 135 million records annually, to support some of these needs. Similar efforts are planned for documents filed by tax-exempt organizations and other entities.

▶ References

Arnsberger, Paul (2006), "Charities and Other Tax-Exempt Organizations, 2003," Internal Revenue Service, *Statistics of Income Bulletin,* Volume 26, Number 2, Washington, DC.

Davitian, Lucy (2005), "Corporation Supercritical Cases: How Do Imputed Returns on the Corporate File Compare to the Actual Returns?" American Statistical Association Proceedings of the Section on Government Statistics.

Internal Revenue Service (2006a), *Statistics of Income 2003 Corporation Income Tax Returns,* Publication 16.

Internal Revenue Service (2006b), *Statistics of Income 2004 Individual Income Tax Returns,* Publication 1304.

Ludlum, Melissa and Mark Stanton (2006), "Private Foundations, Tax Year 2003," Internal Revenue

Service, *Statistics of Income Bulletin,* Volume 26, Number 2, Washington, DC.

McMahon, Paul B. (2002), "Proxies in Administrative Records Surveys," American Statistical Association Proceedings of the Section on Survey Research Methods.

Uberall, Bertrand (1995), "Imputation of Balance Sheets for the 1992 SOI Corporate Program," American Statistical Association Proceedings of the Section on Survey Research Methods.

Wilson, Robert (1988), "Statistics of Income: A By-Product of the U.S. Tax System," Internal Revenue Service, *Statistics of Income Bulletin,* Volume 8, Number 2, Washington, DC.