Old Tabulations, Old Files, and a Brief History of Individual Tax Return Sampling

by Michael E. Weber and David P. Paris, Internal Revenue Service, and Peter J. Sailer, Consultant

The Statistics of Income (SOI) Division of the Internal Revenue Service has published data derived from individual income tax returns filed by taxpayers since 1916. As with most projects of any kind, the focus of this effort is on getting current work completed. Documentation of the work often gets secondary attention. Furthermore, documentation written with history in mind, where it serves not only to remind members of the team what they did, but also to explain to future teams what was done and why, gets even less attention. This paper is the beginning of an attempt to create historical documentation for the data created by SOI based on individual income tax returns.

The authors of this paper have a combined 95 years of SOI experience reaching back to 1966.[1] Unfortunately, only two authors are still employed by SOI. It was the retirement of one of our co-authors, the one who started working with SOI in 1966, that generated the discussions that led to this paper. The retirement brain drain is a serious problem, but not just because it affects the ability of an organization in its day-to-day activities, but also because an organization can also lose its ability to reference the past. The projects discussed below, therefore, have a certain urgency about them.

As noted above, SOI has published data on individual income tax returns since it brought out the 1916 report in 1918. SOI has retained copies of all these publications, which are available in the SOI library in Washington, DC, and at various Federal depository libraries across the country. Using these publications, however, can be difficult due to the fact that publication titles and tabulations and variables within the publications have changed over time. For example, tracing the amount of Sole Proprietor expenses over the course of 50 years could be quite a chore as this data moved between various publications and tabulations throughout the years. Thus, having an index or database of publications and tabulations would save a substantial amount of search time. Ideally, this index or database would be searchable in the sense that someone could enter a query such as “Sole Proprietor expenses” and “1935 to 1985” and receive output that would indicate the applicable publications and tabulations within those publications that satisfy the query. At SOI, we are currently engaged in a project that will create such an index.

SOI is also in the process of scanning all of these publications, principally with the goal of preserving old and deteriorating paper publications but also with the idea of one day making them available online through the IRS Tax Stats Web site. It should be noted that Google is also scanning these publications as part of its Google Books project. No matter which Web site one chooses to access these publications (Tax Stats or Google), they will be accessible through the Internet.

On the one hand, this is a clear advance from days of having to scour a library looking for a collection of dusty old books, but it also presents a few complications. For example, in a library, one could at least expect to find all SOI publications on one bookshelf–thus limiting possible places to search for desired data. In addition, once having found that bookshelf, one would visually know the population of publications in which the data could possibly exist and could thus attempt a reasoned search of those publications for a particular tabulation or variable–even if that meant spreading 20 publications across a table.

[1] Pete Sailer, one of our authors, in his early years in the Division, met an older employee by the name of Winifred Haines. It was rumored that Mrs. Haines started working at SOI in 1920, so that, by association, we have a personal link with SOI back to that year.

When accessing publications through the Internet, one is reliant on the search engine, which may not be powerful or specific enough to lead you to the proper publication, let alone the right page of a 200-page document. Furthermore, one is limited to viewing one page of one document at a time as opposed to the "table covered with books" method. Consequently, we may have traded one problem–getting to a library and physically searching for desired publications–for another–the inability to see large amounts of data from various publications at one time.

The solution lies again in the creation of a searchable index or database that lists all SOI publications, and the tabulations and variables within those publications. This index would be accessible through the Internet, and, in addition, search results would display links to specific scanned publications and specific pages within the publications. Searching for historical data would become a fast and efficient process. The old paper publication is now not only preserved but readily available to interested individuals. Once all of our publications are scanned, SOI will then expand the indexing project to include links to scanned publications.

Another project underway at SOI is the restoration of historical files. SOI has produced Individual Income Tax Return microdata files for many years although we have only retained files back to 1960. Unfortunately, documentation for many of the older files is quite limited. Consequently, a project is underway to fully and consistently document all of our older microdata files. The new documentation will consist of indexed tax form facsimiles which map the fields on the file with actual line items found on the tax returns. While this has been a standard practice at SOI for decades, it was not done, or at the very least not retained, for many of the older files.

Another aspect of the restoration process is to map the fields on the files with published data. Often, a field name and the published table column or stub title do not directly correspond with each other. In addition, published SOI data often represent the summation of multiple individual fields found on microdata files. Finally, uniform sample count and sampling stratifications tables will be developed for years where such tables cannot be found. SOI will be aided in this process by Pete Sailer (a coauthor of this paper), who actually helped to create many of these files early in his SOI career.[2] With the completion of this project, SOI will have readily accessible microdata covering the last 50 years.

The final historical project underway at SOI is perhaps the most ambitious. SOI has produced data for over 90 years and, as just noted, possesses microdata going back 50 years. But what is missing is the story of changes and innovations in the sampling and processing of individual income tax returns for statistical purposes over those years. Our sources for this story are the documentation contained in the various publications and the memories of various participants. A few highlights of this process are discussed below.

The first reference we have to sampling in an SOI program is found in the 1918 publication, which mentions (without further elaboration) that each IRS district office was to supply SOI with a "fair and average sample" of returns with net incomes between $1,000 and $5,000 (returns with net incomes under $1,000 were not included in the statistics). No sampling is mentioned for Tax Years 1919-1924, but, for 1925-1942, returns with incomes under $5,000 (and as low as $1) were again subjected to sampling. Each IRS district office was instructed to select a given percentage of both taxable and nontaxable returns in this income group, with minimum numbers of returns to be selected for each district prescribed as well. In Washington, averages were computed for taxable and nontaxable returns in each income class in each district, then multiplied by counts for these classes supplied by the district offices.

For Tax Year 1943, major changes were made in the methods IRS used for revenue processing purposes. All individual returns were divided into processing groups, based on criteria such as form type (1040 or 1040A), taxability, "assessability" (underwithheld, overwithheld, breakeven), level of adjusted gross income, and amount of business receipts. Within each

[2] We are also very grateful for recollections provided to the authors by former Statistics of Income Director Fritz Scheuren, who also worked on many of these early SOI files.

group, returns were placed into blocks of 100. Blocks were numbered consecutively within each group within each district office. Within the block, returns were assigned serial numbers from 00 through 99. SOI used these groups as the basis of its sampling plan, and used the last two digits of block numbers to instruct district offices to select whole blocks of returns for SOI processing. Any subsampling (first introduced for 1949) was done at the National Office, using serial numbers assigned to returns in each block. For 1952, IRS and SOI abandoned block numbers and serial numbers, and instead used an "account number" (assigned consecutively within each stratum in each district office) for sample selection.

Between the Tax Year 1965 and 1968 programs, SOI phased in the age of computerized sampling. The ending digits of the Social Security number of the primary taxpayer were utilized to select the sample. The number of endings used varied with the sample size of the various stratifications. Interestingly, SOI chose to change the ending digits each year, in order to minimize any bias created by (unproven) special characteristics of returns with certain endings.

For Tax Year 1972, the sample was augmented by a special additional sample, selected manually in the service centers, covering the lower-income strata for the smallest States. Treasury required the expanded sample to produce estimates of grants that would be required to each State under the Federal Revenue Sharing Program. For 1973, sampling at higher rates in the smaller States was incorporated into the original sample design. Also for 1973, the sampling program was changed to sample every nth return processed within each stratum, without using the SSN. Documentation relating to why this change was implemented has not been discovered as of yet.

For 1978, SOI returned to sampling on the basis of the ending digits of the SSN and for 1979 incorporated two Continuous Work History Sample (CWHS) SSN endings that were permanently built into the cross-sectional sample. The CWHS is a system of general multipurpose statistical files designed primarily for socioeconomic analyses and maintained by the Social Security Administration.

The system consists of samples of records of individuals who were ever issued Social Security numbers. SOI chose two CWHS ending digits whose consistent use formed the basis of what today is almost a 30-year longitudinal sample of income tax returns. The major shortcoming of this panel is that marriages and divorces cause certain taxpayers (usually women, who tend to file as secondary taxpayers on joint returns) to enter and leave the sample.

The focus of Federal income tax policy has generally been at the nationallevel of study rather than at the State or local level. Therefore, by the Tax Year 1982 study, Treasury requested that SOI stop selecting samples of individuals at higher rates in the smaller States and leverage resource savings from processing the old sample design in new tax studies needed by Treasury and other SOI customers. As a result, SOI changed its focus to a sampling methodology that would support increasing emphasis placed by Treasury on microsimulation modeling of the revenue and distribution effects of tax law changes. Hence, a "fat year" and "lean year" concept was introduced into SOI sampling.

Fat year samples were designed to capture a larger volume of returns necessary to support the tax policy model. Lean year samples derived the volumes necessary for SOI to continue its legal mandate under IR Code section 6108 of making annual statistics readily available with respect to operation of the internal revenue laws. Also for 1982, SOI began sampling returns on the basis of a transformed SSN. This innovation had the advantage of allowing SOI to simplify the yearly sampling programs because the last five digits of the transformed SSNs were evenly distributed from 1 to 100,000. Therefore, the desired sampling rates for each stratum could be easily adjusted in rather fine gradients, with a minimum of programming.[3] At this time, SOI also chose to generally use the same transformed SSN ranges each year, thus creating a maximum overlap between the set of returns chosen between 2 years.

[3] Whenever sampling rates needed to be adjusted, only one number (the highest transform to be selected) had to be changed for each stratum, instead of a list of four-digit SSN endings.

Over the years, SOI has also become fairly sophisticated at using one return for multiple samples. This concept is implicit in the use of the CWHS SSN endings that SOI began using in the Tax Year 1979 sample, but was taken to a higher level with the 1985 Sales of Capital Assets Study and the 1987 Family Panel Study. Essentially, if roughly the same stratifications are used for multiple studies (panels and annual cross sections), as well as the same range of transformed SSNs within those stratifications, significant sample overlaps will occur.

For example, a return selected for the Tax Year 1987 annual cross-study sample had a high probability of being selected for the 1988 and 1989 cross sections. Consequently, a panel sample selected as a subsample of the 1987 cross section produces a significant overlap between the panel sample and the annual cross-sectional sample. This sampling plan reduces costs and allows higher sample sizes for both files, given a set budgeted cost. From 1985 to 1996, SOI produced the two panel studies mentioned above, the CWHS panel from 1979, and the yearly cross sections using this method. Maximizing this overlap has been a key feature of all Individual Income Tax Return sample designs at SOI since that time.

We hope to produce a more formal and detailed history of the SOI program prior to SOI's 100th anniversary.