Section 2 Description of the Sample

This section describes the 1997 Individual sample design and selection, the method of estimation, the sampling variability of the estimates, and the methodology of computing confidence intervals.

Background

Statistical sampling of individual income tax returns began in 1918. Stratified sampling of individual tax returns was introduced in 1950 and is still used today. Initially, returns were stratified by form, income size, presence or absence of business income and end of year tax payment status. Additional sampling criteria were added in 1968, based on a recommendation made by Dr. W. Edwards Deming in a contracted report for the IRS. The new criteria included largest source of income and size of business receipts. The sample was redesigned in 1982 and was stratified based on the the larger of total income or total loss as well as the size of business plus farm receipts. Since 1991, returns have been stratified based on positive or negative income, whichever is larger, and presence or absence of special forms.

Sampling was initially based on the serial number of the return, which was assigned by the administrative returns processing system. Sampling based on the individual's social security number began in 1967. At that time it was based on the ending digits of the taxpayer's social security number. The redesign in 1982 included a new method of sampling based on permanent random numbers generated by using a mathematical transformation of the social security number.

Domain of Study

The statistics in this report are estimates from a probability sample of unaudited Individual Income Tax Returns, Forms 1040, 1040A, 1040EZ, 1040PC and 1040TEL (including electronic returns) filed by U.S. citizens and residents during Calendar Year 1998.

All returns processed during 1998 were subjected to sampling except tentative and amended returns. Tentative returns were not subjected to sampling because the revised returns may have been sampled later, while amended returns were excluded because the original returns had already been subjected to sampling. A small percentage of returns were not identified as tentative or amended until after sampling. These returns, along with those that contained no income information, were excluded in calculating estimates. This resulted in a small difference between the population total (123,045,360 returns) reported in Table C and the estimated total of all returns (122,421,991) reported in other tables.

The estimates in this report are intended to

Bonnye Walker and Karen Masken designed the sample and prepared the text and tables in this section under the direction of Yahia Ahmed, Chief, Mathematical Statistics Section, Statistical Computing Branch.

represent all returns filed for Tax Year 1997. While about 98 percent of the returns processed during Calendar Year 1998 were for Tax Year 1997, a few were for noncalendar years ending during 1997 and 1998, and some were returns for prior years. Returns for prior years were used in place of 1997 returns expected to be received and processed after December 31, 1998. This was done based on the assumption that the characteristics of returns due, but not yet processed, can be represented by the returns for previous income years that were processed in 1998.

Sample Design and Selection

The sample design is a stratified probability sample, in which the population of tax returns is classified into subpopulations, called strata, and a sample is randomly selected independently from each stratum. Strata are defined by:

1. Nontaxable with adjusted gross income or expanded income of $200,000 or more and no alternative minimum tax.

2. High combined business and farm total receipts of $50,000,000 or more.

3. Presence or absence of special Forms or Schedules (Form 2555, Form 1116, Form 1040 Schedule C, and Form 1040 Schedule F).

4. Indexed positive or negative income. Sixty variables are used to derive positive and negative incomes. These positive and negative income classes are deflated using the Gross Domestic Product Implicit Price Deflator to represent a base year of 1991. (See footnote 1 for details.)

5. Potential usefulness of the return for tax policy modeling. Thirty-two variables are used to determine how useful the return is for tax modeling purposes.

Table C shows the population and sample count for each stratum after collapsing some strata with the same sampling rates. (See references 1 and 2 for details.) The sampling rates range from 0.022 percent to 100 percent.

Tax data processed to the IRS Individual Master File at the Martinsburg Computing Center during Calendar Year 1998 were used to assign each taxpayer's record to the appropriate stratum and to determine whether or not the record should be included in the sample. Records are selected for the sample either if they possess certain combinations of the four ending digits of the social security number, or if their ending five digits of an eleven-digit number generated by a mathematical transformation of the SSN is less than or equal to the stratum sampling rate times 100,000. (See reference 3 for details.)

Data Capture and Cleaning

Data capture for the SOI sample begins with the designation of a sample of administrative records. While the sample was being selected, the process was continually monitored for sample selection and data collection errors. In addition, a small subsample of returns was selected and independently reviewed, analyzed, and processed for a quality evaluation.

The administrative data and controlling information for each record designated for this sample was loaded onto an online database at the Cincinnati Service Center. Computer data for the selected administrative records were then used to identify inconsistencies, questionable values, and missing values as well as any additional variables that an editor needed to extract for each record. The editors use a hardcopy of the taxpayer's return to enter the required information onto the online system.

After the completion of service center review, data were further validated, tested, and balanced at the Detroit Computing Center. Adjustments and imputations for selected fields were used to make each record internally consistent, and the data were then tabulated. Finally, prior to publication, all statistics and tables were reviewed for accuracy and reasonableness in light of provisions of the tax law, taxpayer reporting variations and limitations, economic conditions, and comparability with other statistical series.

Some returns designated for the sample were not available for SOI processing because other areas of IRS needed the return at the same time. For Tax Year 1997, 0.25 percent of the sample returns were unavailable.

Method of Estimation

Weights were obtained by dividing the population count of returns in a stratum by the number of sample returns for that stratum. The weights were adjusted to correct for misclassified returns. These weights were applied to the sample data to produce all of the estimates in this report.

Sampling Variability and Confidence Intervals

The sample used in this study is one of a large number of samples that could have been selected using the same sample design. The estimates calculated from these different samples would vary. The standard error (SE) of an estimate is a measure of the variation among the estimates from the possible samples and, thus, is a measure of the precision with which an estimate from a particular sample approximates the average of the estimates calculated from all possible samples.

The standard error may be expressed as a percentage of the value being estimated. This ratio is called the coefficient of variation (CV). Table 1.4 CV contains estimated CV's for the estimates included in Table 1.4 of this report.

The sample estimate and an estimate of its standard error permit the construction of interval estimates with prescribed confidence that the interval includes the population value. If all possible samples were selected under essentially the same conditions and an estimate and its estimated standard error were calculated from each sample, then:

1. About 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the population value. This is a 68 percent confidence interval.

2. About 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the population value. This is a 95 percent confidence interval.

For example, from Table 1.4, the amount estimate for State Income Tax Refunds, X, is \$14.094 billion, and its related coefficient of variation, CV(X), is 1.13 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$14.094 \times 10^{9}) \bullet (0.0113) \\ &= \$0.159 \text{ billion} \end{aligned}$$

The p percent confidence interval is calculated using the formula:

$$X \pm z \bullet SE(X)$$

where z takes the value 1, 2, or 3 when p is 68, 95, or 99, respectively. Based on these data, the 68 percent confidence interval is from \$13.935 billion to \$14.253 billion, and the 95 percent confidence interval is from \$13.776 billion to \$14.413 billion.

Table Presentation

Whenever a weighted frequency is less than 3, the estimate and its corresponding amount are combined or deleted in order to avoid disclosure of information for specific taxpayers. (The combined or deleted data, if any, are included in the corresponding column totals.) These combinations and deletions are indicated by a double asterisk (**). Estimates based on less than 10 sampled returns are considered to be unreliable. These estimates are noted by a single asterisk (*) to the left of the data unless all of the sampled returns are selected with certainty (at the 100 percent rate).

In the tables, a dash (- or --) in place of a frequency or an amount indicates that either no returns in the population had the characteristic or the characteristic was so rare that it did not appear on any of the sampled returns.

Footnote

[1] Positive and negative income are divided by the ratio of the Gross Domestic Product

Implicit Price Deflator for the fourth quarter of 1997 to the fourth quarter of the base year of 1991. The deflators can be found in U. S. Department of Commerce, Bureau of Economic Analysis, *Survey of Current Business* (December 1997) Vol 77, number 13.

References

[1] Hostetter, S., Czajka, J. L., Schirm, A. L., and O'Conor, K. (1990), "Choosing the Appropriate Income Classifier for Economic Tax Modeling," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 419-424.

[2] Schirm, A. L., and Czajka, J. L. (1991), "Alternative Designs for a Cross-Sectional Sample of Individual Tax Returns: the Old and the New," *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 163-168.

[3] Harte, J.M. (1986), “Some Mathematical and Statistical Aspects of the transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS,” *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 603-608.

SOURCE: IRS, Statistics of Income, Individual Income Tax Returns 1997, Publication 1304 (Rev. 4-2000).