Section 6

Description of the Sample

This section describes the domain of the study, the sample design and selection, data capture and cleaning, the method of estimation, the sampling variability of the estimates, the methodology of computing confidence intervals, and the table presentation.

Domain of Study

The statistics in this report are estimates from a probability sample of unaudited Individual Income Tax Returns, Forms 1040, 1040A, and 1040EZ (including electronic returns), filed by U.S. citizens and residents during Calendar Year 2016.

All returns processed during 2016 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 or frivolous or fraudulent income information when recognized, were excluded in calculating estimates.

The estimates in this report are intended to represent all returns filed for Tax Year 2015. While most of the returns processed during Calendar Year 2016 were for Tax Year 2015, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2014 and 2015.

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 the following characteristics:

Valerie Testa and Tracy Haines designed the sample and prepared the text and the tables in this section under the direction of Tammy Rib, Chief, Mathematical Statistics Section, Corporation Statistics Branch.

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

(2) High business 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 Chain-Type Price Index for the Gross Domestic Product to represent a base year of 1991. (See footnote 1 for details.)

Table C shows the population and sample count for each stratum after collapsing some strata with the same sampling rates [1,2]. (See references 1 and 2 for details.) The sampling rates range from 0.10 percent to 100 percent.

Tax data processed to the IRS Individual Master File at the Enterprise Computing Center at Martinsburg during Calendar Year 2016 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 the five ending 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 [3]. (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 were loaded onto an

online database at the Cincinnati Submission Processing 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.

After the completion of the service center review, data were further validated, tested, and balanced. Adjustments and imputations for selected fields based on prior-year data and other available information were used to make each record internally consistent. 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 2015, about 0.02 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 and were then 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). Tables 1.4 CV, 2.1 CV, and 3.3 CV contain estimated CVs for the estimates included in Tables 1.4, 2.1, and 3.3 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 estimate for State Income Tax Refunds, X, is $31.110 billion, and its related coefficient of variation, CV(X), is 0.65 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$SE(X) = X \times CV(X)$$

$$= (\$31.110 \times 10^9) \times (0.0065)$$

$$= \$0.202 \text{ billion.}$$

The p percent confidence interval is calculated using the formula:

$$p = X \pm z \times SE(X),$$

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 $30.908 billion to $31.312 billion, the 95 percent confidence interval is from $30.706 billion to $31.514 billion, and the 99 percent confidence interval is from $30.504 billion to $31.716 billion.

Table Presentation

Whenever an unweighted 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 (-) 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] Indexing of positive and negative income is done by dividing each by the ratio of the Gross Domestic Product (GDP) Chain-Type Price Index for the third quarter of 2015 to the third quarter of the base year of 1991. The indices were calculated using the GDP Chain-Type Price Index [4].

References

[1] Hostetter, S., Czajka, J. L., Schirm, A. L., and O'Conor, K. (1990), "Choosing the Appropriate Income Classifier

for Economic Tax Modeling," *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.

[4] U.S. Bureau of Economic Analysis, "Price Indexes for Gross Domestic Product," [http://www.bea.gov/] (accessed November 25, 2014).

Table C. Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2015

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	151,238,929	338,857
Form 1040 returns only with adjusted gross income or expanded income of $200,000 and over, with no income tax after credits and no additional tax for tax preferences, total	28,321	28,321
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	326	326
Other returns, total	0	0

Description of the sample strata	Number of returns by type of form attached									
	Form 1040, with Form 1116 or Form 2555		Form 1040, with Schedule C but without Form 1116 or Form 2555		Form 1040, with Schedule F but without Schedule C, Form 1116, or Form 2555		Form 1040, with other Schedules and Forms and Forms 1040A and 1040EZ			
	Population counts	Sample counts	Population counts	Sample counts	Population counts	Sample counts	Population counts	Sample counts		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		
Total	6,598,933	81,221	24,325,981	57,893	1,273,459	6,412	119,011,909	164,684		
Indexed Negative Income [2]										
$10,000,000 or more	407	407	1,192	1,192	155	155	1,421	1,421	3,175	3,175
$5,000,000 under $10,000,000	685	685	1,892	1,892	226	226	2,269	2,269	5,072	5,072
$2,000,000 under $5,000,000	2,893	940	7,102	2,428	978	337	8,905	2,916	19,878	6,621
$1,000,000 under $2,000,000	6,225	965	13,978	2,178	2,586	397	17,887	2,863	40,676	6,403
$500,000 under $1,000,000	13,864	459	31,690	1,052	6,898	215	40,530	1,332	92,982	3,058
$250,000 under $500,000	27,491	290	66,040	657	12,859	129	86,944	828	193,334	1,904
$120,000 under $250,000	46,975	219	125,117	627	19,515	114	178,704	867	370,311	1,827
$60,000 under $120,000	50,553	130	153,157	473	19,881	69	242,470	700	466,061	1,372
Under $60,000	39,550	80	369,929	680	26,136	44	515,886	962	951,501	1,766
Indexed Positive Income										
Under $30,000	570,590	539	10,051,983	9,853	163,037	190	69,131,521	69,384	79,917,131	79,966
$30,000 under $60,000	1,356,615	1,356	5,953,839	5,932	362,648	353	28,958,515	28,754	36,631,617	36,395
$60,000 under $120,000	1,941,798	1,992	4,925,709	4,892	375,401	390	15,115,180	15,303	22,358,088	22,577
$120,000 under $250,000	1,400,587	4,635	1,896,945	6,364	173,183	585	3,732,534	12,462	7,203,249	24,046
$250,000 under $500,000	652,383	4,656	521,734	3,795	72,429	521	732,564	5,242	1,979,110	14,214
$500,000 under $1,000,000	291,519	7,172	150,054	3,787	27,480	638	179,955	4,442	649,008	16,039
$1,000,000 under $2,000,000	116,440	14,161	39,853	4,886	7,526	900	46,969	5,784	210,788	25,731
$2,000,000 under $5,000,000	55,970	18,147	12,554	3,992	2,007	635	15,594	5,094	86,125	27,868
$5,000,000 under $10,000,000	14,789	14,789	2,276	2,276	357	357	2,842	2,842	20,264	20,264
$10,000,000 or more	9,599	9,599	937	937	157	157	1,219	1,219	11,912	11,912

[1] This population includes an estimated 745,666 returns that were excluded from other tables in this report because they contained no income information or frivolous or fraudulent income information when recognized or represented amended or tentative returns identified after sampling.
[2] Positive and Negative Income classes are divided by a Chain-Type Price Index for the Gross Domestic Product of 1.5874 to represent a base year of 1991.
Source: IRS, Statistics of Income Division, Publication 1304, September 2017.