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 2017.

All returns processed during 2017 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 2016. While most of the returns processed during Calendar Year 2017 were for Tax Year 2016, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2015 and 2016.

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 2016. (See footnote 1 for details.)

A sample of 351, 049 returns was taken from a population of 151,014,093. This population includes an estimated 741,936 returns that were excluded from 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. 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 2017 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 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. (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 2016, about 0.03 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 CV's 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 $33.468 billion, and its related coefficient of variation, CV(X), is 0.67 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \cdot CV(X) \\ &= (\$33.468 \times 10^9) \cdot (0.0067) \\ &= \$0.224 \text{ billion.} \end{aligned}$$

The p percent confidence interval is calculated using the formula:

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

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 $33.244 billion to $33.692 billion, the 95 percent confidence interval is from $33.020 billion to $33.916 billion, and the 99 percent confidence interval is from $32.796 billion to $34.140 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.

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] Prior to this year, indexing of positive and negative income would have been done by dividing each by the ratio of the Chain-Type Price Index for the Gross Domestic Product for the third quarter of 2016 to the third quarter of the base year of 1991. The indices were calculated using the Gross Domestic Product (GDP) Chain-type Price Index [4]. For

the current year, the year of comparison was changed to 2016. The deflation index ratio was set to 1.0000.

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.

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

Table C. Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2016

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	151,014,093	351,049
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	28,758	28,758
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	315	315
Other Returns, total	150,985,020	321,976

Description of the sample strata	Number of returns by type of form attached											
	Form 1040, with Form 2555		Form 1040, with Form 1116 but without 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	Population counts	Sample counts		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)		
Total	474,127	16,193	6,067,989	75,366	24,684,886	59,316	1,255,801	6,268	118,502,217	164,833		
Indexed Negative Income [2]												
$15,000,000 or more	7	7	430	430	1,327	1,327	156	156	1,508	1,508	3,428	3,428
$8,000,000 under $15,000,000	17	17	611	611	1,882	1,882	224	224	2,087	2,087	4,821	4,821
$3,000,000 under $8,000,000	115	114	3,133	1,008	7,872	2,673	1,304	442	9,938	3,379	22,362	7,616
$1,500,000 under $3,000,000	261	257	6,306	1,012	15,022	2,341	3,203	492	19,387	3,141	44,179	7,243
$800,000 under $1,500,000	593	224	12,208	399	29,757	972	6,967	227	38,006	1,243	87,531	3,065
$400,000 under $800,000	1,748	166	25,119	242	65,607	687	14,215	127	86,884	875	193,573	2,097
$200,000 under $400,000	4,909	442	37,989	173	114,222	543	20,008	101	165,616	807	342,744	2,066
$100,000 under $200,000	12,634	253	37,332	116	150,847	482	21,753	73	238,183	705	460,749	1,629
Under $100,000	21,856	267	21,471	43	382,765	713	30,229	57	533,745	1,002	990,066	2,082
Indexed Positive Income												
Under $50,000	145,057	1,498	482,934	470	10,449,884	10,337	180,092	210	70,341,781	70,386	81,599,748	82,901
$50,000 under $100,000	115,105	1,162	1,334,182	1,305	6,159,190	6,005	374,256	367	28,547,765	28,596	36,530,498	37,435
$100,000 under $200,000	97,551	1,851	1,829,116	1,823	4,830,962	4,801	354,470	373	14,189,650	14,343	21,301,749	23,191
$200,000 under $400,000	45,249	3,819	1,227,110	4,081	1,739,881	5,752	149,807	526	3,354,914	11,258	6,516,961	25,436
$400,000 under $800,000	19,404	1,618	608,585	4,373	528,805	3,788	66,230	448	730,052	5,120	1,953,076	15,347
$800,000 under $1,500,000	6,165	2,459	249,883	6,233	143,655	3,563	23,199	548	168,946	4,328	591,848	17,131
$1,500,000 under $3,000,000	2,308	897	114,642	14,060	45,337	5,504	7,334	866	51,981	6,206	221,602	27,533
$3,000,000 under $8,000,000	910	904	56,107	18,156	14,700	4,775	1,939	616	17,758	5,833	91,414	30,284
$8,000,000 under $15,000,000	135	135	11,954	11,954	2,137	2,137	272	272	2,629	2,629	17,127	17,127
$15,000,000 or more	103	103	8,877	8,877	1,034	1,034	143	143	1,387	1,387	11,544	11,544

[1] This population includes an estimated 741,936 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 1.0000 to represent a base year of 2016.

SOURCE: IRS, Statistics of Income Division, Publication 1304, September 2018.