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, Form 1040 (including electronic returns) filed by U.S. citizens and residents during Calendar Year 2018.

All returns processed during 2018 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 had no income information, frivolous income information, 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 2017. While most of the returns processed during Calendar Year 2018 were filed for Tax Year 2017, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2016 and 2017.

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 an independent sample is randomly selected from each stratum. Strata are defined by the following characteristics:

(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.)

Table C shows the population and sample count for each stratum. (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 2018 were used to assign each taxpayer's record to the appropriate stratum and to determine whether 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 (SSN), 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

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, SOI Program Support, Statistical Services Branch.

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 considering the 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 2017, 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 sampled returns for that stratum. The weights were adjusted to correct for misclassified returns and were then applied to the sample data to produce all 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 $34.292 billion, and its related coefficient of variation, CV(X), is 0.64 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$SE(X) = X \bullet CV(X)$$
$$= (\$34.292 \bullet 10^{9}) \bullet (0.0064)$$
$$= \$0.219 \text{ billion.}$$

The p percent confidence interval is calculated using the formula:

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

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 $34.073 billion to $34.511 billion, the 95 percent confidence interval is from $33.854 billion to $34.730 billion, and the 99 percent confidence interval is from $33.635 billion to $34.949 billion.

Table Presentation

Whenever an unweighted frequency is less than 3, the estimate and its corresponding amount are either combined or deleted to avoid disclosure of information about 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 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] Indexing of positive and negative income is performed by dividing each by the ratio of the Chain-Type Price Index for the Gross Domestic Product for the third quarter of 2017 to the third quarter of the base year of 2016. The indices were calculated using the Gross Domestic Product (GDP) Chain-type Price Index [4].

References

[1] Hostetter, S., J. L. Czajka, A. L. Schirm, and K. O'Conor, (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 J. L. Czajka, (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 2017

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	153,575,872	352,316
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	25,318	25,318
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	371	371

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			
	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	7,298,266	89,633	25,415,349	62,679	1,245,789	6,432	119,590,779	167,883		
Indexed negative income [2]										
$15,000,000 or more	454	454	1,570	1,570	165	165	1,755	1,755	3,944	3,944
$8,000,000 under $15,000,000	709	709	2,034	2,034	259	259	2,438	2,438	5,440	5,440
$3,000,000 under $8,000,000	3,313	1,070	8,670	2,934	1,366	457	10,637	3,578	23,986	8,039
$1,500,000 under $3,000,000	6,582	1,046	16,023	2,466	3,312	521	19,719	3,120	45,636	7,153
$800,000 under $1,500,000	12,730	408	30,335	1,015	7,171	231	38,539	1,295	88,775	2,949
$400,000 under $800,000	26,171	264	66,194	627	14,490	150	86,570	847	193,425	1,888
$200,000 under $400,000	40,762	171	113,761	579	20,163	99	160,680	789	335,366	1,638
$100,000 under $200,000	47,009	146	148,764	505	21,842	72	227,432	663	445,047	1,386
Under $100,000	40,316	74	393,243	734	29,039	54	481,164	899	943,762	1,761
Grand total										
Under $50,000	688,831	678	10,734,183	10,722	173,356	173	70,531,732	70,175	82,128,102	81,748
$50,000 under $100,000	1,580,746	1,578	6,355,509	6,229	367,336	387	29,033,957	28,888	37,337,548	37,082
$100,000 under $200,000	2,157,790	2,133	4,960,624	4,989	357,757	368	14,496,945	14,554	21,973,116	22,044
$200,000 under $400,000	1,454,213	4,775	1,812,008	5,957	150,482	499	3,474,124	11,460	6,890,827	22,691
$400,000 under $800,000	716,768	5,161	552,404	4,003	65,992	484	767,918	5,490	2,103,082	15,138
$800,000 under $1,500,000	294,184	7,408	152,021	3,787	23,075	513	178,336	4,385	647,616	16,093
$1,500,000 under $3,000,000	135,784	16,566	48,258	5,787	7,547	898	55,306	6,676	246,895	29,927
$3,000,000 under $8,000,000	66,319	21,407	16,158	5,151	1,990	655	18,949	6,293	103,416	33,506
$8,000,000 under $15,000,000	14,397	14,397	2,318	2,318	300	300	2,968	2,968	19,983	19,983
$15,000,000 or more	11,188	11,188	1,272	1,272	147	147	1,610	1,610	14,217	14,217

[1] This population includes an estimated 672,641 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 1.0179 to represent a base year of 2016.
SOURCE: IRS, Statistics of Income Division, Publication 1304, September 2019.