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 2015.

All returns processed during 2015 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 2014. While most of the returns processed during Calendar Year 2015 were for Tax Year 2014, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2013 and 2014.

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.)

(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 0.10 percent to 100 percent.

Tax data processed to the IRS Individual Master File at the Enterprise Computing Center at Martinsburg during Calendar Year 2015 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 2014, 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 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 $30.088 billion, and its related coefficient of variation, CV(X), is 0.69 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

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

$$= (\$30.088 \times 10^9) \cdot (0.0069)$$

$$= \$0.208 \text{ billion.}$$

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 $29.880 billion to $30.296 billion, the 95 percent confidence interval is from $29.672 billion to $30.504 billion, and the 99 percent confidence interval is from $29.464 billion to $30.712 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 (-) 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 Chain-Type Price Index for the Gross Domestic Product for the fourth quarter of 2013 to the fourth quarter of the base year of 1991. The

indices were calculated using the Gross Domestic Product (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," 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/] (accessed November 25, 2014).

Table B. Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2014

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	149,647,908	343,748
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		
$5,000,000 or more Indexed Negative Income or Indexed Positive Income	168	168
Under $5,000,000 Indexed Negative Income or Indexed Positive Income [2]	32,070	30,715
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	339	339
Other Returns, total		

Description of the sample strata	Degree of interest [3]	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)	(9)		
Total		6,675,654	82,968	23,755,976	57,776	1,278,356	7,106	117,905,345	164,676		
Indexed Negative Income [4]											
$10,000,000 or more	All	388	388	1,228	1,228	164	164	1,324	1,324	3,104	3,104
$5,000,000 under $10,000,000	All	692	692	1,833	1,833	256	256	2,304	2,304	5,085	5,085
$2,000,000 under $5,000,000	All	2,934	960	7,004	2,347	1,022	344	8,794	3,007	19,754	6,658
$1,000,000 under $2,000,000	All	6,282	963	13,953	2,181	2,433	382	17,788	2,814	40,456	6,340
$500,000 under $1,000,000	All	14,213	496	32,011	1,067	6,150	205	40,699	1,330	93,073	3,098
$250,000 under $500,000	All	28,458	313	67,440	661	11,440	131	88,680	827	196,018	1,932
$120,000 under $250,000	All	48,914	251	129,050	685	17,434	97	183,375	922	378,773	1,955
$60,000 under $120,000	All	51,990	154	156,314	478	18,030	62	249,903	761	476,237	1,455
Under $60,000	All	37,787	76	372,050	695	23,184	45	532,327	993	965,348	1,809
Indexed Positive Income [3]											
Under $30,000	1	0	0	0	0	0	0	34,735,335	34,935	34,735,335	34,935
Under $30,000	2	241,228	254	3,937,688	3,889	67,548	71	26,540,550	26,290	30,787,014	30,504
Under $30,000	3-4	320,547	283	5,843,729	5,730	87,668	93	6,861,944	6,865	13,113,888	12,971
$30,000 under $60,000	1-2	591,816	601	1,923,476	1,954	131,086	132	21,907,794	21,695	24,554,172	24,382
$30,000 under $60,000	3-4	727,455	757	3,819,709	3,737	222,424	232	6,831,751	6,853	11,601,339	11,579
$60,000 under $120,000	1-3	1,120,119	1,162	2,252,997	2,250	190,979	209	11,707,711	11,753	15,271,806	15,374
$60,000 under $120,000	4	843,053	852	2,545,632	2,536	187,955	179	3,342,170	3,441	6,918,810	7,008
$120,000 under $250,000	1-3	373,526	1,273	422,614	1,413	76,717	263	1,463,871	5,010	2,336,728	7,959
$120,000 under $250,000	4	1,063,822	3,513	1,477,658	4,902	105,563	337	2,321,654	7,660	4,968,697	16,412
$250,000 under $500,000	All	687,466	5,003	537,521	3,942	82,197	597	797,498	5,674	2,104,682	15,216
$500,000 under $1,000,000	All	311,080	7,695	156,059	3,858	33,616	813	199,860	4,933	700,615	17,299
$1,000,000 under $2,000,000	All	123,041	14,936	41,912	5,129	9,424	1,134	50,302	6,118	224,679	27,317
$2,000,000 under $5,000,000	All	56,991	18,494	13,008	4,171	2,478	772	15,619	5,075	88,096	28,512
$5,000,000 under $10,000,000	All	14,379	14,379	2,196	2,196	408	408	2,786	2,786	19,769	19,769
$10,000,000 or more	All	9,473	9,473	894	894	180	180	1,306	1,306	11,853	11,853

[1] This population includes an estimated 951,330 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. The increase in this number for the current tax year was caused by additional processing for returns impacted by identity theft.
[2] A processing error caused 1,355 returns to be excluded from the sample prior to sample selection.
[3] Each population member is assigned a degree of interest based on how useful it is for tax modeling purposes. Degree of interest ranges from one (1) to four (4), with a one being assigned to returns that are the least interesting, and a four being assigned to those that are the most interesting. 'All' refers to income classes for which returns with all four degrees of interest are assigned.
[4] Positive and Negative Income classes are divided by a Chain-Type Price Index for the Gross Domestic Product of 1.5403 to represent a base year of 1991.