Section 2

Description of the Sample

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

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 2014.

All returns processed during 2014 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 from the estimates.

The estimates in this report are intended to represent all returns filed for Tax Year 2013. While most of the returns processed during Calendar Year 2014 were for Tax Year 2013, the remaining returns were mostly for prior years, and a few for noncalendar years ending during 2012 and 2013.

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 (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 B 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 2014 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, 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.

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, SOI selected a small subsample of returns to independently review, analyze, and process 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 Submission Processing

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.

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 processing center completed its review, SOI further validated, tested, and balanced the data. 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, SOI reviewed all statistics and tables 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 2013, some 0.01 percent of the sample returns were unavailable.

Method of Estimation

SOI obtained the weights by dividing the population count of returns in a stratum by the number of sample returns for that stratum, then adjusted the weights 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). 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 $27.849 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:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$27.849 \times 10^9) \bullet (0.0069) \\ &= \$0.192 \text{ billion} \end{aligned}$$

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 $27.657 billion to $28.041 billion, the 95-percent confidence interval is from $27.465 billion to $28.233 billion, and the 99-percent confidence interval is from $27.273 billion to $28.425 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 2012 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 December 5, 2013).

Table B. Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2013

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	147,759,485	332,040
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 Positive Income or Indexed Positive Income	305	305
Under $5,000,000 Indexed Positive Income or Indexed Positive Income [2]	38,811	35,687
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	308	308
Other Returns, total	147,720,061	295,740

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		6,156,223	70,514	23,174,522	56,130	1,295,747	7,242	117,093,569	161,854		
Indexed Negative Income [4]											
$10,000,000 or more	All	384	384	1,144	1,144	166	166	1,337	1,337	3,031	3,031
$5,000,000 under $10,000,000	All	735	735	1,916	1,916	270	270	2,399	2,399	5,320	5,320
$2,000,000 under $5,000,000	All	3,244	1,050	7,264	2,405	1,066	367	9,131	3,160	20,705	6,982
$1,000,000 under $2,000,000	All	7,010	1,087	14,517	2,322	2,499	408	18,248	2,886	42,274	6,703
$500,000 under $1,000,000	All	15,847	518	33,292	1,122	5,914	205	42,825	1,475	97,878	3,320
$250,000 under $500,000	All	32,526	348	71,010	680	11,218	116	93,485	938	208,239	2,082
$120,000 under $250,000	All	57,180	284	136,736	699	17,845	102	197,249	948	409,010	2,033
$60,000 under $120,000	All	61,535	166	166,015	510	19,065	68	271,823	815	518,438	1,559
Under $60,000	All	43,191	74	386,000	732	25,659	50	622,009	1,146	1,076,859	2,002
Indexed Positive Income [3]											
Under $30,000	1							33,686,309	33,739	33,686,309	33,739
Under $30,000	2	251,975	270	3,808,738	3,789	72,874	85	27,512,372	27,719	31,645,959	31,863
Under $30,000	3-4	280,702	285	5,694,949	5,689	93,437	105	6,975,566	6,942	13,044,654	13,021
$30,000 under $60,000	1-2	588,668	569	1,865,355	1,826	141,241	146	21,586,545	21,422	24,181,809	23,963
$30,000 under $60,000	3-4	652,637	614	3,767,147	3,734	227,326	201	6,815,208	6,849	11,462,318	11,398
$60,000 under $120,000	1-3	1,052,156	1,082	2,178,941	2,180	193,101	205	11,454,447	11,499	14,878,645	14,966
$60,000 under $120,000	4	764,428	688	2,504,819	2,549	179,114	171	3,253,257	3,278	6,701,618	6,686
$120,000 under $250,000	1-3	347,974	1,173	405,068	1,341	75,948	281	1,374,800	4,608	2,203,790	7,403
$120,000 under $250,000	4	945,124	3,123	1,419,302	4,750	98,839	324	2,199,154	7,272	4,662,419	15,469
$250,000 under $500,000	All	608,760	4,407	511,606	3,779	80,642	596	732,749	5,243	1,933,757	14,025
$500,000 under $1,000,000	All	271,434	6,722	147,371	3,620	35,937	861	182,243	4,648	636,985	15,851
$1,000,000 under $2,000,000	All	104,733	12,898	38,918	4,794	10,587	1,241	45,129	5,474	199,367	24,407
$2,000,000 under $5,000,000	All	47,386	15,443	11,628	3,763	2,531	806	13,748	4,521	75,293	24,533
$5,000,000 under $10,000,000	All	11,432	11,432	1,949	1,949	342	342	2,367	2,367	16,090	16,090
$10,000,000 or more	All	7,162	7,162	837	837	126	126	1,169	1,169	9,294	9,294

[1] This population includes an estimated 408,186 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 3,124 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 1.5156 to represent a base year of 1991.
Source: IRS, Statistics of Income Division, Publication 1304, August 2015.