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 2013.

All returns processed during 2013 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, were excluded in calculating estimates. This resulted in a small difference between the population total (145,021,073 returns) reported in Table B and the estimated total of all returns (144,928,472) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 2012. While most of the returns processed during Calendar Year 2013 were for Tax Year 2012, the remaining returns were mostly for prior years, and a few for non calendar years ending during 2011 and 2012. Returns for prior years were used in place of 2012 returns received and processed after December 31, 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.

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

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 2013 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 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. (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. The editors use a hardcopy of the taxpayer's return to enter the required information onto the online system.

After the completion of 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 2012, 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. 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 $22.005 billion, and its related coefficient of variation, CV(X), is 0.70 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$22.005 \times 10^9) \bullet (0.0070) \\ &= \$0.154 \text{ billion} \end{aligned}$$

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 $21.851 billion to $22.159 billion, the 95-percent confidence interval is from $21.697 billion to $22.313 billion, and the 99-percent confidence interval is from $21.543 billion to $22.467 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 2011 to the fourth quarter of the base year of 1991. The indices were calculated using the Gross Domestic Product (GDP) Chain-type Price Index. (See reference 4 for details.)

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 05, 2013).

Table B. Number of Individual Income Tax Returns in the Population and Sample, by Sampling Strata, Tax Year 2012

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	145,021,073	338,475
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	35,067	35,067
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	323	323
Other returns, total	144,985,683	303,085

Description of the sample strata	Degree of interest [2]	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		5,876,685	76,739	22,705,982	56,338	1,312,887	7,482	115,090,129	162,526		
Indexed negative income [3]											
$10,000,000 or more	All	428	428	1,119	1,119	176	176	1,353	1,353	3,076	3,076
$5,000,000 under $10,000,000	All	804	804	1,908	1,908	269	269	2,329	2,329	5,310	5,310
$2,000,000 under $5,000,000	All	3,495	1,169	6,929	2,312	1,086	402	8,996	3,140	20,506	7,023
$1,000,000 under $2,000,000	All	7,469	1,166	14,381	2,255	2,442	410	18,433	2,882	42,725	6,713
$500,000 under $1,000,000	All	17,413	580	33,696	1,177	5,777	212	42,735	1,507	99,621	3,476
$250,000 under $500,000	All	35,736	375	73,271	719	10,825	112	95,546	921	215,378	2,127
$120,000 under $250,000	All	65,547	321	143,187	729	17,593	80	204,694	998	431,021	2,128
$60,000 under $120,000	All	72,235	205	174,367	532	19,396	80	289,392	844	555,390	1,661
Under $60,000	All	51,941	101	396,359	720	26,127	59	729,462	1,380	1,203,889	2,260
Indexed positive income [3]											
Under $30,000	1							32,280,106	32,521	32,280,106	32,521
Under $30,000	2	268,131	270	3,760,070	3,686	76,692	81	28,195,074	28,223	32,299,967	32,260
Under $30,000	3-4	242,688	224	5,513,195	5,506	97,524	88	6,931,022	6,975	12,784,429	12,793
$30,000 under $60,000	1-2	631,526	583	1,858,904	1,877	145,843	138	21,166,111	21,027	23,802,384	23,625
$30,000 under $60,000	3-4	575,189	599	3,737,331	3,696	231,232	230	6,713,304	6,683	11,257,056	11,208
$60,000 under $120,000	1-3	1,017,591	1,044	2,122,717	2,105	192,073	214	10,957,433	10,956	14,289,814	14,319
$60,000 under $120,000	4	699,989	661	2,447,927	2,406	179,151	163	3,151,434	3,139	6,478,501	6,369
$120,000 under $250,000	1-3	314,366	1,039	370,972	1,196	75,590	247	1,231,551	4,116	1,992,479	6,598
$120,000 under $250,000	4	871,452	2,883	1,364,891	4,580	98,378	308	2,124,197	7,056	4,458,918	14,827
$250,000 under $500,000	All	556,402	4,090	484,774	3,546	81,765	586	693,768	5,007	1,816,709	13,229
$500,000 under $1,000,000	All	259,913	6,523	143,677	3,557	36,868	909	181,365	4,604	621,823	15,593
$1,000,000 under $2,000,000	All	108,452	13,271	39,829	4,960	10,832	1,311	49,331	6,029	208,444	25,571
$2,000,000 under $5,000,000	All	52,591	17,076	12,951	4,225	2,676	835	17,364	5,707	85,582	27,843
$5,000,000 under $10,000,000	All	13,875	13,875	2,428	2,428	409	409	3,359	3,359	20,071	20,071
$10,000,000 or more	All	9,452	9,452	1,099	1,099	163	163	1,770	1,770	12,484	12,484

[1] This population includes an estimated 92,601 returns that were excluded from other tables in this report because they contained no income information or represented amended or tentative returns identified after sampling.
[2] 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.
[3] 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, July 2014