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 2012.

All returns processed during 2012 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 (145,601,196 returns) reported in Table C and the estimated total of all returns (145,370,240) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 2011. While most of the returns processed during Calendar Year 2012 were for Tax Year 2011, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2010 and 2011. Returns for prior years were used in place of 2011 returns received and processed after December 31, 2012. This was done based on the assumption that the characteristics of returns due, but not yet processed, can best be represented by the returns for previous income years that were processed in 2012.

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.

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.

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 2012 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 was 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 2011, 0.024 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 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 $27.532 billion, and its related coefficient of variation, CV(X), is 0.74 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$27.532 \times 10^9) \bullet (0.0074) \\ &= \$0.204 \text{ billion} \end{aligned}$$

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 $27.328 billion to $27.736 billion, the 95 percent confidence interval is from $27.124 billion to $27.940 billion, and the 99 percent confidence interval is from $26.920 billion to $28.144 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 2010 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 Indexex for Gross Domestic Product," [http://www.bea.gov/] (accessed November 22, 2011).

Table C.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2011

Description of the sample strata	Number of returns	
	Population counts[1]	Sample counts
Grand total	145,601,196	333,106
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	34,663	34,663
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	305	305
Other Returns, total	145,566,228	298,138

Description of the sample strata	Degree of interest [2]	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		Number of returns	
		Population counts	Sample counts	Population counts	Sample counts	Population counts	Sample counts	Population counts	Sample counts	Population counts	Sample counts	Population counts[1]	Sample counts
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)		
Total		435,415	17,624	5,364,481	62,286	22,533,092	53,455	1,343,562	6,794	115,889,678	157,979		
Indexed Negative Income [3]													
$10,000,000 or more	All	9	9	465	465	1,186	1,186	190	190	1,387	1,387	3,237	3,237
$5,000,000 under $10,000,000	All	23	23	851	851	1,932	1,932	286	286	2,420	2,420	5,512	5,512
$2,000,000 under $5,000,000	All	78	76	3,949	1,305	7,222	2,457	1,136	417	9,438	3,192	21,823	7,447
$1,000,000 under $2,000,000	All	228	214	8,264	1,287	14,825	2,335	2,584	413	19,064	3,039	44,965	7,288
$500,000 under $1,000,000	All	615	240	19,485	643	35,169	1,146	6,194	224	44,508	1,502	105,971	3,755
$250,000 under $500,000	All	1,763	153	39,472	403	77,310	749	11,859	121	99,966	964	230,370	2,390
$120,000 under $250,000	All	5,545	489	70,660	330	152,321	784	18,958	106	214,666	1,104	462,150	2,813
$60,000 under $120,000	All	12,347	247	72,943	180	188,482	561	20,603	65	308,030	943	602,405	1,996
Under $60,000	All	16,098	173	45,751	78	420,433	785	35,052	72	839,518	1,588	1,356,852	2,696
Indexed Positive Income [3]													
Under $30,000	1									31,860,343	32,074	31,860,343	32,074
Under $30,000	2	4,973	53	245,085	217	3,717,978	3,775	83,244	83	29,626,199	29,448	33,677,479	33,576
Under $30,000	3-4	88,046	852	147,161	133	5,469,755	5,440	104,195	107	7,060,171	7,022	12,869,328	13,554
$30,000 under $60,000	1-2	5,293	66	639,774	659	1,858,288	1,814	152,958	157	21,215,697	21,298	23,872,010	23,994
$30,000 under $60,000	3-4	95,061	998	479,522	481	3,745,988	3,745	239,428	219	6,633,934	6,818	11,193,933	12,261
$60,000 under $120,000	1-3	9,307	195	1,006,167	1,031	2,111,265	2,081	198,213	213	10,831,664	10,772	14,156,616	14,292
$60,000 under $120,000	4	105,686	2,168	598,432	586	2,428,298	2,481	179,083	177	3,084,045	3,085	6,395,544	8,497
$120,000 under $250,000	1-3	15,410	1,303	280,194	913	344,183	1,148	76,300	270	1,124,819	3,751	1,840,906	7,385
$120,000 under $250,000	4	43,135	3,626	828,222	2,768	1,331,853	4,381	94,786	292	2,085,492	6,948	4,383,488	18,015
$250,000 under $500,000	All	20,789	1,785	505,739	3,662	451,929	3,299	76,455	544	620,840	4,442	1,675,752	13,732
$500,000 under $1,000,000	All	7,574	2,998	225,633	5,552	128,203	3,209	31,237	763	152,067	3,765	544,714	16,287
$1,000,000 under $2,000,000	All	2,368	894	89,185	10,916	33,504	4,212	8,462	1,041	39,148	4,727	172,667	21,790
$2,000,000 under $5,000,000	All	818	813	40,899	13,198	10,344	3,311	1,944	639	12,781	4,209	66,786	22,170
$5,000,000 under $10,000,000	All	169	169	10,201	10,201	1,839	1,839	288	288	2,360	2,360	14,857	14,857
$10,000,000 or more	All	80	80	6,427	6,427	785	785	107	107	1,121	1,121	8,520	8,520

[1] This population includes an estimated 230,956 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 1.4783 to represent a base year of 1991.