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 2003.

All returns processed during 2003 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 (130,540,073 returns) reported in Table C and the estimated total of all returns (130,076,443) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 2002. While about 98 percent of the returns processed during Calendar Year 2003 were for Tax Year 2002, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2003 and 2004. Returns for prior years were used in place of 2002 returns received and processed after December 31, 2003. 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 2003.

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 with adjusted gross income or expanded income of $200,000 or more and no alternative minimum tax.

2. High combined business and farm total 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).

Bonnye Walker, Valerie Testa, and Jana Scali designed the sample and prepared the text and tables in this section under the direction of Yahia Ahmed, Chief, Mathematical Statistics Section, Statistical Computing Branch.

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 0.05 percent to 100 percent.

Tax data processed to the IRS Individual Master File at the Enterprise Computing Center at Martinsburg during Calendar Year 2003 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 2002, 0.13 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). Table 1.4 CV contains estimated CV's for the estimates included in Table 1.4 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 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 amount estimate for State Income Tax Refunds, X, is $23.876 billion, and its related coefficient of variation, CV(X), is 0.85 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$23.876 \times 10^{9}) \bullet (0.0085) \\ &= \$0.203 \text{ billion} \end{aligned}$$

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 $23.673 billion to $24.079 billion, the 95 percent confidence interval is from $23.470 billion to $24.282 billion, and the 99 percent confidence interval is from $23.267 billion to $24.485 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 under essentially the 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 2001 to the fourth quarter of the base year of 1991. The indices were calculated using the Gross Domestic Product (GDP) Chain-type Price Index found in the table titles "Quantity and Price Indexes for Gross Domestic Product" released to the public on November 26, 2002 on the BEA web site (http://www.bea.doc.gov/).

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.

Table C.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2002

Description of the sample strata	Degree of interest [2]	Number of Returns by type of form attached								Number of returns	
		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		All other forms			
		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)		
Grand total										130,540,073	175,566
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										7,109	7,109
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total										174	174
Other Returns, total										130,532,790	168,283
Total		2,829,506	30,941	18,574,439	38,694	1,467,701	4,258	107,661,144	94,390		
Indexed Negative Income [3]											
$10,000,000 or more	All	281	281	767	767	87	87	960	960	2,095	2,095
$5,000,000 under $10,000,000	All	445	445	922	922	187	187	1,380	1,380	2,934	2,934
$2,000,000 under $5,000,000	All	2,000	645	4,002	1,267	734	249	5,526	1,810	12,262	3,971
$1,000,000 under $2,000,000	All	4,266	672	9,082	1,453	1,856	287	11,573	1,814	26,777	4,226
$500,000 under $1,000,000	All	9,698	321	24,198	814	5,125	169	28,001	930	67,022	2,234
$250,000 under $500,000	All	18,493	167	57,753	556	12,371	121	63,966	581	152,583	1,425
$120,000 under $250,000	All	32,079	157	120,725	543	22,537	108	142,922	618	318,263	1,426
$60,000 under $120,000	All	36,378	99	166,108	418	24,436	65	219,790	571	446,712	1,153
Under $60,000	All	37,550	50	432,059	563	45,486	65	1,064,521	1,504	1,579,616	2,182
Indexed Positive Income [3]											
Under $30,000	1							29,713,630	14,781	29,713,630	14,781
Under $30,000	2	160,312	87	2,154,084	1,079	108,422	41	27,251,871	13,671	29,674,689	14,878
Under $30,000	3-4	144,056	136	3,799,078	3,941	156,698	164	5,486,064	5,725	9,585,896	9,966
$30,000 under $60,000	1-2	330,843	167	1,815,237	908	189,277	105	21,551,608	10,550	23,886,965	11,730
$30,000 under $60,000	3-4	276,106	288	3,468,559	3,665	261,152	294	5,632,155	6,133	9,637,972	10,380
$60,000 under $120,000	1-3	481,286	227	1,992,943	1,068	228,745	113	10,531,173	5,178	13,234,147	6,586
$60,000 under $120,000	4	315,781	335	2,377,661	2,457	180,217	155	2,598,580	2,695	5,472,239	5,642
$120,000 under $250,000	1-3	233,313	356	439,009	645	89,886	126	1,512,180	2,173	2,274,388	3,300
$120,000 under $250,000	4	302,904	790	1,115,240	3,235	70,696	196	1,126,707	3,240	2,615,547	7,461
$250,000 under $500,000	All	252,320	1,636	438,282	2,963	51,326	322	520,169	3,492	1,262,097	8,413
$500,000 under $1,000,000	All	114,816	2,848	119,504	2,906	13,606	323	140,624	3,395	388,550	9,472
$1,000,000 under $2,000,000	All	46,348	5,596	28,482	3,518	3,453	431	40,413	4,961	118,696	14,506
$2,000,000 under $5,000,000	All	21,629	7,036	8,537	2,799	1,107	353	13,632	4,529	44,905	14,717
$5,000,000 under $10,000,000	All	5,479	5,479	1,572	1,572	217	217	2,576	2,576	9,844	9,844
$10,000,000 or more	All	3,123	3,123	635	635	80	80	1,123	1,123	4,961	4,961

[1] This population includes an estimated 463,630 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 1.1640 to represent a base year of 1991.