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 2007.

All returns processed during 2007 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 (138,485,355 returns) reported in Table C and the estimated total of all returns (138,139,754) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 2006. While most of the returns processed during Calendar Year 2007 were for Tax Year 2006, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2007 and 2008. Returns for prior years were used in place of 2006 returns received and processed after December 31, 2007. 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 2007.

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, Jana Scali, and Katie Thamert 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.

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 2007 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 2006, 0.17 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 $24.206 billion, and its related coefficient of variation, CV(X), is 0.77 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$24.206 \times 10^9) \bullet (0.0070) \\ &= \$0.169 \text{ billion} \end{aligned}$$

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 $24.037 billion to $24.375 billion, the 95 percent confidence interval is from $23.868 billion to $24.544 billion, and the 99 percent confidence interval is from $23.699 billion to $24.713 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 2005 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 29, 2006 on the BEA web site (http://www.bea.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 2006

Description of the sample strata	Number of returns	
	Population counts [1]	Sample counts
Grand total	138,485,335	321,006
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	13,274	13,274
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	261	261
Other Returns, total	138,471,800	307,471

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		All other forms			
		Population counts	Sample counts	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)	(10)	(11)		
Total		324,044	15,657	4,931,953	70,089	21,266,480	57,767	1,383,973	5,592	110,565,350	158,366		
Indexed Negative Income [3]													
$10,000,000 or more	All	6	6	198	198	609	609	97	97	687	687	1,597	1,597
$5,000,000 under $10,000,000	All	4	4	366	366	961	961	184	184	1,144	1,144	2,659	2,659
$2,000,000 under $5,000,000	All	38	37	1,630	555	3,844	1,279	629	243	4,686	1,533	10,827	3,647
$1,000,000 under $2,000,000	All	107	104	3,585	580	8,707	1,352	1,635	265	9,452	1,449	23,486	3,750
$500,000 under $1,000,000	All	355	143	8,597	281	23,319	766	4,478	151	23,241	770	59,990	2,111
$250,000 under $500,000	All	974	94	17,774	202	57,085	541	10,264	84	54,781	537	140,878	1,458
$120,000 under $250,000	All	2,907	286	30,729	120	124,065	618	17,815	78	124,384	641	299,900	1,743
$60,000 under $120,000	All	7,761	154	31,024	90	168,909	572	18,743	68	187,636	541	414,073	1,425
Under $60,000	All	15,018	151	23,914	40	423,298	762	28,727	58	645,241	1,222	1,136,198	2,233
Indexed Positive Income [3]													
Under $30,000	1									31,663,929	31,670	31,663,929	31,670
Under $30,000	2	5,621	59	231,004	233	2,724,392	2,780	83,508	82	26,320,904	26,129	29,365,429	29,283
Under $30,000	3-4	55,459	582	242,954	361	4,520,526	7,093	115,266	197	5,895,299	9,074	10,829,504	17,307
$30,000 under $60,000	1-2	5,556	63	459,148	449	1,971,856	1,969	168,590	169	21,580,125	21,618	24,185,275	24,268
$30,000 under $60,000	3-4	67,022	680	508,770	815	3,789,084	6,001	256,234	418	6,128,621	9,805	10,749,731	17,719
$60,000 under $120,000	1-3	9,607	218	858,854	865	2,276,452	2,285	220,286	215	11,055,465	10,894	14,420,664	14,477
$60,000 under $120,000	4	72,309	1,507	581,381	851	2,629,243	4,031	190,709	276	2,898,536	4,450	6,372,178	11,115
$120,000 under $250,000	1-3	15,978	1,379	271,042	518	395,554	773	83,773	158	1,195,437	2,395	1,961,784	5,223
$120,000 under $250,000	4	32,632	2,664	765,049	2,494	1,433,341	4,780	91,429	308	1,881,240	6,192	4,203,691	16,438
$250,000 under $500,000	All	20,268	1,693	486,934	3,568	509,536	3,710	64,465	442	638,161	4,598	1,719,364	14,011
$500,000 under $1,000,000	All	7,961	3,134	234,392	5,674	148,820	3,766	19,742	487	178,260	4,494	589,175	17,555
$1,000,000 under $2,000,000	All	2,915	1,159	101,076	12,179	39,596	4,870	5,244	605	52,927	6,438	201,758	25,251
$2,000,000 under $5,000,000	All	1,167	1,161	50,185	16,303	13,405	4,371	1,682	534	19,284	6,173	85,723	28,542
$5,000,000 under $10,000,000	All	257	257	13,921	13,921	2,612	2,612	333	333	4,018	4,019	21,141	21,142
$10,000,000 or more	All	122	122	9,426	9,426	1,266	1,266	140	140	1,892	1,893	12,846	12,847

[1] This population includes an estimated 90,712 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 1.3386 to represent a base year of 1991.