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 2001.

All returns processed during 2001 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 (129,644,980 returns) reported in Table C and the estimated total of all returns (129,367,108) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 2000. While about 98 percent of the returns processed during Calendar Year 2001 were for Tax Year 2000, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2001 and 2002. Returns for prior years were used in place of 2000 returns expected to be received and processed after December 31, 2001. 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 2001.

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 and Valerie Puckett 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 Martinsburg Computing Center during Calendar Year 2001 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 at the Detroit Computing Center. 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 2000, 0.20 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 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 amount estimate for State Income Tax Refunds, X, is $18.310 billion, and its related coefficient of variation, CV(X), is 0.92 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$18.310 \times 10^{9}) \bullet (0.0092) \\ &= \$0.168 \text{ billion} \end{aligned}$$

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 $18.141 billion to $18.478 billion, the 95 percent confidence interval is from $17.973 billion to $18.647 billion, and the 99 percent confidence interval is from $17.806 billion to $18.814 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 1999 to the fourth quarter of the base year of 1991. The indices can be found in U. S. Department of Commerce, Bureau of Economic Analysis, *Survey of Current Business* (January 2001) Vol. 81, number 1.

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 2000

Description of the sample strata										Number of returns	
										Population counts[1]	Sample counts
Grand total										129,644,980	196,149
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[2]										4,114	4,114
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total[3]										1,025	1,025
Other Returns, total										129,639,841	191,010
		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		All other forms			
Description of the sample strata	Degree of interest[4]	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		3,027,283	45,838	17,555,465	39,146	1,498,052	4,572	107,559,041	101,454		
Indexed Negative Income[5]											
$10,000,000 or more	All	123	123	535	535	74	74	673	673	1,405	1,405
$5,000,000 under $10,000,000	All	136	136	669	669	133	133	866	866	1,804	1,804
$2,000,000 under $5,000,000	All	583	213	2,720	847	605	226	3,308	1,060	7,216	2,346
$1,000,000 under $2,000,000	All	1,127	185	6,219	955	1,493	232	6,502	1,094	15,341	2,466
$500,000 under $1,000,000	All	2,392	85	16,715	604	4,171	135	15,394	534	38,672	1,358
$250,000 under $500,000	All	0	0	44,380	417	9,988	115	35,079	317	89,447	849
$120,000 under $250,000	All	7,349	42	84,048	366	17,081	59	73,798	318	182,276	785
$60,000 under $120,000	All	0	0	128,477	323	17,725	44	106,064	308	252,266	675
Under $60,000	All	0	0	328,470	464	32,144	51	411,047	578	771,661	1,093
Indexed Positive Income[5]											
Under $30,000	1							27,785,946	13,699	27,785,946	13,699
Under $30,000	2	160,379	79	1,885,834	1,031	101,811	55	29,080,327	14,672	31,228,351	15,837
Under $30,000	3-4	208,085	212	3,429,032	3,527	157,851	167	6,045,127	6,374	9,840,095	10,280
$30,000 under $60,000	1-2	222,379	117	1,699,023	871	180,104	74	21,194,368	10,500	23,295,874	11,562
$30,000 under $60,000	3-4	338,879	394	3,348,934	3,598	273,876	290	5,749,953	6,216	9,711,642	10,498
$60,000 under $120,000	1-3	449,552	213	1,960,884	1,001	235,555	123	10,685,658	5,233	13,331,649	6,570
$60,000 under $120,000	4	394,887	402	2,332,571	2,367	196,263	189	2,566,216	2,575	5,489,937	5,533
$120,000 under $250,000	1-3	268,140	369	497,036	728	103,647	155	1,783,172	2,562	2,651,995	3,814
$120,000 under $250,000	4	382,565	1,135	1,132,060	3,297	80,401	230	1,113,314	3,232	2,708,340	7,894
$250,000 under $500,000	All	318,792	2,112	476,221	3,217	61,916	396	628,683	4,150	1,485,612	9,875
$500,000 under $1,000,000	All	152,572	3,755	133,491	3,308	17,065	396	185,029	4,380	488,157	11,839
$1,000,000 under $2,000,000	All	67,695	8,218	33,932	4,146	4,311	553	58,474	7,268	164,412	20,185
$2,000,000 under $5,000,000	All	34,750	11,151	10,936	3,598	1,441	478	22,759	7,561	69,886	22,788
$5,000,000 under $10,000,000	All	9,924	9,924	2,236	2,236	266	266	4,911	4,911	17,337	17,337
$10,000,000 or more	All	6,974	6,973	1,042	1,041	131	131	2,373	2,373	10,520	10,518

[1] This population includes an estimated 267,872 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] This population includes 172 Form 1040 returns that were misclassified because of bad data collected during revenue processing.

[3] This population includes 787 Form 1040 returns that were filed and processed during Calendar Year 2002 as a result of legislation for taxpayers affected by the events of September 11, 2001.

[4] 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.

[5] 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.

** Sampling Strata Collapsed.