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, 1040EZ, and 1040PC (including electronic returns) filed by U.S. citizens and residents during Calendar Year 2000.

All returns processed during 2000 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 (127,321,626 returns) reported in Table C and the estimated total of all returns (127,075,145) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 1999. While about 98 percent of the returns processed during Calendar Year 2000 were for Tax Year 1999, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 1999 and 2000. Returns for prior years were used in place of 1999 returns expected to be received and processed after December 31, 2000. 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 2000.

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 2000 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 Service 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 1999, 0.11 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 $17.976 billion, and its related coefficient of variation, CV(X), is 0.97 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE\,(X) &= X \bullet CV(X) \\ &= (\$17.976 \times 10^{9}) \bullet (0.0097) \\ &= \$0.174 \text{ billion} \end{aligned}$$

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 $17.802 billion to $18.15 billion, and the 95 percent confidence interval is from $17.628 billion to $18.324 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 (- or --) 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 1998 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 1999) Vol. 79, 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.

SOURCE: IRS, Individual Income Tax Returns-1999, Publication 1304, Revised 10-2001.

See next page for table

Table C.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 1999

Description of the sample strata										Number of returns	
										Population counts	Sample counts
Grand total										127,321,626	176,966 [1]
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										3,238	3,238 [2]
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total										160	160
Other Returns, total										127,318,228	173,568
		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 [3]	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		2,698,596	36,528	17,272,967	36,746	1,521,415	4,470	105,825,250	95,824		
Indexed Negative Income [4]											
$10,000,000 or more	All	101	101	504	504	65	65	586	586	1,256	1,256
$5,000,000 under $10,000,000	All	86	86	609	609	121	121	750	750	1,566	1,566
$2,000,000 under $5,000,000	All	346	103	2,349	741	533	190	2,673	862	5,901	1,896
$1,000,000 under $2,000,000	All	703	100	5,188	818	1,312	214	5,192	847	12,395	1,979
$500,000 under $1,000,000	All	1,472	54	14,089	498	3,990	123	12,007	401	31,558	1,076
$250,000 under $500,000	All	3,007	35	34,810	310	9,768	78	27,489	258	75,074	681
$120,000 under $250,000	All	5,467	34	75,090	352	17,257	89	58,046	267	155,860	742
$60,000 under $120,000	All	**	**	117,062	292	17,810	36	87,367	224	222,239	552
Under $60,000	All	**	**	321,426	425	33,741	52	327,804	446	682,971	923
Indexed Positive Income [4]											
Under $30,000	1							27,809,524	13,804	27,809,524	13,804
Under $30,000	2	143,649	65	1,874,895	973	108,513	62	29,242,683	14,749	31,369,740	15,849
Under $30,000	3-4	199,772	223	3,464,052	3,586	172,357	188	6,205,425	6,492	10,041,606	10,489
$30,000 under $60,000	1-2	198,137	101	1,686,282	787	184,402	83	20,613,240	10,179	22,682,061	11,150
$30,000 under $60,000	3-4	314,375	373	3,351,363	3,562	281,068	299	5,618,229	6,224	9,565,035	10,458
$60,000 under $120,000	1-3	408,896	191	1,874,804	959	232,413	120	10,025,047	4,905	12,541,160	6,175
$60,000 under $120,000	4	350,365	355	2,274,376	2,361	190,886	161	2,374,629	2,408	5,190,256	5,285
$120,000 under $250,000	1-3	243,101	367	466,388	680	106,656	139	1,584,226	2,346	2,400,371	3,532
$120,000 under $250,000	4	328,531	958	1,085,930	3,115	76,074	198	1,017,036	2,910	2,507,571	7,181
$250,000 under $500,000	All	277,335	1,849	454,376	3,100	61,525	371	567,361	3,727	1,360,597	9,047
$500,000 under $1,000,000	All	128,630	3,105	125,068	2,979	16,675	404	166,746	4,029	437,119	10,517
$1,000,000 under $2,000,000	All	54,290	6,581	31,129	3,767	4,280	542	52,437	6,447	142,136	17,337
$2,000,000 under $5,000,000	All	27,424	8,938	10,170	3,321	1,532	498	20,333	6,545	59,459	19,302
$5,000,000 under $10,000,000	All	7,813	7,813	2,015	2,015	302	302	4,273	4,273	14,403	14,403
$10,000,000 or more	All	5,096	5,096	992	992	135	135	2,147	2,145	8,370	8,368

[1] This population includes an estimated 246,481 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 39 Form 1040 returns that were misclassified because of bad data collected during revenue processing.

[3] 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.

[4] Positive and Negative Income classes are divided by a Chain-Type Price Index for the Gross Domestic Product of 1.1480 to represent a base year of 1991.

** Sampling Strata Collapsed.