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 1999.

All returns processed during 1999 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 (125,037,636 returns) reported in Table C and the estimated total of all returns (124,770,662) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 1998. While about 98 percent of the returns processed during Calendar Year 1999 were for Tax Year 1998, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 1998 and 1999. Returns for prior years were used in place of 1998 returns expected to be received and processed after December 31, 1999. 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 1999.

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 1999 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 1998, 0.08 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 $14.708 billion, and its related coefficient of variation, CV(X), is 0.98 percent. The standard error of the estimate, SE(X), needed to construct the confidence interval estimate, is:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$14.708 \times 10^9) \bullet (0.0098) \\ &= \$0.144 \text{ billion} \end{aligned}$$

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 $14.564 billion to $14.852 billion, and the 95 percent confidence interval is from $14.420 billion to $14.996 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 1997 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* (November 1998) Vol. 78, number 11.

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, Statistics of Income – 1998,
Individual Income Tax Returns, Publication 1304,
Revised 4-01

Table C.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 1998

Description of the sample strata	Number of returns	
	Population counts	Sample counts
Grand total	125,037,636	164,340 [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	2,907	2,907 [2]
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	86	86
Other Returns, total	125,034,643	161,347

Description of the sample strata	Degree of interest [3]	Number of Returns by type of form attached								Population counts	Sample counts
		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		
	(1)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)		
Total		2,465,704	31,283	17,118,658	35,196	1,546,182	4,403	103,904,099	90,465		
Indexed Negative Income [4]											
$10,000,000 or more	All	78	78	484	484	73	73	577	577	1,212	1,212
$5,000,000 under $10,000,000	All	96	96	622	622	121	121	699	699	1,538	1,538
$2,000,000 under $5,000,000	All	332	101	2,247	732	534	182	2,536	820	5,649	1,835
$1,000,000 under $2,000,000	All	671	113	4,932	760	1,313	220	5,073	833	11,989	1,926
$500,000 under $1,000,000	All	1,444	48	13,819	456	4,010	123	11,835	404	31,108	1,031
$250,000 under $500,000	All	2,905	39	34,515	334	10,370	102	26,680	247	74,470	722
$120,000 under $250,000	All	**	**	80,691	356	19,244	82	57,692	265	157,627	703
$60,000 under $120,000	All	**	**	117,403	299	20,446	47	87,359	228	225,208	574
Under $60,000	All	**	**	329,368	431	38,812	60	315,763	407	683,943	898
Indexed Positive Income [4]											
Under $30,000	1							27,033,158	13,427	27,033,158	13,427
Under $30,000	2	74,599	44	1,870,503	970	116,952	55	29,186,440	14,582	31,248,494	15,651
Under $30,000	3-4	251,336	272	3,531,202	3,645	185,113	200	6,939,555	7,322	10,907,206	11,439
$30,000 under $60,000	1-2	167,242	89	1,718,727	827	194,278	91	20,134,368	9,956	22,214,615	10,963
$30,000 under $60,000	3-4	307,440	349	3,365,480	3,611	286,784	299	5,517,314	6,023	9,477,018	10,282
$60,000 under $120,000	1-3	389,111	181	1,817,425	967	230,539	112	9,414,789	4,698	11,851,864	5,958
$60,000 under $120,000	4	322,834	330	2,201,565	2,258	182,835	172	2,182,254	2,207	4,889,488	4,967
$120,000 under $250,000	1-3	227,463	362	436,583	639	106,446	161	1,399,085	2,009	2,169,577	3,171
$120,000 under $250,000	4	284,199	810	1,018,129	2,925	69,351	180	883,073	2,551	2,254,752	6,466
$250,000 under $500,000	All	243,266	1,636	421,230	2,850	57,665	358	493,608	3,298	1,215,769	8,142
$500,000 under $1,000,000	All	112,063	2,676	113,905	2,834	15,408	388	144,754	3,594	386,130	9,492
$1,000,000 under $2,000,000	All	46,392	5,677	27,867	3,462	4,018	513	45,033	5,562	123,310	15,214
$2,000,000 under $5,000,000	All	23,483	7,632	9,191	2,964	1,457	451	17,220	5,522	51,351	16,569
$5,000,000 under $10,000,000	All	6,595	6,595	1,879	1,879	277	277	3,472	3,472	12,223	12,223
$10,000,000 or more	All	4,155	4,155	891	891	136	136	1,762	1,762	6,944	6,944

[1] This population includes an estimated 266,974 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 167 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 1.1403 to represent a base year of 1991.

** Data combined.