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 2004.

All returns processed during 2004 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 (131,291,334 returns) reported in Table C and the estimated total of all returns (130,423,626) reported in other tables.

The estimates in this report are intended to represent all returns filed for Tax Year 2003. While about 98 percent of the returns processed during Calendar Year 2004 were for Tax Year 2003, the remaining returns were mostly for prior years, and a few for non-calendar years ending during 2004 and 2005. Returns for prior years were used in place of 2003 returns received and processed after December 31, 2004. 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 2004.

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.

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).

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

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.

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 2004 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 2003, 0.04 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 $23.425 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:

$$\begin{aligned} SE(X) &= X \bullet CV(X) \\ &= (\$23.425 \times 10^9) \bullet (0.0085) \\ &= \$0.199 \text{ billion} \end{aligned}$$

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 $23.226 billion to $23.624 billion, the 95 percent confidence interval is from $23.027 billion to $23.823 billion, and the 99 percent confidence interval is from $22.828 billion to $24.022 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 2002 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 30, 2003 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.

Source: IRS, Statistics of Income, Individual Complete Report 2003, Publication 1304, October 2005.

Table C.—Number of Individual Income Tax Returns in the Population and Sample by Sampling Strata for 2003

Description of the sample strata	Number of returns	
	Population counts[1]	Sample counts
Grand total	131,291,334	182,810
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	7,161	7,161
Form 1040 returns only with combined Schedule C (business or profession) total receipts of $50,000,000 and over, total	164	164
Other Returns, total	131,284,009	175,485

Description of the sample strata	Degree of interest [2]	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		Form 1040, with other Schedules and Forms and Forms 1040A and 1040EZ			
		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		3,035,704	34,231	19,345,282	40,848	1,451,910	4,555	107,451,113	95,851		
Indexed Negative Income [3]											
$10,000,000 or more	All	281	281	813	813	93	93	1,053	1,053	2,240	2,240
$5,000,000 under $10,000,000	All	502	502	967	967	197	197	1,451	1,451	3,117	3,117
$2,000,000 under $5,000,000	All	2,200	739	4,061	1,278	762	280	5,720	1,953	12,743	4,250
$1,000,000 under $2,000,000	All	4,617	767	9,320	1,512	1,903	304	11,619	1,820	27,459	4,403
$500,000 under $1,000,000	All	10,529	335	24,909	802	5,080	162	28,924	947	69,442	2,246
$250,000 under $500,000	All	20,564	202	59,691	564	11,657	111	66,441	593	158,353	1,470
$120,000 under $250,000	All	36,417	147	125,678	602	20,273	96	149,302	658	331,670	1,503
$60,000 under $120,000	All	41,862	114	173,402	466	22,527	56	231,682	606	469,473	1,242
Under $60,000	All	41,020	58	455,324	640	42,921	65	1,082,502	1,524	1,621,767	2,287
Indexed Positive Income [3]											
Under $30,000	1							30,466,670	15,298	30,466,670	15,298
Under $30,000	2	158,036	84	2,331,855	1,159	100,988	41	26,167,849	13,151	28,758,728	14,435
Under $30,000	3-4	144,301	134	4,007,738	4,229	150,619	158	5,427,135	5,698	9,729,793	10,219
$30,000 under $60,000	1-2	343,966	182	1,852,588	951	178,566	98	21,354,587	10,554	23,729,707	11,785
$30,000 under $60,000	3-4	295,889	319	3,517,419	3,652	256,044	289	5,629,829	6,074	9,699,181	10,334
$60,000 under $120,000	1-3	530,166	262	2,100,607	1,055	228,012	109	10,827,026	5,300	13,685,811	6,726
$60,000 under $120,000	4	344,250	331	2,427,361	2,474	184,010	148	2,540,774	2,546	5,496,395	5,499
$120,000 under $250,000	1-3	210,047	317	404,439	609	89,597	127	1,301,190	1,844	2,005,273	2,897
$120,000 under $250,000	4	370,869	1,024	1,220,586	3,515	78,186	222	1,420,912	4,077	3,090,553	8,838
$250,000 under $500,000	All	270,345	1,727	458,403	3,101	58,414	358	533,713	3,547	1,320,875	8,733
$500,000 under $1,000,000	All	125,287	2,978	127,251	3,167	16,372	395	143,588	3,597	412,498	10,137
$1,000,000 under $2,000,000	All	50,875	6,199	30,948	3,811	4,119	505	41,005	5,009	126,947	15,524
$2,000,000 under $5,000,000	All	23,826	7,674	9,467	3,026	1,254	425	14,092	4,502	48,639	15,627
$5,000,000 under $10,000,000	All	6,173	6,173	1,701	1,701	223	223	2,665	2,665	10,762	10,762
$10,000,000 or more	All	3,682	3,682	754	754	93	93	1,384	1,384	5,913	5,913

[1] This population includes an estimated 867,708 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 1.2297 to represent a base year of 1991.