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 (138,485,466 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:

SE (X) =
$$X \cdot CV(X)$$

= (\$24.206 × 10°) • (0.0070)
= \$0.169 billion

The p percent confidence interval is calculated using the formula:

$$X \pm z \cdot SE(X)$$

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

tied gross income or expanded income of \$200,000 and over, with no income tax after credits at sined Schedule C (business or profession) total receipts of \$50,000,000 and over, total interest [2]	on) total receipts of \$50,000,000 and over, total receipts of \$50,000,000 and over, total with Form 1040, with Form 2555 but with counts counts counts (2) (3) (4) (4) (4) (5) (5) (5) (6) (6) (7) (6) (7) (7) (7) (7) (7) (7) (7) (7) (7) (7	m 1040. m 1040. counts 555 Sample counts (5) 70,082 70,082 70,082 70,082 70,082 860 860 860	tional tax for tax preferences Form 1040, with Schedule C but without Form 1116 or Form 2555 Population Sampt Counts (6) (7) 21,266,480 (7) 21,266,480 (6) 961 961 961	eferences, total eferences, total use C or 116 counts (7) 57,767 609 961 961 4270	Form 10- Schedule F. Schedule C, or Form Population counts (8) (1,383,972	Form 1040, with Schedule F but without Schedule C, Form 116 or Form 2555 population Sample counts (9) (9) (9) (5500	All other forms Population Sa counts 00 (10) (10) (10) (10) (10) (10) (10) (sr forms Sample Counts (10) 158,365	138,485,446 13,388 13,8471,783 1,597 2,658	317,975 10,279 235 307,461
Form 1040, with Form 1040,	Mm 1040, Form 2555 Sample Counts Coun	m 1040, Form 1116 Aut Form 2555 Sample Counts (5) 70,082 198 366 555 555 588	Form 10, with Sched but without Fo or Form 2 or Form 2 counts (6) 21,266,480 951 3,844	40, Iule C 2555 Sample Sample (7) 57,767 609 961 4 270	Form 10. Schedule F. Schedule C. Schedule C. or Form Population Counts (8) 1,383,972	240, with but without Form 1116 m 2555 Sample counts (9) 5,590	All other Population counts (10) (10)	r forms Sample Counts (101) 158,365	1,597	
Poem 1040. With Form 1040. With Form 1040. Degree of interest [2] Population counts cou	Marcology Marc	m 1040, Form 1116 Sample Counts (5) 70,082 70,082 72,082 72,082 556 556 556 580	Form 10 with Sched but without Fo or Form 2 or Form 2 counts (6) 21,266,480 609 961 3,344	40, mm 1116 5555 Sample counts (7) 57,767 609 961 961 9770	Form 10. Schedule F. Schedule F. Schedule F. Schedule for Form or Form counts (8) 1,383,972	240, with but without Form 1116 m 2555 Sample counts (9) 5,590	All other Population counts (10) (10)	sample Sample counts (11) 158,365	1,597	
Degree of interest [2] Prom 1040, with Form 2555 with Form 111 000 All (1) (2) (3) (4) 000 All (1) (3) (4) (4) All (1) (2) (3) (4) (4)	Form 2855 Form 2855 Sample Counts (3) (15,687 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,	Form 1116 out Form 2555 Sample counts (0) (0) (0) (1) 70,082 198 366 555 580 282	but without Fo or Form 2 or Form 2 or Form 2 counts (6) 21,266,480 951 951 951 951 951 951 951 951 951 951	Sample counts (7) 57,767 609 961 977 67	Schedule C, or Form Population counts (8) 1,383,972	Sample counts (9) (5,590	All other Population counts (10) 110,565,350	Sample counts (11) 158,365	1,597	
Degree of Deg	Sample Popu Counts Coun	N 8	Population counts (6) (7) 21,266,480 (609 961 8) 3,844	Sample counts (7) 57,767 609 961	Population counts (8) 1,383,972	Sample counts (9) 5,590	Population counts (10) (10)	Sample counts (11) 158,365	1,597	
(1) (2) (3) (4) (4) (6) (6) (7) (9) (4) (4) (7) (7) (9) (9) (9) (9) (9) (9) (9) (9) (9) (9	(3) (6) (7) (8) (9) (9) (9) (9) (9) (9) (9) (9) (9) (9		(6) 21,266,480 609 961 3,844	(7) 57,767 609 961	(8) 1,383,972	(9)	(10) (110,565,350	(11) 158,365	1,597	4
All 324,044 15,657 4,931,947 All 24,044 15,657 4,931,947 All 386 198 37 1,630 All 2,907 2,86 30,729 All 7,761 15,018 17,774 All 7,761 15,018 1,024 All 7,761 15,018 1,024 All 7,761 15,018 1,024 All 7,761 15,018 1,024 All 7,2309 1,507 581,381 All 20,026 1,697 2,104 All 20,026 1,697 2,104 All 20,268 1,379 271,042 All 20,268 1,379 271,042 All 20,268 1,379 271,042 All 20,268 1,379 271,042 All 24,363 1,374 24,363 All 20,268 1,374 24,363	15,657 4,93 6 6 6 6 7 4 4 4 7 7 104 07 104 07 104 094 1 1 154 1 151 2 181 1 2 181 1 151 1 1 151 1 1 151 1 1 1 1 151 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		21,266,480 609 961 3,844	57,767 609 961 1,700	1,383,972	5,590	110,565,350	158,365	1,597	
1 All 6 6 198 198 190	6 4 6 104 143 143 154 154 157 157		609 961 13,844	609	0			000	1,597	101
D00,000 All 38 0,000 All 355 0,000 All 355 All 7,774 00 All 2,907 All 7,774 15,018 1,774 All 15,018 1,379 1,380	2 4 6 9 4 4 9 9 4 9 9 4 9 9 4 9 9 9 9 9 9		3,844	961		2	007		2,658	
20,000 All 38 37 1,630 20,000 All 107 104 3,585 20,000 All 375 143 3,585 00 All 2,907 2,867 17,74 00 All 7,761 154 31,024 1 1 7,761 151 23,914 1 1 1,508 582 242,964 1 2 5,649 680 231,004 1 3,4 5,549 680 508,770 1 3,4 6,702 680 508,770 1 3,4 6,702 680 508,770 1 3,4 7,379 1,579 27,042 1 4 7,239 1,579 27,042 1 4 7,379 1,579 27,042 1 4 7,381 1,684 36,381 1 4 7,379 2,489,394 36,381	37 104 143 143 286 3 154 151		3,844	1 270	183	183	1.144	1.144		2.658
20,000 All 107 104 3,585 0,000 All 974 94 17,774 00 All 2,907 286 30,729 00 All 7,761 154 31,024 0 All 15,018 151 23,914 1 2 5,624 58 224,964 1 2 5,656 63 499,148 1 3,4 5,566 63 499,148 0 1,3 9,607 218 888,854 0 4 72,309 1,507 861,381 0 4 72,309 1,507 861,381 0 4 7,861 3,134 224,693 5,506 0 4 7,861 1,379 271,042 26,664 765,050 2,664 765,050 2,664 765,050 2,664 765,050 2,664 765,050 2,664 3,334 243,393 5,560 2,643,393 <	104 143 94 1 286 3 154 3		101	6/7:1	629	243	4,686	1,533	10.827	3,647
1,000 Ali 355 143 8,588 1,000 Ali 2,907 2,86 30,729 1,777 1,777 1,777 1,777 1,777 1,000 Ali 7,761 1,54 31,024 1,024 1,23,914 1,24,393 1,24,	143 286 154 151	282	8,707	1,352	1,635	265	9,452	1,450	23,486	3,751
00 All 974 894 17.774 00 All 7.761 154 30.729 00 All 7.761 2.862 23.044 01 15.018 15.104 02 2 5.621 5.82 23.044 03 4 55.489 5.82 242.964 01 1.3 5.86 63 459.148 00 1.3 1.3 5.878 1.379 271.042 00 All 20.268 1.697 281.004 00 All 20.268 1.697 281.004 00 All 20.268 1.697 251.004 00 All 20.268 1.697 251.004 00 All 20.268 1.697 251.042 00 All 20.268 1.693 3.34 55.000	94 286 154 151		23,319	992	4,478	151	23,241	692	59,991	2,111
00 All 2,907 286 30,729 All 15,018 151,024 All 15,018 151 23,914 1 2 5,621 59 231,004 2 2 5,656 63 459,148 1 3 4 55,469 68 63 459,148 0 1 3 4 72,309 1,507 218 0 0 4 72,309 1,379 271,042 0 0 All 20,268 1,984 3,394 0 0 All 20,268 564 3,384 0 0 All 20,268 564 3,384 0 0 All 20,268 569 0 0 All 20,268 569 0 0 All 20,268 569 0 0 All 20,393 55	286 154 151	202	57,085	541	10,265	84	54,781	537	140,879	1,458
0 All 7,761 154 31,024 15,018 15,018 15,018 15,019	154	120	124,065	618	17,815	78	124,384	641	299,900	1,743
All 15,018 151 23,914 1 1 5,621 5,921 2 5,621 5,931 3-4 5,545 682 242,964 1-2 5,556 63 459,148 10 0 1-3 6,072 680 508,770 10 0 4 72,309 1,579 271,042 10 0 All 20,268 1,693 48,934 10 0 All 7,961 3,134 248,934 10 0 All 7,961 3,134 248,393	151	06	168,909	572	18,743	89	187,636	541	414,073	1,425
1 5621 59 231,004 2 5624 59 582 242,954 1 1-2 5565 63 459,148 0 1 24 67,022 680 508,770 0 0 4 72,309 1,507 581,381 0 0 4 32,628 1693 486,934 0 0 All 70,961 3,134 234,393		40	423,298	762	28,727	28	645,241	1,222	1,136,198	2,233
2 5,621 59 231,004 3-4 55,459 63 459,148 1-2 5,556 63 459,148 1-3 67,022 680 508,770 1-3 9,607 218 858,854 4 72,309 1,507 881,381 1-3 15,978 1,379 271,042 4 32,028 1,693 486,934 33 All 7,961 3,134 234,333 5,5										
2 5 621 59 231,004 3-4 55,66 63 242,964 1-2 5,666 83 459,148 3-4 67,022 680 508,770 1-3 9,607 218 858,854 4 72,309 1,507 281,301 1-3 15,978 1,379 271,042 4 32,632 2,664 765,050 2,704,704 All 7,961 3,134 234,333 5,5							31,663,929	31,670	31,663,929	31,670
3.4 55.459 582 242.994 1.2 5.566 63 459,148 3.4 67,022 680 508,770 1.3 9607 2.18 588.894 1.4 72,399 1,507 581.381 1.3 15,978 1,379 271,042 4 32,632 2.664 765,060 All 7,961 3,134 284,393 5,	92	233	2,724,392	2,780	83,508	82	26,320,904	26,129	29,365,429	29,283
1-2 5,556 680 593,148 1-3 6,072 680 508,770 1-3 9,607 218 858,884 1-3 15,978 1,507 581,381 1-4 32,528 1,693 486,94 33,184 All 7,961 3,134 234,393 5,5	582	361	4,520,526	7,093	115,266	197	5,895,299	9,074	10,829,504	17,307
1-3 9,022 080 306,770 1-3 9,077 218 658,884 1-3 15,978 1,507 681,381 4 32,684 76,050 2, 181 20,288 1,693 486,94 33, All 7,961 3,134 294,393 5,	693	9449	1,971,856	1,969	168,590	169	21,580,125	819,12	24,185,275	24,268
4 72,309 1,507 681,381 1.57 881,381 1.57 881,381 1.37 271,042 4.2 32,684 765,090 2. All 20,288 1,693 486,934 3.1 7,961 3,134 294,393 5,5	218	0.00	2,76,452	2.285	220,234	215	11 055 465	10.894	14 420 664	14 477
1-3 15,978 1,379 271,042 2 2 32,684 765,060 2 2 20,041 2 20,058 1,693 486,934 3 3 All 7,961 3,134 224,393 5	1.507	851	2.629.243	4.031	190,709	276	2.898.536	4.450	6.372.178	11.115
4 32.632 2.664 765.050 All 20.268 1,693 486.934 All 7,961 3,134 234.993	1,379		395,554	773	83.773	158	1,195,437	2,395	1.961.784	5,223
All 20,268 1,693 486,934 All 7,961 3,134 234,393	2,664		1,433,341	4,780	91,429	308	1,881,240	6,192	4,203,692	16,438
All 7,961 3,134 234,393	1,693	3,568	509,536	3,710	64,465	442	638,161	4,598	1,719,364	14,011
	3,134	5,675	148,820	3,766	19,742	487	178,260	4,495	589,176	17,557
All 2,915 1,159 101,077	1,159		39,596	4,870	5,244	909	52,927	6,438	201,759	25,252
All 1,167 1,161 50,186	1,161		13,405	4,371	1,682	534	19,284	6,172	85,724	28,542
3,000,000 All 257 257 13,919 1	257		2,612	2,612	333	333	4,018	4,018	21,139	21,139
\$10,000,000 or more All 122 122 9,417 9,417 1,266 1,266 139 139 1,892	122 9,417	9,417	1,266	1,266	139	139	1,892	1,892	12,836	12,836

[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 tentiative returns identified after sample [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 legges 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.