The IRS Population Count: An Update

Peter Sailer and Michael Weber, Internal Revenue Service

In a paper presented at the 1993 Annual Meetings of the American Statistical Association, the authors presented the results of their first attempt to use administrative records available at the Internal Revenue Service (IRS) to count the population of the United States (see Sailer, Weber, and Yau, 1993). In that paper, we noted that a major problem in this use of IRS administrative records was the presence of information documents for deceased individuals. This was problematic since several years could pass between the death of an individual and the closing out of all accounts listed in his or her name. In addition, we were nervous about the accuracy of our gender coding, since it was based entirely on the interpretation of each individual's first name by computer software we had developed. Poor reporting of Social Security numbers (SSN's) of dependents was a further obstacle to getting a correct count.

As will be discussed later, a number of these problems have been dealt with over the last five years, and it appeared to be an opportune time to research whether our processing changes had improved our ability to use IRS records to count the population. This paper covers the results of that research.

Organizationally, this paper is divided into four sections. First, we will demonstrate how administrative records can be used to compute a population estimate. Then, we will discuss the reliability of this estimate. Next, we will compare estimates from our data base, classified by age, sex, and State, to population data published by the Census Bureau. And finally, we will summarize our conclusions and make recommendations for further research.

■ Computation of an IRS Administrative Records Population

Citizens and residents of the United States have numerous opportunities to come to the attention of the Internal Revenue Service. Obviously, the 61 percent of the population who file individual tax returns, either as primary or secondary taxpayers, are easy enough to count. These individuals also report, as exemptions, any children or other individuals they are supporting. In addition, individuals covered by salaries and wages are generally reported to the IRS on Forms W-2; individuals making contributions to Individual Retirement Arrangements (IRA's) or Simplified Employee Pension (SEP) accounts on Forms 5498; individuals receiving gross distributions from IRA's, SEP's, or other pension plans on Forms 1099-R; recipients of interest on Forms 1099-INT; recipients of dividends on Forms 1099-DIV; recipients of original issue discounts on Forms 1099-OID; recipients of patronage dividends on Forms 1099-PATR; recipients of government transfer payments on Forms 1099-G; recipients of Social Security benefits on Forms SSA-1099; sellers of capital assets on Forms 1099-B; sellers of real estate on Forms 1099-S; contractors with the Federal Government on Forms 8596; winners at gambling on Forms W-2G; payers of mortgage interest on Forms 1098; and recipients of many types of non-employment compensation, including prizes, awards, rents, royalties, crop insurance payments, and golden parachute payments on Forms 1099-MISC.

Table 1: Components of the IRS Population Count

(Frequencies in 1,000's)	Weighted Number	Cumulative
Primary Taxpayers (TY 1993)	112,029	112,029
Secondary taxpayers	46,772	158,801
Dependents without information documents	45,868	204,669
Non-filers with information documents	45,257	249,926
Dependents without SSNs	6,674	256,600
Deaths before January 1, 1994	4,331	252,269

Table 1 details how we used all of this information to count the population covered by IRS administrative records. We started, of course, with filers of tax returns for Tax Year 1993 (i.e., returns generally filed on or around April 15, 1994). However, contrary to our usual practice in our Statistics of Income reports, we did not count anybody filing a prior-year return in 1994, since these individuals had a chance of being captured as recipients of information documents. We also excluded anybody filing from a foreign address, since we wanted to compare our results with Census data for 1994, and Census does not count U.S. citizens living abroad. We

counted 112.0 million current-year returns with U.S. addresses.

On joint returns selected for this sample, we counted the secondary taxpayers—a total of 46.8 million. This brought our count to 158.8 million.

We also counted dependents, but not all of them. Dependents with income could be picked up in our sample of information documents or in our sample of tax return filers, so, initially, we counted only those dependents who had SSN's, but for whom a search of our administrative records master files revealed no records. There were 45.9 million such dependents.

To the 204.7 million individuals counted thus far, we added 45.3 million non-filers with information documents. We got these individuals by pulling a simple, random sample of individuals with at least one information document on the Information Returns Master File, and then eliminating all who appeared either as a primary or secondary taxpayer on a tax return. If they appeared on a tax return as a dependent, we left them in, since we were not including dependents with information documents in our count. Again, we eliminated any prior-year documents received by the IRS in 1994, and we did not count documents issued to individuals at foreign addresses.

Unfortunately, our file also contained 6.7 million dependents for whom no SSN was given. This was a major improvement over the 11.4 million dependents for whom no SSN was given for 1989, but still a disappointment. Obviously, in the absence of an SSN, we could neither check the Information Returns Master File (IRMF) for income nor the Year of Birth File for age. We did not have much choice but to count such dependents in the lowest age category and assume that they were not information document recipients. From our Taxpayer Usage Study (IRS, 1994-2), we know that an estimated 3.3 million taxpayers checked a box indicating that the dependent was under age 1 and, therefore, not required to have an SSN. So, for nearly one-half of these dependents, we know that we have the correct ages. Luckily, this problem should pretty much disappear in future years, for IRS is no longer sending out refund checks to taxpayers who fail to provide dependent SSN's, or who provide non-verifiable ones.

At this point, our count is 256.6 million. As mentioned previously, experience has taught us that some of the individuals in this count are deceased. Our big improvement this year was that the Social Security Administration was willing to share the information they gather from various sources on which SSN's belong to the deceased, including the owner's date of death. This meant that we no longer had to make case-by-case decisions as to who in our sample was alive on January 1, 1994—the date of the Census estimates we were using for comparison purposes. Anybody with a date of death prior to January 1, 1994, was simply taken out of the IRS count. This left an IRS "population count" of 252.3 million, or 97.36 percent of the Census estimate of 259.1 million.

■ Evaluation of the Estimate

The estimates presented in Table 1 are based on a highly stratified sample of 104,605 individual income tax returns (Internal Revenue Service, 1995), supplemented by a simple random sample of 45,257 individuals for whom our files contained information documents but no tax returns. Therefore, the estimates are subject to sampling error. Our 95-percent confidence interval is between 251.0 and 253.5 million. So, our estimate lies between 96.88 and 97.84 percent of the Census figure. At this point, it should also be noted that Census admits to an undercount of about 4 million individuals. Assuming that is correct, we have identified between 95.40 and 96.34 percent of the true population in our administrative records file.

The Census figures are updates of the counts from the 1990 Census, using data on births, deaths, immigration, and emigration (U.S. Bureau of the Census, 1998). While they are not subject to sampling variability, they do contain non-sampling errors. The IRS data are subject to non-sampling error as well. While every effort has been made to eliminate incorrect SSN's and substitute correct ones where our historical files provided this information, it is quite likely that our sample still contains incorrect dependent SSN's. These could lead to false matches or false non-matches to information documents. Missing dependent SSN's could only lead to false non-matches, which would have the effect of overstating the IRS data, since we would have no way of detecting whether these dependents were already being

■ Comparisons with Census

Let us now look at the age and sex distribution of individuals in our file of administrative records. As mentioned previously, age and sex were added to our file simply by matching to an extract from the Social Security Administration's (SSA) Year-of-Birth file, which IRS receives for administrative and research purposes. For those few individuals with missing or invalid SSN's, the sex code was generated by matching the first name against a dictionary of gender-coded names. The age was imputed with the help of an algorithm that took into account the individual's sources of income (for example Social Security retirement income), the entry in the "over 65" check-box on the return, and, where available, the ages of the spouse and any dependent children shown on the same return.

As can be seen from Table 2, the overall correspondence between Census and administrative records data

Table 2. Number of Individuals (in 1,000's), January 1, 1994. IRS and Census Estimates

Total Age	IRS	Deaths by 1/1/1994	Adj. IRS	Census	Adj. IRS as % of Census	Census undercount	Adjusted Census	Adj. IRS as % of Adj. Census
Under 15	55,897	35	55,862	57,337	97.43	1,822	59,159	94.43
15 under 25	33,842	179	33,663	35,942	93.66	1,146	37,088	90.77
25 under 35	40,952	46	40,906	41,354	98.92	1,037	42,391	96.50
35 under 45	40,332	85	40,247	41,658	96.61	486	42,144	95.50
45 under 55	29,118	172	28,946	29,870	96.91	46	29,916	96.76
55 under 65	20,430	299	20,131	21,018	95.78	-189	20,829	96.65
65 under 75	19,360	949	18,412	18,712	98.39	-175	18,537	99.32
75 and over	16,670	2,567	14,102	14,446	97.62	-126	14,320	98.48
Total	256,600	4,331	252,269	260,337	96.90	4,047	264,384	95.42
Male								
Age								
Under 15	28,448	16	28,433	29,353	96.86	927	30,280	93.90
15 under 25	17,396	115	17,281	18,347	94.19	603	18,950	91.19
25 under 35	21,181	36	21,145	20,677	102.26	618	21,295	99.29
35 under 45	20,311	65	20,246	20,648	98.05	325	20,973	96.53
45 under 55	14,505	120	14,386	14,591	98.59	59	14,650	98.19
55 under 65	9,955	193	9,762	9,984	97.78	-63	9,921	98.40
65 under 75	8,714	595	8,119	8,290	97.94	-51	8,239	98.54
75 and over	6,294	1,184	5,110	5,185	98.55	-21	5,164	98.95
Total	126,804	2,323	124,481	127,075	97.96	2,398	129,473	96.14
Female								
Age								
Under 15	27,446	19	27,427	27,984	98.01	895	28,879	94.97
15 under 25	16,446	64	16,382	17,595	93.11	543	18,138	90.32
25 under 35	19,771	10	19,761	20,677	95.57	418	21,095	93.67
35 under 45	20,021	20	20,001	21,010	95.20	161	21,171	94.47
45 under 55	14,613	53	14,560	15,279	95.29	-14	15,265	95.38
55 under 65	10,474	106	10,368	11,034	93.96	-126	10,908	95.05
65 under 75	10,646	353	10,293	10,422	98.76	-124	10,298	99.95
75 and over	10,377	1,423	8,954	9,261	96.69	-105	9,156	97.79
Total	129,794	2,048	127,746	133,262	95.86	1,648	134,910	94.69

is extremely good—even better than it was for 1989. The overestimation of the "75 and over" class has disappeared for 1993, thanks to the date-of-death information added to the data base. Some of the apparent difference between coverage of males and females (particularly in the age classes under 25) has been eliminated with the help of sex codes from SSA. The only age/sex class in which IRS shows more individuals than are shown in the Census data is males age 25 to 35. However, the IRS estimate is still slightly below the adjusted Census estimate, so IRS may have been able to account for some young males missed in the 1990 Census.

If the IRS administrative data are to be used in a meaningful way to help Census identify individuals missing from the decennial Census, or even just to make intercensal estimates, it is important that they be classifiable by geographic code. IRS data are somewhat problematical in this regard. Some IRS tax return addresses do not locate the taxpayer's residence—they may represent a tax accountant's address, a business address, a post office box in another town, or a rural route that crosses county lines. The addition of information documents to the data base provides the user with alternative addresses for each taxpayer—unfortunately, in some cases, with several alternatives. With a good deal of research, it may be possible to rank various types of information documents as to their likelihood of showing a residential address. It may also be possible to write algorithms that detect and eliminate addresses which are not residential.

Having admitted these shortcomings, we hasten to add that the vast majority of tax documents do contain addresses which can be used to code the residences of taxpayers. Unfortunately, the data base with which we were working was not designed to produce accurate estimates below the national level. Even at the State level, the estimates tend to show a good deal of sampling error. In order to minimize the error, we derived State estimates through a three-step process: The number of primary taxpayers was taken straight from an IRS Master File Tabulation (Internal Revenue Service, 1994-1) (i.e., it is not subject to sampling error). The count of secondary taxpayers and dependents without income was ratio-adjusted by the same percentage as the number of primary taxpayers (i.e., it is subject to non-sampling error); and the non-filer population was left unadjusted (i.e., it is subject to sampling error). Table 3 shows the

Table 3. Census Adjusted Population by State, as of January 1, 1994 (in 1,000's). Comparison With Census Actual and IRS Estimate

State	Census adjusted	% of Census Adjusted	
		Census	IRS
ALABAMA	4,293	98.29	99.54
ALASKA	614	98.17	94.92
ARIZONA	4,169	97.85	92.66
ARKANSAS	2,495	98.32	92.58
CALIFORNIA	32,250	97.39	93.38
COLORADO	3,731	98.14	98.01
CONNECTICUT	3,296	99.35	97.74
DELAWARE	720	98.31	98.86
DIST. OF COLUMBIA	588	96.36	87.05
FLORIDA	14,218	98.17	96.38
GEORGIA	7,200	98.02	98.60
HAWAII	1,199	98.25	98.30
IDAHO	1,157	98.05	99.16
ILLINOIS	11,873	99.04	96.47
INDIANA	5,783	99.52	96.62
IOWA	2,843	99.60	94.64
KANSAS	2,568	99.32	99.67
KENTUCKY	3,889	98.44	92.38
LOUISIANA	4,410	97.87	93.67
MAINE	1,248	99.27	99.10
MARYLAND	5,101	98.02	94.64
MASSACHUSETTS	6,070	99.52	95.51
MICHIGAN	9,558	99.31	92.25
MINNESOTA	4,587	99.58	97.17
MISSISSIPPI	2,726	97.93	94.59
MISSOURI	5,311	99.40	96.17
MONTANA	875	97.80	88.79
NEBRASKA	1,634	99.37	96.91
NEVADA	1,491	98.06	97.11
NEW HAMPSHIRE	1,144	99.19	99.28
NEW JERSEY	7,948	99.43	100.10
NEW MEXICO	1,704	97.15	103.54
NEW YORK	18,432	98.49	91.56
NORTH CAROLINA	7,196	98.25	97.14
NORTH DAKOTA	643	99.34	94.83
OHIO	11,179	99.33	95.73
OKLAHOMA	3,314	98.27	91.29
OREGON	3,141	98.28	98.33
PENNSYLVANIA	12,098	99.71	97.17
RHODE ISLAND	996	99.85	94.36
SOUTH CAROLINA	3,716	98.04	92.98
SOUTH DAKOTA	730	99.06	96.69
TENNESSEE	5,263	98.34	93.45
TEXAS	18,899	97.43	94.81
UTAH	1,939	98.44	94.92
VERMONT	586	98.92	102.68
VIRGINIA	6,677	98.11	96.62
WASHINGTON	5,430	98.31	98.26
WEST VIRGINIA	1,850	98.60	93.44
WISCONSIN	5,113	99.41	97.84
WYOMING	486	97.93	91.64
Total	264,384	98.47	95.42

comparison of these estimates with Census population figures (adjusted for the undercount) by State. Also shown is a comparison of the official Census count with the adjusted Census count. It shows that, for nine States, the IRS estimate is actually closer to the adjusted Census population than the official Census figure. For 16 more, the Census and IRS estimates are within three percentage points of one another. The low coverage by IRS for New York (91.56 percent) and California (93.38) should not be a sampling variability problem, but may be related to their high rates of immigration. It may take immigrants a while to get into the IRS document systems.

■ Conclusions and Recommendations

In commenting on the authors' earlier paper on this subject, John Czajka et al. wrote that we had "demonstrated the IRS administrative records system provides sufficiently high coverage of the U.S. resident population to be credible as the principal source of data for an enumeration ... with administrative records" (Czajka et al., 1997). We feel that the 1993 data, with improved SSN reporting, better gender codes, and addition of date-of-death information, reconfirm that conclusion. At very least, the Census Bureau, which has access to both tax return and information document files, should be studying how to use this information to identify individuals at addresses missed in the regular decennial Census. Much of the additional research that needs to be done—especially on the quality of address information on various types of documents—will have to be done at the Census Bureau, since only they can start with matched files of Census and administrative data. For our part, we will monitor the effects of improved SSN reporting on this type of data analysis.

■ Acknowledgments

The authors would like to thank Marianne Cooley of the IRS Computing Center in Detroit, Michigan, for help with computer processing. Thanks also to William Wong of the Statistics of Income Division for help with the computation of the coefficient of variation.

■ References

Alvey, W. and Scheuren, F. (1982), "Background for an Administrative Records Census," *Statistics of Income and Related Administrative Record Research,* Washington, DC: U.S. Department of the Treasury, Internal Revenue Service.

Bureau of the Census (1998), Monthly population estimates by age, race, and sex, and annual population estimates are available on the Internet at www.census.gov/population/www/estimates.

Czajka, John L.; Moreno, Lorenzo; and Shirm, Allen L. (1997), *On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population,* Washington, DC: Mathematica Policy Research.

Internal Revenue Service (1994-1), *Statistics of Income Bulletin, Fall 1994,* Washington, DC: U.S. Government Printing Office.

Internal Revenue Service (1994-2), "Taxpayer Usage Study, 1993," an unpublished weekly report of the Statistics of Income Division, distributed electronically and on paper during each filing season.

Internal Revenue Service (1995), *Statistics of Income—1993, Individual Income Tax Returns,* Washington, DC: U.S. Government Printing Office.

Panel To Evaluate Alternative Census Method (1993): *A Census that Mirrors America,* Washington, DC: National Academy Press.

Sailer, Peter; Weber, Michael; and Yau, Ellen (1993), "How Well Can IRS Count the Population?" *Proceedings, Section on Survey Research Methods*, American Statistical Association.