HOW WELL CAN IRS COUNT THE POPULATION?

Peter Sailer, Michael Weber, and Ellen Yau, Internal Revenue Service
(O:S) P.O. Box 2608, Washington, DC 20013-2608

KEY WORDS: Administrative records, Census, Age and Sex Distributions

The following paper is an outgrowth of research performed with a data base of merged individual income tax returns and information documents. Tax Year 1989 was the first year for which such a data base was created and perfected. Traditionally, the Statistics of Income (SOI) Division of the Internal Revenue Service has interpreted its mandate to produce "statistics reasonably available with respect to the operations of the internal revenue laws" [1] as meaning tabulating data shown on tax returns. In recent years, with the computerization of the millions of information documents prepared by employers, banks, stock brokers, payers of pensions, etc., data from these documents have increasingly become "reasonably available." Data from information documents, when matched to tax returns, can be used to serve as a check on the data shown on individual income tax returns, as well as to provide an indication of how much of the income on a joint return belongs to the husband and how much to the wife. In addition, it is possible to pull a sample of information documents that do not match to tax returns, and use them to tally data about non-filers.

The data base used for this paper was created as a tool to compare tax return data to data gathered from information documents. It includes a sample of tax returns matched to information documents, as well as unmatched tax returns and unmatched information documents. The age of each individual in the sample was determined by matching his or her social security number (SSN) to the Year of Birth file, which contains information supplied at the time the SSN is applied for. While the data base used for this paper was set up primarily for tax analysis purposes, it is also a rich source of information with which to evaluate recent proposals for a greater use of administrative records in structuring Censuses and inter-censal estimates. [2] This paper is presented as a modest first step in performing the proposed research on the population covered by administrative records in the possession of the Internal Revenue Service.

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 results from the 1990 Census. And finally, we will summarize our conclusions and make some 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 64 percent of the population that files individual tax returns, either as primary or secondary taxpayers, is 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; recipients of pensions on Forms W-2P; [3] individuals making contributions to Individual Retirement Arrangements (IRAs) or Simplified Employee Pension (SEP) accounts on Form 5498; individuals receiving gross distributions from IRAs, SEPs, or other pension plans on Form 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 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 1989 (i.e., returns generally filed on or around April 15, 1990). However, contrary to our usual practice in our Statistics of Income reports, [4] we did not count anybody filing a prior-year return in 1990, 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 to those from the 1990 Census, and Census does not count U.S. citizens living abroad. We counted 109.0 million current-year returns with U.S. addresses.

On joint returns selected for this sample, we counted the secondary taxpayers--a total of 46.9 million. This brought our count to 155.9 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 only counted those dependents who had SSNs, but for whom a search of our administrative records master files revealed no records. There were 36.3 million such dependents.

To the 192.2 million individuals counted thus far, we added 43.7 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 a secondary taxpayer on a tax return. If they appeared on a tax return as a dependent, we left them in, since we were not counting dependents with income in the Third column of table 1. Again, we eliminated any prior-year documents received by the IRS in 1989, and we did not count documents issued to individuals at foreign addresses.

Unfortunately, our file also contained 11.4 million dependents for whom no SSN was given. 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. 1989 was only the third Tax Year for which any dependent SSNs were requested on tax returns, and the first on which they were requested for dependents between the ages of two and five. It seems IRS was still having a bit of a problem trying to convince taxpayers to get SSNs for their young dependents. According to data available from our Taxpayer Usage Study, [5] some 4.1 million taxpayers checked a box indicating that the dependent was under age 2, and therefore not required to have an SSN. Based on U.S. vital statistics, as many as 3.9 million more of these dependents may have been under age 2, although the box was not checked. An additional 2.5 million dependents had the words "applied for" entered in the SSN space. So we are reasonably confident that the vast majority of these

11.4 million dependents were very young and had no income. Therefore, we decided it was appropriate to include all of them in our population estimate, and to count them in the lowest age bracket.

At this point, our count is at 247.3 million, or 99.4 percent of the number counted in the 1990 Census (they counted 248.7 million), which is, of course, extremely impressive. The only trouble is, when we distributed these taxpayers by age, our counts in the top age brackets--age 65 and over--exceeded Census's count by about 3.2 million--even after we had made allowances for all deaths between the beginning of Tax Year 1989 and the 1990 Census. It is our current working hypothesis that a number of accounts remain active--and therefore generate information documents--even after the beneficiary has died. This is particularly true of joint accounts where the taxpayer listed as primary beneficiary has died. If the surviving spouse fails to file the needed paperwork, he or she can keep on using the account, even though it is issuing information documents to the deceased spouse.

IRS does not currently have any in-house information on deceased non-filers. We are in the midst of negotiations with Disclosure Officers at IRS and the Social Security Administration. We would like SSA to help us identify any deceased individuals who got into our sample--or at least provide us with statistics on how many of the individuals involved are deceased. In the meantime, we are using as a proxy for the deceased those aged taxpayers who show no evidence of any earned or retirement income--in other words, all they had for Tax Year 1989 was some account bearing unearned income (usually interest or dividends),

and, of course, no tax return was filed in their name. Our files showed an estimated 3.0 million such information document recipients in the upper age brackets. As is shown in the sixth column of table 1, we considered them all to have been deceased prior to 1989, and therefore removed them from our population estimates.

Of course, at that point, our sample still contained tax returns and information documents for individuals who were alive in 1989, but had died by the time of the 1990 Census. On the other hand, the 1990 Census included infants born during the first three months of 1990, who would have been excluded from our administrative records system. We therefore used data on vital statistics to adjust for deaths during 1989 and births and deaths between January 1 and March 31, 1990. This brought our bottom line estimate to 242.6 million, or 97.54 percent of the number counted by Census.

Evaluation of the Estimate

Obviously, the estimates presented in Table 1 are subject to both sampling and non-sampling error. In regards to the latter, it can be taken as given that the number of taxpayers (both primary and secondary) is reasonably solid, given the legal sanctions against fraudulent multiple filings. However, all of the remaining administrative records estimates are valid only to the extent that reporting of social security numbers (SSNs) is accurate, both on the tax return and on the information document side of the equation. For example, a mistake in an entry for a dependent SSN may well have caused that dependent not to match up with his or her information documents, or to have matched up with somebody else's information documents. If there were multiple information documents for the same taxpayer, but only one had the wrong SSN, the same person might be counted twice in this system, if neither SSN matched to a tax return.

We have not completed our research on incorrect SSNs yet. However, we are in the middle of an extensive verification effort of all SSNs in the file as part of a family panel study begun for Tax Year 1987. So far, we have verified approximately 60 percent of the SSNs, those that were common to the 1987, 1988, and 1989 samples. We have found only minor problems. As expected, primary SSNs are almost always correct, since they are verified during mainline IRS processing. Only about .02 percent needed to be corrected. About 1 percent of the secondary SSNs were incorrect. Since 1989, IRS has started verifying these as well, so we should do better in the future. The biggest problem was with dependent SSNs, which are verified only on a sample basis during mainline processing. About 2 percent of those checked were in error (representing nearly 3 percent of the population when the data were weighted). As a result of the corrections, the proportion of dependents matching to information documents went down slightly--about two-thirds of the incorrect SSNs matched to information documents, about one third of the corrected SSNs matched. This in turn raised the SOI coverage minimally, since we counted only dependents who did not have information documents in this tally. The most common pattern of incorrect dependent SSNs occurs when more than one individual in a family uses the same SSN--either a dependent using the same SSN as the parent, or two or more dependents using the same SSN. Even if all of these individuals are, in fact, receiving income using the same SSN, they will be counted only once in the "non-filer with information documents" group.

It should also be noted that the estimate of 11.4 million dependents without SSNs is a troublesome aspect of this analysis. To the extent that these dependents really had SSNs and were receiving income that was reported on information documents, we would be double-counting them. However, the evidence points, not to fraud, but to simple failure to obtain an SSN on time. When we matched our sample returns with missing SSNs to the primary taxpayer's return for the following year, about 40 percent of these tax filing units showed an increased number of reported dependent SSNs in the following year.

In regards to sampling variability, the administrative records population estimate is based on a sample of 106,628 tax returns and 8,220 non-filers with information documents. The coefficient of variation of the total estimate (242.6 million) is .8635 percent at the 95 percent level of confidence. The true value of our administrative records population estimate therefore lies between 240.5 million and 244.7 million, or between 96.7 and 98.4 percent of the Census count.

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 95.2 and 96.8 percent of the true population in our administrative records file. Obviously, the coefficients of variation are correspondingly higher for the subtotals shown in table 1.

Comparisons to Census

Let us now look at the age and sex distribution of individuals in our file of administrative records. As mentioned previously, age was 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. SSA also has data on the gender of individuals with social security numbers; however, since IRS has no administrative need for this information, SSA does not provide it to us. Therefore, sex codes had to be generated based on the first name of the individual. Since it was know from previous studies that over 95 percent of married couples filing jointly show the husband as the primary taxpayer, we assumed that any first name associated largely with primary taxpayers on joint returns was male, any name associated largely with secondary taxpayers female. A manual review of the resulting dictionary of names revealed no discernible errors.

The dictionary of names was then applied to all taxpayers, dependents, and information document recipients. The dictionary coded 89 percent of the individuals in the data base. The remainder were assigned sex codes randomly within each age category. While future refinements of the dictionary, with the help of experts on a number of foreign languages, will reduce the number of randomly coded individuals, they should not change the results of the following analysis appreciably.

As can be seen from Table 1, the overall correspondence between Census and administrative records data is extremely good from age 35 on--actually from about age 30, if the data are presented in smaller age breaks--through age 75. Any differences can be explained by sampling variability on the administrative records side, by reporting differences at various stages of life (individuals reported their ages to Census in 1990, to SSA when the first applied for an SSN), and by the fact that the Census figures are unadjusted for the undercount. However, there is definitely undercoverage in the lower age brackets (especially among women), as well as an overcoverage in the 75 and over class (especially among men).

In all probability, the single largest category we are missing is children and young mothers on welfare. Presumably, once you get into the money economy, you tend to get into and stay in the administrative records system. Even if you lose your job, your unemployment compensation is, after all, covered both on information documents and on tax returns. It does appear that the 845,000 adjustment for presumed pre-1989 deaths among male information document recipients age 75 and over fell short of the mark. Or, perhaps, surviving widows are filing tax returns using their husbands' SSNs, while receiving information documents under their own SSNs, and are thus being counted twice, once as males and once as females. More research is needed in this area.

Conclusions and Recommendations

The data from this first attempt at counting the population by using administrative records are very encouraging--certainly encouraging enough to warrant further research. The Internal Revenue Service, by itself, can do a very good job of counting working age residents of the United States. We are not quite as good at counting young people, but some other agency

(perhaps the Census Bureau) might be able to fill in the gaps by gaining access to data on the Aid to Families with Dependent Children (AFDC) program from the various states. In the top age brackets, there is some evidence that individuals stay in one of our administrative records systems for a while after they are deceased. This problem could also be solved by matching our records to those of another agency (for

example, SSA) that has better mortality data. The authors hope that the findings presented here will stimulate additional research throughout the Federal statistical community.

Acknowledgments

The authors would like to extend their thanks to Marianne Cooley, Lorne Woodworth, and Antoine Stone 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

[1] ***Internal Revenue Code***, Section 6108(a).

[2] Panel to Evaluate Alternative Census Methods: ***A Census that Mirrors America.*** Washington, DC: National Academy Press, 1993. For a much earlier call for this type of research, see Alvey, W., and Scheuren, F., "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, 1982, pp. 47-65.

[3] Form W-2P was in use for 1989; it has since been replaced by an expanded Form 1099-R.

[4] ***Statistics of Income--Individual Income Tax Returns.*** Washington, DC: U.S. Government Printing Office, annually, 1916 to present.

[5] The Taxpayer Usage Study is an unpublished weekly report of the Statistics of Income Division, distributed electronically and on paper during each filing season.

TABLE 1 -- RECONCILIATION OF ADMINISTRATIVE RECORDS ESTIMATES TO THE 1990 CENSUS OF THE POPULATION

AGE	Primary Taxpayers	Secondary Taxpayers	Dependents W/ No Docs	Nonfilers w. Info docs.	Deps Without SSNs	Adjustment For Pre-1989 Deaths	All Individuals (SOI)	Adjustments For 1989-90 Births & Deaths	Administrative Records Estimate	All Individuals (Census)	Admin Records as % of Census Pop
Under 15	1,417,965	92,680	29,447,805	6,740,000	11,394,000	0	49,092,450	930,125	50,022,575	53,568,000	93.38
15 Under 25	19,315,836	2,048,265	5,817,466	7,080,000	0	0	34,261,567	(45,625)	34,215,942	36,775,000	93.04
25 Under 35	25,950,677	11,375,978	363,915	5,055,000	0	0	42,745,570	(75,750)	42,669,820	43,176,000	98.83
35 Under 45	21,820,774	12,137,799	130,528	4,240,000	0	0	38,329,101	(101,125)	38,227,976	37,579,000	101.73
45 Under 55	14,420,683	8,446,143	124,543	2,725,000	0	0	25,716,369	(147,750)	25,568,619	25,224,000	101.37
55 Under 65	11,294,094	6,845,084	167,508	2,640,000	0	0	20,946,686	(325,000)	20,621,686	21,147,000	97.52
65 Under 75	8,786,975	4,368,907	141,318	6,000,000	0	(775,000)	18,522,200	(601,500)	17,920,700	18,107,000	98.97
75 and Over	5,970,480	1,577,106	154,580	9,240,000	0	(2,265,000)	14,677,166	(1,320,625)	13,356,541	13,136,000	101.68
Total	108,977,484	46,891,962	36,347,663	43,720,000	11,394,000	(3,040,000)	244,291,109	(1,687,250)	242,603,859	248,712,000	97.54

MALE AGE	Primary Taxpayers	Secondary Taxpayers	Dependents W/ No Docs	Nonfilers w. Info docs.	Deps Without SSNs	Adjustment For Pre-1989 Deaths	All Individuals (SOI)	Adjustments For 1989-90 Births & Deaths	Administrative Records Estimate	All Individuals (Census)	Admin Records as % of Census Pop
Under 15	779,693	1,522	16,071,393	3,375,000	5,832,704	0	26,060,312	460,125	26,520,437	27,422,000	96.71
15 Under 25	11,175,267	48,647	3,019,613	3,695,000	0	0	17,938,527	(34,000)	17,904,527	18,779,000	95.34
25 Under 35	18,220,163	334,397	201,169	2,820,000	0	0	21,575,729	(55,250)	21,520,479	21,573,000	99.76
35 Under 45	16,363,859	346,149	62,366	2,220,000	0	0	18,992,374	(69,125)	18,923,249	18,594,000	101.77
45 Under 55	11,014,465	154,733	44,165	1,500,000	0	0	12,713,363	(93,500)	12,619,863	12,326,000	102.38
55 Under 65	8,720,793	135,732	59,715	1,155,000	0	0	10,071,240	(199,500)	9,871,740	9,981,000	98.91
65 Under 75	6,280,833	102,692	50,597	2,320,000	0	(410,000)	8,344,122	(348,000)	7,996,122	7,942,000	100.68
75 and Over	3,216,133	55,501	50,597	3,170,000	0	(845,000)	5,647,231	(553,000)	5,094,231	4,624,000	110.17
Total	75,771,206	1,179,373	19,559,615	20,255,000	5,832,704	(1,255,000)	121,342,898	(892,250)	120,450,648	121,241,000	99.35

FEMALE AGE	Primary Taxpayers	Secondary Taxpayers	Dependents W/ No Docs	Nonfilers w. Info docs.	Deps Without SSNs	Adjustment For Pre-1989 Deaths	All Individuals (SOI)	Adjustments For 1989-90 Births & Deaths	Administrative Records Estimate	All Individuals (Census)	Admin Records as % of Census Pop
Under 15	638,272	91,158	13,376,412	3,365,000	5,561,296	0	23,032,138	470,000	23,502,138	26,146,000	89.89
15 Under 25	8,140,569	1,999,618	2,797,853	3,385,000	0	0	16,323,040	(11,625)	16,311,415	17,996,000	90.64
25 Under 35	7,730,514	11,041,581	162,746	2,235,000	0	0	21,169,841	(20,500)	21,149,341	21,603,000	97.90
35 Under 45	5,456,915	11,791,650	68,162	2,020,000	0	0	19,336,727	(32,000)	19,304,727	18,985,000	101.68
45 Under 55	3,406,218	8,291,410	80,378	1,225,000	0	0	13,003,006	(54,250)	12,948,756	12,898,000	100.39
55 Under 65	2,573,301	6,709,352	107,793	1,485,000	0	0	10,875,446	(125,500)	10,749,946	11,166,000	96.27
65 Under 75	2,506,142	4,266,215	90,721	3,680,000	0	(365,000)	10,178,078	(253,500)	9,924,578	10,165,000	97.63
75 and Over	2,754,347	1,521,605	103,983	6,070,000	0	(1,420,000)	9,029,935	(767,625)	8,262,310	8,512,000	97.07
Total	33,206,278	45,712,589	16,788,048	23,465,000	5,561,296	(1,785,000)	122,948,211	(795,000)	122,153,211	127,471,000	95.83