RESEARCH INTO THE VALIDATION OF SOCIAL SECURITY NUMBERS COLLECTED BY THE UNITED STATES BUREAU OF THE CENSUS

James B. Treat and Theresa F. Leslie, U.S. Bureau of the Census*
James B. Treat, U.S. Bureau of the Census, DSSD, Room 1657-3, Washington, DC 20233-0001

KEY WORDS: Administrative Records, Quality, Demographic Data

1. INTRODUCTION

Social security number (SSN) has become a widely used personal identifier even though this use is not endorsed by the Social Security Administration (SSA). The reality is that most programs use SSN to identify participants. Since the SSN is such a widely used personal verifier, the Census Bureau decided to conduct research dealing with the collection and use of SSN.

The purpose of this paper is to present the methodology and the results of the 1992 National Census Test I SSN Validation project conducted by the United States Bureau of the Census in cooperation with the SSA. For this project, the SSA validated the SSNs collected during the 1992 National Census Test I. There were two purposes for this validation project. The first purpose was to evaluate the quality of the SSNs collected. The second purpose was to look at the data elements/variables associated with the SSN on the SSA file to determine how the content of this administrative record compares to the census data. This paper will address these issues.

2. BACKGROUND

The Census Bureau has been exploring the potential uses of administrative records to determine their effects on reducing both the differential undercoverage and the cost of the census. The SSA file potentially could be used in the Census Bureau administrative records program. Therefore, the Census Bureau and SSA entered into a cooperative work agreement to conduct research into the use and collection of SSN. The Census Bureau sent the tape containing the 1992 National Census Test I data to the SSA for validation. The SSA completed the validation and the results were returned to the Census Bureau.

The 1992 National Census Test I enumerated 4,694 persons within the SSN panel. The file that the Census Bureau sent to the SSA for the validation project included 4,039 records (86 percent of the persons enumerated). Six hundred fifty five (14 percent) of the person records were not sent to validation for the following reasons: SSN was left blank on 643 person records (98.0 percent); the respondent refused to provide SSN for five person records (0.8 percent); and a complete 9-digit number was not provided for seven person records (1.2 percent). Before transmitting the file for validation, the Census Bureau recoded the answers to the race and Hispanic origin questions. The recoding was performed to standardize the census data to SSA standards.

3. METHODOLOGY

The SSA Office of Research and Statistics has developed a validation procedure to work with the Census Bureau on joint research programs. During the validation, a one-time pass is made of all records linked to a SSN. The match code is based on a composite of demographic data that best matches the input data, in this case the 1992 National Census Test I file. During the pass, the first record for a SSN is checked. If it matches the input record exactly, it is considered an exact match; otherwise, the program searches each additional record linked to the SSN to find demographic data that most closely matches the input file. Based on the search of the records linked to a SSN, a 2-digit match code is assigned to each person record. Table 1 shows the results of validation based on the 2-digit match code, by sex as reported on the 1992 National Census Test I file. Variables are considered a match if the data is the same on both files. In addition, if the data is missing on both files the variables are considered a match; for example, if sex is not reported on the census record and the SSA record, sex is considered to match because it is the same on both files, namely, blank. The variables matched between the two files were name (last and first), date of birth (DOB), sex and race.

4. RESULTS

4.1 Validation

This section contains a brief analysis of each of the categories outlined in the Table 1. The denominator for the results is 4,039--the number of person records sent to SSA for validation.

1. All Items Verified - The SSA was able to verify about 73.9 percent (2,986) of the 1992 National Census

Test I person records sent to validation where the 1992 National Census Test I and SSA data matched identically. The data items matched include: SSN, full name (first and last), DOB, sex and race.

2. Name only Differed - There were 75 (1.9 percent) name only differences. Of these, 59 records were females. Of these, the first names matched for 50 cases. Of the remaining 16 cases, most of the differences were due to keying errors either in misspelling the first initial of the last name, transposition of the first and last name on the census or SSA record, and different handling of last names such as O'My. This type of name appears as "O My" on the census record and "OMy" on the SSA record.

3. DOB only Differed - Of the 39 cases that fell into this category, 16 were for males and 23 for females. Based on the results, the SSA file could have been used to supplement responses for 19 records for which Census did not capture DOB during the 1992 National Census Test I. The remaining large number of differences in the reporting of DOB was the year of birth entry. For two persons which lived in the same housing unit, both had 3/30/92 entered as their DOB. It is not clear from the census data file when they completed their census form but given the dates reported on the SSA file and given that the 1992 National Census Test I was conducted near the date reported on the census as their DOB, the dates entered on the census file seem incorrect for this household. If you add these two persons to the 19 mentioned earlier, we could have used the SSA file to provide a DOB for 21 persons (53.8 percent).

Of the records that vary by five or more years, seven of these have earlier years of birth reported on the census record than the SSA record. One problem that SSA informed us of is that some persons completing an application for SSN fill in the year they apply for the SSN rather than their actual year of birth. After looking at this classification of data, it seems clear that all other demographic data for these persons clearly matched between the two files.

4. Name and DOB Differed - Of the 28 cases that fall into this category, at least six cases seem to have been miscoded by SSA during validation. These six cases also had different sexes attached to them. Looking at the demographic data associated with the SSN that pulled the two records together, it is not possible to say these data represent the same persons.

5. Sex only Differed - 42 cases that fall into this category. For the 29 of the 31 persons where the census did not have a response for sex, the SSA file recorded their sex as males. Based on further examination of the names for these cases, the SSA assignment of sex seems appropriate. Therefore, we could have completed responses for 29 cases that did not have census responses. Of the seven records where the entry for sex varied between the two records, six of the seven SSA records had sex designated as male and the census record showed female. Given name information associated with the record, it appears that the census designation is correct in all but one of these cases.

6. Name and Sex Differed - There were no cases in this category.

7. DOB and Sex Differed - There were six cases for which DOB and sex both differed. Of these six records, four records were for two households. In the first case, it appears that the SSN for the head of household and their spouse were switched. In the second case, it appears that the respondent mixed the SSN of his/her children. Looking at the census and SSA data, the demographic information for each of these persons matches. The problem is that the SSN was not attached to the correct person within the household. Of the remaining two cases, the first case appears to be a possible match. The names, SSNs and races are the same. The census did not collect sex for this person and the DOB differs by only one month. The last case does not seem to be a match. The only similar demographic characteristic is the last name. The first names, sex and DOB are different.

8. Name, DOB and Sex Differed - There were 12 cases that fell into this category. In all cases, the census record contained some data and the SSA record had "unknown" entered in each field.

9. Race only Differed - There were 683 cases (16.9 percent) in this category. There are several reasons for the disparity between the race categories maintained by the two agencies. Prior to 1977/1978, the SSA maintained paper files. When a client filed for Social Security benefits, the SSA pulled the original Form SS-5, Application for SSN, from the SSN Application files and replaced it with a Substitute Form SS-5. The original Form SS-5 was sent to the Benefits department for processing. The substitute SS-5 was placed back into the application files. The substitute SS-5 does not include race data, therefore, we will find many cases where race is "unknown" for persons 70 years of age and older because the substitute SS-5 was keyed. Also, prior to the early 1980's, the race categories available on Form SS-5 included: White, Black, Other and Unknown. In the early 1980s, the Office of Management and Budget issued guidelines for all federal agencies to use when collecting race data. Following these guidelines, the SSA began collecting the expanded race classification data, i.e., White, Black or African American, Other, Asian Pacific Islander, Hispanic, North American Indian/Eskimo, and Other.

Race data collected prior to the issuance of the new regulations were not recoded.

There is another problem emerging with the race data maintained by SSA. As cited by Duncan, Jabine and de Wolf (1993), the SSA has entered into a cooperative venture with many of the state vital statistics offices for joint issuance of birth certificates and SSNs. As a result of this agreement, the SSA no longer receives race-ethnic information for most births. Information on the race-ethnic status of the parents is recorded on birth certificates, but it is not being made available to SSA. This creates long run limitations on the Census Bureau use of the SSA data. The Census Bureau will have difficulty providing reliable intercensal population estimates by race. This will also affect decisions relating to greater reliance on administrative records in conducting the decennial censuses of population. For these reasons, it will be very difficult or impossible to use the SSA file to complete race if it is missing from the census form.

10. Name and Race Differed - There were 28 cases that fell into this category; 18 were for females, nine for males and one of unknown sex. Of the female records, there were 10 persons where the difference in name was only the last name and the difference in race was that the SSA reported race as either white, black, or other and the census reported race as either Hispanic or Asian Pacific Islander. There was one female where the middle and last name on the SSA record were combined to form the last name on the census record. Again, the census race was Hispanic. For one female the first and last names were reversed on the two records and the SSA record did not have a race entry. There were three females where the last name differed and the census did not have an entry for race. Finally there were four females where the last name differed and there was no race data on the SSA record.

Of the nine male records that differed, there were two persons where the last and first names were reversed between the two records and there was no entry for race on the census record. There was one male where the first and last names were reversed but there was no entry for race on the SSA record. For two males, the last name on the census record was a combination of the middle and last name on the SSA record. There were two males where the last name differed and the race was Hispanic on the census record but white, black, or other appeared on the SSA record. There was one male where the last name was the same but spaced differently on the two records; for example, the name appeared as such "McNugget" on the SSA record and as "Mc Nugget" on the census record. Finally, there was one male in which the entire name differed and does not appear to be the same person.

11. DOB and Race Differed - There were 23 cases that fell into this category. In all but one case, the race data on the census record is more complete with regard to reporting of Hispanic origin than the race data on the SSA record. The SSA records all have race entries of white, black, other or unknown. The census records have entries of Hispanic, other, white and black. The SSA record could be used to complete the missing DOB item for six persons. There were two census records that had year of birth entries of 1992. Given the other demographic data associated with this person, the SSA data seem more appropriate. Therefore, the SSA data could have been used to provide DOB data for eight census records that did not have a census response for DOB.

12. Name, DOB and Race Differed - There were 36 cases in this category. In all cases, the census record was complete and the SSA record was incomplete; "unknown" appeared for each demographic characteristic.

13. Sex and Race Differed - There were 24 cases in this category. Of these, 16 were clearly males and five were clearly females. In 11 of the 16 male cases, the census did not have an entry for sex but the SSA did. For seven of these same cases, the census had an entry for race but the SSA did not. For five cases, the race entry on the census record was more complete in terms of the reporting of Hispanic origin than the race entry on the SSA record. There were two male cases coded incorrectly in this category. Looking at the sex data, they matched on both records. The only variable that differed was race and in both cases the census had more complete race data with respect to the reporting of Hispanic origin. For all five female records, the census had more complete sex and race data than the SSA record. There were two cases where the sex on the census record was male and the sex on the SSA record was female. In one case, we cannot tell which is correct and in the other case, the first name also differs which suggests that they are different people. There was one additional case where female appeared on the census record and male appeared on the SSA record. Looking at the name, it appears that the census record is incorrect.

14. Name, Sex and Race Differed - There were two cases in this category. In both cases, the census record had data but the SSA record had "unknown" entered in the sex, race and DOB fields. The name field is blank. In both cases, the SSA considered the DOB a match however, no data are entered for the DOB.

15. DOB, Sex and Race Differed - There were nine cases in this category. Six were for females, three for males. Again, all nine records had at least some data on the census record and none on the SSA record. In

one case, the SSA considered the name a match; however, no name was provided on the SSA record.

16. All Items Differed - There were 23 cases in this category. All information differed on the census and SSA record but the SSN associated with the census record is a valid number.

17. Invalid SSN - There were 23 cases in this category. The SSN supplied on the census record is not an issued SSN by the SSA.

18. Duplicate SSN - The SSA assigns a code of 18 if the SSN represented is a duplicate of another SSN validated. There were no cases in this category.

4.2 Future Uses of the SSA File

The Census Bureau has made a policy decision not to ask for the SSN during the 2000 census. Research has shown that SSN is a very good match key and is available on many administrative records files. As a result of this decision, the SSA becomes the primary source to obtain a SSN for persons or to validate SSNs collected on administrative records. This decision raises three important questions for the administrative records program. These are:

1. What are the good match keys in order to obtain SSN from the SSA?

Determining the best person matching keys between two files require ways of uniquely identifying persons. Four data characteristics are maintained on both the 1992 National Census Test I and the SSA data files. These data characteristics are name, DOB, sex and race. Sex is too general of a characteristic for matching since there are only two possible categories; male and female. The differences associated with the race categories between the Census Bureau and the SSA make it very difficult to use race when matching. Therefore, sex and race are not good matching keys. The two remaining characteristics, name and DOB, were examined as possible matching keys. Name was examined three ways; last name only, last name and first initial, and last and first name. In addition, we examined DOB on its own and in combination with the three name categories.

From the 1992 National Census Test I, the Census Bureau obtained data for 24,719 people, all five panels. Using these data we examined name and DOB as a way of uniquely identifying people. When last name was used as the only characteristic to uniquely identify a person, 3,385 persons (13.7 percent) were not matched to any other person within the 1992 National Census Test I file. The remaining 21,334 persons were matched to at least one other person within the 1992 National Census Test I file. When last name and first initial were used, 15,710 persons (63.6 percent) were uniquely identified. Finally, when last and first name were used, 23,085 persons (93.4 percent) were uniquely identified. The analysis of name indicates what we expected; as more of the person's name is used more of the people can be uniquely identified.

The next step was to examine DOB only. DOB uniquely identified 10,414 persons (42.1 percent). This was more than three times the number of people that last name only uniquely identified.

The last step was to examine the combination of DOB with name. When DOB and last name were used together, 23,973 persons (97.0 percent) were uniquely identified. DOB and last name showed a 3.6 percentage points higher rate than last and first name to uniquely identify persons. Adding first initial to DOB and last name uniquely identified 24,485 persons (99.1 percent). Finally, when DOB, last name and first name were used, 24,583 persons (99.4 percent) were uniquely identified. Therefore, there is relatively little improvement in the rate between first initial and first name when used with DOB and last name.

From the SSA, the Census Bureau obtained data for 4,039 people. Using this data we examined name and DOB as a way of uniquely identifying people. For the three variations on name, 922 (22.8 percent), 3,139 (77.7 percent) and 3,844 (95.2 percent) persons were uniquely identified based on last name only, last name and first initial, and last and first name, respectively. The next step was to examine DOB only. DOB uniquely identified 3,446 persons (85.3 percent). This rate was higher than the rate observed using the 1992 National Census Test I data. This could be due to the smaller number of people on the SSA data file.

Finally, we examined the combination of DOB with name. From the analysis, 3,912 (96.9 percent), 3,926 (97.2 percent) and 3,934 (97.4 percent) persons were uniquely identified based on DOB with last name, with last name and first initial, and with last and first name, respectively. As with the 1992 National Census Test I analysis, there is relatively little improvement in the rate between first initial and first name when used with DOB and last name.

In order to determine the best variables for matching, we matched the uniquely identified persons in the 1992 National Census Test I file to the uniquely identified persons in the SSA file by name and/or DOB. The objective is to maximize the number of matches between the two files. A match by last name only between the 3,385 persons on the 1992 National Census Test I file with the 922 persons on the SSA file resulted in a match of only 368 persons. This represented only 39.9 percent of the 922 persons on the SSA file. The highest match rate between the two file results when DOB and last name were used. The match resulted in

3,136 persons matched between the 23,973 persons on the 1992 National Census Test I file and 3,912 persons on the SSA file. This represented 80.2 percent of the 3,912 persons on the SSA file.

After the matching, we compared the SSN from both files of the matched persons to determine if the match was correct. Under all seven matching criteria over 93 percent of the matches between the 1992 National Census Test I and the SSA file had the same SSN.

When DOB and any of the name variations (last name only, last name and first initial, and last and first name) were matched 99.9 percent of the persons had the same SSN on both files. Therefore, using DOB and last name we uniquely identified 23,973 persons (97.0 percent) from the 1992 National Census Test I file and 3,912 persons (96.9 percent) from the SSA file. Performing a match of these persons by DOB and last name resulted in 3,136 persons matched between the two files. This represented 80.2 percent of the persons on the SSA file. Comparing the SSNs from the two file resulted in 3,133 persons (99.9 percent) of the matched persons having the same SSN on both files. Only 3 persons (0.1 percent) did not have the same SSN. Therefore, DOB and last name are the best match keys in order to obtain SSN from the SSA.

However, there is a limitation on the match rates presented in this section. The calculated match rates represent an overestimate the true match rate. This is due to the fact that the SSA used special validation procedures to obtain the SSA data for the Census persons. The only way to do a true test of this would be to do a search by an alphabetic sort of name. The SSA Office of Research and Statistics has proposed this future research; however, it is not presently planned. We encourage that the Census Bureau consider doing this project after the 1995 Census Test.

2. Can we use the SSA file to obtain census item nonresponse for data items such as DOB, sex and middle name?

In order to answer this question, we examined the census item nonresponse rates for DOB, sex and middle name for the 4,039 persons with SSN. The census item nonresponse rate for DOB, sex and middle name was 0.9 percent (36 persons), 3.2 percent (128 persons) and 100 percent (4,039 persons), respectively. Note that middle initial was collected in the census and not middle name. This explains the 100 percent census item nonresponse rate for middle name. Using the SSA data, we determined the number and percent of cases that the SSA could supply the missing data item. The SSA file contained DOB for 29 of 36 persons (80.6 percent) with missing DOB on the census file. The SSA file contained sex for 120 of the 128 persons (93.8 percent) with missing sex on the census file. Finally, the SSA file contained middle name for 2,955 of the 4,039 persons (73.2 percent) with missing middle name on the census file. Therefore, data from the SSA file could be used to obtain census item nonresponse for DOB, sex and middle name.

3. What role could the SSA file play in the 1995 Census Test?

The proposal for the 1995 Census Test would require the Census Bureau to compile an administrative records data base from a variety of federal, state and local level files. SSN would be a variable on the administrative records data base. The collected SSNs would be validated in order to determine their accuracy. This would be accomplished using the same validation procedures which were used to validate the SSNs collected in the 1992 National Census Test I. In a post-census test evaluation, the administrative records staff would evaluate the accuracy of the validation results. This will involve comparing the persons data collected on the administrative records data base to the SSA data file and to the census test data. This may help shed light on the rules used to assign priority to administrative records files during data base development.

5. CONCLUSION

Based on the analysis, we can draw the following conclusions about the potential benefits of validating SSN and using the SSA data:

- Those who report a SSN report an accurate SSN. There were only 23 (0.6%) cases in which the SSN reported by the respondent was an invalid number. Likewise, there were an additional 23 cases in which the SSN was valid but the demographic information associated with the number differed completely between the two records (SSA and Census).
- The SSA data file should be used to supplement and complete responses when a census question has not been answered.
- If the DOB entry on the census form includes the same year as the year the census is conducted, we recommend using the SSA data to edit the entry. Results from this analysis indicates that for these cases the SSA is more complete.
- Given the limitations in reporting of race on the SSA file, we recommend that the census race always be used as the preferred race. SSA race should only be used as a last resort.
- Based on the analysis, we recommend that the Census Bureau not use the SSA data if the following validation codes accompany the data: 4 - Name and DOB Differed; 8 - Name, DOB and Sex Differed; 12 - Name, DOB and Race Differed; 14 - Name, Sex and

Race Differed; 15 - DOB, Sex and Race Differed; 16 - All Items Differed; 17 - Invalid SSN; or 18 - Duplicate SSN. Based on this study, this represents 133 (3.3 percent) of the validated records.

- The Census Bureau should continue to negotiate with the SSA to validate SSNs collected while building the administrative records data base for the 1995 Census Test. This will allow us to evaluate the quality of other administrative records data and to continue research on matching keys. In addition, the SSA file should be evaluated after the 1995 Census Test to determine if it can be used to supplement and complete responses when a census question has not been answered.

6. REFERENCE

Duncan, George T., Thomas B. Jabine, and Virginia A. De Wolf, eds. 1993. Private Lives and Public Policies. National Academy Press.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 1

Match Code		Male	Female	Not Reported	Total Number	Total Percent
Records sent to SSA		1839	2072	128	4039	100.0
1.	All Items Verified	1391	1536	59	2986	73.9
2.	Name only Differed	13	59	3	75	1.9
3.	DOB only Differed	15	22	2	39	1.0
4.	Name & DOB Differed	18	10	0	28	0.7
5.	Sex only Differed	4	7	31	42	1.0
6.	Name & Sex Differed	0	0	0	0	0.0
7.	DOB & Sex Differed	2	3	1	6	0.1
8.	Name, DOB & Sex Differed	3	9	0	12	0.3
9.	Race only Differed	320	350	13	683	16.9
10.	Name & Race Differed	9	18	1	28	0.7
11.	DOB & Race Differed	10	12	1	23	0.6
12.	Name, DOB & Race Differed	23	11	2	36	0.9
13.	Sex & Race Differed	5	8	11	24	0.6
14.	Name, Sex & Race Differed	1	1	0	2	0.0
15.	DOB, Sex & Race Differed	0	6	3	9	0.2
16.	All Items Differed	12	11	0	23	0.6
17.	Invalid SSN	13	9	1	23	0.6
18.	Duplicate SSN	0	0	0	0	0.0

DOB indicates date of birth
SSN indicates Social Security Number