Census 2000 Testing,
Experimentation,
and Evaluation Program
January 13, 2003

# Analysis of the Social Security Number Validation Component of the Social Security Number, Privacy Attitudes, and Notification Experiment

FINAL REPORT

This research paper reports the results of research and analysis undertaken by the U.S. Census Bureau. It is part of a broad program, the Census 2000 Testing, Experimentation, and Evaluation (TXE) Program, designed to assess Census 2000 and to inform 2010 Census planning. Findings from the Census 2000 TXE Program reports are integrated into topic reports that provide context and background for broader interpretation of results.

Linda Brudvig
Planning, Research, and
Evaluation Division

USCENSUSBUREAU
*Helping You Make Informed Decisions*

# CONTENTS

# LIST OF TABLES

# EXECUTIVE SUMMARY

The possibility of using administrative records from other federal government agencies to supplement census data has been investigated for some time at the Census Bureau. The use of administrative records could potentially increase completeness of measurement by reducing respondent burden with shorter questionnaires and improve data quality by eliminating memory/respondent errors. The Social Security number has become a widely used personal identifier for identifying program participants and, for this reason, the Census Bureau conducted research about its collection and use. The purpose of the Social Security number, Privacy Attitudes, and Notification Experiment is to assess the effects of Social Security number requests and different notifications of administrative record use on census response behavior.

This study is one of three analytical components of the Social Security number, Privacy Attitudes, and Notification Experiment. It examines the accuracy of respondents' reported Social Security numbers by comparing them to the Census Numident File and assessing overall accuracy, differences between high coverage and low coverage areas, and differences between Social Security number request strategies and administrative records use notification strategies. The results are further compared to the results of the 1992 Simplified Questionnaire Test.

The analysis of the experimental data (Social Security number validation rates) was conducted by using a simple test for statistical significance and by measuring the pairwise differences in the validation rates among the panels. The pairwise analysis was designed so that statements about the significance of treatment effects (i.e., differences in validation rates) can be made about all tests simultaneously while maintaining a 90 percent confidence level.

This paper shows that, if reported, Social Security numbers are accurate. There is a small but statistically significant difference between high and low response areas. There is no evidence that requesting Social Security number for Person 1 only, all persons in household, or inclusion of general or specific notification of administrative records use affects the quality of the reported Security number. More specifically:

- The respondent-provided Social Security numbers are accurate. The overall validation rate (based on matching Social Security number, name, gender, and year of birth to the Census Numident File) is 94.77 percent.

- Valid Social Security number rates are high for both high response areas (95.15 percent) and low response areas (92.80 percent). The difference is small (2.35 percent) and statistically significant.

- The rate of valid Social Security numbers for Person 1 in each panel is high, ranging from 96.01 to 96.93 percent. The results of the pairwise comparisons show no statistical significance when the panel requesting Social Security number for Person 1 only is compared to the panel requesting Social Security number for all household members. Further, the distinction between general and specific notification of administrative records

use has no measurable influence on valid responses.

- Social Security number validation rates generally show a steady decline in each panel from Person 2 through Person 6.

- The 1992 Simplified Questionnaire Test supported the high validation rates shown in this study when considering comparable panel design and matching variables. Overall validation rates were 91.00 percent and 94.83 percent respectively.

Overall, the responses that we receive to requests for Social Security number are accurate. However, to assess the usefulness of such requests within the unique environment of a decennial census, we must also consider the reduction in mail response when Social Security number is requested, increased Social Security number item nonresponse, the extent to which having a valid Social Security number allows us to link to other files, and the extent to which we are able to obtain correct information from such a link. To this end, we recommend the following research:

- **Look at the cumulative nonresponse to requests for Social Security number** (unit nonresponse, Social Security number item nonresponse, and Social Security number invalid rates) to obtain an indicator of the extent to which matching to administrative records could take place.

- **Conduct analysis of the characteristics of households** that provide and do not provide the Social Security number, the accuracy of households reconstructed from administrative records, and the effect of having and not having the Social Security number in household reconstruction. In conjunction with the accuracy of the Social Security number, this analysis would be an indicator of the quality of data we might expect.

- **Use the cumulative nonresponse to Social Security number and the results of other research to conduct a cost-benefit analysis to determine the extent to which we could use administrative records data to complete household enumeration** (assuming decennial census population and housing unit universes within decennial census time constraints).

- **Conduct focus groups or research similar to the Survey of Privacy Attitudes later in the decade.** This Survey of Privacy Attitudes examined patterns of attitude change in the public's privacy concerns between 1996, 1999, and 2000. Given the profound impact of September 11, 2001, we may expect changing views on privacy attitudes.

- **Continue to work closely with the Privacy Office and privacy advocates to stay informed of trends in the privacy and confidentiality arena.**

# 1. BACKGROUND

## 1.1 Overview

The Census Bureau undertakes a program of experimentation during decennial censuses to measure the effectiveness of new techniques, methodologies, and technologies in the special environment that a decennial census generates, such as mass temporary hiring, promotion and outreach in coordination with local governments, the national paid advertising campaign, and the nationwide distribution of public use forms. Results from experiments form recommendations for subsequent testing and ultimately help design the next decennial census (Neugebauer, 1999).

Decennial censuses beginning in 2010 may rely on expanded use of administrative records information obtained from other Federal agencies (Neugebauer, 1999). The use of administrative records could potentially increase completeness of measurement by reducing respondent burden with shorter questionnaires and improve data quality by eliminating memory/respondent errors (Guarino, Hill, and Woltman, 2001). The Social Security number (SSN) has become a widely used personal identifier for identifying program participants. For example, citizens are required to use SSN as the taxpayer identification number. Likewise, many other federal, state, and local government agencies collect and use SSN to administer their programs. Since the SSN is such a widely used personal identifier, the Census Bureau conducted research dealing with its collection and use (Leslie and Treat, 1994).

The Social Security Number, Privacy Attitudes, and Notification (SPAN) experiment consists of three major components to achieve the research objectives. The first component uses a list-assisted random digit dial (RDD) telephone survey to collect data on the public's privacy concerns; it is referred to as the Study of Privacy Attitudes in 2000 (SPA2000). The second component analyzes the effects of different notifications, two strategies for obtaining SSN information, and notification combined with the SSN request on response behavior and is called the SSN notification component. The third analytical component involves the validation of SSNs collected from four experimental panels that request it. It examines what percentage of SSNs obtained in the experiment are valid by panel (Neugebauer, 1999). This report is the third component. It contains a full analysis of the SSN validation.[1]

The goal of the SSN validation component is to examine the validation rates of the SSNs collected from the four panels that request it. It is the first empirical research to measure the effects of an SSN request or public notification of administrative record use on the validity of the SSNs provided by respondents in a decennial census environment. This report examines verifying the SSNs collected from the four panels that request it against the Census Numerical Identification (Numident) file. It defines and examines what percentage of SSNs obtained in the experiment are valid (direct and indirect matches) and invalid by panel (Neugebauer, 1999). Note that planned analysis of the characteristics of households that provide and do not provide the SSN, the accuracy of households reconstructed from administrative records, and the effect of having and not having the SSN in household reconstruction was not undertaken because of decennial resource considerations.

[1]A related paper, "The Effect of Administrative Record Use Notification on SSN Reports," focuses specifically on response rates to the SSN item at the person level (Stapleton, 2002).

## 1.2 Past research

### *1.2.1 Willingness to provide SSNs and accuracy of SSNs*

Past studies in the privacy and confidentiality realm show that people who are most concerned with privacy participate less in surveys and censuses than those who are not concerned (Kulka, Holt, Carter, and Dowd, 1991; Singer, Mathiowetz, and Couper, 1993; Gates and Bolton, 1998). To study this phenomenon, qualitative and quantitative analyses have been conducted to assess public opinion and response behavior to SSN requests on census forms (Guarino et al., 2001). Response behaviors include mail response rates, data quality as suggested by form completeness, SSN item response, and the validity of reported SSNs.

Qualitative research such as the 1992 focus groups indicated extreme negative reaction to an SSN request; however a mailout/mailback test [the 1992 Simplified Questionnaire Test (SQT)] showed that there was a small, but significant, actual decrease (-3.4 percent) in mail response rates. It also indicated that asking for SSN seemed to lower the response in the Low Response Area (LRA) stratum more than the High Response Area stratum. Additionally, it showed that respondents do a good job reporting accurate SSNs when they choose to report an SSN and that SSNs for Person 1 were provided more accurately than Person 6 (Leslie and Treat, 1994). These findings were unexpected and seemingly contradicted the anticipated extent to which respondents would resist providing an accurate identifier with data linking implications (Guarino et al., 2001).

Dillman, Sinclair, and Clark (1993) also found that asking SSN lowers survey completion rates. These drops could be due to the respondent's objections to providing personally identifying information, or to the difficulty in obtaining this information for some household members.

For further investigation, a question asking respondents' willingness to provide their SSNs on census forms was included in a series of surveys aimed at measuring privacy attitudes of U.S. residents over time. Singer (Singer, VanHoewyk, Tourangeau, Steiger, Montgomery, and Montgomery, 2001) reports that the percentage of respondents willing to provide their SSN on a census form declined from 68 percent in 1996 to 55 percent in 1999 and 56 percent in 2000. The drop in willingness was significant between 1996 and 1999, but there was no further significant change between 1999 and 2000.

### *1.2.2 Access difficulty*

Bates (1992) analyzed response to SSN by person number on the questionnaire. The results indicate that the reporting of SSN becomes more difficult beyond person number two. There is some evidence that failure to provide SSN is not always due to unwillingness. Nonresponse may sometimes be a result of the lack of availability, or inaccessibility of the information to the respondent. Previous research indicates that the first person on the census form (Person 1) is usually the respondent for the entire household (DeMaio and Bates, 1990). Relationship to Person 1 may be an indicator of item response to SSN and reflective of the respondent's inability

to provide SSN for household members.

Bates (1992) cites focus group evidence that providing SSNs for children might be more difficult because SSNs are not routinely used before a certain age. She found that item nonresponse to the SSN item increased as person number increased. Presumably, this reflects the difficulty in providing SSN for children or unrelated household members (as household members are often listed in order by age or by relationship to the respondent). A report by Dillman, Sinclair, and Clark (1993) noted that item nonresponse to SSN for children under age 17 was 25 percent, substantially higher than for other questions. Dillman, Reynolds, and Rockwood (1991) report that a focus group investigation revealed that even though some people had no objection to providing SSNs, finding this information, especially for children and unrelated household members, might be difficult.

### *1.2.3 SSN Request and Notification of Administrative Record Use*

Before Census 2000, no empirical research measured and assessed response behavior (such as mail response, questionnaire item nonresponse, nonresponse to the SSN item, and the validity of the SSNs that were received from respondents) to a particular type of SSN request (SSN requested for all household members versus only for the person completing the form) and notification of administrative record use (general notification that the Census Bureau may use statistical data from other federal agencies versus more specific notification where agencies are named) in a decennial census environment. However, some research was conducted during mid-cycle tests. Bates studied item nonresponse to SSN in the 1992 Simplified Questionnaire Test (SQT). She found that asking SSN significantly lowered unit response rates overall and for the 1990 LRA groups. Mail completion rates for the SSN form were 6.2 percentage points lower for households from LRAs. However, the item nonresponse rate from residents of 1990 LRAs was not significantly different than respondents from other areas.

Past research on notification of administrative record use is qualitative in nature and therefore does not indicate the effect of notification on census response or accuracy of the SSNs that are reported. However, Singer (1978) investigated the effects in face-to-face interviews of more (versus less) information about sensitive subject matter in survey introductions. She found no effect of varying information about content on response rates. Other findings reveal that focus group participants are generally unsure about what effect notification will have on census response (Guarino et al., 2001). Some believe that notification of administrative record use will have no effect on response, while others believe that notification will decrease response. With regard to the type of notification, focus group administrators note that many of the participants did not understand the task of rating which notification was most persuasive in increasing participation, and instead rated the notification specimens by which use of records they felt was most justifiable (Aguirre International, 1995).

Interestingly, current research during Census 2000 (Guarino et al., 2001) shows that: SSN request slightly decreases response and there is no differential effect of SSN request between low coverage areas (LCA) and high coverage areas (HCA); SSN request increases return of

incomplete forms; general notification of administrative record use slightly decreases response while the inclusion of specific notification does not; and specific/general notification increases response to the Person 1 SSN.

Additionally, Stapleton (2002) indicates that, for Person 1, the type of SSN request does not affect response to the SSN item. However, for Person 1 and Persons 2 - 6, including notification of administrative record use (regardless of type) does significantly increase the odds of a respondent providing their SSN.

### *1.2.4 Summary*

In summary, a review of previous and current research indicates that: asking SSN decreases survey response rates; increases incomplete forms; results in high SSN item nonresponse; and there is no differential effect of SSN request between HCAs and LCAs. There is some evidence that there is no difference on response rates and SSN item response rates when respondents were provided more information about a sensitive subject (notification of SSN use). Other research indicates that respondents have difficulty providing this information for children and unrelated household members. Lastly, previous research shows that when respondents do provide an SSN, it is accurate.

## 1.3 Hypotheses

Relying upon prior research, three hypotheses were developed concerning the validation rates of reported SSNs overall, by coverage area, and by person number. In the absence of past quantitative studies regarding the effects of notification upon the validity of SSN responses, two hypotheses were developed based on expectations from privacy research.

1. The SSN validation rate will be high when SSN is reported.

2. There will be little difference in validation rates between low coverage areas and high coverage areas.

3. SSN validation rates will steadily decrease by Person number. That is, Person 2 will have higher SSN validation rates than Person 3 and so on through Person 6.

4. Notification of administrative record use will cause small but significant drops in SSN validation rates, with specific notification (including agency names) having a stronger effect than general notification.

5. Requesting SSN in the absence of general or specific notification will yield higher validation rates for Person 1 when SSN is requested only for Person 1 as compared to all household members.

The SSN validation component will provide a better understanding of the potential ramifications of requesting SSN on behavior regarding accuracy of SSNs in a (limited) decennial census environment.

# 2. METHODS

## 2.1 Panel design

The experimental treatments for the SPAN experiment are implemented within ten panels. The SSN validation component involves the four panels where SSN is requested. Households selected for this experiment were randomly assigned to each panel (Guarino et al., 2001). Two short form panels have forms modified with an SSN request either for all household members or for only the person completing the form (i.e., "Person One"). Notification, beyond the statement informing respondents that providing SSN is voluntary, is not a part of these panels. Two short form panels combine the notification aspect and SSN request for all household members.

Specifically, the four experimental groups are:

Panel 1: All (household members) SSN Request
Panel 2: One (Person 1) SSN Request
Panel 3: All SSN Request, General Notification
Panel 4: All SSN Request, Specific Notification

Each panel receives the full complement of census mailout materials in the same sequence and timing as the official Census 2000 schedule. Experimental letters and forms *are* the official census forms received by the sampled households (Guarino et al., 2001).

As noted, the two notifications are referred to as "general" and "specific." The notification is written in the letters accompanying the questionnaires and describes how and why the Census Bureau may use administrative records data from other Federal agencies. The general notification mentions the Census Bureau's possible use of statistical data from other Federal agencies, while the specific notification goes further to name the Federal agencies. The general notification is:

> To improve the quality of census statistics, the Census Bureau sometimes uses records from other government agencies. Using other agencies' records helps make the census more complete. By making better use of government records that already exist, the Census Bureau may be able to ask you fewer questions in the census.

The specific notification wording is:

> To improve the quality of census statistics, the Census Bureau sometimes uses records from other government agencies, such as the Social Security Administration, the Internal Revenue Service, or agencies providing public housing assistance. Using other agencies' records helps make the census more complete. By making better use of government records that already exist, the Census Bureau may be able to ask you fewer questions in the census.

Because providing the SSN is voluntary, the cover letter for all four panels with the SSN request contained an additional statement:

To improve the quality of census statistics, the Census Bureau sometimes uses records from other government agencies. For that purpose, we are asking for your social security number; however, providing your social security number is voluntary.

## 2.2 Sample design

The sample of households was taken from the July 1999 version of the Decennial Master Address File (DMAF) mailout/mailback universe of over 92 million addresses. This universe excludes samples for the Accuracy and Coverage Evaluation (A.C.E.) Listing, the contamination evaluation, congressional addresses, list/enumerate areas, and update/leave areas.

The sample was equally allocated to two strata that reflect expected difference in the population composition by race, tenure, and anticipated Census 2000 mail return rates (taken from previous census experience). These strata are based on 1990 census tract level race and tenure data and referred to as low and high coverage areas (LCA and HCA, respectively). The LCA stratum was expected to contain a high proportion of African-American and Hispanic populations and renter occupied housing units. When selection of the sample households was conducted, nearly 81 percent of the total DMAF universe consisted of households within the HCA stratum. Oversampling of the LCA occurred to equally allocate the sample across the two strata.

Approximately 52,000 households were selected and randomly assigned to each experimental panel. The mailout sample size for each of the ten panels consisted of a little over 5,200 addresses, equally allocated to the HCA and LCA strata (i.e., around 2,600 addresses per stratum). Specific details about address omissions such as undeliverables and duplicates and replacements can be found in Guarino et al., 2001.

The total number of households selected in this experiment (i.e., Panels 1 through 4) was 20,998. For this paper, we examine the accuracy of the SSNs that were reported[2] for all persons in these housing units that returned their questionnaires by mail. There were a total of 21,745 reported SSNs as shown by panel in Table 1.

**Table 1. Number of SSN responses by panel**

| All Panels | Panel 1 (All SSNs) | Panel 2 (One SSN) | Panel 3 (All SSNs, general notification) | Panel 4 (All SSNs, specific notification) |
|---|---|---|---|---|
| 21,745 | 6,348 | 2,713* | 6,367 | 6,317 |

*In Panel 2, SSN was requested for Person 1 only; in Panels 1, 3, and 4, SSN was requested for Persons 1 - 6.

Figures in this paper will be weighted to account for oversampling of the LCA stratum. The inverse of the sampling interval for each stratum with an experimental group is the weight for each case contained in the panel and stratum (Guarino et al., 2001).

[2]Cases with a reported SSN that is less than nine digits or is missing are counted as missing values. Item nonresponse rates for SSN are presented in a separate report.

## 2.3 SSN verification procedures

The unedited person records with SSN entries on the questionnaires from housing units within the mailout/mailback experimental sample Panels 1, 2, 3, and 4 were matched to the Census Numident file to determine the validity of the SSN that was provided. The Census Numident file is provided by the Social Security Administration (SSA) and is reformatted for Census use. It includes the following information, if available: SSN, first, middle, and last name, year of birth, gender, alternate or previous names, and race (Administrative Records Research Staff, 2000). See the Appendix for an abbreviated version of the verification procedures.

## 2.4 Analytic procedures

The analysis of the experimental questionnaire data is conducted by measuring the pairwise differences in SSN validation rates among the panels. The analysis is designed so that statements about the significance of treatment effects (i.e. differences in SSN validation rates) can be made about all tests simultaneously while maintaining a 90 percent confidence level (the Census Bureau Standard). For a more complete discussion on pairwise analysis, refer to the SSN notification component (Guarino et al., 2001).

## 2.5 Variance estimation

Since the analysis is done at the person level, a clustering effect at the household level must be considered because each household has one respondent for all household members. This is done within WesVar, treating the survey as a one-stage sample design, and treating each household as a primary selection unit (PSU). Since this creates well over the maximum number of PSUs that WesVar can handle, clusters were grouped randomly into 256 "pseudo" clusters. To take into account the stratified sample design in the data analysis, WesVar was used to compute standard errors for all estimates using a stratified jackknife approach. This replication option is suitable for stratified designs with two or more PSUs per stratum (Stapleton, 2002).

## 2.6 Applying quality assurance procedures

We applied quality assurance procedures throughout the creation of this report. They encompassed how we determined study methods, created specifications for project procedures and software, designed and reviewed computer systems, developed clerical and computer procedures, analyzed data, and prepared this report (U.S. Census Bureau, 2000).

# 3. LIMITS

While this experiment was conducted during a decennial census, it was conducted using a small sample of the entire population. If extended to the entire nation, results of such a request would likely be very different because the media attention concerning the privacy issues of asking for SSN would likely be magnified.

The planned analysis of the characteristics of households that provide and do not provide the SSN, the accuracy of households reconstructed from administrative records, and the effect of having and not having the SSN in household reconstruction was not undertaken because of decennial resource constraints. In conjunction with the accuracy of the SSNs, this analysis would be an indicator of the quality of data and usefulness of collecting SSN in future surveys and censuses.

# 4. RESULTS

## 4.1 Are the SSNs provided accurate?

Accurate SSNs for this study are the valid SSNs–those where the SSN and name provided by the respondent match an SSN, name, and, as needed, year of birth and gender on the Census Numident file. There are five categories of valid outcomes that resulted from the SSN verification matching process. For each, the SSN matches and the:

- parsed name matches the Numident. A parsed name is one that is placed into first, middle, and last name fields, if available. This match is flagged as a "1."

- parsed name matches the alternate name on the Numident. An alternate name is any previous name such as a maiden name or a name before a name change . This match is flagged as a "2."

- parsed name matches the concatenated name on the Numident. A concatenated name is one where all the letters in the first, middle, and last name are merged together with no spaces. This match is flagged as a "3."

- standardized name matches the Numident. A standardized name is a more formal version of a name, for example, Debbie is standardized to Deborah and Jim to James. This match is flagged as a "5."
- standardized name matches the concatenated name on the Numident. This match is flagged as a "7."

Note that direct matches are defined as flags 1 and 2 and, for this study, are considered stronger matches than indirect matches, flags 3, 5, and 7.

Invalid (not accurate) SSNs are:

- invalid entry. The last two digits of the SSN are non-numeric. (Records with fewer than nine digits or all nine digits of the SSN blank are not in universe.) These are flagged as a "0."

- the SSN is not in the Numident. These are flagged as an "A."

- the SSN is in the Numident but the name doesn't match. These are flagged as a "B."

Table 2 confirms our expectation that the SSN verification rates for valid SSNs are high overall and for each panel.

**Table 2. SSN verification rates by panel**

| Verification Status | All Panels | Panel 1 (All SSNs) | Panel 2 (One SSN) | Panel 3 (All SSNs, general notification) | Panel 4 (All SSNs, specific notification) |
|---|---|---|---|---|---|
| **SSN Valid:** | | | | | |
| **Direct match:** | | | | | |
| 1 = parsed name matched Numident | 93.65% | 93.66% | 94.57% | 92.84% | 94.05% |
| 2 = parsed name matched alternate name on Numident | 0.68% | 0.63% | 0.83% | 0.73% | 0.60% |
| **Indirect match:** | | | | | |
| 3 = parsed name matched concatenated name on Numident | 0.17% | 0.20% | 0.26% | 0.11% | 0.16% |
| 5 = standardized name matched Numident | 0.26% | 0.34% | 0.40% | 0.28% | 0.11% |
| 7 = standardized name matched concatenated name on Numident | 0.01% | 0.00% | 0.00% | 0.00% | 0.03% |
| **SSN Invalid:** | | | | | |
| 0 = Invalid entry | 0.16% | 0.08% | 0.15% | 0.31% | 0.09% |
| A = SSN not in Numident | 2.03% | 2.04% | 0.93% | 2.53% | 2.00% |
| B = SSN in Numident, name didn't match | 3.04% | 3.04% | 2.87% | 3.20% | 2.97% |

As shown in Table 3, when examining the direct and indirect matches of the valid SSN entries, it is clear that most are direct matches with an overall match rate of 94.33 percent compared to the overall indirect match rate of 0.44 percent. Invalid SSNs are quite low with an overall rate of 5.23 percent.

**Table 3. Verification rates for valid and invalid SSNs by panel**

| Verification Status | All Panels | Panel 1 (All SSNs) | Panel 2 (One SSN) | Panel 3 (All SSNs, general notification) | Panel 4 (All SSNs, specific notification) |
|---|---|---|---|---|---|
| **SSN Valid:** | **94.77%** | **94.83%** | **96.05%** | **93.96%** | **94.94%** |
| Direct match | 94.33% | 94.29% | 95.40% | 93.57% | 94.65% |
| Indirect match | 0.44% | 0.54% | 0.65% | 0.39% | 0.29% |
| **SSN Invalid** | **5.23%** | **5.17%** | **3.95%** | **6.04%** | **5.06%** |

## 4.2. Are there differences in SSN verification rates between HCAs and LCAs?

Table 4 shows that the valid SSN rates for both HCA and LCA are very high at over 90 percent. As expected, the difference between them is small; the rate for HCA is 2.35 percentage points higher than LCA and this difference is statistically significant.

**Table 4. SSN verification rates by mailout/mailback areas in all experimental panels**

| Verification Status | All Panels | HCA | LCA |
|---|---|---|---|
| **SSN Valid:** | **94.77%** | **95.15%** | **92.80%** |
| **Direct match:** | | | |
| 1 = parsed name matched Numident | 93.65% | 94.04% | 91.66% |
| 2 = parsed name matched alternate name on Numident | 0.68% | 0.68% | 0.64% |
| **Indirect match:** | | | |
| 3 = parsed name matched concatenated name on Numident | 0.17% | 0.14% | 0.31% |
| 5 = standardized name matched Numident | 0.26% | 0.28% | 0.18% |
| 7 = standardized name matched concatenated name on Numident | 0.01% | 0.01% | 0.01% |
| **SSN Invalid:** | **5.23%** | **4.85%** | **7.20%** |
| 0 = Invalid entry | 0.16% | 0.15% | 0.20% |
| A = SSN not in Numident | 2.03% | 1.88% | 2.82% |
| B = SSN in Numident, name didn't match | 3.04% | 2.82% | 4.18% |

## 4.3 Are there differences in the valid SSN rates between the two strategies for obtaining SSN and the two differently worded notifications of administrative records use?

In order to examine for significance the differences in the valid SSN rates between the two strategies for obtaining SSN (SSN requested for all household members versus only for the person completing the form) and the two differently worded notifications of administrative records use (general notification that the Census Bureau may use statistical data from other federal agencies versus more specific notification where agencies are named), pairwise comparisons of Person 1 valid SSN rates are considered among the panels for which this information is requested (Panels 1 - 4). Since Panel 2 requested SSN only for Person 1, our pairwise comparison is limited to Person 1. Each of these four panels receives some degree of notification of the possibility of administrative record use due to the statement in the cover letter explaining the request for SSN. As shown in Table 5, the rate of valid SSNs for Person 1 in each

panel is very high, ranging from 96.01 percent to 96.93 percent, a difference of less than one percentage point.

**Table 5. Valid SSN rates for Person 1 by panel**

| Panel | Validation Rate |
|---|---|
| Panel 1 (all SSNs) | 96.93% |
| Panel 2 (one SSN) | 96.06% |
| Panel 3 (all SSNs, general notification) | 96.14% |
| Panel 4 (all SSNs, specific notification) | 96.01% |

Table 6 shows the results of the pairwise comparisons. There is no significant difference in the valid SSN rate when the panel requesting only one SSN is compared to the panel requesting all SSNs. From the perspective of Person 1, these forms do not differ in their request for SSN and, therefore, no difference in response to this item is expected. Further, the distinction between general and specific notification has no measurable influence on valid responses to the SSN item for Person 1.

These results were unexpected; however, they confirm the findings from the 1992 SQT, discussed in the next section.

**Table 6. Multiple comparisons of valid SSN rates for Person 1 by panel**

| Pairwise Comparison | Difference | SE of Difference |
|---|---|---|
| Panel 1 (all SSNs) - Panel 2 (one SSN) | 0.88% | 0.54 |
| Panel 1 (all SSNs) - Panel 3 (all SSNs, general notification) | 0.79% | 0.52 |
| Panel 1 (all SSNs) - Panel 4 (all SSNs, specific notification) | 0.92% | 0.57 |
| Panel 2 (one SSN) - Panel 3 (all SSNs, general notification) | -0.09% | 0.59 |
| Panel 2 (one SSN) - Panel 4 (all SSNs, specific notification) | 0.04% | 0.65 |
| Panel 3 (all SSNs, general notification) - Panel 4 (all SSNs, specific notification) | 0.14% | 0.62 |

Table 7 shows the valid SSN rates for Persons 2 - 6 for the panels that requested it. Again, valid SSN rates for each person are similar in each panel. For example, the Person 2 valid SSN rates are all over 95 percent. We see that each of the panels show high validation rates for all persons and, as expected, for most, there is a steady decline in each panel from Person 2 through Person 6 (the exception is a slight increase for Person 5 in Panel 4).

**Table 7. Valid SSN rates for Persons 2 - 6 by panel***

| Panel | Person 2 | Person 3 | Person 4 | Person 5 | Person 6 |
|---|---|---|---|---|---|
| All Panels | 95.45% | 92.90% | 89.08% | 87.53% | 82.80% |
| Panel 1 (all SSNs) | 95.34% | 93.87% | 89.82% | 85.33% | 84.38% |
| Panel 3 (all SSNs, general notification) | 95.03% | 91.93% | 86.60% | 86.46% | 80.23% |
| Panel 4 (all SSNs, specific notification) | 95.98% | 93.15% | 90.75% | 91.07% | 83.48% |

*Panel 2 requested SSN for Person 1 only

## 4.4 How do the SSN validation rates for the Census 2000 SPAN compare to the 1992

### SQT SSN validation rates?

As part of the 1992 National Census Test I, the 1992 SQT SSN validation project included a panel that requested the SSN for each person living in the housing unit. The SSA validated the SSNs collected during the test by matching them to their NUMIDENT (Numerical Identification) file. The SSA's NUMIDENT file is a transaction file of SSN applications. It contains a record for each update made to persons' data associated with an SSN. This includes full name, date of birth, race, and gender.

The 1992 SQT panel is most like the Census 2000 SPAN Panel 1 where SSN was requested for all household members and general or specific notification of administrative record use was not included. The Census Bureau's Numident file match used matching variables similar to the 1992 SQT match (SSN, name, date of birth, and gender), but the verification process was different. For example, the 1992 SQT matched on full name and the Census 2000 SPAN Panel 1 match for name included parsed, alternate, standardized, and concatenated names in its match. These differences limit the extent to which the 1992 SQT results can be compared to the Census 2000 SPAN Panel 1 results.

Table 8 shows that, generally, the valid SSN rates are high and the invalid rates are low for both the 1992 SQT and Census 2000 Panel 1. Based on these results, we can say that respondents do a good job reporting accurate SSNs when they choose to report an SSN.

**Table 8. Comparison of 1992 SQT and Census 2000 Panel 1 valid and invalid SSN rates**

| Verification Status | 1992 SQT* | Census 2000 Panel 1 |
|---|---|---|
| **Valid** (SSN and name matched; date of birth and gender matched as required by matching process) | 91.00% | 94.83% |
| **Invalid** (SSN, name, or other non-match) | 9.00% | 5.17% |

*Race is not included as a matching variable.

## 5. RECOMMENDATIONS

This paper shows that, if reported, SSNs are accurate. There is a small but statistically significant difference between HCA and LCA. There is no evidence that requesting SSN for Person 1 only, all persons in household, or inclusion of general or specific notification of administrative records use affects the quality of the reported SSN.

However, to assess the usefulness of requesting SSN within the unique environment of a decennial census, we must also consider the reduction in mail response when SSN is requested, increased SSN item nonresponse, the extent to which having a valid SSN allows us to link to other files, and the extent to which we are able to obtain correct information from such a link. To this end, we recommend the following research:

- Look at the cumulative nonresponse to SSN (unit nonresponse, SSN item nonresponse, and SSN invalid rates) to obtain an indicator of the extent to which matching to administrative

records could take place.

- Conduct analysis of the characteristics of households that provide and do not provide the SSN, the accuracy of households reconstructed from administrative records, and the effect of having and not having the SSN in household reconstruction. In conjunction with the accuracy of the SSNs, this analysis would be an indicator of the quality of data we might expect.

- Use the cumulative nonresponse to SSN and the results of other research to conduct a cost-benefit analysis to determine the extent to which we could use administrative records data to complete household enumeration (assuming decennial census population and housing unit universes within decennial census time constraints).

- Conduct focus groups or research similar to the Survey of Privacy Attitudes later in the decade. This Survey of Privacy Attitudes examined pattens of attitude change in the public's privacy concerns between 1996, 1999, and 2000. Given the profound impact of September 11, 2001, we may expect changing views on privacy attitudes.

- Continue to work closely with the Privacy Office and privacy advocates to stay informed of trends in the privacy and confidentiality arena.

## Acknowledgments

I wish to thank Debbie Bolton and Randall Neugebauer for their initiation, design, and implementation of this Census 2000 experiment; Henry Woltman, Joan Hill, Jennifer Guarino, Courtney Stapleton, and Jason Machowski for the statistical expertise and guidance; Dean Judson and Debbie Wagner for processing the SSN verification data; and George Train and Jacques Wilmore for processing the verification data for analysis.

## References

Administrative Records Staff. "Statistical Administrative Records System, 1999–Social Security Number Verification Programming Specification," U.S. Census Bureau; Planning, Research, and Evaluation Division. October 12, 2000.

Aguirre International. "Public Concerns About the Use of Administrative Records," Contractor's report, July 1995.

Bates, N. "Revised Item Nonresponse Results for Social Security Number From the Simplified Questionnaire Test (SQT)," Memorandum to Robert D. Tortura and Susan M. Miskura, dated June 18, 1992.

DeMaio, T.J., and Bates, N.A. (1990) "Who Fills Out the Census Form?," *Proceedings of the Survey Research Methods Section of the American Statistical Association.*

Dillman, Don A., Sinclair, Michael D., and Clark, John R. (1993) "Effects of Questionnaire

Length, Respondent Friendly Design, and a Difficult Question on Response Rates for Occupant-Addressed Census Mail Surveys," *Public Opinion Quarterly*, Vol. 57, pp 289-304.

Gates, G.W., and Bolton, D. (1998) "Privacy Research Involving Expanded Statistical Uses of Administrative Records," *Proceedings of the Section on Government Statistics and Section on Social Sciences of the American Statistical Association*.

Guarino, J.A., Hill, J.M., and Woltman, H.F. "Analysis of the Social Security Number - Notification Component of the Social Security Number, Privacy Attitudes, and Notification Experiment." U.S. Census Bureau; Planning, Research, and Evaluation Division. July 3, 2001.

Kulka, R.A., Holt, N.A., Carter, W., and Dowd, K.L. (1991) "Self-Reports of Time Pressures, Concerns for Privacy, and Participation in the 1990 Mail Census," *Proceedings of the Annual Research Conference*.

Leslie, Theresa F., and Treat, James B. "Results From the Verification of Social Security Numbers Collected During the 1992 National Content Test I." U.S. Census Bureau; Decennial Management Division and Decennial Statistical Studies Division. December 14, 1994.

Neugebauer, Randall, "Census 2000 Experimentation Program Master Plan: The Social Security Number, Privacy Attitudes, and Notification Experiment." U.S. Census Bureau; Planning, Research, and Evaluation Division. November 5, 1999.

Singer, E. 1978. "Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys," *American Sociological Review*, Vol. 43, Issue 2, pp 144-162.

Singer, E., Mathiowetz, N.A., and Couper, M. (1993) "The Impact of Privacy and Confidentiality Concerns on Survey Participation: The Case of the 1990 U.S. Census," *Public Opinion Quarterly*, Vol 57, pp 465-482.

Singer, E., Van Hoewyk, J., Tourangeau, R., Steiger, D.M., Montgomery, M., Montgomery, R., "1999 - 2000 Surveys of Privacy Attitudes." U.S. Census Bureau; Planning, Research, and Evaluation Division. December 10, 2001.

Stapleton, Courtney N. (2002) "The Effect of Administrative Record Use Notification on Social Security Number Reports," *Proceedings of the Survey Research Methods Section of the American Statistical Association*.

U.S. Census Bureau. "Census 2000 Evaluation Program Quality Assurance Process." Decennial Statistical Studies Division and Planning, Research, and Evaluation Division. July 31, 2000.

## Appendix: Abbreviated Social Security Number Verification Procedures[3]

Using the last two digits of the SSN, invalid entries (non numeric) are flagged and not run through the verification program.

The remaining input records go through the following process:

- Input records are standardized, that is, the names are:

  1) parsed (placed into first, middle, and last name fields, if available)

  2) standardized names are added (for example, Debbie is standardized to Deborah; Jim to James)

- SSN for the input record is matched to the Numident. Nonmatches are assigned an A flag.

- SSN on the input file matches the Numident file. The input record is run through a series of matches until it is flagged. (Each subsequent match implies that the input record was a nonmatch at the previous step. For example, a record that does not match the Numident name, goes to the match for the Numident alternate name. If it matches the Numident alternate name, it is assigned a 2 flag and does not go to next match.)

  - The parsed name on the input record is matched to the Numident as follows:

    - Matched to the Numident name–a match on all or part of the name goes to a decision logic table that includes birth year (+-2) and gender. If certain conditions are met, it is assigned a 1 flag

    - Matched to the Numident alternate (previous) names–a match on all or part of the name goes to a decision logic table that includes birth year and gender. If certain conditions are met, it is assigned a 2 flag.

    - The parsed name on the input record is concatenated (that is, all names are joined together with no spaces) and matched to the Numident–a match on all or part of the name goes to a decision logic table that includes birth year. If certain conditions are met, it is assigned a 3 flag.

[3]Source: Administrative Records Staff. "Statistical Administrative Records System, 1999–Social Security Number Verification Programming Specification," U.S. Census Bureau; Planning, Research, and Evaluation Division. October 12, 2000.

    - The parsed name on the input record is concatenated and matched to the Numident alternate names–a match on all or part of the name goes to a decision logic table that includes birth year. If certain conditions are met, it is assigned a 4 flag.

- The standardized name on the input record is matched to the Numident as follows:

    - Matched to the Numident name–a match on all or part of the name goes to a decision logic table that includes birth year and gender. If certain conditions are met, it is assigned a 5 flag

    - Matched to the Numident alternate (previous) names–a match on all or part of the name goes to a decision logic table that includes birth year and gender. If certain conditions are met, it is assigned a 6 flag.

    - The standardized name on the input record is concatenated (that is, all names are joined together with no spaces) and matched to the Numident–a match on all or part of the name goes to a decision logic table that includes birth year. If certain conditions are met, it is assigned a 7 flag.

    - The standardized name on the input record is concatenated and matched to the Numident alternate names–a match on all or part of the name goes to a decision logic table that includes birth year. If certain conditions are met, it is assigned an 8 flag.

- If there is no successful name match, the record is assigned a B flag.