Did Proxy Respondents Cause Age Heaping in the Census 2000?[i]

Kirsten K. West[1], J. Gregory Robinson[2], and Michael Bentley[2]
National Academies of Sciences, Washington, DC, kkwest@nas.edu[1]
U.S. Census Bureau[2]

Keywords: Proxy respondents, age, data quality, Master Trace Sample Data file

1. Introduction

Misstatement of age is a common example of content error in censuses and surveys. Different cultures have different social values attached to age. In the United States, most respondents know their age, and provide date of birth and age correctly. However, there are some who chose not to reveal their true age and therefore, they do not report date of birth, or they misstate their actual age. For example, among the young and the elderly, age might be overstated. Other respondents may understate their age or gravitate towards one favorite age such as 29. If age is obtained from a proxy respondent, the response might be either an approximation or a guess. The next-door neighbor might say that John Doe is about 50, and the true age might be 47 or 52, or something entirely different. When the true age is unknown or misstated, there is a tendency to quote ages in round numbers, such as the nearest even number, or one that ends in 0 or 5. From a demographic perspective, this creates age heaping (U.S. Census Bureau, 1973).

The census data serve many purposes where it is important to have an accurate age distribution. The Population Division's Intercensal Estimates Program, for example, uses the census as the base population for producing national population estimates. Major federal government surveys such as the Current Population Survey and the American Community Survey use the intercensal population estimates as controls. The estimates are derived by taking the population counts by age, sex, race and ethnicity from the preceding decennial census and updating them throughout the next decade with births, deaths, and international migration. If there is age heaping in the census age distribution, the error will be carried forward for a decade, and every year there will be an overstatement of some cohorts and an understatement of others. Age heaping on ages ending in 0 and 5 in 2000 would affect ages with end digits of 1 and 6 in 2001, of 2 and 7 in 2002 and so on. It is therefore important to examine where in the age distribution and in what population subgroups the misreporting occurs, and with what magnitude. If the problem is found to be severe or concentrated disproportionately in certain population subgroups, it may be necessary to undertake a systematic correction of the base population file. .

It is difficult to circumvent the tendency for respondents who do not provide a birth date to concentrate their answers on a preferred age or for enumerators or other proxy respondents to pick an age such as 25 or 50 when "guessing" the ages of non-respondents. Age heaping may always be present in the data, but it is possible to control and minimize the error through edit and quality assurance procedures. To address solutions to fix the problem, we need to know what is causing it. We suspect that the heaping on ages with digits 0 and 5 stems primarily from proxy reporting, but we do not know if this suspicion is empirically valid.

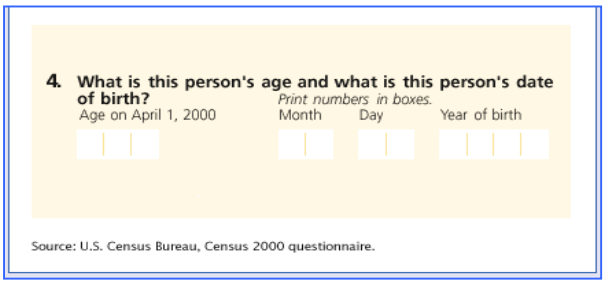
In this study, we use an extract from the Master Trace Sample Database (MTSD) to identify age heaping on enumerator completed returns in Census 2000 by respondent type and close-out status. We examine the occurrences by place of processing and type of data collection area. We examine the data before and after edits and imputation procedures have been implemented. If we find proxy respondents to provide data of an unacceptable quality, if the occurrences differ by data collection area or place of data processing, or the outcomes look different before and after edit and imputation, we can take steps to ensure that this particular issue does not affect the quality of the next census or the surveys controls.

2. Background

Several research and evaluation efforts have already been undertaken to examine the census age distribution. The age question itself was enhanced in Census 2000 compared to the 1990 census (Figure 1). In Census 2000, age data were collected from a question which asked for age on April 1, 2000 (the 1990 census asked for age at last birthday), and complete date of birth, i.e., month, day, and year (the 1990 question asked for only year of birth). It appears that these enhancements to the census form improved the overall quality of the Census 2000 age data compared to the 1990 census age data, especially for the population less than one year old at the time of the

census (West, 2003).

Figure 1. Excerpt from Census 2000 Questionnaire



Similarly, evaluations of Census 2000 operations judged the edit and imputation procedures involving age to have performed well (U.S. Census Bureau, 2004a and 2004b). The procedures do not appear to have introduced systematic bias in the data. In Census 2000, respondents also seem to have been closer to the April 1 reference date (Census Day) when responding about their age than in the 1990 census (Carter, 2002). Finally, a census brief examined the Census 2000 data by age, and found no evidence of age anomalies (U.S. Bureau of the Census, 2001).

Age heaping was the specific focus of internal U.S. Census Bureau reports examining the single year of age distribution of the Census 2000 Modified Race Data file (West, 2003; West, 2004). The analyses used traditional demographic indices constructed to quantify digit preference. Based on these indices, there was no indication that 0 or 5 are preferred digits, nor evidence of substantial avoidance of any terminal digit in the universe where the data were completed by the respondents or where date of birth was provided. However, heaping was present for returns (1) completed by enumerators and (2) where the only age information came from the "Age on April 1, 2000" question ('age only'). The analysis was done by sex, race and Hispanic origin and by different levels of geography (nation, state, county).

Word and Robinson (2002) similarly examined census data by type of enumeration method and also concluded that there is no evidence of age heaping when the forms are filled out through self-response. For this universe the age distribution looks reasonable. However, for the Nonresponse Followup (NRFU) universe, the age distribution showed evidence of heaping when age, but no date of birth, was provided on the form. They found a systematic overage in ages divisible by 5, starting at age 25. They estimated that the age distribution for 2-3 percent of the census population might be affected.

Table 1 summarizes their findings. The analysis was restricted to the ages from 23 to 82. This universe was further divided into two groups: 23 to 52 and 53 to 82. The youngest ages (0-22) were excluded because there is no sign of age heaping for these ages. The oldest ages (over 82) were excluded because the age distribution is less stable.

In the mailout/mailback universe (MO/MB), for the age category 23-52 the distribution of males in ages divisible by 5 is 19.3 percent if age is obtained from date of birth and age, and 23.8 percent if it is assigned based on age only (Column 3). The same percentages for females are 20.1 and 23.2 percent (Column 6). In the NRFU universe the distribution for males is 20.6 percent and 35.3 percent (Column 3). For females, the percentages are 20.4 and 35.1 (Column 6). Since 20 percent is the expected result, the finding of 35 percent for ages 0 or 5 is indicative of heaping generic to the population in the NRFU universe.

Similar patterns are observed for the age category 53-82, but the differences are larger in the NRFU universe compared to the MO/MB universe. Gender does not appear to influence the pattern (Column 3 compared to Column 6).

In the present analysis we focus on the age only Nonresponse Followup universe. We compare data obtained from household members with data obtained from proxy respondents. If a household member at a followup address could not be reached, the NRFU enumerators were allowed to obtain the information from a knowledgeable non-household member (proxy). The enumerator had to attempt at least three personal visits and three telephone contacts before resorting to a proxy respondent (U.S. Census Bureau, 2004a).

When 95 percent of the NRFU workload was completed in a crew leader district, final attempt procedures were implemented in that area. During this phase of NRFU, enumerators made one final visit to each remaining NRFU address to obtain a complete interview or, at a minimum, the unit status and the population count for the unit.

Table 1. Total Household Population. Proportion of The Population by Enumeration Method, Response to Age Question, and In Ages Divisible by Five by Age Group and Sex.

	Enumeration Method[1] (%) (1)	Assignment of Age[2] (%) (2)	In ages divisible by 5 (%) (3)	Enumeration Method[1] (%) (4)	Assignment of Age[2] (%) (5)	In ages divisible by 5 (%) (6)
		Ages 23-52				
		Male			Female	
MO/MB	72.0			74.0		
dob/age		99.0	19.3		99.0	20.1
age only		1.0	23.8		1.0	23.2
NRFU	28.0			26.0		
dob/age		92.0	20.6		93.0	20.4
age only		8.0	35.3		7.0	35.1
		Ages 53-82				
		Male			Female	
MO/MB	84.0			84.0		
dob/age		99.0	19.6		99.0	19.7
age only		1.0	24.2		1.0	24.8
NRFU	16.0			16.0		
dob/age		94.0	20.0		95.0	20.0
age only		6.0	39.6		5.0	40.6

[1]Data captured in mailout/mailback (MO/MB) or during Nonresponse Followup (NRFU)
[2]Age assigned from date of birth (dob) and age data or age only data

Intuitively, we expect proxy respondents to provide data of lesser quality (more age heaping) than data collected from a household member. Similarly, we expect data collected in the close-out stage to be of lesser quality (more age heaping) than data collected before close-out.

3. Methodology

3.1 Data Source

Census 2000 data from multiple sources were merged in the Master Trace Sample Database (MTSD) to form a relational database. These sources include the Census 2000 address frame, collection, enumeration, capture, processing, and response and coverage files. A total of approximately 1.5 million Census 2000 housing unit records, from both a simple random sample and an additional block sample, was obtained from this merged universe. The database contains information on all census returns for each housing unit in the sample, including housing unit level and person data. This allows researchers to trace response and operational data such as housing unit person counts or unit status codes through stages of Census 2000 processing (Hill and Machowski, 2003).

For this study, the analysis examines age distributions in data prior to the full editing and allocation process -- the Decennial Response File (DRF2) data. These

results are compared to post-editing distributions using the Hundred Percent Edited File (HCEF) data.[1]

The primary focus was to present a series of tables that show the percentage of people on NRFU returns with only age reported, and the percentage of these with age digits 0 or 5 (e.g., age 30, 35, 40, 45, etc.). For the calculation of standard errors, we controlled for the clustering of people within households. This was done using the jackknife replication variance estimation method with random groups formed from the households. T-statistics were computed for each pair-wise comparison to test for statistical significant differences.

Results are provided at the national, as well as several sub-levels where there is enough sample to support the analysis (i.e., processing office, regional offices, and metro areas such as Los Angeles, Chicago, New York, Atlanta, and Miami).

3.2 Techniques To Measure Age Heaping in the Data

It is difficult to measure digit preference in the age distribution, because a precise distinction cannot be made between errors due to digit preference, other errors and real fluctuations in birth cohort size. However, indices have been developed to capture deviations from assumed rectangular distributions. Software programs such as the SINGAGE developed by the Population Division's International Programs Center (IPC) perform this type of analysis (Arriaga, 1994).

In the indices, the population aged 23 to 62 is in scope. This age interval excludes the youngest and the oldest population groups where errors other than digit preference are prevalent. The program allows the calculation of three indices: Whipple, Myers and Bachi. Here we use the Whipple and the Myers indices.

Whipple's index detects a preference for ages ending in 0, 5 or both. If age reporting is free of preferences, the index is fluctuating slightly around 1. The higher the value of the index, the higher the preference will be for the digits 0 or 5. An index value of 5 indicates that only 0 and 5 are reported. The index is constructed as follows:

$$\frac{\Sigma(P_{25} + P_{30} + P_{35} + P_{40} + P_{45} + P_{50} + P_{55} + P_{60})}{1/5\ \Sigma(P_{23} + P_{24} + P_{25} + \ldots\ldots\ldots + P_{60} + P_{61} + P_{62})}$$

The Myers index shows the excess or deficit of people in ages ending in any of the 10 digits expressed as percentages. The theoretical range of Myers' index is from 0 to 90. The larger the value of the indices, the more preference there is for certain digits. Values close to zero indicate no heaping, and 90 would result if all ages were reported only in ages ending in a single digit, say zero.

The single year age distribution is depicted by the calculation of these indices. It is the assumption that the population is equally distributed among the ages, i.e., in the absence of known shifts in the annual number of births, deaths and immigration, the population size of adjacent ages should be rather similar. The indices are used to assess this assumption.

4. Limitations

The MTSD is not intended to provide official totals or point estimates. Any limitations, anomalies, or errors present in the original census files remain in the MTSD. For general references on the MTSD limitations, see Hill and Machowski, 2003. The limited sample size, given the 1 in 200 MTSD sample plus the NRFU universe plus the population with age only reported, yields low statistical power. Furthermore, multiple comparison procedures were not attempted. This increases the chance of compounded error. Finally, for ease of comparison between the DRF2 data and the HCEF data, all age restrictions are based on the HCEF data.

5. Findings

5.1 Age Heaping before and after Edit and Imputation

First, we assess the age data before and after the full editing and allocation process. The age distribution before editing and allocation for the selected NRFU universe is shown in Figure 2. When respondents provide an age only, no day, month or year of birth, there is clearly heaping on ages ending in 0 or 5. Age 25, age 30 (not 29!), 35, 40, 45 and so on stand out as "favorite" responses. Note also the absence of conspicuous heaping at ages under 23 and over 80.

[1] Due to edit and imputation, there are about 1.7 percent more cases on the edited file than the unedited file.

Figure 2. Age Distribution before Edit and Imputation

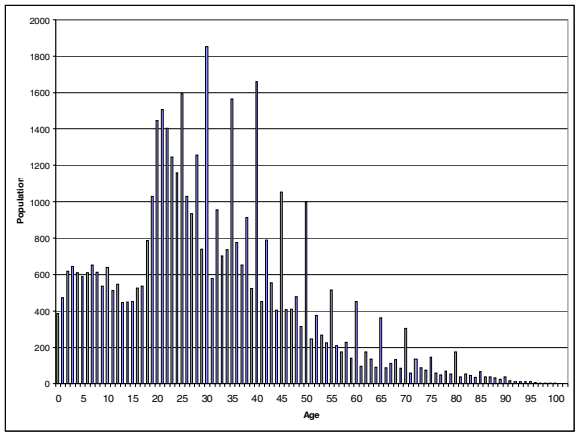


Figure 3. Age Distribution after Edit and Imputation

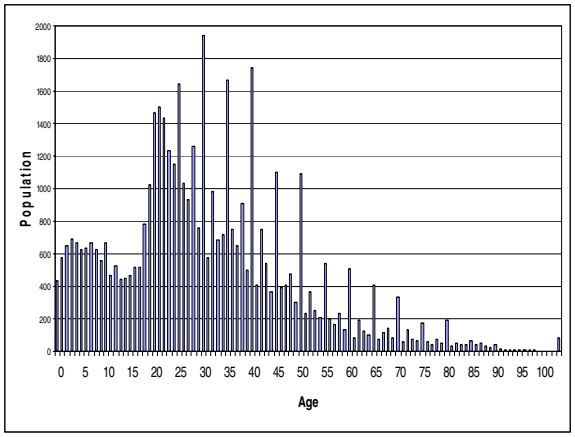


Figure 3 shows the data after editing and allocation. The visual image looks very similar to the one obtained from inspecting Figure 2. Age heaping is evident for the ages ending in 0 or 5. The edits and the imputation process did not eliminate or exacerbate the outcome.

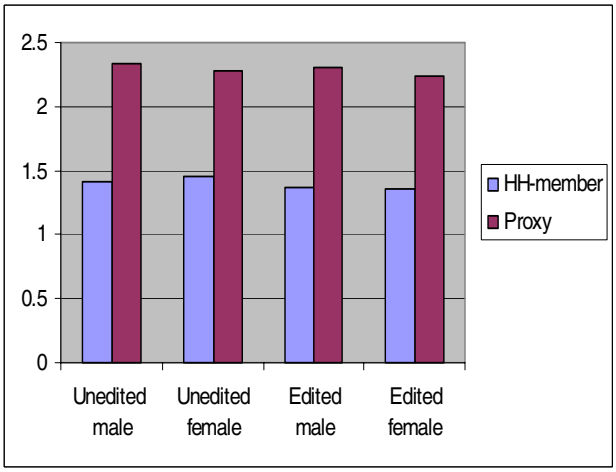
In the MTSD NRFU age only sample, the Whipple index has a value of 1.80 (before edit and imputation) and 1.79 (after edit and imputation). On a scale from 0 to 5, with 0 being no age heaping, these scores confirm the age heaping. For comparison, in the national population (when computed for all respondents in the census), the score was 1.02 after edit and imputation. For the NRFU age only universe, the MTSD sample and the national results agree (West, 2003).

Similarly, using the NRFU age only MTSD sample, the Meyers index scores are 29.48 before edit and imputation and 26.54 after edit and imputation, respectively. For comparison, the score for the same national universe in Census 2000 is 27.9 (West, 2003). (The scale ranges from 0 to 90, with 0 if no heaping and 90 if all responses concentrated in one digit).

5.2 Age Heaping by Type of Respondent

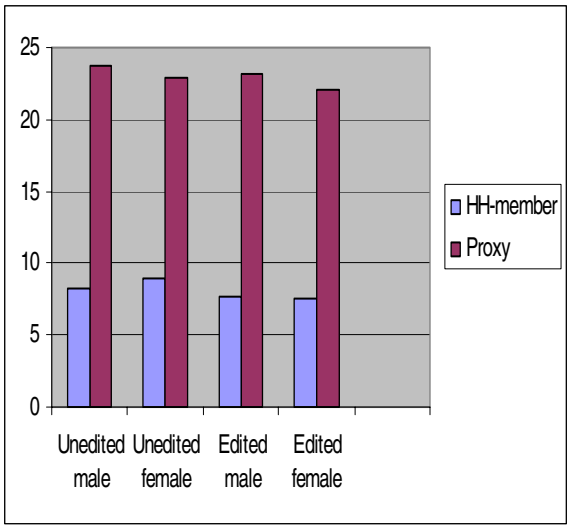
Next, we examine the NRFU 'age only' universe age distribution by type of respondent. It is our hypothesis that proxy respondents are more likely to provide age only responses than householder members (non-proxy-respondents) and that age only responses from proxy respondents are more likely to reflect age heaping than those from household members.

Figure 4. Whipple's Index. Age Heaping by Type of Respondent, Edit Status and Sex



Figures 4 and 5 present the scores on the age heaping indices. Both indices support the finding that proxy respondents are more likely to provide answers that heap on 0 or 5 than household members. Sex of respondents does not matter and edit and imputation procedures do not change this outcome.

Figure 5. Meyers' Index. Age Heaping by Type of Respondent, Edit Status and Sex



5.3 Age Heaping and Close-out Status

We also compared the data collected before and after close-out. Again, we focus here on the age only NRFU universe where the enumerators were successful in collecting a response to the age only on April 1 part of the question to age. It is the hypothesis that there would be more age heaping in the close-out data than in the non-close-out universe.

Our findings show that in both the younger and the older age groups, there is significantly more age only cases in the close-out than in the non-close-out universe – around 11 to 13 percent compared to around 6 to 7 percent. However, the age heaping results differ by age. There are no statistically significant digit preference differences between close-out and non-close-out cases for the younger ages. Statistically significant differences emerge for the older ages. Here, there are around 44 percent in ages divisible by 5 (compared to an expected 20 percent) for the close-out cases. There is also age heaping in the non-close-out cases with around 35 to 36 percent in ages divisible by 5.

The unedited and edited distributions do not show much variation. There is no indication that age heaping disappears in the edit and imputation processes. The Whipple and Myers index scores are consistent with these findings. Close-out cases have more age heaping than non-close-out cases and edit and imputation procedures do not change that picture. The sex of the respondents plays no role in these patterns.

5.4 Age Heaping by Respondent Type for Non-close-out Cases

Next, we examined the data on age only responses and age heaping by respondent type restricted to the non-close-out cases. (We did not have enough cases in the sample to conduct the same type of analysis for the close-out universe). Even within the non-close-out universe, proxy respondents are more likely to provide age only data than household members. On average, 4 percent of the data collected from a household member had age only compared to more than 25 percent when collected from a proxy. Proxy respondents are also more likely to provide an age ending in 0 or 5. The results are very similar to the results for the close-out cases.

The Whipple and Myers index scores confirm the observations. Even in cases that are not close-out, proxy respondents, regardless of their sex, are more likely to respond with an age ending in 0 or 5.

5.5 Age Heaping and Place of Processing

We examined age only responses and age heaping for each of the four Census 2000 data capture centers. This was done to help determine whether any possible differences in the form processing may have contributed to heaping on age digits 0 or 5.

In the pair-wise comparisons, we find that for the age group 0-22, Jeffersonville and Pomona are not statistically different from each other when it comes to amount of cases with assignment of age only data. Edit and imputation procedures do not change this relationship. All other comparisons are statistically significant. At ages 0-22, there is little evidence of age heaping in any processing center (the distributions are close to the expected 20 percent). For the 23-82 age group, Pomona has the highest amount of age only cases, but the largest amount of age heaping is found in Baltimore and Jeffersonville. The amount in the two

offices is not statistically different from each other, though they are different from Phoenix and Pomona.

In general, though certain differences are found to be statistically significant, the patterns do not suggest that errors during the data capture process caused one processing office to produce unusual results compared to the other processing offices. Furthermore, as noted in earlier comparisons, the age 0-22 universe is much closer to the expected 20 percent distribution than the age 23-82 universe.

5.6 Age Heaping and Metropolitan Area of Data Collection

Results were also examined for several metropolitan areas. For ages 0-22, there are no significant differences in the amount of age only data and no indication of heaping in any data. For ages 23-82, there are more age only cases in Los Angeles and Miami than in the other metropolitan areas. Atlanta has the least amount of age heaping and Miami the most. The edit and imputation procedures do not change these relationships.

5.7 Age Heaping and Regional Census Office

Finally, we looked at all 12 regional census offices. The offices vary in the amount of age only data, but for all offices we note that there is not much evidence of age heaping for the younger ages. However, age heaping occurs for ages 23-82. New York has significantly more age heaping than Kansas, Charlotte, Dallas and Denver. Dallas emerges with significantly less heaping than any other office, but still show 28.6 percent in ages divisible by 5 in the edited data, down from 30.1 percent in the unedited data.

Overall, we see a clear pattern of universally higher indices of age heaping at ages 23-82 than ages 0-22, regardless of place of processing, metropolitan areas, or regional census office.

6. Conclusion

In this analysis, we posed the specific research question: Did proxy respondents contribute to the distinct pattern of age heaping found for enumerator completed questionnaires with 'age only' data? The answer is yes.

We used a database from the Census 2000 specifically designed for research purposes: the Master Trace Sample Database (MTSD). The sample allows the tracing of responses through the data capture and processing stages. We selected cases from the Nonresponse Followup universe of 'age only' cases, because the prevalence of age heaping had already been established by other studies for this universe.

We hypothesized that proxy respondents would not only have more 'age only' data than household members. They would also be more likely to provide an age ending in the digits 0 or 5. The data led us to accept both hypotheses. We further noted that age heaping is observed in proxy data collected early as well as in close-out operations, i.e, regardless of stage of data collection, the outcome is likely to be age heaping on digits ending in 0 or 5.

We wanted to see if respondents provide the same answers in different geographic locations. Within the selected universe, we examined data from five major metropolitan areas compared to all other areas. The phenomenon appears to be more of a factor in the Los Angeles and the Miami metropolitan areas than in other metropolitan areas. However, when we focused on the regional office level, we were unable to detect a regional effect. We compared the 12 regional offices and did not find enough of a pattern to suggest differences between the regions. The age heaping in the proxy data was prevalent everywhere.

We could not isolate a data capture or a data processing effect. The edit and imputation did nothing to change this picture. We compared data from the four processing offices. Even though there were differences and some of them statistically significant, one office did not seem worse or better than any other.

Results based on the MTSD are not definitive. However, they provide much more specific information about potential reasons for the age heaping experienced in Census 2000 than we currently have. Age heaping appears to be a respondent issue, especially a proxy respondent issue. Our findings may not be strong enough to recommend new procedures for the 2010 census minimizing the collection of proxy data for response error-prone data items such as date of birth; or to develop expanded editing procedures that "blank" and modify proxy data of poor quality. Yet, this analysis provides a good 'case study' of the potential deleterious data quality effects of proxy-collected information for a specific item (single years of age).

We suggest that checks for age heaping, where possible, become part of census data quality reviews. Especially, we recommend that a check for age heaping be instituted before the census file is adopted as the base file for the intercensal population estimates. Depending on the severity of the problem, decisions

can then be made to smooth the age distribution or to inform data users of age anomalies in the data.

Finally, we note that the major findings of this study were based on analysis made possible by exploiting the rich detail of the Master Sample Database. We hope other researchers will utilize this new data source.

References

Arriaga, Eduardo E. (1994), "Population Analysis With Microcomputers." U.S. Census Bureau.

Carter, Nathan. (2002), "Data of Reference for Age and Birth Date Used by Respondents of Census 2000." Census 2000 Evaluation (H.10). U.S. Department of Commerce. U.S. Census Bureau, Washington, D.C. November 14, 2002.

Hill, Joan M. and Jason D. Machowski. (2003), "Master Trace Sample." Census 200 Evaluation B6. September 29, 2003.

U.S. Bureau of the Census. (2004a), "Data Collection in Census 2000." Census 2000 Topic Report No. 13. Issued March 2004.

----------. (2004b), "Content and Data Quality in Census 2000." Census 2000 Topic Report No. 12. Issued March 2004.

----------. (2001), "Age. 2001." Census Brief by Julie Meyer. Issued October 2001.

----------. (1973), "The Methods and Materials of Demography." Shryock, Henry S., Jacob S. Siegel and Associates. U.S. Government Printing Office, Washington, D.C.

West, Kirsten K. (2003), "Age Heaping in Census 2000." Internal Census Bureau Memorandum for Signe I. Wetrogan, Population Division, November 10, 2003.

West, Kirsten K. (2004), "Evidence of Age Heaping in Census 2000 County Level Data." Internal Census Bureau Memorandum for Signe I. Wetrogan, Population Division, February 17, 2004.

Word, David and Gregg Robinson. (2002), "Specifications for Allocating Age in Census 2000 MARS File." Internal Population Division Memorandum. May 28, 2002.

Acknowledgement

A special thank you to Joan Marie Hill, Jason Machowski, and Jennifer Guarino Tancreto, staff members in the Planning, Research, and Evaluation Division (PRED) for providing the Master Trace Sample Database (MTSD) extract with preliminary analysis results. The MTSD was constructed to help the U.S. Census Bureau and its customers answer 2010 research questions that go beyond those addressed by the Census 2000 evaluations and experiments.

[i] This report is released to informed interested parties of research and to encourage discussion. The views expressed on operational issues are those of the authors and not necessarily those of the Census Bureau.