Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

MISSING DATA RESULTS FOR THE CENSUS 2000 ACCURACY AND COVERAGE EVALUATION [1]

Mary Frances E. Zelenak, David E. McGrath, Nganha T. Nguyen, and Patrick J. Cantwell
Mary Frances E. Zelenak, Bureau of the Census, DSSD, Washington, DC 20233-7600
mary.f.zelenak@census.gov

KEY WORDS: Noninterview Adjustment, Characteristic Imputation, Imputation Cell Estimation, A.C.E.

1. Introduction

The Census Bureau conducted the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) survey to measure the accuracy of Census 2000 and to adjust the person count for coverage errors. To calculate dual system estimates of the population, we resolved three types of missing data in the A.C.E. First, we conducted a noninterview adjustment within noninterview cells to compensate for housing-unit nonresponse. Second, we imputed missing demographic characteristics needed to assign A.C.E. people to estimation post-strata. Third, we assigned a probability to people with unresolved resident, match, or enumeration status.

This paper provides results of the missing data procedures. All data are from the United States, that is, results from Puerto Rico are not included. Due to the extensive efforts of the field staff and the high level of cooperation from the respondents, all rates of missing data were low. We show interview rates and noninterview adjustment factors for Census Day and the day of the A.C.E. interview. We display imputation rates for five characteristics: age, sex, tenure, race, and Hispanic origin by proxy/non-proxy response status and mover status. For unresolved status, we show the proportion of residents, matches, and correct enumerations among resolved cases in each imputation cell, that is, the probability assigned to unresolved people within the cell.

The levels of missing data in the 2000 A.C.E. are similar to those in the 1990 Post Enumeration Survey (PES). Among occupied housing units, the unweighted interview rate was higher in the A.C.E. (97.0 percent for Census Day and 98.9 percent for A.C.E. Interview Day) compared to 98.4 percent for Interview Day in the PES. The characteristic imputation rates were slightly higher for age and sex, and slightly lower for tenure and race in 2000. For unresolved status, only 1.2 percent (weighted) of the P Sample had unresolved match status, compared to 1.8 percent in the 1990 PES. About 2.6 percent (weighted) of the E Sample had unresolved enumeration status in the A.C.E. while 2.3 percent were unresolved in the PES.

In Section 2, we give a brief summary of the A.C.E. procedures. Section 3 covers the household-level noninterviews in the P Sample. In Section 4, we discuss missing demographic characteristics (age, sex, tenure, race, and Hispanic origin) used to assign people to a post-stratum. Section 5 addresses the unresolved resident, match, and correct enumeration status.

More detailed results are given in Cantwell et al. (2001). For additional details about the missing data procedures, see Ikeda and Cantwell (2001).

2. Three Types of Missing Data in the A.C.E.

The Accuracy and Coverage Evaluation (A.C.E.) used dual system estimation to determine Census population estimates. During A.C.E. operations, the Census Bureau obtained a roster of the A.C.E. sample blocks independently of the Census. We then interviewed the people in these blocks asking who lived there at the time of the interview, and who was a resident there on Census Day. We gathered information to identify people who had moved in or out of the residence since Census Day. Using this information, we then matched the independent roster (P Sample) to the list of census enumerations (E Sample).

We used the results of the matching to estimate the number of people missed in the census. From the E Sample, we estimated the proportion of census enumerations that were indeed correct enumerations. We calculated population estimates separately within estimation domains called post-strata and then determined a coverage correction factor within each post-stratum to be applied to all people enumerated in the census within that post-stratum. We calculated adjusted counts for geographic areas by summing the adjusted counts of people in that area and applied an appropriate rounding

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

method to produce integer counts of people at all levels. For more information about estimation methods and results, see Haines and Davis (2001) and Davis (2001), respectively.

For each component of the dual system estimator, certain required data were not collected for some people or housing units in the A.C.E. We encountered three types of missing data in the A.C.E., and used appropriate procedures to correct for them:

- household-level noninterviews were addressed by a noninterview adjustment within small groups, generally, the same type of housing unit in the same neighborhood,
- values for missing demographic items were imputed using hot-deck procedures and selections from frequency distributions, and
- people with unresolved resident, match, or enumeration status were assigned probabilities through imputation cell estimation.

Note that the term "missing data" applies after all follow-up attempts are complete. In the following sections, we summarize the missing data procedures used in the A.C.E. See Ikeda and Cantwell (2001) for further details.

3. Noninterview Adjustment

A small number of occupied housing units in the A.C.E. were not interviewed. In a majority of these, the household could not be contacted or the interview was refused. Two noninterview adjustments were performed on the P Sample--one for Census Day and another for A.C.E. Interview Day; however, there was none performed on the E Sample. Each of the two noninterview adjustments generally spreads the weights of household noninterviews among households that were interviewed in the same noninterview adjustment cell, defined as the block cluster crossed with the type of basic address. For purposes of this adjustment, there were three types of basic address: single-family units, units with multiple residences--such as apartments and condominiums--and all others.

Two rosters were created for each household--one for each day. To accomplish this, A.C.E. interviewers asked questions to determine who currently lives in the household and who lived in the household on Census Day. The Census Day housing-unit status for P-Sample units was used to compute the Census Day noninterview adjustment, which was then applied (at the appropriate level) to the person weights of non-movers and out-movers. Similarly, A.C.E. Interview Day housing-unit status was used to compute the A.C.E. Interview Day noninterview adjustment, which was then applied to the person weights of in-movers. For the definition of mover status and other details, see Ikeda and Cantwell (2001).

A.C.E. noninterview rates were extremely low for both Census Day and A.C.E. Interview Day. Table 1 shows the unweighted A.C.E. household interview status counts and noninterview rates for Census Day and for A.C.E. Interview Day. The noninterview rate is the number of noninterviews divided by the sum of interviews and noninterviews.

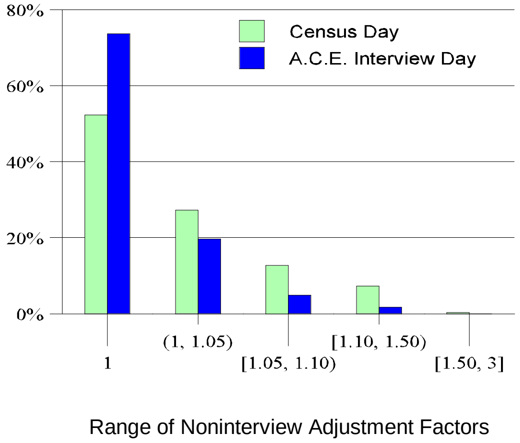
In the A.C.E., we attempted to interview residents at 300,913 addresses located in sampled block clusters. Of the 261,969 housing units occupied on Census Day, 7,794 (3.0 percent) were noninterviews. For Interview Day, 3,052 housing units (1.1 percent of the 267,155 occupied housing units) were noninterviews. The weighted noninterview rates were 2.9 and 1.2 percent, respectively. It is not surprising that the noninterview rates for A.C.E. Interview Day were lower than those for Census Day, as interviewers more often speak with the current housing-unit residents for A.C.E. Interview Day; therefore, we expect a better chance of obtaining an interview.

Table 1. Status of A.C.E. Household Interviews for Census Day and for A.C.E. Interview Day (Unweighted)

	Census Day		A.C.E. Interview Day	
	Number	Percent	Number	Percent
Total Housing Units	300,913	100.0 %	300,913	100.0 %
Interviews	254,175	84.5 %	264,103	87.8 %
Noninterviews	7,794	2.6 %	3,052	1.0 %
Vacants	28,472	9.5 %	29,662	9.9 %
Deletes	10,472	3.5 %	4,096	1.4 %
Noninterview rate	3.0 %	N/A	1.1 %	N/A

Due to the high response in the A.C.E., most noninterview adjustment factors were close to 1. Chart 1 shows the distribution of noninterview adjustment factors for Census Day and A.C.E. Interview Day. Of the 254,175 interviewed housing units on Census Day, 52.3 percent had a noninterview adjustment factor of 1, indicating that all housing units in the initial noninterview cell (usually, block cluster by type of basic address) were interviewed. For A.C.E. Interview Day, 73.6 percent of the 264,103 interviewed housing units had a noninterview adjustment factor of 1. Because of the higher response rate for A.C.E. Interview Day, fewer housing units had noninterview adjustment factors greater than 1.10 for A.C.E. Interview Day (1.8 percent) than for Census Day (7.7 percent).

Chart 1. Distribution of Noninterview Adjustment Factors for Census Day and A.C.E. Interview Day



4. Characteristic Imputation

Some people in the P and E Samples were missing one or more of the following characteristics on the A.C.E. questionnaire or in the census: age, sex, tenure, race, or Hispanic origin. When missing, each of these items had to be imputed so that the person could be assigned to a post-stratum for dual system estimation. Characteristic imputation is generally not carried out for other missing variables (with the exception of the unresolved status items discussed later) as they are not needed to determine post-stratification. The imputation methods for the P Sample and the E Sample differ, as each has different sources of data available to use for imputation.

Before imputation began, age and sex distributions were calculated nationally using the P-Sample data. Missing age or sex was then drawn from the appropriate conditional distribution. Tenure was imputed using a hot-deck procedure. Race and Hispanic origin were imputed using a hot-deck procedure combined with random selection from within the household. For these characteristics in the E Sample of the 2000 A.C.E., we matched the E-Sample person record to its counterpart on the edited file for the entire 2000 Census, and extracted the characteristic. Thus, the E-Sample imputation rates (shown in Table 2) derive from the census enumeration, which was done mostly by mailout/mailback.

Table 2 shows that characteristic imputation rates for the P Sample were very low for all five characteristics, ranging from 1.4 to 2.4 percent (weighted). The unweighted rates were very similar. The distributions for the five imputed characteristics after imputation remained about the same as before imputation. Other breakouts by proxy status and mover status are also shown.

Due to a processing error, some data reported by respondents from a small subset of the P Sample were not properly stored. This affected the variables tenure, sex, and Hispanic origin. These data were not lost. We recovered the data for people whose resident status code indicated they were in-movers or "removed from the P Sample."

Because of the timing, we did not recover the data for some non-movers, out-movers, and people with unresolved mover status. For these people, their tenure, sex, and Hispanic origin are "missing" on the P-Sample input files, even though they reported the data; we imputed the three characteristics as if the data were never reported. The consequence is that the imputation rates for tenure, sex, and Hispanic origin in the P Sample are greater than one would determine from the interviews. That is, these rates reflect two components: 1) the actual level of respondents' failure to report the information, and 2) the contribution of the processing error.

Table 2. Percent of Characteristic Imputation in the P Sample (also by Proxy and Mover Status) and in the E Sample (Weighted)

	Percent of people with imputed characteristic				
	Age	Sex	Tenure	Race	Hispanic origin
P Sample	2.4 %	1.7 %	1.9 %	1.4 %	2.3 %
Proxy status					
Non-proxy	2.1 %	1.5 %	1.7 %	1.0 %	1.8 %
Proxy	7.9 %	4.2 %	5.2 %	8.7 %	11.0 %
Mover status					
Non-mover	2.3 %	1.7 %	1.9 %	1.2 %	2.1 %
In-mover	2.3 %	0.4 %	0.4 %	1.3 %	0.8 %
Out-mover	6.0 %	3.4 %	2.4 %	8.0 %	9.0 %
E Sample	2.9 %	0.2 %	3.6 %	3.2 %	3.4 %

5. Imputation of Resident, Match, and Correct Enumeration Status

After all follow-up activities were completed, there remained a small fraction of A.C.E. sample people for whom we still did not have enough information to compute the components of the dual system estimator. For some respondents in the P Sample, we were unable to determine their resident status (whether or not the person was living in the block cluster or the associated extended search area on Census Day) or their match status (whether or not the person matched to someone enumerated in the census in the same block cluster or the extended search area). Determining resident status is important for P-Sample people because Census Day residents are used to estimate the number of matches in the P Sample. Similarly, for some people in the E Sample, there was not enough information to determine whether the person was correctly enumerated in the Census.

Such cases where status cannot be determined are said to be "unresolved." Table 3 displays the distribution of the statuses before imputation. The weighted rates of unresolved resident status (2.2 percent), match status (1.2 percent), and enumeration status (2.6 percent) were low. The analogous unweighted rates were 2.3 percent, 1.2 percent, and 3.0 percent, respectively. We used imputation cell estimation to assign probabilities for P-Sample people with unresolved match or Census-Day resident status, and for E-Sample people with unresolved enumeration status.

All P-Sample and E-Sample people--resolved and unresolved--were separated into groups called imputation cells based on operational and demographic characteristics. We used different variables to define cells for P- and E-Sample people, and, among P-Sample people, to define cells for resolving resident and match status. Within each imputation cell the weighted proportion of residents (or matches or correct enumerations) among the cases with resolved status was calculated, and that value was imputed for all unresolved people in the cell. See Tables 4, 5, and 6 for those values.

For Tables 3 and 4, the *resident rates* were determined by dividing the number of confirmed residents by the number of resolved cases--all confirmed residents and nonresidents. When calculating these rates, we only included people with mover status of non-mover and out-mover. By definition, non-movers and out-movers should both be Census Day residents; however, we create the mover-status variable prior to field follow-up work. This work may reveal that a non-mover or out-mover was not actually a Census Day resident. For example, a person may report he or she lived in the housing unit since March 20. Preliminary operations would label this person a non-mover. However, follow-up operations may confirm this person moved into the housing unit on April 20. Therefore, this person becomes a confirmed nonresident for Census Day.

In the same way, for *match rates* in Tables 3 and 5, we considered only Census Day confirmed residents and people with unresolved resident status. That is, we excluded confirmed nonresidents while calculating match probabilities. The *enumeration rates* in Tables 3 and 6 were also computed in an analogous manner.

Tables 4, 5, and 6 show the imputation cells used for cases with unresolved resident, match, and enumeration status, respectively, as indicated in Ikeda and Cantwell (2001). We created these cells with combinations of operational and demographic variables, such as match status, tenure, race, ethnicity, mover status, and the number of variables that were imputed. All imputation cells (36 for resident status, 7 for match status, and 28 for enumeration status) contained a sufficient number of people with resolved status, thus allowing accurate estimates of the probabilities used for imputation.

Table 3. Final Resident and Match Status for the P Sample (also by Mover Status) and Enumeration Status for the E-Sample (Weighted)

	Final status			Average rate for resolved cases
	Yes	No	Unresolved	
Resident ?	96.1 %	1.7 %	2.2 %	98.2 %
Mover status				
Non-mover	96.9 %	1.5 %	1.6 %	98.5 %
Out-mover	75.7 %	7.1 %	17.2 %	91.5 %
Match ?	90.6 %	8.2 %	1.2 %	91.7 %
Mover status				
Non-mover	91.4 %	7.7 %	0.9 %	92.2 %
Out-mover	68.0 %	21.7 %	10.3 %	75.8 %
Correct Enumeration ?	93.4 %	4.0 %	2.6 %	95.9 %

For P-Sample resident status and E-Sample enumeration status, we divided match code group 3, (partial household nonmatches needing follow-up) into two parts: 3a includes those in group 3 who are 18-29 years of age and are children of the reference person; 3b includes all other people in group 3. This variable tried to isolate people, many of whom were college students or military personnel, who should have been enumerated in a college dorm or other group quarters.

Each cell in these tables contains the weighted rate for resolved cases (the actual probability assigned to unresolved people in the cell). For the frequencies of resolved and unresolved cases in each cell, see Cantwell et al. (2001).

5.1 Resident Status

The weighted resident rate over all confirmed people in the P Sample was 98.2 percent, while the average weighted rate assigned to people with *unresolved* status was 77.4 percent. This difference follows because the frequencies of unresolved cases within cells were not proportional to those of the resolved cases. In fact, there were many more unresolved cases in those cells with lower probabilities (e.g., match code group 8).

Note that the probability imputed for unresolved people in group 7 (insufficient information for matching) is the weighted average over groups 1 – 5 and 8. In general, people in group 7 were not followed up in the field; thus there were no resolved cases. We note that groups 7 and 8 had the largest number (unweighted) of people with unresolved resident status (7,510 and 2,324, respectively).

The match code groups discriminated well among people with resolved resident status. The resident rates were significantly lower for potentially fictitious people and people said to be living elsewhere on Census Day (group 8), 13.9 percent. For most of the imputation cells, neither the race/ethnicity variable nor the tenure (owner/non-owner) variable discriminated very well. An exception was for the group 3a, where the probabilities assigned ranged from 75.5 percent to 92.8 percent in the four cells. This group discriminated well for resident status regardless of race/ethnicity status or tenure. Only 83.6 percent of these people were residents of the housing unit on Census Day, compared with 96.4 percent of other people (group 3b) in partially-matched households.

Table 4. Imputation Cells and Probabilities Assigned to P-Sample People with Unresolved Resident Status (Weighted)

Match code group	Owner		Non-owner	
	Non-Hisp. White only	Others	Non-Hisp. White only	Others
1. Matches needing follow-up	98.2 %	98.6 %	99.3 %	99.1 %
2. Possible matches	97.3 %	96.8 %	96.6 %	97.2 %
3a. Partial household nonmatches needing follow-up, age 18-29 and child of reference person	75.5 %	90.1 %	88.3 %	92.8 %
3b. Partial household nonmatches needing follow-up, others not in 3a	95.6 %	97.1 %	95.9 %	96.9 %
4. Whole household nonmatches needing follow-up, not conflicting households	92.0 %	94.3 %	91.1 %	91.4 %
5. Nonmatches from conflicting households	91.0 %	92.7 %	94.5 %	95.4 %
6. Resolved before follow-up	99.3 %	99.0 %	99.0 %	98.8 %
7. Insufficient information for matching	81.3 %	86.7 %	84.4 %	87.2 %
8. Potentially fictitious or said to be living elsewhere on Census Day	11.9 %	12.3 %	17.7 %	15.7 %

5.2 Match Status

The main problem with assigning probabilities to people with unresolved match status is that we have very little information about these cases. In fact, for only 1.7 percent of the 7,826 unresolved matches did we have sufficient information for matching (a valid name and certain other characteristics). Further, only 3.8 percent of these cases were sent to follow-up to gather more information. Therefore, we were essentially restricted to using the few characteristics that were known for each unresolved case.

To create the imputation cells for match status, we used a) mover status (non-mover or out-mover), b) the number of imputed characteristics among age, sex, tenure, race, and Hispanic origin, and c) whether or not the person's housing-unit address had matched to the census address list in early 2000. We combined this information to form seven cells as depicted in Table 5.

The average weighted match rate among all people with resolved status was 91.7 percent, while the average weighted match rate assigned to people with *unresolved* status was 84.3 percent. The majority of people with unresolved match status were non-movers from housing unit matches. The match rates vary greatly by imputation cell. Non-movers (92.2 percent) had much higher match rates than out-movers (75.8 percent). The housing-unit address match code discriminated well for match status. A person's chance of being a match was much lower in a housing unit that failed to match. People who supplied incomplete data, indicated by having one or more imputed characteristics, tended to have lower match rates than those who provided complete data.

Table 5. Imputation Cells and Probabilities Assigned to P-Sample People with Unresolved Match Status (Weighted)

	Housing-unit address match code			
	Housing unit match		Housing unit nonmatch or conflicting household	
Mover status	No imputes	1 or more imputed characteristics	No imputes	1 or more imputed characteristics
Non-mover	94.5 %	90.1 %	69.0 %	56.7 %
Out-mover	79.8 %	79.1 %	51.6 %	

5.3 Enumeration Status

The average weighted rate of correct enumerations among all resolved people was 95.9 percent. The average weighted rate assigned to *unresolved* people was 76.2 percent. One observes that the match code groups for resolving enumeration status in Table 6 are similar to those used for resolving resident status (Table 4), but slightly more detailed. Groups 4 and 7 had the greatest number (unweighted) of people with unresolved enumeration status (4,813 and 3,881, respectively). For people resolved before follow-up, almost all (99.2 percent weighted) were correct enumerations. Among resolved people, about 6.4 percent of those determined to be potentially fictitious during follow-up (group 11) were correct enumerations while about 22.5 percent of those said to be living elsewhere on Census Day (group 12) were correct enumerations.

Table 6. Imputation Cells and Probabilities Assigned to E-Sample People with Unresolved Enumeration Status (Weighted)

Match code group	No imputes		1 or more imputed characteristics
1. Matches needing follow-up	97.7 %		97.7 %
2. Possible matches	96.8 %		96.8 %
3a. Partial household nonmatches, age 18-29 and child of reference person	87.1 %		90.8 %
3b. Partial household nonmatches, others not in 3a	97.4 %		96.0 %
4. Whole household nonmatches where the housing unit matched; not conflicting households	Non-Hisp. White 96.5 %	Others 97.4 %	95.8 %
5. Nonmatches from conflicting households; housing unit *not* in regular nonresponse follow-up	97.5 %		96.5 %
6. Nonmatches from conflicting households; housing unit *in* regular nonresponse follow-up	91.4 %		92.6 %
7. Whole household nonmatches, where the housing unit did not match in housing unit matching	Non-Hisp. White 95.9 %	Others 94.7 %	95.0 %
8. Resolved before follow-up	Non-Hisp. White 99.5 %	Others 99.0 %	97.9 %
9. Insufficient information for matching	0.0 % [2]		
10. Targeted extended search people [3]	92.8 %		85.8 %
11. Potentially fictitious	5.8 %		8.8 %
12. Said to be living elsewhere on Census Day	22.9 %		21.0 %

Within partially matched households (group 3), separating out the persons aged 18 to 29 who were children of the reference person discriminated well for enumeration status. Of these people, only 87.6 percent (weighted) were correct enumerations, compared to 97.2 percent for all others in partially-matched households. Other than groups 3a, 11, and 12, the match code groups did not discriminate as well as they did when assigning resident probabilities. The other two variables used to form imputation cells--the number of variables imputed and race/ethnicity--discriminated only minimally with respect to enumeration probabilities.

References

Cantwell, P., McGrath, D., Nguyen, N., and Zelenak, M.F. (2001). “Accuracy and Coverage Evaluation: Missing Data Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-7*.

Childers, D. (2000). “The Design of the Census 2000 Accuracy and Coverage Evaluation,” DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1.

Davis, P. (2001). “Accuracy and Coverage Evaluation: Dual System Estimation Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-9*.

Haines, D. and Davis, P. (2001). “Coverage Measurement Results from the Census 2000 Accuracy and Coverage Evaluation Survey,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, in press.

Ikeda, M. and Cantwell, P. (2001). “Missing Data Procedures for the Census 2000 Accuracy and Coverage Evaluation Sample,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, in press.

[2] By definition, all unresolved enumerations found in group 9 are assigned a correct enumeration probability of 0.

[3] For information about the Targeted Extended Search operation and this cell, see Ikeda and Cantwell (2001).