ACCURACY AND COVERAGE EVALUATION REVISION II MISSING DATA METHODOLOGY AND RESULTS

Michael Beaghen, Robert D. Sands
U.S. Census Bureau
4700 Silver Hill Rd., Washington, DC 20233-9001

1 Background

This paper presents an overview of the missing data methodology and results for the Accuracy and Coverage Evaluation (A.C.E.) Revision II[1]. The A.C.E. was the Census Bureau's program for measuring coverage in the Census 2000 (US Census Bureau 2001). The A.C.E. Revision II revised the A.C.E. and represents the Census Bureau's best estimates of coverage in the Census 2000 (US Census Bureau 2003).

The A.C.E. comprised two samples, a population (P) sample to measure census omissions and an enumeration (E) sample to measure census erroneous enumerations. The P sample was obtained by independently listing and conducting a person interview in a sample of block clusters. The E sample consisted of the census enumerations in those sample blocks. The P sample was matched to the census listings. P-sample people not matching the census listing were identified as census omissions; certain P-sample non-matches were sent to an additional interview, the A.C.E. Person Followup (PFU), to confirm their Census Day residency status. E-sample enumerations that had been matched to P-sample people were counted as correct enumerations. Non-matched E-sample people were followed up in the A.C.E. PFU interview to determine whether they were correctly enumerated.

The Census Bureau's evaluation program determined that the A.C.E. was flawed because it failed to detect many erroneous enumerations due to duplication, that is, people enumerated twice (Kostanich 2002, Chapter 1). The A.C.E. Revision II corrected this flaw in two ways: first, it incorporated into the coding of correct enumeration status and Census Day residency status the results of an evaluation followup (EFU) interview. The EFU was conducted in a one in five sample of A.C.E. block clusters, with some subsampling within block clusters. Note that the A.C.E. Revision II subsample of the A.C.E. is referred to as the Revision II sample, and the new coding operation in this sample as the Revision II coding. Second, it conducted an automated search for duplicated people from the A.C.E. E sample and P sample to census enumerations throughout the nation. The A.C.E. Revision II missing data program handled mostly coding issues in this A.C.E. Revision II coding. It did not attempt to solve any missing data problems encountered in the automated search for duplicates.

Missing data arises because we do not obtain interviews for all sample cases or obtain answers to all interview questions. To put the A.C.E. Revision II missing data methods in perspective it is worth reviewing briefly what missing data the A.C.E. adjusted for. For the A.C.E. P sample we made the following adjustments;

- we applied household non-interview adjustments to compensate for non-interviewed households;
- we imputed missing demographic characteristics such as age, race, Hispanic origin, sex and owner/renter;
- where we were unable to assign a definitive Census Day residency status or match status, we assigned probabilities of match and probabilities of residency.

For the A.C.E. E sample there was no non-interview adjustment, nor was there an imputation for missing characteristics as the census imputations were used. However, for E-sample cases with unresolved enumeration status we assigned probabilities of correct enumeration. See Ikeda & McGrath (2001) for details on the A.C.E. missing data methodology.

The Revision II P sample used the same imputations for missing characteristics that the A.C.E. did with the one exception of age imputation. It was necessary to impute age again because the A.C.E. Revision II post-strata had different age groupings. However, the Revision II measurement methodology differed from the A.C.E. measurement methods in ways that required developing new missing data methods. In particular, the Revision II recoding used information from an additional interview, the EFU. Thus the A.C.E. Revision II confronted three general types of missing data problems:

[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.

- new non-interviewed households: Revision II P-sample households that were considered interviews in the A.C.E. were identified as non-interviews in the Revision II coding when it was determined that none of the P-sample people there were valid Census Day residents according to the EFU;
- Revision II E-sample and P-sample cases could now have unresolved match, enumeration, or residency status because of incomplete or ambiguous interview data from either the PFU or the EFU, or both;
- Revision II E-sample or P-sample cases could have conflicting enumeration or residency status because contradictory information was collected in the PFU and the EFU interviews and it could not be determined which was valid.

2 Age Imputation for the A.C.E. P Sample

For the A.C.E., P-sample people with missing age were assigned to age categories as defined by the post-stratification plan. The A.C.E. Revision II P-sample post-stratification subdivided the A.C.E. post-stratification group of 0-17 years old into groups of 0-9 and 10-17 years old. Those people with missing age who had been assigned to the 0-17 group were reassigned to either the 0-9 or the 10-17 group. This reassignment assumed that the age distribution of people missing age was uniform within the 0-17 age grouping. Other people with unresolved age remained in the age group they had been originally assigned to. Note that the new age categories were applied to both the A.C.E. Revision II P-sample and the original A.C.E. sample.

3 The Household Non-Interview Adjustment for the Revision II P Sample

The A.C.E. household non-interview adjustment generally spread the weights of P-sample non-interviewed housing units over interviewed housing units in the same block cluster with the same housing unit structure type. Housing units were determined to be non-interviews in two ways; first, an interview was not conducted during the A.C.E. person interview; second, based on the results of the A.C.E. PFU it was determined that a whole household of P-sample people should not have been listed in the first place, and that another household may have been residents at that housing unit. Separate household non-interview adjustments were implemented for Census Day and A.C.E. Interview Day.

The A.C.E. Revision II non-interview adjustment methodology for A.C.E. Interview Day was essentially unchanged from that of the A.C.E. There was, however, an important change from the A.C.E. methodology for the non-interview adjustment for Census Day residency. In A.C.E. Revision II we defined a cell for new non-interviews due to whole households of A.C.E. non-movers who were determined to be inmovers or non-resident outmovers by the Revision II coding. The new non-interview cell spread the weights of these non-interviewed units over housing units with at least one person who indicated he/she lived at another address or who was identified as potentially fictitious in the A.C.E. We assumed that people in these new non-interviews would have both a low match rate and a low residency rate similar to this cell. Otherwise the non-interview adjustment for Census Day used methodology similar to that of the A.C.E.

In the A.C.E. Revision II there were 3,264,389 weighted non-interviewed housing units. Of these, 127,279 were new non-interviewed housing units assigned to the new non-interview cell. See Ikeda (2002) for a more detailed account of the results of the A.C.E. Revision II non-interview adjustment.

4 Revision II E-Sample and P-Sample Assignment of Probabilities of Correct Enumeration, Census Day Residency and Match Status

In the A.C.E., P-sample people with unresolved Census Day residency or match status came about in one of two ways. First, the A.C.E. person interview may not have provided sufficient information for match and followup. Second, the A.C.E. PFU may not have collected adequate information to allow us to determine a person's Census Day residency status or their match status. Analogously, for A.C.E. E-sample people the PFU may not have collected adequate information to allow us to determine a person's enumeration status. The A.C.E. Revision II also encountered these types of unresolved cases. However, new unresolved cases arose because of the EFU.

In the A.C.E. Revision II, as in the A.C.E., we used imputation cell estimation to assign probabilities for P-sample people with unresolved match or Census Day residency status, and for E-sample people with unresolved enumeration status. Unresolved P-sample and E-sample people were separated into groups called imputation cells based on operational and demographic characteristics. Similarly, resolved people with those same operational and demographic characteristics were

associated with each imputation cell. For each imputation cell the weighted proportion of matches (or residents or correct enumerations) among the cases with resolved status was calculated, and that value was imputed for all unresolved people in the cell. This method is illustrated in *Section 4.2*.

4.1 Imputation for Revision II P-Sample People with Insufficient Information for Match and Followup

The Revision II P-sample people with insufficient information for match and followup tended to be the same people who had insufficient information for match and followup in the A.C.E., except for some rare cases with coding changes. Note that people who had insufficient information in the A.C.E. were neither sent to PFU nor to EFU. There were about three million weighted people with insufficient information for match and followup in both the A.C.E. and the Revision II samples.

In the A.C.E., P-sample people with insufficient information for match and followup were assigned a probability of Census Day residency equal to the residency rate of P-sample people who went to PFU. In the A.C.E. Revision II we improved upon this by defining finer imputation cells that took into account whether or not the housing unit was matched, non-matched, or had a conflicting household. A conflicting household was said to exist when the P-sample household had no people in common with the E-sample household.

The probability of match assigned was the overall match rate, divided into cells based on mover status and housing unit match status as was done in the A.C.E., and additionally on conflicting household status. **Table 1** summarizes the overall results for P-sample people with insufficient information for match and followup. Note that there were a small number of possibly matched people. These had unresolved match status and residency status and were treated in the same framework as the insufficient information people and

Table 1 Census Day Residency Rates and Match Probabilities for A.C.E. People with Insufficient Information for Match and Followup

Weighted Recipients	Unweighted Recipients	Overall Residency Rate	Overall Match Rate
3,110,487	1,777	0.8004245	0.81126

are also included in the results in **Table 1**. For more details on the results of the imputation for insufficient information for match and followup see Beaghen & Sands (2002a).

4.2 Imputation for P-Sample and E-Sample People with Incomplete or Ambiguous Followup

The residency status for Revision II P-sample people and the correct enumeration status for Revision E-sample people often changed from the A.C.E. to the Revision II coding. These statuses changed because the Revision II coding processed not just the original information from the PFU, but also the new information from the EFU. Thus while the EFU information resolved many cases that were unresolved in the A.C.E. on account of the PFU, EFU cases with incomplete or ambiguous information were a new source of unresolved cases in the Revision II coding.

The original A.C.E. missing data plan based the imputation cells on information obtained before any followup was conducted. An ad hoc fix to the A.C.E. missing data methodology was effected by using information from the person followup (Cantwell & Childers, 2001). Based on the keyed PFU data we created the after-followup cells for 'potential fictitious' and 'lived elsewhere on Census Day'. The new cells used information highly relevant to residency or enumeration status. Further, they showed greater discrimination in assigning probabilities of correct enumeration and residency. In the A.C.E. Revision II we fully exploited the keyed after-followup information. We abandoned the before-followup imputation cells and defined our cells based on after-followup information. This change was the single most important improvement in the A.C.E. Revision II missing data methodology.

To define the after-followup cells we employed the keyed responses to the PFU and EFU questionnaire checkboxes and the 'Why' codes. Why codes took into account both the responses in the questionnaire checkboxes and the handwritten notes (Adams & Krejsa, 2002). Using the keyed results and the Why codes we identified the following:

- unresolved cases with the same history, i.e., the recipient cells;
- the resolved followup cases that shared that history up to the point of being unresolved, i.e., the donor pool.

We defined PFU after-followup cells for those cases that were unresolved as a result of the PFU, and EFU after-followup cells for those cases unresolved on

account of the EFU. It was necessary to define separate cells for the PFU and EFU because their interviews and questionnaires were different. However, the same after-followup cells were employed for the P-sample and E-sample unresolved cases, as the PFU and EFU questions about Census Day residency were the same as the EFU and PFU questions about enumeration status.

It is useful to make a distinction between what we call uninformative and informative unresolved cases:

- uninformative unresolved; the PFU or EFU was a non-interview or an incomplete interview, though there was no evidence of an erroneous enumeration or non-resident.
- informative unresolved; a followup interview was conducted and there was evidence of an erroneous enumeration or non-resident.

Note that when either the PFU or EFU interview was uninformative unresolved, but the other interview was resolved, the Revision II coding chose (i.e., the code was based on) the resolved interview. On the other hand, when the unresolved interview was informative, the Revision II coding could choose the unresolved interview over the resolved one. See Adams & Krejsa (2002) for details of the Revision II coding.

It often happened that both the PFU and the EFU interviews were unresolved. In that case in order to assign a cell for imputation the missing data processing chose the unresolved interview that was more informative. When both interviews had the same level of information we usually chose the EFU over the PFU because we believed the EFU questionnaire questions were more detailed.

At this point we give an example of an after-followup cell. One cell of unresolved E-sample people was defined as people with evidence from the EFU interview that they had moved in after Census Day, or moved out before Census Day, though the EFU interview did not provide the address they moved to or from. We could not determine the enumeration status of these people since we did not know whether the Census Day address was in the A.C.E. cluster. The corresponding donor pool consisted of those resolved people who indicated in the followup that they moved in after Census Day or moved out before Census Day; these were generally people who provided the mover address in the EFU. Note we characterize this cell as informative because the followup provided evidence of an erroneous enumeration.

Table 2 and **Table 3** show the nine EFU after-followup cells. The nine PFU after-followup cells were similar and can be found in Beaghen & Sands (2002b). People who moved in after Census Day or moved out before Census Day were the largest informative after-followup cell. Another important informative after-followup cell consisted of people who, according to the followup, had another residence such as a vacation home, though the followup did not indicate whether the other residence or the sample address was the Census Day residence. The non-interview cells and "didn't answer other residence questions" cell were the larger uninformative cells.

Some of the larger EFU after-followup cells were subdivided by A.C.E. operational variables such as whether or not the household went to PFU, or whether the household was conflicting. The uninformative after-followup cells tended to have imputed probabilities of correct enumeration or residency close to one, typically in the range of 0.9 and higher, whereas the informative after-followup cells had lower probabilities, with several less than 0.3.

Table 2 EFU Informative Cells

The followed up person 'Lived elsewhere' or at an 'other residence', but the address was not given
Followed up person moved in after Census Day or out before Census Day, but Census Day address not given
Respondent indicated the followed up person 'Never lived here' at the sample address, but did not provide the Census Day address
The followed up person had an 'Other residence', but did not indicate whether sample address or the other residence was the Census Day residence
Followed up person moved in or moved out, but no move dates given

Table 4 presents a summary of the imputation for E-sample people with unresolved enumeration status. Recipients refer to the people with unresolved status. There were about the same number of E-sample unresolved cases in the Revision II as in the A.C.E., more than six million, with about half of these representing new unresolved cases. See Cantwell et al (2001) for a summary of A.C.E. missing data results.

In **Table 6** we see the Revision II coding generated 4.2 million P-sample unresolved cases. This was substantially more than the A.C.E. 2.7 million (Cantwell et al 2001). We saw this increase because all the Revision II P-sample except those with insufficient information went to EFU, including two groups of

Table 3 EFU Uninformative Cells

The respondent indicated the followed up person 'Lived here' at the sample residence, but did not answer the other residence question
Noninterview (1); the respondent answered the current residence question, but did not answer the group quarters and other residence question
Noninterview (2); the respondent did not answer the usual residence question, nor the group quarters and other residence questions
Potentially fictitious person, no respondents knew of the followed up person

people who generally did not go to PFU: matched people and whole households of non-matched people. These people were usually assumed in the A.C.E. to be resolved and became unresolved because of EFU information.

For illustrative purposes **Table 5** shows the two largest imputation cells for E-sample unresolved cases. For all of the cells see Beaghen & Sands (2002a). While in **Table 4** we see an overall correct enumeration rate of about 0.74, in **Table 5** we see quite a bit of discrimination, 0.28 to 0.99. This degree of discrimination is a sign of a successful imputation plan.

Table 6 presents a summary of the imputation for P-sample people with unresolved Census Day residency status. Recipients refer to the people with unresolved status.

Table 4 Summary of Imputation for E-sample People with Unresolved Enumeration Status

E-sample Unresolved	Weighted Recipients	Unweighted Recipients	Overall Correct Enumeration Rate
Total	6439382.51	4671	0.7420261
EFU After Followup Cells	5613177.25	3814	0.7511079
PFU After Followup Cells	826205.26	857	0.6803254

Table 5 Two-Largest EFU Imputation Cells for the E-sample Unresolvedly

Cell Description	Weighted Recipients	Weighted Donor Correct Enumerations	Weighted Total of Donor Enumerations	Proportion Correct
Moved in After Census Day or Moved Out before Census Day	1537389	472549	1701178	0.27778
Didn't Answer Other Residence Questions? : Non-Conflicting Household	1966332	212000021	214431817	0.98866

Table 6 Summary of Imputation for P-sample People with Unresolved Census Day Residency Status

P-sample Unresolved	Weighted Recipients	Unweighted Recipients	Overall Residency Rate
Total	4126545.34	1958	0.6859972
EFU After Followup Cells	3541054	1388	0.7060697
PFU After Followup Cells	585491.34	570	0.5645987

5 Imputation for Revision II E-Sample and P-Sample Conflicting Coding Cases

When the A.C.E. person followup (PFU) and the evaluation followup (EFU) interviews had contradictory information and we could not determine which was correct, the Revision II coding assigned the case a code of conflicting (conflicting coding is not to be confused with conflicting households, which was described in *Section 4.3.1*). All cases found to be conflicting in the Revision II automated coding were sent to analysts for clerical review. There were some cases where the interviews appeared to be of equal quality, such as when both respondents were household members or both respondents were of equal caliber proxy. For these conflicting cases, the interviews seemed equally likely to be correct based on the expertise of the analysts. Therefore, probabilities of 0.5 were assigned both for correct enumeration status of Revision II E-sample conflicting cases and for Census Day residency status of Revision II P-sample conflicting cases. It should be noted that the recoding of the Revision II samples resulted in about 100,000 conflicting cases.

6 References

Adams, T. & Krejsa, E. (2002): *'A.C.E. Revision II Measurement Subgroup Documentation,'* DSSD A.C.E. Revision II Memorandum Series #PP-6

Beaghen, M. & Sands, R. (2002a): *'Revised A.C.E.: Results from the Imputation for Unresolved Enumeration, Residency and Match Status,'* DSSD A.C.E. Revision II Memorandum Series #PP-57

Beaghen, M. & Sands, R. (2002b): *'A.C.E. Revision II: Specifications for the Probability of Enumeration Status, Census Day Residency and Match Status,'* DSSD A.C.E. Revision II Memorandum Series #PP-23

Cantwell, P. & Childers, D. (2001): '*Accuracy and Coverage Evaluation Survey: A Change to the Imputation Cells to Address Unresolved Resident and Enumeration Status,'* DSSD Census 2000 Procedures and Operations Memorandum Series Chapter Q-44

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

Ikeda, M., (2001): *Accuracy and Coverage Evaluation Survey: Some Notes Related to Accuracy and Coverage Evaluation Missing Data Procedures*, DSSD Census 2000 Procedures and Operations Memorandum Series Chapter Q-77

Ikeda, M. & McGrath, D. (2001): *'Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures; Revision of Q-25,'* DSSD Census 2000 Procedures and Memorandum Series Chapter Q-62

Ikeda, M. (2002): *'A.C.E. Revision II - Results from the Noninterview Adjustment,'* DSSD A.C.E. Revision II Memorandum Series #PP-56

Kostanich, D. (2002): *'A.C.E. Revision II: Design and Methodology,'* DSSD A.C.E. Revision II Memorandum Series #PP-30

US Census Bureau (2001): *'Census 2000: A.C.E. Methodology, Volume 1,'* http://www.census.gov/dmd/www/DetailSpec.htm

US Census Bureau (2003): 'Technical Assessment of A.C.E. Revision II,' A.C.E. Revision II Results