Chapter 2: Summary of Accuracy and Coverage Evaluation Revision II Methodology

David C. Whitford and Donna Kostanich
U.S. Bureau of the Census

1. Correction of Measurement Error in the A.C.E. Revision II E & P Samples

The original A.C.E. estimates were found to be unacceptable because they failed to detect significant numbers of erroneous census enumerations. There were also suspicions that the A.C.E. may have included residents in its P sample that were actually non-residents. Thus, the major goal in revising the A.C.E. estimates included a correction of these measurement errors. One aspect of these corrections involved correcting a subsample of the A.C.E. data. Another aspect, discussed later, involved correcting measurement errors that could not be detected with the information available in the subsample. (These additional errors were identified via a duplicate study.)

1.1 Background

The steps leading to the corrected A.C.E. results were as follows:

1. The A.C.E estimates produced in March 2001 were based on the full E and P samples, which were probability samples of over 700,000 persons in 11,000 block clusters.

2. The Matching Error Study (MES) and the Evaluation Follow-up (EFU) were two programs that evaluated the March 2001 A.C.E. estimates. These evaluations were conducted in a subsample of 2,259 block clusters selected from the original 11,000 block clusters. A further subsample of persons within these block clusters was done for the EFU evaluation. The probes used for EFU were designed to capture unusual living situations.

3. The PFU/EFU Review occurred next; it was not part of the planned evaluations. It was done in order to resolve major discrepancies in enumeration status between the EFU and PFU results. Thus, the Review E sample was a subsample of the EFU E sample.

4. The A.C.E. Revision II E and P samples were then developed for purposes of producing A.C.E. Revision II estimates. These samples were essentially the same as the evaluation E and P samples for EFU, but the data have undergone a major recoding to correct for measurement error.

5. These data along with other measurement error corrections identified by the duplicate study were used to adjust the full E and P samples to produce A.C.E. Revision II estimates.

1.2 Correcting Measurement Error in A.C.E. Revision II Samples

In general, the original A.C.E. person interview (PI) and person follow-up (PFU), the evaluation follow-up interview (EFU), the matching error study (MES), and the PFU/EFU review results were used to correct for measurement error in the enumeration status, the residence status, the mover status, and the matching status for subsamples of the full A.C.E., called the A.C.E. Revision II samples.

The A.C.E. Revision II samples underwent extensive recoding using all available data indicated above. This recoding included the original interview and matching results, the evaluation interview and matching results, as well as the recoding done for the PFU/EFU review.

The A.C.E. Revision II recoding operation was an extension of the PFU/EFU Review clerical recoding, which was used to examine discrepancies between enumeration status in the original A.C.E. and the Evaluation Follow-up (EFU). Given the information available, the recoding that was done on the 17,500 case Review E sample was considered to have negligible error since these data were reviewed and recoded by expert matchers using rules consistent with census residence rules.

An automated coding algorithm based on specific responses to the PFU and the EFU questionnaires was used to determine an appropriate code for each case. This was done for both the PFU interview and the EFU interview. The automated coding also assigned a "Why" code which described the reason why the particular code was assigned.

A three-step process was followed to assign final codes to each case:

- Validation – Determine for categories of "Why" codes if the automated coding was of high quality based on level of agreement with the Review data.
- Targeting – Target only those "Why" code categories that had codes produced by automated coding that had low levels of agreement with the Review data.
- Clerical Coding – Clerically recode only cases in the targeted "Why" code categories. The clerical recoding took advantage of hand written interviewer comments.

In general, cases did not go to clerical review if both the PFU and EFU automated codes agreed and the mover statuses also agree and the why code category was deemed to be of high enough quality.

After the A.C.E. Revision II recoding operation corrected for enumeration, residence, and mover status, the results of the Matching Error Study (MES) were used to correct for false matches and false nonmatches. Some matching errors were a result of incorrect residence status coding and had been corrected as part of the recoding operation discussed above. To determine the correct match status, each of the possible combinations of match status was reviewed to determine the appropriate match status for each type of case. In general, the MES match status was assigned when there were changes from a match to a nonmatch or changes from a nonmatch to a match. For other situations the match status from the EFU coding was assigned.

2. Adjustment for Missing Data

As with all survey data it is not possible to obtain interviews for all sample cases nor is it possible to obtain answers to all interview questions. For the full A.C.E. E and P Samples, household noninterview adjustments were used to adjust for noninterviewed households and imputation methods were used to adjust for missing characteristics such as age or tenure as well as enumeration, residency and match status. These missing data adjustments for the full A.C.E. E and P Samples were essentially unchanged from those used to produce the March 2001 A.C.E. estimates.

For the A.C.E. Revision II E and P samples, there were three new types of missing data to deal with:

- Non-interviewed households: A.C.E. Revision II P-sample households that were considered interviews in the A.C.E. full E and P samples but were identified as non-interviews in the A.C.E. Revision II coding because it was determined that there were no valid census day residents;
- A.C.E. Revision II E or P sample cases with unresolved match, enumeration, or residency status because of incomplete or ambiguous interview data;
- A.C.E. Revision II E or P sample cases with conflicting enumeration or residency status because contradictory information was collected in the A.C.E. PFU and the EFU interviews and it could not be determined which was valid.

2.1 Household Non-Interview Adjustment for the A.C.E. Revision II P Sample

For the original March 2001 A.C.E. estimates, the household non-interview adjustment generally spread the weights of the full P-sample non-interviewed housing units over interviewed housing units in the same block cluster with the same housing unit structure type.

The methodology for the A.C.E. Revision II P sample household non-interview adjustment for interview day was essentially unchanged from that used for the full P sample. There was, however, an important change for the non-interview adjustment for census day residency. A separate cell was defined for new non-interviews due to whole households of persons determined to be inmovers or nonresident outmovers based on the recoding that was done to correct for measurement error.

2.2 Imputation for A.C.E. Revision II E or P Sample Unresolved Cases

In the full A.C.E. P sample, persons with unresolved census day residency or match status came about in two ways. First, the person interview (PI) may not have provided sufficient information for matching and follow-up. Second, the person follow-up (PFU) may not have collected adequate information to determine a person's census day residency status or their match status. The imputation method differed by how the case came to be unresolved.

The A.C.E. Revision II P sample persons with insufficient information for matching and follow-up tended also to have had insufficient information in the original coding of the full P sample, except for some rare coding changes. These persons with insufficient information were not sent out for an evaluation follow-up interview.

For the A.C.E. Revision II P-sample, the imputation of census day residency was improved upon by defining finer imputation cells that included whether or not the housing unit was matched, not matched, or had a conflicting household. The probability of a match was imputed based on the overall match rate for five groups defined by mover status, housing unit match status as in the original A.C.E., and also on conflicting household status.

For the P and E A.C.E. Revision II sample persons who were unresolved because of ambiguous or incomplete follow-up information, the situation was more complicated because there were two follow-up interviews to consider, the PFU and EFU.

For the full E and P samples, imputation cells were based mostly on information obtained before any follow-up was conducted. For the A.C.E. Revision II E and P samples, imputation cells relied on the after follow-up information. This change was the single most important improvement in the missing data methodology.

2.3 Imputation for A.C.E. Revision II E or P Sample Conflicting Cases

When the A.C.E. PFU and the evaluation follow-up EFU interviews had contradictory information, the case was assigned a code of conflicting. All cases determined to be conflicting based on the automated recoding were sent to analysts for further clerical review. By examining the handwritten notes of interviewers, the analysts could often determine which of the interviews was the better and appropriately assign a code. There were some cases where the interviews appeared to be of equal quality, such as both respondents were household members or both respondents were of equal caliber proxy. For these conflicting cases, the interviews seemed equally valid based on the expertise of the analysts. Therefore, probabilities of 0.5 were imputed for correct enumeration for A.C.E. Revision II E-sample conflicting cases and for census day residency for A.C.E. Revision II P-sample conflicting cases.

3. Further Study of Person Duplication

Evaluations of the March 2001 A.C.E. coverage estimates indicated the A.C.E. failed to detect a large number of erroneous census enumerations. One type of these census erroneous enumerations is duplicate census enumerations: census enumerations included in the census two or more times. The A.C.E. was not specifically designed to detect duplicate census enumerations beyond the A.C.E. search area. However, the expectation was that the A.C.E. would detect that these E-sample enumerations had another residence and that roughly half the time this other place was the usual residence. This did not happen in many cases.

For purposes of A.C.E. Revision II estimates, this study used matching and modeling techniques to identify duplicate links between the full E and P samples to census enumerations including group quarters, reinstated, deleted and E-sample eligible records throughout the entire nation. The matching algorithm used statistical matching to identify linked records. Statistical matching allowed for the matching variables not to be exact on both records being compared. Because linked records may not refer to the same individual even when the characteristics used to match the records were identical, modeling techniques were used to assign a measure of confidence, the duplicate probability, that the two records refer to the same individual.

3.1 Matching Algorithm

The matching algorithm consisted of two stages. The first stage was a national match of persons using statistical matching. Statistical matching links records based on similar characteristics or close agreement of characteristics. Statistical matching allowed two records to link in the presence of missing data and typographical or scanning errors. The second stage of matching was limited to matching persons within households that contained a link from the first stage.

In the first stage national match of persons six characteristics common to both files, called matching variables, were used to link records in the full E and P samples with records in the census. These characteristics were: first name, last name, middle initial, month of birth, day of birth, computed age.

The second stage of matching was limited to matching persons within linked households. The first stage established a link between two housing units. The second stage was a statistical match of all the household members in the sample housing unit to all of the household members in the census housing unit. The second-stage matching variables were the same as the first-stage; however, the matching parameters differed. A key difference is that there was considerably less weight on last name agreement since this was a within-household match.

3.2 Modeling Techniques

The set of linked records consists of both duplicated enumerations and person records with common characteristics. Using two modeling approaches, the probability that the linked records were the same person was estimated. One approach used the results of the statistical matching and relied on the strength of multiple links within the household to indicate person duplication. The second relied on an exact match of the census to itself and the distribution of births, names and population size to indicate if the individual link was a duplicate. These two approaches were combined to yield an estimated duplicate probability for the linked records from the statistical matching of the full E and P samples to the census.

4. The A.C.E. Revision II DSE Formula

The DSE formula using version C for movers with different post-strata for the E & P Samples is:

$$DSE^{C}_{ij} = (Cen'_{ij} - II'_{ij}) \frac{\left[\frac{CE_i}{E_i}\right]}{\left[\frac{M_{nm,j} + \left[\frac{M_{om,j}}{P_{om,j}}\right] P_{im,j}}{P_{nm,j} + P_{im,j}}\right]}$$

The A.C.E. Revision II DSE formula, using version C for movers, separate E & P post-strata, measurement error corrections from the E & P A.C.E. Revision II Samples and Duplicate Study results is:

$$DSE^{C}{}_{ij} = (Cen'_{ij} - II'_{ij}) \frac{\left[\frac{CE_i^{\,ND} f_{1,i'} + \tilde{CE}_i^{D}}{E_i}\right]}{\left[\frac{M_{nm,j}^{ND} f_{2,j'} + \tilde{M}_{nm,j}^{D} + \left[\frac{M_{om,j} f_{3,j'}}{P_{om,j} f_{4,j'}}\right]\left(P_{im,j} f_{5,j'} + g\left(P_{nm,j}^{D} - \tilde{P}_{nm,j}^{D}\right)\right)}{P_{nm,j}^{ND} f_{6,j'} + \tilde{P}_{nm,j}^{D} + P_{im,j} f_{5,j'} + g\left(P_{nm,j}^{D} - \tilde{P}_{nm,j}^{D}\right)}\right]}$$

Recall that the II' term includes the late census adds.

Notation		
Terms	CE	Correct enumerations
	E	E-Sample total
	M	Matches
	P	P-Sample total
	F's	Adjusts for measurement error
	G	Adjusts nonmovers to movers due to duplication
Subscripts	*I,j*	Full E and P post-strata
	i',j'	A.C.E. Revision II E and P post-strata
	nm, om, im	nonmover, outmover, inmover
Superscripts	C	DSE version C for movers
	ND	Not a duplicate to census enumeration outside search area
	D	Duplicate to census enumeration outside search area
	~	Includes probability adjustment for residency given duplication

5. Adjustment for Measurement Error Using the A.C.E. Revision II E & P Samples

The A.C.E. Revision II E and P Samples are subsamples of the full E and P Samples. They are each comprised of over 70,000 sample persons. These A.C.E. Revision II samples have been subjected to an additional field interview and/or rematching operation as part of the original A.C.E. evaluation program. In support of the A.C.E. Revision II program, the A.C.E. Revision II samples have undergone extensive recoding using all available interview data and matching results. Missing data adjustments have also been applied to the A.C.E. Revision II sample data. This recoded data from the A.C.E. Revision II samples were used to correct for measurement error in the original full E and P Samples.

The ratio adjustments that correct for measurement error were based on the P or E A.C.E. Revision II Sample and were a ratio of an estimate using the A.C.E. Revision II coding to the an estimate using the original coding. These adjustments were done by measurement error correction post-strata *i'* or *j'* and are denoted by the *f* 's in the A.C.E. Revision II DSE formula.

The term g adjusts the number of inmovers for those full P-sample nonmovers who are determined to be nonresidents because of duplicate links. Some of these nonresidents are nonresidents because they are inmovers and should be added into the count of inmovers. The term:

$P^{D}_{nm,j} - \tilde{P}^{D}_{nm,j}$ is an estimate of nonresidents among nonmovers with duplicate links.

6. Adjustment for Duplicates using the Duplicate Study

Next we turn our attention towards adjusting for those cases that have a duplicate link to a census enumeration outside the A.C.E. search area. P and E sample cases with duplicate links were assigned a nonzero probability of being a duplicate. P and E sample cases without duplicate links were assigned a probability of zero.

When estimating terms in the A.C.E. Revision II DSE involving nonduplicates, those indicated by a superscript *ND*, it was necessary to include the probability of not being a duplicate in the tallies. This probability of not being a duplicate was included in all of the terms involving the *ND* superscript.

6.1 Adjustments for Duplicates

Although the duplicate study identified E and P sample cases linking to census enumerations outside the A.C.E. search area, this study could not determine which component of the link was the correct one since there were no additional data collected to determine this. On the E sample side, this study does not identify whether the linked E sample case is the correct enumeration. On the P sample side, this study does not identify whether the linked P sample case is a resident on Census day. Thus, it is necessary to estimate two conditional probabilities, which are reflected for the E sample in $\tilde{CE}^{D}_{i}$. In the P sample, these probabilities are reflected in the terms $\tilde{P}^{D}_{nm,j}$ and $\tilde{M}^{D}_{nm,j}$ (summed over the nonmover [nm] cases).

7. Adjustment for Correlation Bias using Demographic Analysis

Next the A.C.E. Revision II DSE estimates are adjusted to correct for correlation bias. Correlation bias exists whenever the probability that an individual is included in the census is not independent of the probability that the individual is included in the A.C.E. This form of bias generally has a downward effect on estimates, because people missed in the census may be more likely to also be missed in the A.C.E. Estimates of correlation bias are calculated using the "two-group model" and sex ratios from Demographic Analysis (DA). The sex ratio is defined as the number of males divided by the number of females. This model assumes no correlation bias for females or for males under 18 years of age; and that Black males have a correlation bias, which is different than the relative correlation bias for Nonblack males. The correlation bias adjustment is also done by three age categories: 18-29, 30-49, and 50 and over. This model further assumes that relative correlation bias is constant over male poststrata within age groups. The Race/Hispanic Origin Domain variable is used to categorize Black and Nonblack.

The DA totals are adjusted to make them comparable with A.C.E. Race/Hispanic Origin Domains. Black Hispanics are subtracted from the DA total for Blacks and added to the DA total for Non-blacks. This is done because the A.C.E. assigns Black Hispanics to the Hispanic domain, not the Black domain. The second adjustment deletes the group quarters (GQ) people from the DA totals using Census 2000 data. The reason for making this adjustment is that the GQ population is not part of the

A.C.E. universe. A final adjustment that could be made would be to remove the Remote Alaska population from the DA totals, since it too is not part of the A.C.E. universe. Since this population is small, the DA sex ratios would not be affected in any meaningful way.

8. Synthetic Estimation

The coverage correction factors for detailed post-strata ij were calculated as:

$$CCF_{ij} = \frac{DSE_{ij}}{Cen_{ij}}$$

where:

$\tilde{\text{R}}\text{e}\ DSE_{ij}$'s are the correlation bias adjusted DSEs for post-strata ij.

Cen_{ij}'s are the census counts for post-strata ij.

A coverage correction factor was assigned to each census person excluding persons in group quarters or in Remote Alaska (effectively these persons have a coverage correction factor of 1.0). Recall that in dealing with duplicate links to group quarters persons, the person in the group quarter was treated as the correct enumeration or that this was their correct residence on census day. A synthetic estimate for any area or population subgroup *b* is given by:

$$\tilde{N}_b = \sum_{ij \in b} Cen_{b,ij}\, CCF_{ij}$$