UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

January 2, 2003

MASTER FILE

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-35

MEMORANDUM FOR Documentation

From: Donna Kostanich DK
Chair, A.C.E. Revision II Planning Group

Subject: A.C.E. Revision II: Summary of Methodology

The attached document was compiled as the Accuracy and Coverage Evaluation Revision II (A.C.E. Revision II) methodology was being developed. Many persons working on various aspects of A.C.E. Revision II contributed significantly to this document. The purpose of this document is to highlight the major features of A.C.E. Revision II. As such, it is not intended to be comprehensive in either scope or detail. For a more in-depth discussion of A.C.E. Revision II see, "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-30. More detail can be found in the specifications and requirements contained in this memorandum series.

Summary of A.C.E. Revision II Methodology

To provide an overview of the key features used to prepare A.C.E. Revision II estimates of Census 2000 coverage.

Contents

I. Correction of Measurement Error in the Revision 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 nonresidents. Thus, the major goal for the A.C.E. Revision II estimates includes a correction of these measurement errors. One aspect of these corrections involves correcting a subsample of the A.C.E. data. Another aspect, discussed later, involves correcting measurement errors that cannot be detected with the information available in the subsample. (These additional errors are identified via a duplicate study.)

Background

To understand this, it is important to be familiar with the various sources of information available. These are summarized in the following chart.

Chart 1. Overview of A.C.E. Revision II Data Sources

Program	Sample	Sample Size	What & When
Decennial Census			Spring 2000
A.C.E.	Full E and P Samples	E & P: About 700,000 persons in 11,000 block clusters	A.C.E. Person Interviewing (PI), Summer 2000 A.C.E. Person Followup (PFU), Fall 2000
Matching Error Study (MES)	Evaluation E and P Samples	E & P: About 170,000 persons in 2,259 block clusters	Rematching Operation, December 2000
Evaluation Followup (EFU)	EFU E and P Samples[1]	E: About 77,000 persons in 2,259 block clusters P: About 61,000 persons in 2,259 block clusters	Evaluation Person Followup (EFU), Winter 2001
PFU/EFU Review	Review E Sample	E: About 17,500 persons in 2,259 block clusters	Recoding Operation, Summer 2001
A.C.E. Revision II	Revision E and P Samples	E: About 77,000 persons in 2,259 block clusters P: About 61,000 persons in 2,259 block clusters	Recoding Operation, Summer 2002

[1] The number of sample cases included in the evaluation followup is less than those selected to be in this sample. Cases were excluded from followup for certain situations such as insufficient information or a duplicate enumeration.

The A.C.E estimates produced in March 2001 were based on the full E and P samples, which are probability samples of over 700,000 persons in 11,000 block clusters. The Matching Error Study (MES) and the Evaluation Followup (EFU) were two programs that had been planned to evaluate 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. The PFU/EFU Review 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 is a subsample of the EFU E sample. The revision E and P samples are referred to as such for purposes of producing A.C.E. Revision II estimates. These samples are 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. 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.

Correcting Measurement Error in Revision Samples

In general, the original A.C.E. person interview (PI) and person followup (PFU), the evaluation followup 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 revision samples.

The revision samples have undergone extensive recoding using all available data indicated above. This includes the original interview and matching results, the evaluation interview and matching results, as well as the recoding done for the PFU/EFU review. The recoding operation for A.C.E. Revision II is 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 Followup (EFU). The recoding operation for A.C.E. Revision II was extended to include all of the persons in the Revision E sample as well as the entire Revision P sample. Note that the PFU/EFU Review did not involve any recoding of P sample cases, except those that matched to a review E-sample case. The A.C.E. Revision II recoding operation was also modified to include automated recoding for all cases supplemented by clerical recoding by analysts for cases identified as problematic as described below.

There were coding errors in both the A.C.E. PFU and the EFU resulting from limitations of their respective interviews (Bean, 2001 and Adams and Krejsa, 2001, respectively). Furthermore, the EFU did not strictly follow census residence rules. Given the information available, the recoding that was done on the 17,500 Review E sample is considered to have negligible error since this 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 describes the reason why the particular code was assigned. There were over 60 different

possible "Why" code categories. These detailed codes can be summarized by the following broad groupings:

- No followup
- Noninterview
- Geocoding issues
- Mover issues
- Other residence issues
- Group quarter issues
- Died before census day or born after census day
- Lived there, no unusual living situations noted

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

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

In general, cases did not go to clerical review if both the PFU and EFU automated codes agree and the mover statuses also agree and the why code category was deemed to be of high enough quality. In some instances cases were exempt from clerical review because they could be coded based on information available in data files. For many of these situations, consistent and complete data were obtained from both the PFU and EFU interviews.

Some cases automatically went to clerical review; such as cases in the Review that had resulted in a conflicting status or cases such as noninterviews or cases where mover dates could not be read in the EFU form. Some of the cases that went to clerical recoding did so because the original A.C.E or PFU results did not agree with the EFU results. Most of the cases went to clerical recoding because the automated coding process was not reliable for that "Why" code category.

After the A.C.E. Revision II recoding operation corrects 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 have been corrected as part of the recoding operation discussed above. To determine the correct match status, each of the possible combinations of match status were 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.

II. 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 are essentially unchanged from those used to produce the March 2001 A.C.E. estimates.[2] The one exception being that it was necessary to impute age again for the full P sample because the A.C.E. Revision II post-strata use different age groupings.

For the revision E and P samples, there were three new types of missing data to deal with:

- noninterviewed households: revision 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 revision coding because it was determined that there were no valid census day residents;
- revision E or P sample cases with unresolved match, enumeration, or residency status because of incomplete or ambiguous interview data;
- revision 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.

Age Imputation for the Full P Sample

For the original A.C.E. full P sample, persons with missing age were assigned to age categories as defined by the post-stratification plan. The A.C.E. Revision II full P Sample post-stratification divided the original post-stratification group of 0-17 years old into groups of 0-9 and 10-17 years old. Those persons 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 missing ages was uniform from 0-17. Other persons with unresolved age remained in the age group they had been originally assigned to.

Household Non-Interview Adjustment for the Revision 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. Housing units were determined to be non-interviews if an interview for a housing unit was not conducted or if a whole household of P-sample people should not have been include and the people who might have lived there on census day were never interviewed. The household non-interview adjustment was done separately for interview day and census day.

[2] Note also that the coding of the A.C.E. full E and P samples remains as it was for the March 2001 estimates. Corrections for measurement error using the Revision Samples are accomplished via estimation using double sampling ratio adjustments.

The methodology for the revision 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. These new non-interviewed units had their weights spread over housing units with at least one person who indicated they lived at another address or who was identified as potentially fictitious. It was assumed that these new non-interviews, which are now representing persons who might have lived at the address on census day, would have had both a low match and residency rate similar to this group. Otherwise, the non-interview adjustment for census day used methodology similar to that used for the A.C.E. full P sample.

Imputation for Revision 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 followup. Second, the person followup (PFU) may not have collected adequate information to determine a person's census day residency status or their match status. Persons in the full E sample with unresolved enumeration status arose only the second way, that is, because the PFU did not collect enough information to determine their enumeration status. The imputation method differs by how the case came to be unresolved.

Imputation for Revision P sample with Insufficient Information

The revision P sample persons with insufficient information for matching and followup 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 followup interview.

In the A.C.E. full P-sample, persons with insufficient information for matching and followup were imputed a probability of census day residency equal to the residency rate of P-sample persons who went to PFU. For the revision 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.

Imputation for Revision P & E Sample with Incomplete or Ambiguous Followup

For the P and E revision sample persons who were unresolved because of ambiguous or incomplete followup information, the situation becomes more complicated because there are two followup interviews to consider, the PFU and EFU. On the one hand the EFU resolved many cases that were previously unresolved in the A.C.E. after the PFU interview. On the other hand, EFU cases with incomplete or ambiguous information were a new source of unresolved cases in the Revision samples. In particular, there were revision P sample cases that were whole households of

matches and non-matches that were sent to EFU; but, in general, were never required to go to PFU. These cases were originally assumed to be resolved and could have become unresolved because of EFU.

For the full E and P samples, imputation cells were based mostly on information obtained before any followup was conducted. For the revision E and P samples, imputation cells rely on the after followup information. This change is the single most important improvement in the missing data methodology. The PFU and EFU interview results and "Why" codes were used to identify:

- unresolved cases with the same history (recipient or imputation cells);
- resolved followup cases sharing that history up to the point of being unresolved (donor pool).

After followup groups are broken down by PFU or EFU groups because the questionnaires are different. However, for a PFU or an EFU group, the after followup groups were defined the same way for the revision P and E samples, because the questions about census day residency are the same as the questions about enumeration status for a given PFU or a given EFU followup interview. There are two general types of after followup groups:

- uninformative, where the interview was incomplete, though there was no evidence of an erroneous enumeration or nonresident.
- informative, where an interview was conducted and there was evidence of an erroneous enumeration or nonresident.

In all, there were nine PFU after followup groups and nine EFU after followup groups. Some of the larger EFU groups were subdivided by variables such as whether or not the household went to PFU, or whether the household was conflicting. A brief summary of the nine PFU and nine EFU groups follows.

PFU After Followup Groups by Degree of Information

Informative Groups:

1 ›Lived elsewhere‹ or at ›other residence‹; address not given

2 Moved in after or out before census day; census day address not given

3 ›Did not live here‹; other address, group quarters and other residence questions not answered

4 ›Other residence‹; usual residence not given

Uninformative Groups:

5 ›Lived here‹; other residence question not answered

6 Usual residence question answered; group quarters and other residence questions not answered

7 ›Lived here‹ question is DK/refused; group quarters and other residence questions not answered

8 Blank questionnaire

9 Potentially fictitious person; no one knew person

EFU After Followup Groups by Degree of Information

Informative Groups:

1 ›Lived elsewhere‹ or at ›other residence‹; address not given

2 Moved in after or out before census day; census day address not given

3 ›Never lived here‹; census day address not given

4 ›Other residence‹; census day address not given

5 Moved in or moved out; mover dates not given

Uninformative Groups:

6 ›Lived here‹; other residence question not answered

7 Current residence question answered; group quarters and other residence questions not answered

8 Usual residence; group quarters and other residence questions not answered

9 Potentially fictitious person; no one knew person

Consider EFU after followup group 2 above, this cell consists of unresolved persons who the followup interview indicated they moved out before census day or moved in after census day, but the followup interview did not provide the address they moved to or from. The donor pool consisted of those resolved persons who indicated in the followup that they moved out before census day or moved in after census day, and did provide the mover address in the followup.

The probability of being correctly enumerated (or a resident on census day) for a group of unresolved cases is imputed as:

$$\Pr(CE) = \frac{\sum_{donors} weighted\ CEs}{\sum_{donors} weighted\ Es} \qquad \Pr(\text{Res}) = \frac{\sum_{donors} weighted\ \text{Res}idents}{\sum_{donors} weighted\ Ps}$$

Persons who moved out before census day or moved in after census day were the largest informative after followup group. Another one was persons who had another residence such as a vacation home, but the followup interview did not indicate whether the other residence was the census day residence.

Imputation for Revision E or P Sample Conflicting Cases

When the A.C.E. PFU and the evaluation followup EFU interviews had contradictory information, the case was assigned a code of conflicting (these conflicting codes are not to be confused with conflicting households, which are households that include an entirely different set of people in the P and E samples.) 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 revision E-sample conflicting cases and for census day residency for revision P-sample conflicting cases. It should be noted that the recoding of the revision samples resulted in considerably less conflicting cases than the PFU/EFU Review sample. The weighted number of conflicting cases in the PFU/EFU Review sample was about 2.6 million in contrast to only about 100,000 in the Revision samples.

III. 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. The matching algorithm used statistical matching to identify linked records. Statistical matching allows 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 are identical, modeling techniques were used to assign a measure of confidence, the duplicate probability, that the two records refer to the same individual.

This study does not identify which enumeration is in the correct location. A component of the A.C.E. Revision II estimation methodology is the determination of the conditional probability that the sample case is in the correct location given that it has a duplicate link to a census enumeration outside the A.C.E. search area. These conditional probabilities are referred to as z_t's and h_t's for the E and P samples respectively. A discussion about these conditional probabilities can be found in Section VII.

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 allows 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.

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 are:

- First Name
- Last Name
- Middle Initial
- Month of Birth
- Day of Birth
- Computed Age

Matching parameters, associated with each matching variable, measure the degree to which the matching variables agree between the two records, ranging from Full Agreement to Full Disagreement. The measurement of the degree to which each matching variable agrees is called the variable match score. The overall match score for the linked records is the sum of the variable match scores.

Full agreement of at least four characteristics was required to be considered a duplicate link at the first stage. Imposing such a requirement limits the power of statistical matching, however, this was considered critical because this study did not have the benefit of a clerical review of duplicate links.

The search for duplicate links between the full E and P samples and the census is limited to those pairs that exactly agree on certain values or blocking criteria. Blocking criteria are sort keys and are used to increase the computer processing efficiency by searching for links where they are most likely to be found. For instance, if we wanted to search only for duplicates within county, both the sample and the census files would be sorted by First name, Last name, the blocking criteria. Then, all possible pairs within the First name, Last name are searched for duplicate links. False nonmatches can occur by using blocking criteria. The plan is to use four sets of blocking criteria. Multiple sets of blocking criteria minimize the number of missed matches. The blocking criteria are:

- First name, Last name
- First name, First initial of last name, Age groupings (0 - 9, 10 - 19, 20 - 29, etc.)
- Last name, First initial of first name, Age groupings (0 - 9, 10 - 19, 20 - 29, etc.)
- First initial of first name, First initial of last name, Month of birth, Day of birth

Second Stage Match of Persons within Households

The second stage of matching was limited to matching persons within linked households. A household was included in the second stage multiple times if the household had persons with links to multiple households in the first stage. If an E or P sample case linked to a person record in a group quarters, the case did not go to the second stage. 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 is a within-household match.

Only one set of blocking criteria was used at the second-stage, the household. The A.C.E. sample records were allowed to link only with one census record within the household. Each link had an overall second-stage match score.

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 are 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 is a duplicate. These two approaches are referred to as the statistical match modeling and the exact match modeling, respectively. 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.

Statistical Match Modeling

After the second-stage matching, each full E or P sample record within a household had a match score based on the attempted match with a census household. So, for each sample household, a set of match scores is observed. For any resulting set of match scores, a probability of not observing this set of match scores was estimated for each link within the household. The higher this probability, the more likely that the set of linked records in the household are duplicates.

The estimate of the probability of not observing this set of match scores assumed independence of the individual match scores within each household on the basis of the low weight given to last name within the second-stage matching. The probability of observing the individual match scores was estimated from the empirical distribution of individual match scores resulting from the entire second-stage matching. Further, this measure accounted for the number of times that a single sample household was matched to different census households within a given level of geography: within block, within tract (outside block), within county (outside tract) within state (outside county) and different state.

The probability of not observing this set of match scores was translated into 1/0 “statistical match” duplicate probability based on critical values which varied by geographic distance of the link mentioned above.

Exact Match Modeling

"Exact match" duplicate probabilities were modeled by doing an exact match of the census to itself. The methodology took into account the overall distribution of births, frequency of names and population size in a specific geographic area. Duplicate probabilities were computed separately by links within county, links within state and different county, and different states. Duplicate probabilities are greater for links within county, gradually decreasing for links within states and links across states. Further, duplicate links were modeled separately by how common the last name is as well as separately by Hispanic names. Similarly, duplicate probabilities decrease as the number of links increases.

Combining the two approaches

The duplicate probability for the links to group quarters in the first stage and one-person household links were from the exact match modeling. For all other links, the duplicate probability was the larger of the two model estimates. For non-exact matches, this was always from the statistical match modeling. For exact matches, adjustments were made to account for the integration of these two methods.

The results of this matching and modeling provide, for each full E and P sample person who links to a census person outside the A.C.E. search area, the probability that they are in fact the same person. These probabilities, referred to as p_t, were used in obtaining A.C.E. Revision II estimates.

IV. The 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 DSE formula for A.C.E. Revision II, using version C for movers[3], separate E & P post-strata, measurement error corrections from the E & P Revision Samples and Duplicate Study results is written:

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

Both Cen' and II' terms exclude 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'	Revision E and P measurement error correction 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

[3] The formula using version A treatment for movers is given in Appendix 1.

V. The Full E & P Samples

The full E & P Samples with the original coding results that were used to produce the March 2001 estimates of Census coverage provide the basis of the A.C.E. Revision II estimates. The original estimates were determined to be unacceptable because of the presence of large amounts of measurement error. These full samples are comprised of over 700,000 sample persons each. Instead of one post-stratification, the A.C.E. Revision II estimates include separate post-strata for the full E & P Samples indicated by i and j, respectively.

For the full P Sample, the post-strata are nearly identical to those used for the March 2001 estimates. The 0 to 17 age group has been split into two groups, 0 to 9 and 10 to 17, which has resulted in some collapsing differences. Therefore, the full P-Sample is now defined by **480** post-strata based on the following characteristics (as opposed to the previous 416 post-strata):

- Race/Hispanic Origin Domain
- Tenure
- Size of Metropolitan Statistical Area
- Type of Census Enumeration Area
- Tract Return Rate (Low vs. High)
- Region
- Age
- Sex

For further information see Tables 1. and 5. and Appendix 2.

For the full E Sample, the post-strata have undergone major revisions. Some of the original post-stratification variables have been omitted and additional variables have been added. The full E-Sample is now defined by **525** post-strata based on the following characteristics:

- Proxy Status
- Race/Hispanic Origin Domain
- Tenure
- Household Relationship
- Household Size
- Type of Census Return (Mailback vs. Nonmailback)
- Date of Return (Early vs. Late)
- Age
- Sex

For further information see Tables 2. and 3. and Appendix 2.

VI. Adjustment for Measurement Error using the Revision E & P Samples

The Revision E and P Samples are subsamples of the full E and P Samples. They are each comprised of over 70,000 sample persons. These revision 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 revision samples have undergone extensive recoding using all available interview data and matching results. Missing data adjustments have also been applied to the revision sample data. This recoded data from the revision samples are used to correct for measurement error in the original full E and P Samples.

The ratio adjustments that correct for measurement error are based on the P or E Revision Sample and are a ratio of an estimate using the revised coding (indicated by *) to the an estimate using the original coding. These adjustments are done by measurement error correction post-strata *i'* or *j'* .

Adjustments to E-Sample Correct Enumerations

The measurement error adjustment for correct enumerations that are not duplicates is given by:

$$f_{1,i'} = \frac{CE_{i'}^{ND*}}{CE_{i'}^{ND}}$$

where *i'* are defined by: proxy status, domain, household relationship, age & sex

The i' are always a subset of a full E Sample post-stratum i. Consequently, ratio adjustment factors are also calculated for collapsed age and sex groups. See Table 4. for additional information.

Adjustments to P-Sample Nonmovers

The measurement error adjustment for nonmover matches and nonmovers (not duplicates) is given by:

$$f_{2,j'} = \frac{M_{nm,j'}^{ND*}}{M_{nm,j'}^{ND}} \qquad f_{6,j'} = \frac{P_{nm,j'}^{ND*}}{P_{nm,j'}^{ND}}$$

where j' for nonmovers are defined by: domain, tenure, age & sex

Adjustments to P-Sample Movers

The measurement error adjustment for the mover terms is given by:

$$f_{3,j'} = \frac{M_{om,j'}^{*}}{M_{om,j'}} \qquad f_{4,j'} = \frac{P_{om,j'}^{*}}{P_{om,j'}} \qquad f_{5,j'} = \frac{P_{im,j'}^{*}}{P_{im,j'}}$$

where j' for movers are defined by: tenure

The j' are always a subset of a full P Sample post-stratum j. Consequently, ratio adjustment factors are also calculated for collapsed age and sex groups. See Table 6. for additional information.

Each of the terms in the above double sampling ratio adjustments are weighted tallies from the revision sample. The weights reflect the inverse of the probability of selection, adjustments for household noninterviews, missing data imputations, and targeted extended search.

Adjustment for movers due to Duplicates

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.

This term gets multiplied by g, which is an estimate of the proportion of originally coded nonmovers with duplicate links who are true nonresidents that have moved in since Census day. g is estimated using the revision sample and both the original A.C.E. and the revised coding as follows:

$$g = \frac{P^{D}_{nm,im*}}{P^{D}_{nm,nr*}}$$

where:

$P^{D}_{nm,im*}$ is an estimate of persons (using the revision P sample) with a duplicate link who were originally coded as a nonmover but the revision coding determined them to be inmovers, which are of course a subset of nonresidents.

$P^{D}_{nm,nr*}$ is an estimate of persons (using the revision P sample) with a duplicate link who were originally coded as a nonmover but the revision coding determined them to be nonresidents.

A couple of important assumptions are:

- If the revision coding determined a person was a nonresident, they really are a nonresident; i.e., revision-coded nonresidents are a subset of true nonresidents.
- The rate of inmovers for revision-coded nonresidents is the same as that for true nonresidents.

VII. 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. The duplicate study used computer-based record linkage techniques to match the full P and E samples to census enumerations outside the search area. The

census enumerations included those enumerations that were added too late to be included in the E sample as well as those enumerations that were determined to be duplicates and were never included in the census. P and E sample cases with duplicate links were assigned a nonzero probability of being a duplicate, p_t. P and E sample cases without duplicate links were assigned a p_t of zero. This probability is usually 0 or 1 for E and P sample cases, but some duplicate links have a value in between indicating less confidence that the link is representing the same person. These probabilities are also transferred to the E and P revision samples.

Adjustments for Nonduplicates

When estimating terms in the DSE involving nonduplicates, those indicated by a superscript ND, it is necessary to include the probability of not being a duplicate, $1-p_t$, in the tallies. This probability of not being a duplicate is included in all of the following terms:

For full E and P terms: CE_i^{ND} $M_{nm,j}^{ND}$ $P_{nm,j}^{ND}$

For revision E and P terms: $CE_{i'}^{ND*}$ $CE_{i'}^{ND}$ $M_{nm,j'}^{ND*}$ $M_{nm,j'}^{ND}$ $P_{nm,j'}^{ND*}$ $P_{nm,j'}^{ND}$

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. Assuming that the linked person does exist, the goal is to determine which of the two locations is the appropriate place to count the person. Since linked persons may be geographically close or far apart, this has implications for the degree of synthetic error.

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 addition conditional probabilities:

- z_t is the probability that an E sample case is a correct enumeration given that it is a duplicate to another census enumeration outside the A.C.E. search area.

- h_t is the probability that a P sample case is a resident on Census day given that it links to a census enumeration outside the A.C.E. search area.

<u>E Sample Links</u>

From the duplicate study, an estimate of correct census enumerations can be derived by considering the situation of the linked enumerations as well as assuming that each link represents one correct enumeration. This assumes of course that the link consists of true duplicates. These assumptions are

used to estimate the contribution to correct enumerations from full E sample cases with duplicate links, including those originally coded as correct as well as those originally coded as erroneous. This contribution to correct enumerations is given by the term: $\widetilde{CE}^{D}{}_{i}$. To estimate this term, the E sample links are first classified according to the characteristic of the linked situation and the original coding of the E sample. Table 7. summarizes this classification and the rules for assigning z_t's.

First, linked situations are identified where one component of the link is thought to be correct and the other incorrect. If a person in a housing unit links with a person in a group quarters, such as a college dormitory, the person in the housing unit is taken to be incorrect and assigned a z_t of zero. See "Linked Situation" 1. in Table 7. If a linked person 18 years of age or older is listed in only one of the households as a child of the reference person, this person is assumed to be incorrectly included with their parents and correctly included in the other household unless A.C.E. had already determined them to be an erroneous inclusion. An example of this might be a college student that was listed with their parents and also listed in an apartment off campus. This is represented by "Linked Situations" 2a. and 2b. in Table 7.

For other "Linked Situations" the choice of which person is correct is not clear. Consider links between whole households where all household members are duplicated. ("Linked Situation 3.) This includes families that might have moved some time around census day and were inadvertently included at both places or this might involve households with multiple residences with a helpful, but perhaps, uninformed proxy respondent. Another situation, "Linked Situation" 4., involves children ages 0 to 17, perhaps of divorced parents, that are linked between two different households. For these and all other situations, it is assumed that only half of these census enumerations with duplicate links are correct. To estimate the conditional probability, z_t, that the E-sample person is the correct enumeration, controls cells are defined for "Linked Situations" 3., 4. and 5. as shown in Table 7. by:

- 3 Race/Hispanic Origin Domain
- Tenure

These resulting control cells are given in Table 8. Within each control cell the z_t's are determined such that duplicate E-sample cases originally coded correct or unresolved will weight up to one half the number of census duplicates identified including the erroneous enumerations. This was calculated as:

$$\hat{z}_t = \frac{0.5\sum_t W_{.t}\ p_{.t}}{\sum_t W_{.t}\ p_{.t}\ \Pr(CE)}$$

The summations are over the links in a control cell regardless of the original E sample coding.

The z_t's are then included in the weighted tallies along with the p_t's as given by:

$$\widetilde{CE}^{D}{}_{i} = \sum_t W_{.t}\ p_{.t} z_{.t}\ \Pr(CE)$$

The sum is over all E-sample persons with duplicate links in post-stratum i and W_t is the person's final weight.

P-Sample Links

Unlike the E-Sample side, the duplicate study does **NOT** provide an estimate of the number of correct census day residents in the P sample. In order to estimate h_t, the probability that a P-sample case is a resident on Census day given that it links to a census enumeration outside the search area, it is necessary to borrow the resulting z_t's from the E-sample links. Table 7. summarizes how the h_t's borrow information from the z_t's.

First, the P-sample links to census enumerations outside the search area are identified for situations where it can be determined which component of the link is the correct residence. The "Linked Situations" and rules for assigning h_t's are the same as used for comparable types of E-sample links. For example, consider a P-sample person 18 years of age or older listed as a child of the reference person who links with a census enumeration in a household where they are not listed as a child, this P-sample person would be assigned an h_t of zero regardless of how A.C.E. coded this person. Thus, it is assumed that this person should not have been included in the P sample.

For the other "Linked Situations" 3., 4., and 5., there once again is no information to determine whether the P sample had the person at the correct location or whether the census had them at the correct location. Additionally, there is no reasonable assumption about how many of these linked P-sample persons should be at the correct location. To overcome this obstacle, it is assumed that the error in identifying correct residence is similar to the error in identifying correct enumeration for similar situations. Therefore, the h_t for P-sample persons is set equal to the z_t determined for the E sample for comparable linked situations as identified by the control cells in Table 8.

The h_t's are then included in the weighted tallies, along with the p_t's, to calculate the duplicate contribution to the full P-Sample nonmovers and nonmover matches as indicated by:

$$\tilde{P}^{D}_{nm,j} = \sum_t W_{.t}\ p_{.t} h_{.t} \Pr(\mathrm{Re}\, s) \qquad \text{where the summation is over nonmovers.}$$

$$\tilde{M}^{D}_{nm,j} = \sum_t W_{.t}\ p_{.t} h_{.t} \Pr(M) \Pr(\mathrm{Re}\, s) \qquad \text{where the summation is over matched nonmovers}$$

VIII. Adjustment for Correlation Bias using Demographic Analysis

Next the 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 relative 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 post-strata 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 Nonblacks. 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. The resulting DA sex ratios for the three age groups by Black and Nonblack domain are shown in Table 9.

In general the correlation bias adjustment factor, c_κ , is defined for the three k age groups such that:

$$E\,[c_\kappa DSE^m{}_k] = \text{True male population for age group } k.$$

where:

$DSE^m{}_k$ is the sum of DSEs over male post-strata in age group k.

Since the purpose of this adjustment is to reflect persons missed in both the census and the A.C.E., the value of c_κ will not be allowed to be less than one.

Correlation Bias Adjustment for Black and Nonblack Males 18 Years and Older:

The correlation bias adjustment for Black and Nonblack males 18 years and older is done so that the A.C.E Revision II sex ratios will agree with the DA sex ratios for Blacks and Nonblacks. This correlation bias adjustment is calculated as:

$$c_{R,\kappa} = \left(\frac{\sum_{ij \in k} DSE_{ij}^{Rf}}{\sum_{ij \in k} DSE_{ij}^{Rm}} \right) r_{DAR,\kappa}$$

where:

DSE_{ij}^{Rf} = DSE for race, R=Black or Nonblack, female post-strata ij.

DSE_{ij}^{Rm} = DSE for race, R=Black or Nonblack, male post-strata ij.

$r_{DAR,k}$ = DA sex ratio for race, R=Black or Nonblack, for age group *k* as given in Table 9.

The sum over the ij post-strata includes only the intersection of those post-strata with age group *k*.

DSEs adjusted for Correlation Bias:

A correlation bias-adjusted DSE for a male 18+ post-stratum *ij* in the age-race group *k* is calculated as:

$$\widetilde{DSE}_{ij}^{m} = c_k \, DSE_{ij}^{m}$$

For all remaining post-strata, which includes female post-strata as well as post-strata for persons under 18 years of age, no correlation bias adjustment is done. Thus:

$$\widetilde{DSE}_{ij}^{f} = DSE_{ij}^{f}$$

The $\widetilde{DSE}_{ij}$'s are then used to form the synthetic estimates.

IX. Synthetic Estimation

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

$$\widetilde{CCF}_{ij} = \frac{\widetilde{DSE}_{ij}}{Cen_{ij}}$$

where:

$\widetilde{DSE}_{ij}$'s are the correlation bias-adjusted DSEs for post-stratum ij.

Cen_{ij}'s are the census counts for post-stratum ij. Note that this Cen_{ij} includes late census adds.

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:

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

Note that the coverage correction factor can be expressed as:

$$\tilde{CCF}_{ij} = \left(\frac{DD_{ij}}{Cen_{ij}} \right) \left(\frac{r_{ce,i}}{r_{m,j}} \right) c_k$$

where:

$r_{ce,i}$ is the correct enumeration rate component of the DSE, varying over i post-strata.

$r_{m,j}$ is the match rate component of the DSE, varying over j post-strata.

c_k is the correlation bias adjustment factor, varying over the Black and Nonblack groups and k age cells.

$\frac{DD_{ij}}{Cen_{ij}}$ is the data-defined rate, varying over the ij post-strata.

Table 1. Full P-Sample Post-Stratum Groups and Number of Age and Sex Groupings (j)

Race/Hispanic Origin Domain Number*	Tenure	MSA/TEA	High Return Rate				Low Return Rate			
			NE	MW	S	W	NE	MW	S	W
Domain 7 Non-Hispanic White or "Some other race"	Owner	Large MSA MO/MB	8	8	8	8	8	4	8	4
		Medium MSA MO/MB	8	8	8	8	4	8	8	8
		Small MSA & Non-MSA MO/MB	8	8	8	8	4	8	8	8
		All Other TEAs	8	8	8	8	8	8	8	8
	Non-Owner	Large MSA MO/MB	8				8			
		Medium MSA MO/MB	8				8			
		Small MSA & Non-MSA MO/MB	8				8			
		All Other TEAs	8				8			
Domain 4 Non-Hispanic Black	Owner	Large MSA MO/MB Medium MSA MO/MB	8				8			
		Small MSA & Non-MSA MO/MB All Other TEAs	8				8			
	Non-Owner	Large MSA MO/MB Medium MSA MO/MB	8				8			
		Small MSA & Non-MSA MO/MB All Other TEAs	8				4			
Domain 3 Hispanic	Owner	Large MSA MO/MB Medium MSA MO/MB	8				8			
		Small MSA & Non-MSA MO/MB All Other TEAs	8				8			
	Non-Owner	Large MSA MO/MB Medium MSA MO/MB	8				8			
		Small MSA & Non-MSA MO/MB All Other TEAs	8				4			
Domain 5 Native Hawaiian or Pacific Islander		Owner	4							
		Non-Owner	4							
Domain 6 Non-Hispanic Asian		Owner	8							
		Non-Owner	8							
American Indian or Alaska Native	**Domain 1** (On Res.)	Owner	8							
		Non-Owner	8							
	Domain 2 (Off Res.)	Owner	8							
		Non-Owner	8							

Table 2. Full E-Sample Post-Stratum Groups and Number of Age and Sex Groupings (i)

Proxy Status & Domain		Tenure	Relationship	HH Size	Early Mail-back	Late Mail-Back	Early Non-Mailback	Late Non-Mailback
***Proxy:* Domain 7** Non-Hispanic White or "Some Other Race"					8			
***Proxy:* Domain 4** Non-Hispanic Black					8			
***Proxy:* Domain 3** Hispanic					8			
***Proxy:* Domain 5** Native Hawaiian or Pacific Islander					1			
***Proxy:* Domain 6** Non-Hispanic Asian					4			
***Proxy:* Domain 1** American Indian or Alaska Native On Reservation					4			
***Proxy:* Domain 2** American Indian or Alaska Native Off Reservation					1			
Non-Proxy: **Domain 7** Non-Hispanic White or "Some Other Race"		Owner	HHer/Nuclear	2-3	8	8	8	8
				4+	8	8	4	8
			Other	1	2	2	1	2
				2-3	8	8	2	4
				4+	8	8	4	8
		Non-Owner	HHer/Nuclear		8	8	8	8
			Other		8	8	8	8
Non-Proxy: **Domain 4** Non-Hispanic Black		Owner	HHer/Nuclear		4	4	2	4
			Other		8	8	4	8
		Non-Owner	HHer/Nuclear		8	8	8	8
			Other		8	8	8	8
Non-Proxy: **Domain 3** Hispanic		Owner	HHer/Nuclear		8	8	4	8
			Other		8	8	4	8
		Non-Owner	HHer/Nuclear		8	8	8	8
			Other		8	8	8	8
Non-Proxy: **Domain 5** Native Hawaiian or Pacific Islander		Owner & Non-Owner	HHer/Nuclear		2	2	2	2
			Other		2	2	1	2
Non-Proxy: **Domain 6** Non-Hispanic Asian		Owner & Non-Owner	HHer/Nuclear		8	8	4	4
			Other		4	4	2	4
Non-Proxy: American Indian or Alaska Native	**Domain 1** On Reservation	Owner & Non-Owner	HHer/Nuclear		8			
			Other		8			
	Domain 2 Off Reservation	Owner & Non-Owner	HHer/Nuclear		2	2	2	2
			Other		2	2	1	2

Table 3. E-Sample Age and Sex Groupings

Age	8 Groups		4 Groups		2 Groups		1 Group	
	Male	Female	Male	Female	Male	Female	Male	Female
0 – 9								
10 – 17								
18 – 29								
30 – 49								
50+								

Table 4. Revision E-Sample Post-Strata (i')

Proxy Status & Domain	Relationship	Age	8 Groups		4 Groups	2 Groups	1 Group
			Male	Female			
Proxy: **Domain 7** Non-Hispanic White or "Some Other Race" **Domain 4** Non-Hispanic Black **Domain 3** Hispanic **Domain 5** Native Hawaiian or Pacific Islander **Domain 6** Non-Hispanic Asian **Domain 1** American Indian or Alaska Native On Reservation **Domain 2** American Indian or Alaska Native Off Reservation							
Non-Proxy: **Domain 7** Non-Hispanic White or "Some Other Race" **Domain 4** Non-Hispanic Black **Domain 3** Hispanic **Domain 5** Native Hawaiian or Pacific Islander **Domain 6** Non-Hispanic Asian **Domain 2** American Indian or Alaska Native Off Reservation	HHer/Nuclear	0 – 9					NA
		10 – 17					
		18 – 29					
		30 – 49					
		50+					
	Other	0 – 9					
		10 – 17					
		18 – 29					
		30 – 49					
		50+					
Non-Proxy: **Domain 1** American Indian or Alaska Native On Reservation							

Table 5. P - Sample Age and Sex Groupings

Age	8 Groups		4 Groups		1 Group*	
	Male	Female	Male	Female	Male	Female
0 – 9						
10 – 17						
18 – 29						
30 – 49						
50+						

* The 1 Group is not used for the Full P post-strata (j), only the Revision P post-strata (j').

Table 6. Revision P-Sample Post-Strata (j')

Mover Status & Domain	Tenure	Age	8 Groups		4 Groups		1 Group
			Male	Female	Male	Female	
Movers: **Domains 1 thru 7**	Owner						
	Non-Owner						
Non-Movers: **Domains 2 thru 7**	Owner	0 – 9					NA
		10 – 17					
		18 – 29					
		30 – 49					
		50+					
	Non-Owner	0 – 9					NA
		10 – 17					
		18 – 29					
		30 – 49					
		50+					
Non-Movers: **Domain 1** American Indian or Alaska Native On Reservation							

Table 7. Rules for Assigning z_t & h_t for Full P & E Sample Duplicate Links

The "Linked Situations" and assignment of z_t's and h_t's occur in the order in which they are listed below.

"Linked Situation" (E or P) ⇔ (Census)		Original E Coding	z_t	Original P Coding	h_t
1.	(Person in a housing unit) ⇔ (Person in a group quarters)	EE	0	NonRes	0
		CE/UE	0	Res/UE	0
2a.	(Person 18+, child of reference person) ⇔ (Person 18+, not child of reference person)	EE	0	NonRes	0
		CE/UE	0	Res/UE	0
2b.	(Person 18+, not child of reference person) ⇔ (Person 18+, child of reference person)	EE	0	NonRes	0
		CE/UE	1	Res/UE	1
3.	(All persons in a housing unit) ⇔ (All persons in another housing unit)	EE	0	NonRes	0
		CE/UE	$\hat{z}_1$	Res/UE	$\hat{z}_1$
4.	(Child 0-17) ⇔ (Child 0-17)	EE	0	NonRes	0
		CE/UE	$\hat{z}_2$	Res/UE	$\hat{z}_2$
5.	All Remaining Linked Situations	EE	0	NonRes	0
		CE/UE	$\hat{z}_3$	Res/UE	$\hat{z}_3$

EE is erroneous enumeration.
CE is correct enumeration.
UE is unresolved.
Res is resident on Census day.
NonRes is not a resident on Census day.

Table 8. Control Cells for Linked E Sample

Race/Hispanic Origin Domain	Tenure	"Linked Situation"	Control Cell
Domain 4 Non-Hispanic Black	Owner	3.	
		4.	
		5.	
	Non-Owner	3.	
		4.	
		5.	
Domain 3 Hispanic	Owner	3.	
		4.	
		5.	
	Non-Owner	3.	
		4.	
		5.	
Domain 7 Non-Hispanic White or "Some Other Race" **Domain 5** Native Hawaiian or Pacific Islander **Domain 6** Non-Hispanic Asian **Domain 1** American Indian or Alaska Native On Reservation **Domain 2** American Indian or Alaska Native Off Reservation	Owner	3.	
		4.	
		5.	
	Non-Owner	3.	
		4.	
		5.	

Table 9. Correlation Bias Adjustment Groupings and Factors

Race/Hispanic Origin Domain	Age	DA Sex Ratios	Adjustment Factor
Black: **Domain 4** Non-Hispanic Black	18 - 29	0.8972	
	30 - 49	0.8890	
	50+	0.7603	
Nonblack: **Domain 3** Hispanic **Domain 7** Non-Hispanic White or "Some Other Race" **Domain 5** Native Hawaiian or Pacific Islander **Domain 6** Non-Hispanic Asian **Domain 1** American Indian or Alaska Native On Reservation **Domain 2** American Indian or Alaska Native Off Reservation	18 - 29	1.0437	
	30 - 49	1.0060	
	50+	0.8561	

The DA Sex Ratios exclude the group quarters population and have been adjusted so that Black Hispanics are included in the "Nonblack" grouping rather than in the "Black" grouping above. This latter adjustment defines Black as those persons who checked Black alone or Black in combination with any other race in Census 2000 (Model 2).

Appendix 1. DSE with Mover Treatment A

The DSE formula that uses version A for movers with different post-strata for the E & P Samples is:

$$DSE^{A}_{ij} = (Ce\ n'_{ij} - II'_{ij}) \frac{\left[\frac{CE_{i}}{E_{i}}\right]}{\left[\frac{M_{nm,j} + M_{om,j}}{P_{nm,j} + P_{om,j}}\right]}$$

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

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

Both Cen' and II' terms exclude the late census adds.

This version of the DSE is only used when the sample size for outmovers in the Full P-Sample is strictly less than 10. This occurs for 93 of the P-sample post-strata.

Appendix 2. Full Sample Post-Stratum Variable Definitions

Variable	E or P	Definition
Race/Hispanic Origin Domain	E & P	Hierarchical assignment of persons to a Race & Hispanic Ethnicity group accounting for multiple race responses and some geography: • Domain 1 – American Indian or Alaska Native on Reservations • Domain 2 – American Indian or Alaska Native off Reservations • Domain 3 – Hispanic • Domain 4 – Non-Hispanic Black • Domain 5 – Native Hawaiian or Pacific Islander • Domain 6 – Non-Hispanic Asian • Domain 7 – Non-Hispanic White or "Some other race"
Tenure	E & P	Owner includes persons in owned housing units. All other persons are considered Non-Owners.
MSA/TEA	P	Persons in housing units are classified by size of Metropolitan Statistical Area(MSA) and Census Type of Enumeration Area (TEA): • Large MSA MO/MB; mailout/mailback areas in 10 largest MSAs • Medium MSA MO/MB; mailout/mailback areas in MSAs (excluding 10 largest) with population of at least 500,000 • Small MSA & Non-MSA MO/MB; remaining mailout/mailback areas. • All other TEAs;
High/Low Return Rate	P	Persons in housing units are associated with a tract-level census return rate. High and low return rate indicators are assigned to Owners and Non-Owners in Domains 3, 4, and 7 based on a 25^{th} percentile cutoff value. A low (high) return rate indicator value is assigned to people at or below (above) their designated cutoff value.
Region	P	Persons in housing units classified by Census Region: • NE is Northeast • MW is Midwest • S is South • W is West
Proxy Status	E	Non-proxy includes household members who lived there on Census Day and responded to the census (other than via an Enumerator Questionnaire).
Relationship	E	The HHer/Nuclear relationship category includes persons in housing units consisting only of the householder with spouse or own children (l7 or younger). The "Other" relationship category consists of single-person households and persons in housing units with any other type of relationship, including unrelated persons.
HH Size	E	Household size, or number of persons residing in the housing unit.
Early/Late Mailback	E	Persons in mailback housing units with an earliest form processing date: • on or before March 24 are early • after March 24 are late
Early/Late Non-mailback	E	Persons in non-mailback housing units with an earliest form processing date: • on or before June 1 are early • after June 1 are late