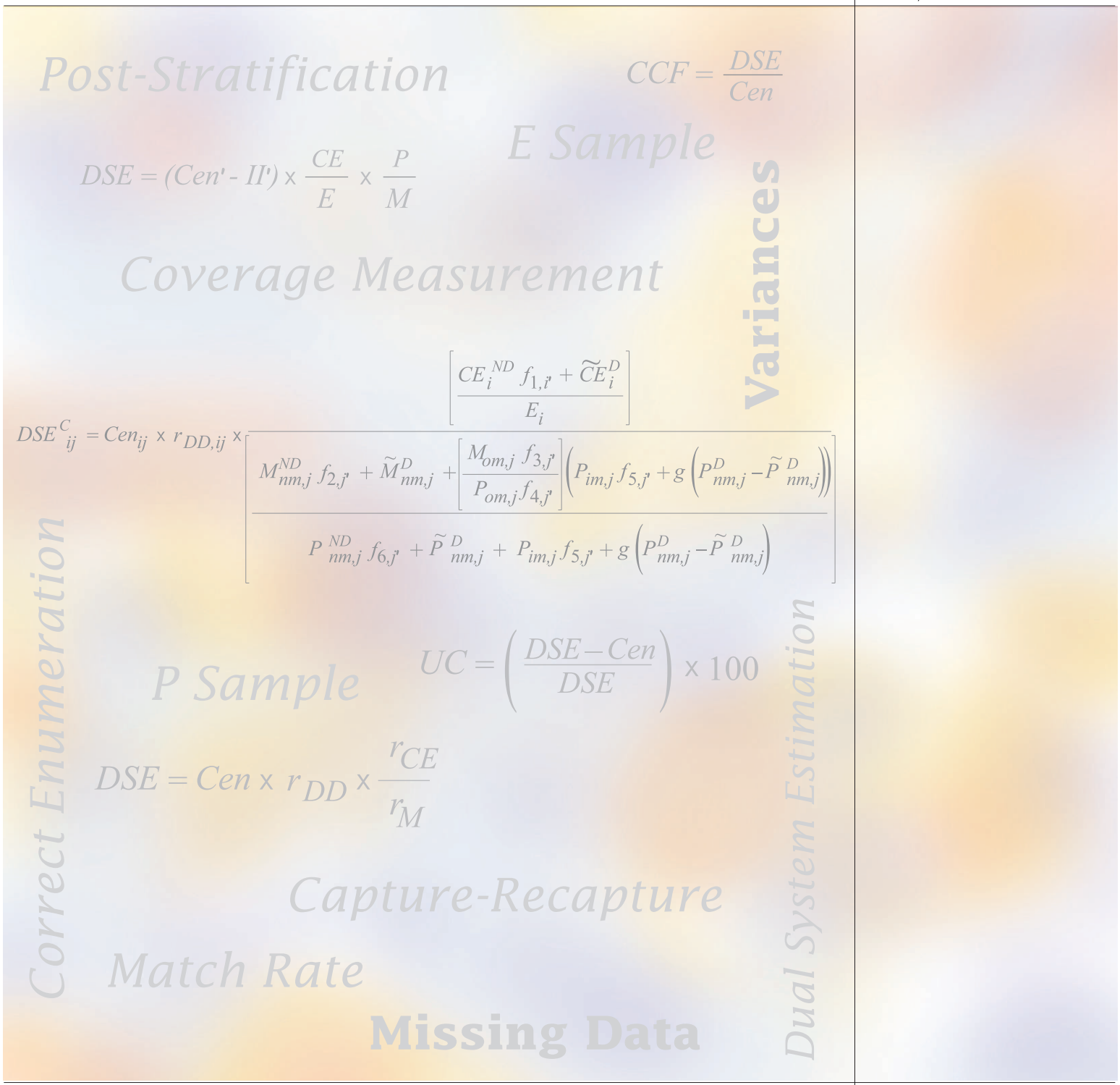
# Accuracy and Coverage Evaluation of Census 2000: Design and Methodology

Issued September 2004

DSSD/03-DM




U S C E N S U S B U R E A U

*Helping You Make Informed Decisions*

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

# ACKNOWLEDGMENTS

This technical document was prepared under the direction of **Donna Kostanich,** Assistant Division Chief for Sampling and Estimation, Decennial Statistical Studies Division. The overall management and coordination of the review was conducted by **Dawn Haines** and **Douglas Olson**. The combined efforts of numerous U.S. Census Bureau staff have culminated in the publication of this document. Some staff members wrote chapters, while others reviewed chapters. In some cases, staff members filled both capacities.

Contributing to the March 2001 portion of Accuracy and Coverage Evaluation of Census 2000: Design and Methodology were **Patrick Cantwell, Inez Chen, Danny Childers, Peter Davis, James Farber, Deborah Fenstermaker, Richard Griffin, Dawn Haines, Howard Hogan, Michael Ikeda, Donna Kostanich, Vincent Thomas Mule, Mary Mulry, Alfredo Navarro, Douglas Olson, J. Gregory Robinson, Robert Sands,** and **Michael Starsinic. Joseph Waksberg,** of Westat, Inc., reviewed these chapters for readability and consistency.

Contributing to the A.C.E. Revision II section of Accuracy and Coverage Evaluation of Census 2000: Design and Methodology were **Tamara Adams, Michael Beaghen, William Bell, Patrick Cantwell, Deborah Fenstermaker, Richard Griffin, Dawn Haines, Michael Ikeda, Donna Kostanich, Elizabeth Krejsa, Vincent Thomas Mule, Mary Mulry, Rita Petroni, Robert Sands, Eric Schindler, Bruce Spencer,** of Northwestern University, and **David Whitford. Rhonda Geddings** provided administrative support.

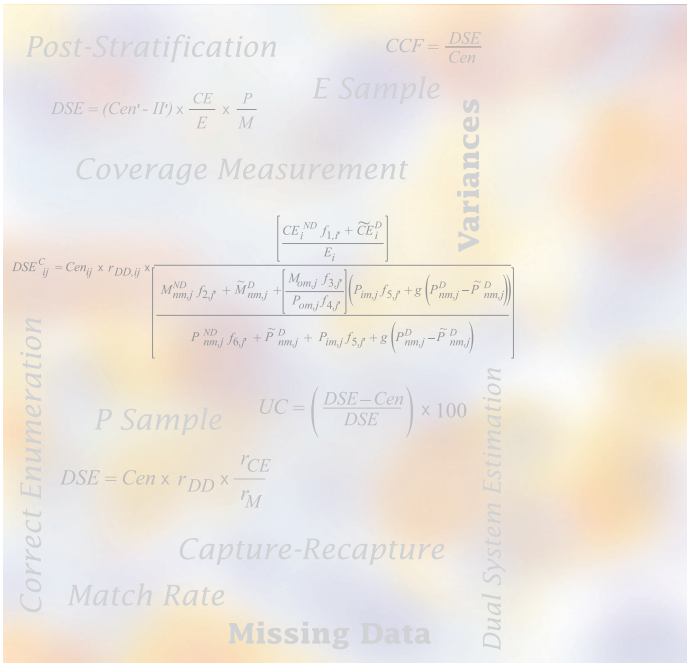
**Bernadette Beasley, Meshel Butler, Helen Curtis, Susan Kelly,** and **Kim Ottenstein** of the Administrative and Customer Services Division, **Walter Odom,** Chief, provided publications and printing management, graphics design, and composition and editorial review for print and electronic media. General direction and production management were provided by **James Clark,** Assistant Division Chief.

**Margaret Smith** of ACSD provided assistance in placing the electronic version of this document on the Internet (see www.census.gov/dmd/www/refroom.html).

We are grateful for the assistance of the individuals listed and all others who contributed but are not specifically mentioned. The preparation and publication of this document was possible because of their invaluable contributions.

# Accuracy and Coverage Evaluation of Census 2000: Design and Methodology

Issued September 2004
DSSD/03-DM




**U.S. Department of Commerce**
**Donald L. Evans,**
Secretary

**Theodore W. Kassinger,**
Deputy Secretary

**Economics and Statistics Administration**
**Kathleen B. Cooper,**
Under Secretary
for Economic Affairs

**U.S. CENSUS BUREAU**
**Charles Louis Kincannon,**
Director

SUGGESTED CITATION

FILES: Census 2000,
Accuracy and Coverage
Evaluation of Census 2000:
Design and Methodology
U.S. Census Bureau, 2004


**Economics and Statistics Administration**

**Kathleen B. Cooper,**
Under Secretary
for Economic Affairs


**U.S. CENSUS BUREAU**

**Charles Louis Kincannon,**
Director

**Hermann Habermann,**
Deputy Director and
Chief Operating Officer

**Vacant,**
Principal Associate Director
and Chief Financial Officer

**Vacant,**
Principal Associate
Director for Programs

**Preston Jay Waite,**
Associate Director
for Decennial Census

**Nancy M. Gordon,**
Associate Director
for Demographic Programs

**Cynthia Z.F. Clark,**
Associate Director
for Methodology and
Standards

**Marvin D. Raines,**
Associate Director
for Field Operations

**Arnold A. Jackson,**
Assistant Director
for Decennial Census

# Foreword

The U.S. Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) survey to measure the coverage of the population in Census 2000. The A.C.E. was designed to serve two purposes: (1) to measure the net coverage of the population, both in total and for major subgroups, and (2) to provide data that could serve as the basis for correcting the census counts for such uses as Congressional redistricting, state and local redistricting, funds allocation and governmental program administration. The A.C.E. survey provides critical information that can be used to improve the census-taking process. However, the design, methodology, operations and data collection efforts are extremely complex and not widely understood. The work described in this publication was a major undertaking, and the technical documentation is intended to increase awareness and knowledge, and subsequently improve the 2010 Census and coverage measurement techniques.

Despite the fact that coverage measurement techniques have been utilized by the Census Bureau for several decades, this is the first comprehensive documentation of its kind. This technical document describes the methodologies that were used to produce estimates of Census 2000 coverage error from the A.C.E. The first part of this document discusses the entire survey design used to produce the original estimates of net undercount released in March 2001. Analysis and evaluations indicated that there were serious errors in the March 2001 A.C.E. Research efforts to fix the detected errors resulted in improved coverage estimates referred to as A.C.E. Revision II. The second part of this document describes the methodology used to correct for errors in the March 2001 A.C.E.

After extensive analysis and consideration, the Census Bureau ultimately decided not to use the A.C.E. - neither the March 2001 nor the Revision II results - to correct the Census 2000 counts or any other data products. A.C.E. Revision II, the superior of the two results, provides useful coverage measurement information that can be used for research purposes. All of these results, decisions, supporting analyses, technical assessments, and limitations can be found on the Census Bureau's Web site at *www.census.gov/dmd/www/EscapRep.html.*

This document is intended to promote knowledge and encourage collaboration on coverage measurement issues. As such, we welcome comments and suggestions from colleagues on technical issues and also on the value of this document.

Charles Louis Kincannon
Director, U.S. Census Bureau

CONTENTS

**Section I: A.C.E. March 2001**

Chapters



Appendixes



**Section II: A.C.E. Revision II March 2003**

Chapters

# Accuracy and Coverage Evaluation of Census 2000: Design and Methodology

## Section I

## A.C.E. March 2001

# Chapter 1. Introduction to the A.C.E.

## INTRODUCTION

The U.S. Census Bureau conducted the Accuracy and Coverage Evaluation (A.C.E.) to measure the coverage of the population in Census 2000 and to allow for the possibility of correcting the census results for the measured undercount. It also provides a wealth of information on the census process and may, thus, enable improvement in future censuses. This document is written to provide a clear and permanent record of the methods and operations used in this project.

The current chapter presents the objectives and scope of the A.C.E., and discusses limitations of what it was attempting to accomplish. It includes a brief history of the evolution of the statistical and operational methods upon which the A.C.E. is based. Chapter 2 presents an overview of the various statistical steps necessary to produce estimates of census coverage and how they are tied into the operation of the survey. The sequence of major activities and their timing is given. Subsequent chapters discuss in detail A.C.E. sampling, interviewing, processing, and estimation steps.

### Goals

The evaluation of the completeness of census enumeration has been an integral part of the decennial census since the 1950 census. This evaluation has taken on many forms including demographic analysis, administrative record checks, matches to independent surveys, and dependent record rechecks and reinterviews.

The evaluation of the five censuses from 1950 to 1990 clearly showed that each of the traditional decennial censuses undercounted the total population, and further, missed certain identifiable population groups at greater rates than others. Specifically, these evaluations clearly showed that undercounts were not merely random occurrences, but predictable biases in the census taking process. The undercount has been consistently higher for the African-American population than for the rest of the population, and while the data set is not so extensive, the evidence also pointed to consistently higher undercounts for Hispanics, Asians, Pacific Islanders, and American Indians than for the White non-Hispanic population. The undercount was also related to socioeconomic status, chiefly measured by home ownership, with renters having consistently higher undercounts. The U.S. Census Bureau designed the Accuracy and Coverage Evaluation to measure this differential undercount and, if possible, correct the counts, thereby making the census more accurate.

As mentioned earlier, the A.C.E. was designed to serve two purposes. One goal was to measure coverage of the population, both total and in various major subdivisions such as race/ethnicity, sex, major geographical areas, and socioeconomical groupings. These measurements indicate whether changes made in enumeration methods in the 2000 census were successful in improving the census and show where improvements may be necessary in future censuses. Another goal was to provide data that could serve as the basis for correcting the census counts. In planning the A.C.E., the Census Bureau focused on the accuracy of population totals for both geographic areas and demographic groups. Consideration was given to the possibility of both improving the population totals (numeric accuracy) and population shares (distributive accuracy). Although early planning considered using dual system estimation to produce a "one number census," after the Supreme Court ruled on the use of sampling for congressional apportionment in 1999, the survey was redesigned and refocused on non-apportionment uses. One important use was congressional redistricting. Thus an important consideration in the design was to improve the accuracy of congressional districts, which average around 650,000 people. The U.S. Census Bureau also recognized other uses, including state and local redistricting, funds allocation, and program administration. The traditional goals of coverage evaluation to inform users and aid in the planning of the next census continue to be important. These goals greatly influenced the sample and estimation design.

### The A.C.E. Defined

The A.C.E. is a post-enumeration survey, based on the theory of dual system estimation. The results of the dual system estimation can be used with model-based estimation to produce census files adjusted for the measured net undercount (or net overcount). The design involved comparing (matching) the information from an independent sample survey to initial census records.

In this process, the Census Bureau conducted field interviewing and computerized and clerical matching of records. Using the results of this matching, the Census Bureau applied dual system estimation to develop estimates of coverage for various population groups. The initial plans were to apply correction factors to the census files that could be used to produce all required Census

2000 tabulations, other than apportionment. The correction aspect of Census 2000 tabulations was later abandoned. The A.C.E. can be summarized as follows:

- Select a stratified random sample of blocks for the A.C.E.
- Create an independent list of housing units in the sample of A.C.E. blocks.
- Begin conducting telephone interviews of housing units that mailed in a completed questionnaire and that could be clearly linked to a telephone number.
- After the initial census nonresponse follow-up, conduct a personal visit interview at every housing unit on the independent list not already interviewed by telephone.
- Match the results of the A.C.E. interview to the census and vice versa.
- Search the census records for duplicates.
- Resolve cases that require additional information for matching by conducting a personal visit follow-up interview.
- Use the information from other, similar people to impute missing information.
- Categorize the A.C.E. data by age, sex, tenure, race/ethnicity and other appropriate predefined variables into estimation groupings called post-strata.
- Calculate the coverage correction factors for each post-stratum using the dual system estimator.
- If appropriate, apply the coverage correction factors to correct the initial census data using a model-based estimator and tabulate the statistically corrected census results.

There are a number of assumptions inherent in the A.C.E. Proper application of the dual system estimation (DSE) model requires the A.C.E. be conducted independently of the census and that the rules used to determine correct enumerations are the same as the rules used to determine cases eligible for matches. The DSE model can be sensitive to measurement errors. It is important to obtain consistent reporting of Census Day residence. Inclusion of fictitious persons and errors in matching can directly influence the DSE. There are other assumptions necessary in developing models for handling nonresponse and other missing information. The A.C.E. design was based very much on the theoretical concepts discussed and publicly presented by the Census Bureau in advance of the census. These concepts included careful attention to statistical independence, a strict application of the concepts of sufficient information, and careful attention to balancing the concepts used to measure census misses, as well as census erroneous inclusions. For a more detailed discussion of this approach see Hogan (2000).

### Design Limitations of the A.C.E.

The A.C.E. was designed to measure the household population for large social, economic, ethnic, racial and geographic groups and compare them with the census counts. The results provide a measure of net undercount and a mechanism to correct that net undercount, if that appears advisable. Although the goal of the A.C.E. was to measure the net undercount, it also provides information on the separate components of the net undercount such as omissions and various types of erroneous enumerations in the census. Measures of gross error cannot be obtained directly and exclusively from these components because of the strict definition of "correct" that is needed to implement the dual system estimator. For example, A.C.E. treats census enumerations as not correctly enumerated if they lacked sufficient information for accurate matching. This requirement allows for more precise matching, but increases both the number of nonmatching cases and the number of cases coded as erroneous. A similar strict rule on correct block location of an address also increases both the non-matches and erroneous enumerations. These rules may be inapplicable in the census outside the DSE context.

The design of the A.C.E. does not provide information on very local or unique errors in the census process. Specifically, the A.C.E. was not designed to correct for particular errors made by, say, a census taker or a local census manager, or to correct for local errors in the census address list. The Census Bureau had other programs in place to deal with these issues, such as the quality assurance process, the coverage improvement follow-up, and the local update of census addresses. The A.C.E. was designed, rather, to correct for large systematic errors in census taking, most especially the historic differential undercount.

Finally, the A.C.E. was not designed to measure the undercount for some special population groups such as the group quarters population (including college dormitories, institutions, and military barracks), the population that uses homeless shelters and/or soup kitchens, or the remote areas of Alaska. The Census Bureau instituted specialized procedures for these groups in order to achieve the best count possible. Extending the A.C.E. methods to all of these populations would have been very costly and difficult to implement properly.

## HISTORY

Starting with 1950, every census has included a formal study of the coverage of the population. The 2000 Accuracy and Coverage Evaluation (A.C.E.) is very much a continuation of that tradition.

### 1950 through 1970

The U.S. Census Bureau conducted its first post-enumeration survey, or PES, as part of the 1950 census. The essential elements in a post-enumeration survey are a

second attempt to enumerate a sample of households and, using case-by-case matching, to determine the number and characteristics of people not included in the first census enumeration. This first PES was not based on dual system estimation.

During the next two decades the Census Bureau experimented with alternative coverage measurement methods based on case-by-case matching including a "Reverse Record Check," administrative record checks, and a match to the Current Population Survey. In addition, there were various alternative versions of PES designs.

Soon after the completion of the 1950 census, methods of aggregate demographic analysis for coverage analysis were developed at Princeton University by Ansley Coale and colleagues. See Coale (1955), Coale and Rives (1973), and Coale and Zelnick (1963) for details. Demographic analysis (DA) is the construction of an estimate of the "true" population using birth, death, migration and other data sources. This methodology can provide independent measures of the census net undercount by age, sex, and Black/non-Black; however, it is subject to its own limitations and uncertainties. An important limitation is the lack of data to independently estimate the Hispanic, Asian, and American Indian populations or other detailed demographic groups, such as homeowners or renters. Nor can demographic analysis provide estimates for geographic areas below the national level. In addition, the level of emigration and undocumented immigration must be estimated using indirect methods. Since the U.S. only had reasonably complete birth registration since 1935, sophisticated analysis was needed in 1950 for the population over age 15. Early studies were restricted to the native-born White population, but with time were expanded to include the native-born African-American population as well.

Later work at the U.S. Census Bureau by Jacob Siegel and colleagues expanded the estimates to the total population, with the first official estimates being issued in conjunction with the 1970 census (Siegel, 1974). The 1970 estimates recognized the need to address the problem of "race misclassification in the complete count." By the time of the 1970 census, the population covered by birth registration included those under age 35, with tests of birth registration completeness having been conducted in 1940, 1950, and the mid-1960's. Medicare data now provided a basis for estimates for those over age 65.

However, the difficulty of measuring migration, an important component of DA, gained attention. These studies noted "The figures on net immigration for the 1960 to 1970 decade should be considered as estimates subject to considerable error." Importantly, the estimates did "not include any allowance for...unrecorded alien immigration, particularly illegal immigration." See Siegel (1974) for more details.

During these same decades, the methods of dual system estimation were being refined for use in the human population. Although introduced over a century ago for use in animal populations, dual system estimation was first used with human populations in an important article by Sekar and Deming (1949) that applied the technique to measuring births. Dual system estimation was widely used to measure births and deaths in developing countries during the 1970's in conjunction with important operational and theoretical work. The ideas from dual system estimation soon applied to post-enumeration surveys. See most importantly Marks (1979).

### 1980

The design of the A.C.E. traces most directly to the 1980 Post-Enumeration Program (PEP). This was the first large scale post-enumeration survey to use dual system estimation. In addition, it included several important innovations, as well as important lessons on the design of a PES.

The 1980 PEP was based on a match of people included in the April and the August Current Population Survey to the 1980 census. This match was used to determine the proportion of people counted in the census. It was a sample of people known to exist and be residents of the U.S., and was labeled the Population or P sample.

All matching was done by clerks and technicians. In order to make it possible to do the matching, each person's address needed to be assigned the correct census geographic code (geocoded). This process was slow and error prone.

In addition, a separate sample of census records was drawn. This was known as the Enumeration or E sample. The census records included in the E sample were checked in the office to see if they were duplicated, followed by a field operation to determine whether the people were real, lived at the address on Census Day, and whether the unit was assigned the correct census geographic code (correctly geocoded).

One important concept introduced in 1980 was that of sufficient information for matching. "Sufficient information for matching" means that a record, from either the P or E sample, contains sufficient information, including most importantly a name, to allow accurate matching and follow-up. Records that lack this information are removed from matching, processing and estimation. For the E sample, this exclusion is done in two parts: census imputed records ("non-data-defined") are excluded from the sampling frame, and then sampled "data-defined" records are reviewed for name and other necessary information.

Another concept used earlier but made explicit in 1980 was that of "search area." A person was only considered correctly enumerated if he/she was counted in a specific,

defined area that included the address where he/she should have been enumerated. This "search area" was to be applied to both the P and the E samples.

The 1980 PEP was also, very importantly, the first PES to be, itself, carefully evaluated (Fay et al., 1988). This evaluation proved invaluable to the design of the 1990 PES. Among the important findings were:

- Sampling variances were very high.
- Geocoding a sample of housing units was costly and error prone.
- Drawing independent P and E samples made it very hard to apply the same concepts, especially that of search area.
- Levels of missing data needed to be reduced and methods to account for the missing data needed to be refined.
- Matching needed to be made more accurate and faster.
- An independent sample of people living in institutions proved nearly impossible to match and process, both because the interviews relied on the same set of administrative records and because administrators often refused to give names, even to the Census Bureau.

By 1980, the precision of demographic analysis benefited from the fact that the part of the population not covered by either adequate birth registration data or Medicare data was now reduced to only those 45 to 65 (in 1980). However, immigration, especially illegal/undocumented/unauthorized immigration, remained a problem. Early demographic estimates for 1980, which again did not contain an allowance for illegal immigration, showed a net overcount of the population. However, pathbreaking work by Jeff Passel and colleagues produced the first estimates of the number of illegal immigrants counted in the census. This work was generally validated when data from the Immigration Reform and Control Act (IRCA) produced similar numbers of immigrants applying for legalization.

Although the 1980 PEP was not explicitly designed to correct the census for measured undercount, it was the first PES to be considered in this context. Increased use of census results for congressional, state, and local redistricting, as well as for federal funds allocation highlighted the importance of census accuracy. The voting rights cases of the 1960's (Baker v. Carr (1962), Reynolds v. Simms (1964)) had greatly increased the importance of census data in redistricting. General Revenue Sharing funds, distributed in part based on census data, became an important source of local government revenue in the mid 1970's. The legal and statistical questions were discussed in academic journals and as part of several lawsuits, including influential suits by the City of Detroit and the City of New York. The U.S. Census Bureau's position was that the 1980 PEP was not of sufficient accuracy for this purpose, and this decision was upheld.

### 1990

Building on the knowledge gained in 1980, the Census Bureau made major design changes for the 1990 PES. Important changes included:

- Excluding institutional population and military ships/barracks from the universe.
- The use of a block sample tied to census geographic codes, with the same sample of blocks used for both the P and the E sample.
- Repeated call-back to reduce nonresponse and missing data.
- A computer and computer-assisted clerical matching operation.
- A model to account for missing data taking into account the important covariates.

The design of the estimation cells (post-strata) was completely changed. Following the advice of John Tukey and others, the estimation cells were not restricted to a single state, but allowed to cross state lines. Thus, Hispanics living in Utah could be combined with Hispanics living in Colorado and other mountain states to form one estimation cell, rather than being combined with non-Hispanics living in Utah. A "smoothing model" was used to combine information within Census Region.

The 1990 PES was explicitly designed so that it could be used to adjust the census results. Specifically, model-based methods were developed to carry the estimates down to the smallest census geographic units (blocks) and to include positive or negative whole person records to account for the measured net undercount or overcount. This complete file could then be aggregated to obtain data that was consistent for all geographical levels.

Many lessons were learned in 1990, many having to do with the need for tight operational control and testing. One important statistical lesson concerned the use of the statistical smoothing methods. These methods became highly controversial and became the focus of much statistical analysis and debate. They were not well understood and the U.S. Census Bureau decided to drop the use of smoothing and instead recompute the results with fewer and thus larger estimation cells.

Demographic analysis estimates went very smoothly in 1990 with birth registration and Medicare data covering all but those age 55 to 65. The IRCA data and the work of Jeff Passell and others (see Fay et al., 1988, Chapters 2

and 3) provided an allowance for undocumented immigrants. Further, for the first time, the Census Bureau produced explicit allowances for the uncertainty in the demographic analysis estimates. This analysis showed that the "preferred" or "point" demographic analysis estimates tended to fall at the lower end of the uncertainty range. However, this method of expressing the uncertainty range came under criticism from outside the Census Bureau. Limitations of this method are documented in Robinson et al. (1993) and Himes and Clogg (1992).

The 1990 Demographic Analysis estimates were in general agreement with the results of the 1990 PES. At the national level the two estimates were very close — 1.8 percent undercount for demographic analysis (later revised to 1.7 percent) and 1.6 percent for the PES. At more detailed levels, differences emerged, especially the tendency for the PES to greatly underestimate the undercount for adult African-American males. Taking into account what was known about the biases and uncertainties of each, it seemed clear that both were measuring a real differential undercount even though PES was underestimating the amount for adult African-American males.

### 2000

In the early 1990s, task forces and National Academy of Science Panels suggested that the differential undercount in the census could not be reduced without elaborate enumeration and matching procedures, which are too costly to be carried out except on a sample of the population. In the 1995 and 1996 Census Tests, an alternative "Census Plus" methodology was compared to the DSE. The performance of the DSE was better and subsequent research efforts focused on improving the DSE. Consequently, most of the A.C.E. design can be seen as a continuation and refinement of the 1990 PES design. Among the important refinements are:

- Much larger and better designed block sample.
- Earlier interviewing, including the use of early telephone interviewing.
- Computer-assisted (laptop) telephone and personal interviewing.
- More refined estimation cells (post-strata).
- Explicit collapsing rules to account for small cell size.
- Explicit weight trimming rules in case of extraordinary (outlier) cells.

The survey universe was restricted to the housing unit/household population. All group quarters, not just military and institutional, populations were excluded. Consequently, the A.C.E. estimate of coverage error will be underestimated to the extent there were errors in the group quarters population.

Another concern is the treatment in the DSE of cases involved in the Housing Unit Duplication Operation (referred to as late census adds) and the level of whole person imputations in the census. These records were not included in the A.C.E. matching, processing, or follow-up processes. They were also excluded from the DSE, although properly accounted for in computing the net undercount. It is possible that, had these records been included in the A.C.E. and the DSE, the estimated undercount would have differed. The number of excluded records is much larger than it was in 1990. If the ratio of matches to correct enumerations is the same for the excluded and included cases, the DSE expected value should be nearly the same. However, if the people referred to in the correct cases were either much more likely to have been included in the A.C.E. or much less likely to have been included, then excluding these cases from the A.C.E. would have changed the level of correlation bias and affected the A.C.E. For more detail, see Hogan (2001).

There was a change in the treatment of people who had moved between April 1 and the time of the PES interview. In 1980 and 1990, these "movers" were sampled at their current (i.e. PES Interview Day) address. In the A.C.E., they were sampled at their Census Day, April 1, address.

Although conceptually much the same, the implementation of the "search area" was very different. In 1990, the entire search area was always to be searched for all cases in order to find matches or duplicates, and all cases were "map-spotted" to determine whether they were inside the search area. In 2000, the search of the surrounding blocks was restricted by both targeting and sampling. First, the surrounding block was searched for only certain kinds of cases, specifically cases where there was a likelihood of geocoding error in the basic census process. In addition, a stratified sub-sample was taken for this search, with only some of the initial sample blocks subjected to this extended search. This process was known as Targeted Extended Search, or TES.

Because of the difficulty in explaining and defending the 1990 smoothing methods, smoothing models were not employed. Instead, the A.C.E. relied upon a larger sample size and a more refined set of estimation cells to produce estimates.

Finally, although this was not a separate step, the A.C.E. was subjected to much more exacting specification, documentation and testing than any previous coverage measurement study. Much of the operational success of the A.C.E. can be traced to the care and attention given to documentation and testing.

This document is then very much part of the overall A.C.E. process. It attempts to document, concisely and clearly as well as precisely and accurately, the A.C.E. design.

# Chapter 2.
# Accuracy and Coverage Evaluation Overview

## INTRODUCTION

The Accuracy and Coverage Evaluation Survey (A.C.E.) was designed primarily to measure the net undercoverage or overcoverage in the census enumeration. The methodology used was dual system estimation that requires two independent systems of measurement. The P sample or "Population sample" measured the housing unit population, as did the census, but was conducted independently of the census. This was done by selecting a sample of block clusters, geographically contiguous groups of blocks, and interviewing housing units that were obtained by independently canvassing each block cluster. The results of the P sample were matched to census enumerations to determine the omission rate in the census. Additionally, a sample of census enumerations, the E sample, was selected to measure the erroneous enumeration rate in the census. The E sample was comprised of census enumerations in the same sample block clusters as the P sample. These overlapping samples reduced variance on the dual system estimator, reduced the amount of field activities and their cost, and resulted in efficient data processing.

There were considerable challenges in the implementation of the A.C.E. One of the requirements of the A.C.E. was to produce measures of net undercount or overcount shortly after the census counts were compiled. This was a daunting task because the requirement for independence meant that A.C.E. activities could not interfere, or in any way affect the results of the census enumerations, or vice versa. As with most surveys, the A.C.E. consisted of designing a sample, creating a frame, selecting the sample, conducting the interviews, dealing with nonresponses and missing information, as well as producing the estimates. In addition, the A.C.E. had several matching and field follow-up activities. In order to accomplish these tasks and meet the goals of the A.C.E. in a timely manner, its design was uniquely built around census operations. Additionally, to ensure quality with such a compressed time schedule, it was essential that software systems be written and thoroughly tested prior to the start of an activity.

One census operation that had major influence on the A.C.E. design and estimation plan was the Housing Unit Duplication Operation. As the census questionnaires were being processed, the Census Bureau suspected that there was a significant number of duplicate addresses in the census files. To address the suspected housing unit duplication, the Housing Unit Duplication Operation was introduced in the fall of 2000. See Nash (2000) for further details. The primary goal of this census operation was to improve the quality of the census; however, its design allowed the A.C.E. operations to proceed. Essentially, suspected duplicate housing units were temporarily removed from the census files, while further analysis was done for these cases. Approximately 5.9 million person records were in these suspected duplicate housing units, which were: 1) out-of-scope for the E-sample component of the A.C.E., 2) not available for the person matching including the identification of person duplicates in the E sample, and 3) excluded from the census component in the dual system estimates. Approximately 2.3 million person records were reinstated into the census after the E sample was selected and were reflected in the net coverage estimates. Hogan (2001) showed that excluding these person records from the A.C.E. would not affect the dual system estimates, if the number of P-sample matches was reduced proportionately to the number of E-sample correct enumerations.

This chapter summarizes the major activities of the A.C.E. and indicates their relationship to the census. Subsequent chapters go into considerably greater detail about the methodology of the A.C.E. and are organized as follows:

- Chapter 3. Design of the A.C.E. Sample
- Chapter 4. A.C.E. Field and Processing Activities
- Chapter 5. Targeted Extended Search
- Chapter 6. Missing Data Procedures
- Chapter 7. Dual System Estimation
- Chapter 8. Model-Based Estimation for Small Areas

The intent of this chapter is to provide a broad context for the design of the A.C.E. Here we give a sequential accounting of these activities. Table 2-1 gives the order in which the A.C.E. activities occurred and maps the activities to the chapter where each is discussed in further detail. This table shows the substantial integration of the sampling and operational activities. Figure 2-1 shows the flow of the major activities.

Table 2-1. **Sequence of A.C.E. Activities**

| Activity | Description | Chapter(s) |
|---|---|---|
| 1 | First-phase sampling | 3 |
| 2 | Independent listing | 4 |
| 3 | Second-phase sampling | 3 |
| 4 | Initial housing unit matching/field follow-up | 4 |
| 5 | Targeted extended search | 4 & 5 |
| 6 | Subsampling within large block clusters | 3 |
| 7 | A.C.E. person interviewing | 4 |
| 8 | E-sample identification | 3 & 4 |
| 9 | Person matching and field follow-up | 4 |
| 10 | Missing data processing | 6 |
| 11 | Dual system estimation | 7 |
| 12 | Model-based estimation for small areas | 8 |

Table 2-2 further illustrates the integration of the sampling activities and operations by summarizing the sample size at each phase of sampling and the operations for which the sample is an input. The data collected from each operation is input to the next sampling operation. For example, the first phase of sampling resulted in 29,136 sample areas with almost 2 million housing units. Independent address lists were created for these areas. The results of the independent listing were used in the second phase of sampling.

### Activity 1. First-Phase Sampling

Timing: March through June, 1999; prior to the creation of the census address list.

At the time of the January, 1999 Supreme Court ruling against the use of sampling for apportionment, the Census Bureau was heavily involved in the first phases of sampling for the Integrated Coverage Measurement (ICM). The goal of the ICM was to produce reliable estimates of coverage of each state's total population, and this required a very large sample – a 750,000 housing unit sample was planned. As a result of the Supreme Court ruling, state population estimates for apportionment were no longer key estimates of the coverage survey; instead, the goal was to measure census coverage for national and subnational population domains having different census coverage properties. These estimates could be measured with sufficient precision with a sample of about 300,000 housing units.

Rather than abandoning the effort, i.e., software development, etc., that had already been invested in the ICM, it was more efficient, particularly from a software quality perspective, to complete the sampling for the ICM, and then select a subsample for the A.C.E. The infrastructure for the field staff was being deployed in preparation for the first field operation that started in September, 1999, and the development of the sampling system that was scheduled to begin production in March, 1999 was well underway. There was not adequate time to redesign the A.C.E. sample allocation entirely, select the sample, produce the different listing materials including maps, conduct the listing as scheduled, and ensure a high level of quality in a revised software system. Consequently, the A.C.E. sample design was derived from the ICM design using a double sampling approach. The entire ICM sample was selected as originally planned and then reduced through various steps to yield the A.C.E. target housing unit sample.

The first-phase sampling consisted of:

- Forming primary sampling units.
- Stratifying primary sampling units.
- Systematic sampling of primary sampling units.

The A.C.E. primary sampling unit was the block cluster, a group of one or more geographically contiguous census blocks. To make efficient field workloads, the target size of block clusters was about 30 housing units, although block clusters varied in size. Within each state, block clusters were stratified by size using housing unit counts from a preliminary census address list: small (0 to 2 housing units), medium (3 to 79 housing units), and large (80 or more housing units). Some states included a separate sampling stratum for American Indian Reservations. Within each sampling stratum, a systematic sample of block clusters was selected with equal probability.

This phase of sampling yielded 29,136 block clusters with an estimated 2 million housing units in the 50 states and the District of Columbia.

Table 2-2. **Sample Sizes by Sampling Phase and Operation**

| Sampling phase | Sample size | | Operations |
|---|---|---|---|
| | Areas | Housing units | |
| First-phase | 29,136 | 1,989,000 | Independent listing |
| Second-phase | 11,303 | 844,000 | Initial housing unit matching/follow-up |
| Subsampling within large cluster (P-sample) | 11,303 | 301,000 | A.C.E. person interviewing, person matching/follow-up, dual system estimation |
| E-sample identification | 11,303 | 311,000 | Person matching/follow-up, dual system estimation |

### Activity 2: Independent Listing

Timing: September through early-December, 1999; well before census enumeration began.

Field staff visited the sample block clusters and created an independent address list of all housing units, including housing units at special places. The goal of this operation was to create an independent address frame of all the housing units that were likely to exist on Census Day, April 1, 2000. Since this operation occurred prior to Census Day, any potential housing unit structures were included on the independent address list. Later, during housing unit follow-up, these structures were visited to confirm that they actually contained housing units on Census Day. Since housing units could not be added to the independent address frame in this later operation, but could be removed, it was important to include structures with questionable housing unit status during the independent listing.

This listing consisted of approximately 2 million housing units or potential housing units in the 50 states and the District of Columbia.

### Activity 3: Second-Phase Sampling

Timing: December, 1999 through February, 2000; prior to mailing the census questionnaire.

The second phase of sampling selected block clusters from the first phase to be the final A.C.E. sample areas. Block clusters were stratified using two housing unit counts: 1) a count from the independent listing operation, and 2) a count from the updated census address list as of January, 2000. It was important to reduce the first-phase sample before the next operations, the housing unit matching and field follow-up, to reduce the number of clusters going into those operations. The stratification of the block clusters was done separately by first-phase sampling strata: 1) medium and large strata, and 2) small strata. All first-phase clusters from the American Indian Reservation stratum were retained in the second-phase sample.

**Medium and large strata.** The resulting national sample allocation was roughly proportional to state population with some differential sampling within states. The two goals of the differential sampling were: 1) to provide sufficient sample to support reliable estimates for several sub-populations, and 2) to reduce the variance contribution due to clusters with the potential for high omission or erroneous enumeration rates. These clusters were identified and put into separate sampling strata by comparing the consistency of housing unit counts between the independent list and the updated census list for each cluster.

**Small cluster stratum.** Conducting interviews and follow-up operations in small block clusters is much more costly per housing unit than in medium or large block clusters. Lower sampling rates were, therefore, used in this stratum. However, two considerations were taken into account in establishing the lower rates. One goal was to avoid having small clusters with an overall probability of selection much lower than the probability of selection of other clusters in the sample. A second goal was to have higher probabilities of selection for small clusters in which the number of housing units was greater than the expected 0 to 2 housing units. These two goals attempted to reduce the contribution of small clusters to the variance of the dual system estimates. Small block clusters with the potential for high erroneous enumeration or nonmatch rates were retained at higher rates. The second-phase sample contained 11,303 block clusters for the 50 states and the District of Columbia.

### Activity 4: Initial Housing Unit Matching and Field Follow-Up

Timing: February through April, 2000; prior to census nonresponse follow-up.

The objectives of these operations were:

1. Create a list of confirmed A.C.E. housing units in order to:
   - obtain the best list of housing units to facilitate person interviewing in later activities.
   - have better control of the final A.C.E. housing unit sample size.
2. Establish a link between the A.C.E. and census housing units in order to:
   - identify the A.C.E. housing units eligible for telephone interviewing.
   - facilitate overlapping P and E samples.
3. Identify potential geocoding errors in order to:
   - establish the targeted extended search sampling frame.
   - identify sample areas for which the creation of a new independent address list, or relisting, was necessary.

**Housing unit matching.** The housing units on the census address list in January, 2000 were matched to the A.C.E. independent address list. First, the addresses were computer matched. The computer matching was followed by a clerical review of the computer match results in an automated environment intended to find additional matches using supplemental materials. There was also a clerical search, limited to the block cluster, for duplicate housing units during this phase of the matching. Possible duplicates in both the A.C.E. and the census were identified.

**Housing unit follow-up.** In some cases, the computer and clerical matching were not able to determine the status of a housing unit. Field staff visited these cases to get more information about these housing units. After matching, the cases which were not matched, possibly matched, or possible duplicates were sent to the field for follow-up interviews. Some of the matched cases were also sent for additional information. The field follow-up was designed to determine if a housing unit existed, if it existed in the block cluster, or if different addresses were referring to the same housing unit.

### Activity 5: Targeted Extended Search

Timing: May, 2000.

The targeted extended search was designed to improve the accuracy of the dual system estimate by searching for matches, correct enumerations and duplicates one ring beyond the sample block cluster. The operation was implemented in a subset of A.C.E. block clusters selected through a combination of certainty and probability sampling.

There are census geocoding errors of exclusion and inclusion in the A.C.E. sample block clusters. Census geocoding errors of exclusion (i.e., housing units miscoded in the census so they appear to be outside the A.C.E. block cluster) affect the P-sample match rate. Census geocoding errors of inclusion (i.e., housing units miscoded in the census to appear inside the block cluster) affect the erroneous enumeration rate in the census or E sample. If the census housing unit is omitted from the sample block cluster, the P-sample household can not be matched. This yields a lower match rate. On the E-sample side, if a housing unit is included in the sample block cluster due to a geocoding error, the E-sample people will be considered erroneously enumerated.

The primary motivation for using an extended search area was to reduce the sampling variance of the dual system estimates due to census geocoding error. Even though the extended search allowed more P-sample people to be matched and more E-sample people to be converted to correct enumerations, the expected value of the dual system estimate should not be affected as long as the two samples were treated equally with respect to the search area. Another benefit is that the extended search makes the dual system estimate more robust by protecting against potential bias due to P-sample geocoding error.

Previous census evaluations have shown that geocoding errors are highly clustered. The targeted extended search was designed to take advantage of the distribution of geocoding errors by focusing on those clusters that contain the most potential geocoding errors. The implementation of this operation resulted in dual system estimates with more precision.

The initial housing unit matching results were used to identify the A.C.E. housing unit nonmatches and census housing unit geocoding errors. Clusters without A.C.E. housing unit nonmatches or census geocoding errors were out-of-scope for the targeted extended search sampling. Changes to the census inventory of housing units after January, 2000 were not reflected in the housing unit matching used to identify targeted extended search clusters.

Only whole households of nonmatched people were eligible for the extended search during person matching. Partial household nonmatches (i.e., some household members were matches) were not as likely to indicate that the housing unit was a geocoding error.

### Activity 6: Subsampling Within Large Block Clusters

Timing: April and May, 2000; during census nonresponse follow-up.

Subsampling was used in large block clusters for the final selection of housing units to participate in the P sample. The objective was to reduce costs and yield manageable field workloads without seriously affecting the precision of the A.C.E. by taking advantage of the high intra-class correlation expected in large block clusters. Since the large block clusters had a higher initial probability of selection than medium block clusters, the reduction in sample size had a fairly minor effect on the precision of the A.C.E. estimates. The subsampling of housing units within large clusters brought the overall probability of selection of these housing units more in line with housing units in the medium clusters.

Any block cluster with 80 or more confirmed A.C.E. housing units, based on the initial housing unit match, was eligible for this housing unit reduction. The reduction of housing units within a large block cluster was done by forming groups of adjacent housing units, called segments, and selecting one or more segments for A.C.E. person interviewing. The segments had roughly equal numbers of housing units within a block cluster. Segments of housing units were used as the sampling unit in order to obtain compact interviewing workloads and to facilitate the identification of an overlapping E sample. The A.C.E. housing units that were retained after all of the subsampling comprise the P sample.

After the reduction of housing units within large block clusters was completed, the A.C.E. interview sample size for the 50 states and the District of Columbia was approximately 300,000 housing units.

### Activity 7: A.C.E. Person Interviewing

Timing: April through mid-June, 2000 for the telephone phase; Mid-June through mid-September, 2000 for the personal visit phase; after census enumeration was complete.

The goal of the A.C.E. person interview was to provide a list of persons who lived at the sample address on Census Day, as well as those who lived at the address at the time of A.C.E. interviewing. The A.C.E. person interview was conducted using a Computer Assisted Personal Interview (CAPI) instrument.

To get an early start on interviewing, a telephone interview was conducted at households for which the census questionnaire was data-captured and included a telephone number. Both households with mail returns and enumerator-filled questionnaires were eligible for telephone interviews. Certain types of housing units, such as those without house number and street name, were not eligible for a telephone interview. All remaining interviews following the telephone operation were conducted in person. However, some nonresponse conversion operation interviews and interviews in gated communities or secured buildings were conducted by telephone.

The person interview was conducted only with a household member during the first 3 weeks of interviewing. If an interview with a household member was not obtained after 3 weeks, an interview with a nonhousehold member was attempted. This was called a proxy interview. Proxy interviews were allowed during the remainder of the interviewing period. During the last 2 weeks of interviewing a nonresponse conversion operation was attempted for the noninterviews using interviewers who were considered to be the best available.

### Activity 8: E-Sample Identification

Timing: October, 2000.

The E sample consisted of the census enumerations in the same sample areas as the P sample. All data-defined census person records in the A.C.E. block clusters were eligible to be in the E sample.[1] To be a census data-defined person, the person record must have two 100-percent data items filled. Name was not required for the person record to be considered data-defined, but could be one of the two items required to be data-defined. Like the P sample, it was sometimes necessary to subsample the census housing units in a cluster when it contained a large number of census housing units. The goal of the E-sample identification was to create overlapping P and E samples in an effort to reduce person follow-up workloads. An overlapping P and E sample is not necessary, but improves both the cost effectiveness of the subsequent operation and the precision of the dual system estimates.

If a block cluster had fewer than 80 census housing units, then all of the census housing units in the block cluster were in the E sample. For block clusters with 80 or more census housing units, the within-cluster segments of adjacent housing units defined for the P-sample reduction were mapped on to the census records. This was possible when a link between the census and A.C.E. housing unit was established during the initial housing unit matching. Using specific rules, census housing units that did not have this link were assigned to a segment. The segment selected for the P sample was selected for the E sample. If the sample segment contained 80 or more census housing units with no established link to an A.C.E. housing unit, then a systematic sample of these housing units was selected to reduce the E-sample person follow-up workloads.

This resulted in approximately 311,000 census housing units in the E sample for the 50 states and the District of Columbia.

### Activity 9: Person Matching and Field Follow-Up

Timing: October and November, 2000.

**Insufficient information for matching.** Rules were established for determining which person records had sufficient information for matching. These rules were established and applied before the start of the matching operation to avoid introducing potential bias into the matching results. Both the P and E samples used the same rules. Each person record required a complete name and two other characteristics.

**Person matching.** All P-sample persons who lived at each sample housing unit on Census Day were matched to the people enumerated in the census to estimate the match rate. Census persons in the E sample who matched to the P sample were considered to be correctly enumerated. The E-sample person records that did not match to the P sample were interviewed during field follow-up operations to classify them as correctly or erroneously enumerated. This matching was a computer operation with clerical review. Variables such as name, address, date of birth, age, sex, race, Hispanic origin, and relationship were used to identify matches between the P sample and census enumerations. Duplicates were identified in both the P sample and E sample. If a case qualified for targeted extended search, the search for matches and duplicates was extended to the ring beyond the sample block cluster.

**Person follow-up.** The person follow-up interview collected additional information that was sometimes necessary for the accurate coding of the residence status of the nonmatched P-sample people and the enumeration status of the nonmatched E-sample people. The goal of this operation was to confirm that ambiguous P-sample nonmatches actually lived in the sample block cluster on Census Day. Thus, follow-up interviews for P-sample nonmatched cases were carried out when there was a possibility the residence status was not correct. Similarly,

[1] Excludes data-defined person records temporarily removed from the census.

E-sample nonmatch cases were subject to follow-up interviews to determine if they were correctly or erroneously enumerated in the block cluster. Possible matches were interviewed to resolve their match status. There were also other cases sent to follow-up, such as matched people with unresolved residence status and other types of cases considered to have the potential for geographic errors in the P sample. The person follow-up interview used a paper questionnaire. Interviewers gathered information that permitted each person to be coded as a matched resident/nonresident or a nonmatched resident/nonresident of the block cluster on Census Day. There was considerable emphasis on obtaining a knowledgeable respondent before the follow-up questions were asked. After the follow-up interview was completed, the results were reviewed by clerks who assigned final status to these cases using an automated system.

### Activity 10: Missing Data Processing

Timing: December, 2000 through the early part of January, 2001.

Since the results of the matching operation were to be used in the estimation phase of the A.C.E., it was necessary to determine the match, correct enumeration and residence status of all sample cases. When these could not be resolved through computer and clerical matching or through field follow-up interviews, the match, correct enumeration, or residence probabilities were imputed based on the distribution of outcomes of the resolved follow-up interviews. Also, as in the census, some respondents did not answer all the questions in the A.C.E. interview which were needed for estimation. If the variables tenure, sex, race, Hispanic origin, or age were blank for P-sample individuals, the missing information was imputed based on the distribution of the variable within the household, the overall distribution of the variable, or using hot-deck methods, depending on the variable. Imputation for missing information in the E sample was resolved in the census processing. Finally, a noninterview adjustment was made to account for the weights of households that should have been interviewed in A.C.E., but were not.

### Activity 11: Dual System Estimation

Timing: Late January, 2001.

Dual system estimation was used to estimate the net undercount or overcount of the household population included in the census. Coverage estimates of persons living in group quarters or in Remote Alaska areas were not made.

The term dual system estimation is used because data from two independent systems are combined to measure the same population. After matching to the census, the P sample was used to measure the omission rate in the census. The E sample was used to measure the erroneous enumeration rate in the census. The dual system estimator assumes that all persons have the same probability of being captured in the census. This is obviously an oversimplification of the existing situation. Post-stratification sharply reduced the likelihood that this assumption would bias the results, since it only requires equal capture probabilities within post-strata.

**Post-stratification.** Dual system estimation was used to calculate the proportion of persons missed in each of a number of relatively homogeneous population groups called post-strata. The post-strata for the Census 2000 A.C.E. were defined by the variables: race/Hispanic origin domain, age/sex, tenure, census region, metropolitan statistical area size/type of enumeration area, and census return rate. A complete cross-classification of these variables would have unnecessarily increased the variances of the estimates due to small expected sample sizes in many of the post-strata. Consequently, many of the detailed cells were combined. In the United States, there were 448 potential post-strata which were collapsed to 416 post-strata on the basis of small observed sample sizes or high coefficients of variation.

**The dual system estimate.** The dual system estimate (DSE) for each post-stratum was defined by:

$$\hat{\mathrm{DSE}} = \mathrm{DD} \times \frac{\mathrm{CE}}{\mathrm{N}_e} \times \frac{\mathrm{N}_p}{\mathrm{M}}$$

where DD was the number of data-defined persons in the census at the time of A.C.E. matching,[2] CE was the weighted estimate of the number of people in the census who were correctly enumerated, $N_e$ was the weighted estimate of the number of people in the census, $N_p$ was the weighted estimate of the number of people found by the independent A.C.E. collection procedures, and M was the weighted estimate of the number of persons found by the independent A.C.E. collection procedures who were matched to persons enumerated in the census.

### Activity 12: Model-Based Estimation for Small Areas

Timing: February, 2001.

Activities 1 through 11 were designed to provide estimates of net coverage for Census 2000. These estimates can serve two purposes. One purpose was to provide information on the quality of the census so that analysts can make more intelligent use of the data, and to help the Census Bureau improve procedures for future censuses. The second purpose was to have a basis for adjusting the census counts for net coverage, if deemed appropriate. The sample sizes used in the A.C.E. provided adequate

[2]The data-defined persons term excludes cases temporarily removed from the census.

reliability for such estimates for the U.S. as a whole, and for major geographical areas. However, the sample sizes were too small to provide reliable estimates for most states, counties, cities, and the thousands of other municipalities that normally make use of census data. As a result, model-based estimation was used in these areas.

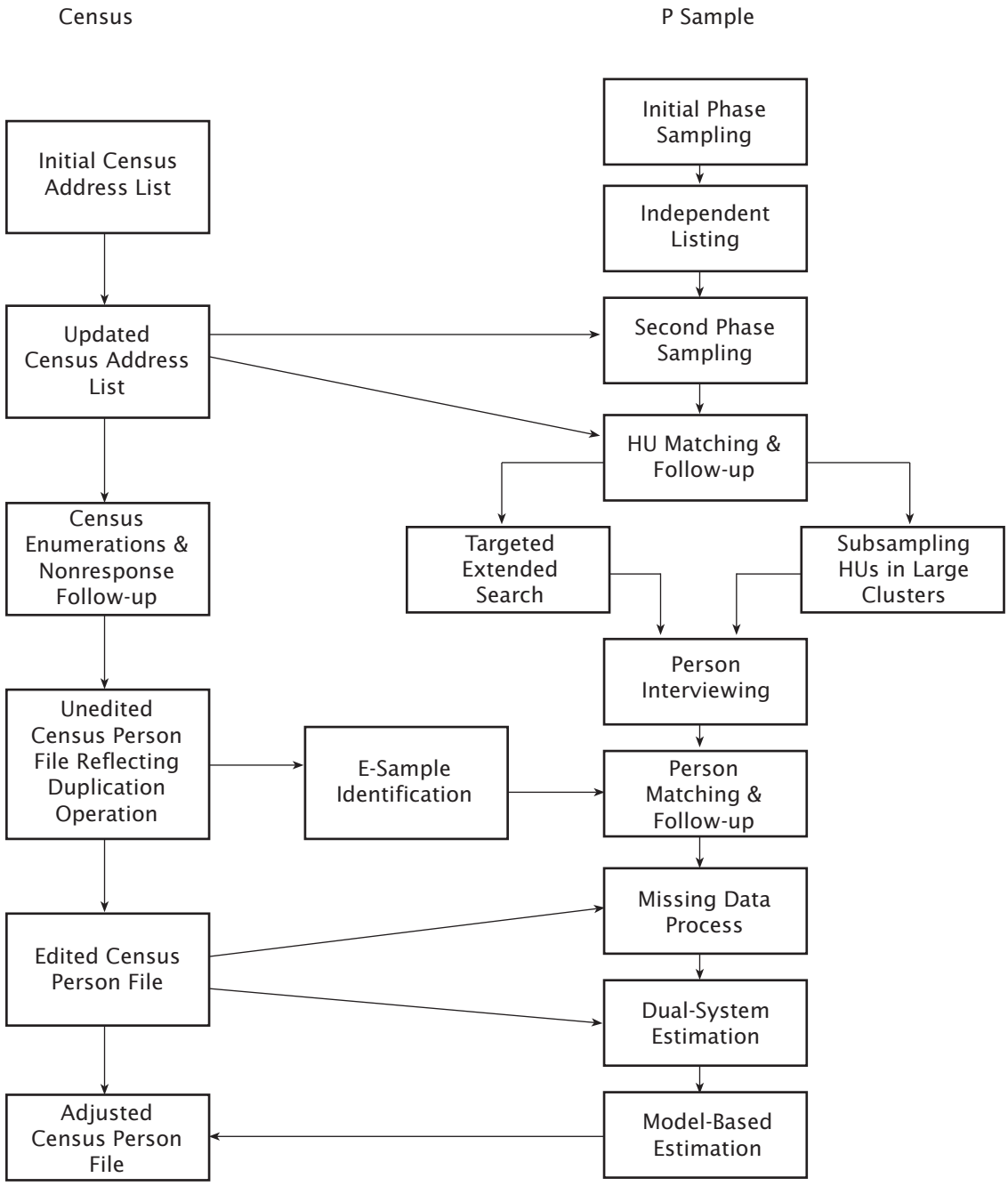
Model-based estimation treats the coverage correction factors as uniform within a given post-stratum. Another way of saying this is that the coverage error rate for a given post-stratum is assumed to be the same within all geographic areas. This assumption is obviously an oversimplification, and small errors are introduced. However, the model-based estimates provide a consistent set of estimates in which the sum of the population counts for small areas are equal to the dual system estimates of much larger areas (e.g., the U.S. total, regions, etc.).

Coverage correction factors were obtained by dividing the dual system estimates by the census counts of persons in housing units. Persons in group quarters were not adjusted for net coverage. Coverage correction factors for population groups that generally had good coverage were close to 1.00. Population groups with poor coverage had coverage correction factors higher than 1.00, while coverage correction factors less than 1.00 in a post-stratum occurred when erroneous enumerations rates in the census exceeded omission rates.

A coverage correction factor was calculated for each post-stratum. If a post-stratum was estimated to have more persons than the census count, within each block a random sample of the appropriate size of census people in the post-stratum was selected. The data of the selected people were replicated in their blocks with a weight of +1. If a post-stratum was estimated to have fewer people than the census count, within each block a random sample of the appropriate size of people in the post-stratum was selected. The data of the selected people were replicated in their blocks with a weight of -1. Under this procedure no reported data for any individual was removed from the Census 2000 data files. A controlled rounding procedure was used to produce integer-valued model-based estimates at various geographic levels.

Estimates were made at various levels by aggregating the data from the appropriate blocks and/or post-strata.

Figure 2-1
**Flow of Major A.C.E. Activities With Respect to Major Census Activities**



# Chapter 3.
# Design of the A.C.E. Sample

## INTRODUCTION

The A.C.E. sample design was a multiphase, national sample of 301,000 housing units. Its development was heavily influenced by its planned predecessor, the Integrated Coverage Measurement survey (ICM). Initial plans for Census 2000 were for a one-number census corrected for coverage based on the ICM. A primary purpose of the ICM was to produce direct state estimates of coverage with sufficient reliability for apportionment population counts. This called for a state-based design and a much larger overall national sample of 750,000 housing units.

The January 1999 Supreme Court ruling against the use of sampling for apportionment resulted in a change of plans for the Census 2000 coverage survey for which the primary goal became the production of reliable national census coverage estimates, and of selected sub-populations. This did not require as large a sample.

The A.C.E. sample design was derived from the ICM sample design. By the time the change of plans for the Census 2000 coverage survey occurred, many operational plans for the ICM were too far advanced to make significant changes required for a newly conceived sample design plan. The implementation plans and software systems for creation of the sampling frame and selection of the ICM sample were moving along and almost ready to start. Much of the field office infrastructure and staffing was being put in place for the first field operation under the ICM sample plan. It was critical to proceed as planned in order to meet schedules.

The A.C.E. sampling plan was thus developed as a multiphase design. The much larger ICM sample was first selected. Field staff canvassed the sample areas to create an independent address list. Then, using updated measures of size from the field canvass, the ICM sample was re-stratified and reduced with differential probabilities of selection to create the A.C.E. sample design.

Sections on the A.C.E. sample and its design are directed to a general audience. They provide results of the A.C.E. sample along with a broad overview of the sample design. Later sections of this chapter provide a more in-depth description of the A.C.E. design and are available for readers who desire greater detail.

## A.C.E. SAMPLE OVERVIEW AND RESULTS

The A.C.E. consisted of two parts. The Population Sample, P sample, and the Enumeration Sample, E sample, have traditionally defined the samples for dual system estimation. Both the P sample and the E sample measured the same household population. However, the P-sample operations were conducted independent of the census. The E sample consists of census enumerations in the same sample areas as the P sample. After matching with the census lists and reconciliation, the P sample yields an estimated rate at which the population was missed in the census whereas the E sample yields an estimated rate at which enumerations were erroneously included in the census. Combining them yields an A.C.E. estimate of net census coverage of the household population.

The Accuracy and Coverage Evaluation had three sampling phases:

1. **First-phase sample.** The selection of the ICM sample, comprising a large number of sample areas for which a list of housing unit addresses was created independent of the census.
2. **Second-phase sample.** The reduction of the first-phase sample which resulted in the A.C.E. sample areas.
3. **Third-phase sample.** The reduction of housing units by subsampling within unusually large A.C.E. sample areas.

Table 3-1 summarizes the A.C.E. sample size after each phase of sampling for the United States. The dates given in the table are the production dates. The housing unit counts are approximate, based on the best known information at the time of the particular sampling phase.

Table 3-1. **Census 2000 A.C.E. Sample Sizes by Sampling Phase**

| Start and finish date | Sampling phase | Sample areas | Estimated housing units |
|---|---|---|---|
| March, 1999 thru June, 1999 | First-phase | 29,136 | 1,989,000 |
| December, 1999 thru February, 2000 | Second-phase | 11,303 | 844,000 |
| April, 2000 thru May, 2000 | Third-phase | 11,303 | 301,000 |
| | within-cluster reduction | 3,153 | 106,000 |
| | no within-cluster reduction | 8,150 | 195,000 |

## SURVEY CHARACTERISTICS AND THE A.C.E. SAMPLE DESIGN

### Main Characteristics of the A.C.E. Sample

**The A.C.E. sample:**

- Is a probability sample of 301,000 housing units in 11,303 sample areas for the United States.
- Yields estimates of net census coverage of persons in households and housing units for the nation excluding Remote Alaska.
- Has independent samples in each state and the District of Columbia, but there are no state-based design criteria.
- Has total state sample sizes roughly proportional to population size with the exception that the smaller states have additional sample; these smaller states have similar sample sizes.
- Uses some differential sampling within states for areas that may contribute disproportionately to total variance or have higher concentrations of historically undercounted population groups.
- Has a separate sample of American Indian Reservation and other associated trust lands.
- Uses updated measures of size at each phase of sampling.
- Balances operational limitations such as field workloads and statistical issues such as weight variation.

### Overview of the Design

The A.C.E. uses a multiphase sample to measure the net coverage for the household population in Census 2000. The national sample, 301,000 housing units in 11,303 sample areas, was distributed among the 50 states and the District of Columbia roughly proportional to population size except for the smaller states that had their samples increased.

**Primary sampling unit.** The block cluster was the Primary Sampling Unit (PSU) for the A.C.E. Each block cluster consisted of one or more geographically contiguous census blocks. Each block cluster contained on average 30 housing units, which was an efficient interviewer workload. An important block cluster characteristic was well-defined, physical boundaries. Ambiguous block cluster boundaries could potentially lead to errors of omission or erroneous inclusion in the A.C.E. sample.

**Phases of the A.C.E. sample.** Three phases of the A.C.E. sampling were:

1. Selection of an initial sample of approximately 30,000 block clusters for which the field staff developed an independent list of housing unit addresses.
2. Selection from the initial sample results of a subsample of block clusters for the A.C.E. sample based on the results of the independent list.
3. Selection of a subsample of housing units within large block clusters.

**First phase consisted of the selection of a systematic sample in each state.** In the first phase of the A.C.E. sampling, block clusters in each state were classified by size into four mutually exclusive groups known as sampling strata: (1) clusters with 0 to 2 housing units (small stratum), (2) clusters with 3 to 79 housing units (medium stratum), (3) clusters with 80 or more housing units (large stratum), and (4) clusters on American Indian Reservations with three or more housing units (American Indian Reservation stratum). Block clusters with 80 or more housing units were selected with higher probability than medium clusters in this phase because housing units in large clusters were subsampled in a later operation, bringing the overall probability of selection–the inverse of the sampling weight–for housing units in these clusters more in line with the overall selection probabilities of housing units in medium clusters. Within each sampling stratum, clusters were sorted and a systematic sample was selected with equal probability.

**Second phase involved the reduction of the ICM first-phase sample to the level desired for the A.C.E.** In the second phase, the block clusters from the medium and large sampling strata were re-stratified based on the estimated demographic composition of the block clusters and the relationship between the housing unit count from the independent list and the January 2000 updated census address list. This was done separately for

the medium and large strata within each state. These substrata are referred to as "reduction strata." Within each reduction stratum, the clusters were sorted, and a systematic sample was selected with equal probability within each reduction stratum. This reduction used different selection probabilities across the reduction strata within a state and across states.

Next, using housing unit counts from the independent list and the January 2000 updated census address list, the small block clusters were stratified within each state by size, and systematic samples were selected from each stratum with equal probability. All clusters from the small sampling stratum with 10 or more housing units based on the updated information were retained. All clusters from the small sampling stratum that were on American Indian land as well as List/Enumerate clusters were also retained. The second phase of sampling was not done for the American Indian Reservation sampling stratum.

**The third phase consisted of the sample reduction of housing units within large block clusters.** In the third phase of A.C.E. sampling, a subsample of housing units was selected within large clusters. If a cluster contained 79 or fewer housing units, all the housing units were included in the A.C.E. sample. In clusters with 80 or more housing units, a subsample was selected to reduce the cost of data collection. This phase of sampling resulted in lower variation of selection probabilities for housing units within the same reduction stratum because the large clusters had a higher probability of selection at the first phase. This subsampling was done by forming groups of adjacent housing units, called segments. A systematic sample of segments within each cluster was selected. All housing units in the selected segments were included in the A.C.E. sample.

**The P sample and the E sample.** The P sample consisted of the households used for the A.C.E. interviews that were conducted in these selected block clusters and block cluster segments. The E sample was the set of census enumerations in these same block clusters and block cluster segments.

### Measures of Size

As stated earlier, the A.C.E. sample design used updated measures of size at each phase of sampling.

**First-phase sample.** The block cluster measure of size for the first-phase sample was based on preliminary census files existing in the spring of 1999. Ideally, the source of the block cluster measure of size would have been the Decennial Master Address File, the base file of census addresses for the decennial programs. However, the first version of this file was not available until the summer of 1999, too late for use in the block clustering. Instead, the first-phase measure of size was typically the higher of the preliminary census housing unit count or the 1990 census address count for a block cluster containing city-style addresses, house number and street name. For block clusters with non-city-style addresses, the measure of size was the preliminary 2000 census housing unit count. The rules for determining which housing units on the preliminary 2000 census files would eventually move forward to the Decennial Master Address File had not been defined, so the block cluster measure of size was based on a reasonable set of criteria, but not the final set.

**Second-phase sample.** For the second phase of sampling, the block cluster measure of size was the count of housing units on the list of housing unit addresses created independently of the census in the fall of 1999. The reduction of the medium and large block clusters used a preliminary count of these housing units, which was a clerical tally of housing units from the listing sheets. The small block cluster reduction used the count of housing units from the independent listing sheets after the addresses had been keyed. For the most part, the preliminary and the keyed counts for each block cluster were identical, but for some clusters there were differences. Using a preliminary count was necessary because the medium and large cluster reduction had to be completed before the keying of the independent listing sheets was done.

**Third-phase sample.** For the third phase of A.C.E. sampling, the block cluster measure of size was the housing unit count resulting from the housing unit matching and follow-up operation. This operation confirmed the count resulting from the independent listing and removed any nonexistent addresses from the sampling frame.

## FIRST PHASE OF THE A.C.E. SAMPLE DESIGN

The sample selection during the first phase consisted of three major steps:

1. Definition of the primary sampling units.
2. Stratification and allocation of the primary sampling units within each state.
3. Selection of the primary sampling units within each state.

### Defining the Primary Sampling Unit

The Primary Sampling Units (PSUs) for the A.C.E. were block clusters. The PSUs were delineated in such a way that they encompass the entire land area of the United States, except for extremely remote areas of Alaska. Each block cluster consisted of a census block or several geographically contiguous census blocks. They contained an average of 30 housing units. The land area for each PSU was made reasonably compact so it could be traversed by an interviewer in the field without incurring unreasonable costs.

**Why the block cluster?** A basic design decision, which was a continuation from the 1990 Post-Enumeration Survey, was that the PSU would be a block cluster, a single block or a group of adjacent blocks established for the collection of Census 2000 information. These blocks may be standard city blocks or irregularly shaped areas with identifiable political or geographic boundaries. Using block clusters as PSUs, instead of counties or county groups that are more commonly used in national surveys, improved the precision considerably with only a modest increase in costs.

An alternative sample design was considered that would have defined PSUs by segmenting whole blocks into smaller components (roughly one-half of a block.) The alternative design would likely have resulted in reduced sampling error, but was rejected because it would increase costs (primarily due to increased matching workloads and interviewer travel) and probably would have resulted in matching errors due to problems in identifying (spatially) the PSU boundaries.

**Goals of block clustering.** Block clusters were formed to meet both statistical and operational goals. In the Census 2000 Dress Rehearsal, a small census block was by definition a single block cluster. This rule led to a large number of small block clusters that could potentially exert undue influence on the final population and variance estimates. One feature of block clustering under the Census 2000 A.C.E. procedure was to combine small census blocks with adjacent census blocks, if the neighboring block contained one or more housing units. This change in the treatment of small census blocks had an enormous impact on the number of small block clusters, which was reduced by approximately 65 percent as seen in Table 3-2. Still, many block clusters contained zero housing units. Roughly 70 percent of the zero housing unit blocks occurred in sparsely populated areas. Without populated neighboring blocks, these zero housing unit blocks remained stand-alone zero block clusters.

The two operational goals of forming block clusters were to increase listing efficiency and to reduce the chance of listing error. The first goal was met by collapsing census blocks to produce block clusters that were geographically compact and which averaged about 30 housing units, a manageable workload. The second goal was to create block clusters that were well defined to minimize the chance that the cluster would be listed incorrectly. For example, a listing error may result when a census block has an invisible or nonphysical boundary such as city limits making it unclear where the block boundary was. As a result, census blocks separated by invisible boundaries were always combined.

**Limitations.** As mentioned earlier, the block cluster measure of size for the first phase was based on preliminary census address counts. Some census operations that helped build the census address list were not available at the time block clustering started. Instead, a snapshot of the best known information was used. This presented some limitations with the data used for block clustering.

- Address limitations: The results of the Block Canvassing and Local Update of Census Addresses (LUCA) operations were not incorporated into the census address list in time for block clustering. Block Canvassing was a Census 2000 field operation in mailout/mailback areas (mostly city-style addresses). The Census Bureau sent staff into the field to canvass their assignment areas and provide updates to the address list such as corrections, adds, or deletes. Local Update of Census Addresses was also a Census 2000 program that provided an opportunity for local and tribal governments to review and update address information in the census address list.

Table 3-2. **Accuracy and Coverage Evaluation: Block Cluster Summary Statistics**[1]

| | Preliminary number of housing units | | | |
|---|---|---|---|---|
| | 0 - 2 | 3 - 79 | 80+ | Total |
| Number of census blocks[2] | 2,969,000 | 4,009,000 | 245,000 | 7,223,000 |
| Number of block clusters | 1,029,000 | 2,486,000 | 252,000 | 3,767,000 |
| Number of blocks per cluster[3] | 1.3 | 2.2 | 1.5 | 1.9 |
| Number of housing units per cluster | 0.3 | 29.2 | 181.9 | 31.5 |

[1]The United States and Puerto Rico are included in these summary statistics.

[2]Count of census collection blocks before clustering and before block suffixing. Does not include water blocks or census blocks in Remote Alaska.

[3]These numbers are not the first row divided by the second row. They are the number of census blocks in each block cluster size category divided by the number of block clusters in each category. For example, if two census blocks with 40 housing units collapse to form an 80 housing unit block cluster, those two census blocks are counted in the 80+ category for the number of blocks per cluster computation. Block clustering can combine across categories; therefore, the first and second rows are not consistent.

- Geographic limitations: Each block in the census address list had a Type of Enumeration Area (TEA) assignment. For Census 2000, TEA is a classification that identified both the census enumeration method and the method used to compile the census address list. The block clustering operation occurred concurrently with the census review of TEA assignments to ensure the most complete coverage of the area. This review process sometimes changed the TEA assignment of blocks after the block cluster was defined. On a few occasions, this resulted in a block cluster consisting of blocks that had different methods for compiling the census address list. For example, a block cluster consisted of three blocks, and all three blocks had a TEA assignment of "Block Canvassing and Mailout/Mailback" at the time of block clustering. After the census TEA review, one of those blocks was converted to an "Address Listing and Update/Leave" TEA assignment. For a complete list of TEAs for Census 2000, see the attachment or visit http://www.geo.census.gov/mob/homep/teas.html.

**General rules for defining block clusters.**

- Block clusters were formed by combining neighboring Census 2000 blocks.
- Block clusters did not cross specific geographical boundaries. Among these were county, interim census tract, Local Census Office, TEA group, military area, and American Indian Country. For "TEA groups," blocks from certain TEAs could be clustered together if the TEAs had the same method for compiling the address list. American Indian Country refers, collectively, to lands that are American Indian Reservation or other trust lands, tribal jurisdiction statistical areas (now known as Oklahoma Tribal Statistical Areas), tribal designated statistical areas, and Alaska native village statistical areas.
- Blocks separated by an invisible boundary, a city line, for example, were clustered except for the situations described above.
- Whenever possible, small census blocks, those with fewer than three housing units, were clustered with neighboring census blocks containing housing units to reduce the total number of small block clusters. If there were no neighboring census block with housing units, the small census block was a cluster by itself.
- To prevent block clusters from becoming too large with respect to housing unit size, census blocks with 80 or more housing units were generally not clustered with other census blocks.
- In addition to the criteria of unit size, any block larger than 15 square miles was generally a block cluster by itself.

These rules produced 3.8 million block clusters, about half the 7.2 million non-suffixed census blocks. The block clusters had an average of 29.2 housing units per medium block cluster and an average of 31.5 overall. The number of small block clusters also decreased from nearly three million to about one million, an approximate 65 percent reduction from the Census 2000 Dress Rehearsal rules of defining a small block to be a cluster by itself. However, since about 70 percent of small blocks occurred in less populated areas with little or no population to combine, many single zero-housing unit block clusters were formed.

### Stratifying and Allocating the Primary Sampling Units

**Stratifying the first-phase sample.** Prior to sampling, block clusters were stratified according to the expected number of housing units and the American Indian Reservation (AIR) status of the block cluster. The four sampling strata and their definitions are presented in Table 3-3.

**Allocating the first-phase sample.** As stated earlier, the Census Bureau was preparing to conduct the ICM, a much larger coverage measurement survey of 750,000 housing units, when the use of sampling for apportionment counts was disallowed by the Supreme Court in January, 1999. To keep the coverage measurement survey on schedule, the Census Bureau went ahead with the plans to select the ICM sample and create independent address lists. This was followed by the subsampling of the first-phase sample to produce the A.C.E. sample design.

The first-phase sampling plan was a national sample of 30,000 block clusters: 25,000 medium and large block clusters and 5,000 small block clusters. Included in the 25,000 block clusters was a separate sample of block clusters for American Indian Reservations.

It is important to point out that the allocation of the 25,000 medium and large block clusters was dependent on the ICM sample design and under the assumption of roughly 30 housing units per block cluster. The allocation of the 5,000 small block clusters to the states and the separate American Indian Reservation sample to the states was done prior to defining block clusters for all states, since the first-phase sampling was done on a state-by-state flow-basis. This means that the first-phase sample was selected for some states before the block clusters had been defined for other states. As a result, we used the best information we had at the time to carry out the allocation.

**Medium and large block clusters.** The 25,000 medium and large block clusters were allocated to the states to meet the ICM sample requirements (Schindler, 1998) with some minor modifications. Most states had between 300 to 500 block clusters and the very largest states had an allocation of between 1,000 and 2,000 block clusters.

Table 3-3. **First-Phase Sampling Strata**

| First-phase sampling stratum | Definition |
|---|---|
| Small | 0 to 2 housing units |
| Medium | 3 to 79 housing units |
| Large | 80 or more housing units |
| American Indian Reservation | 3 or more housing units and on American Indian Reservations |

Within each state, the block cluster sample was proportionally allocated to the medium and large sampling strata based on the number of housing units in the sampling stratum:

$$c_{state,k} = C_{state} \times \frac{H_{state,k}}{H_{state}}$$

where,

- $k$ = medium or large sampling stratum;
- $c_{state,k}$ = target number of clusters in sampling stratum k within state;
- $C_{state}$ = target number of A.C.E. first-phase medium and large sample clusters for state;
- $H_{state,k}$ = number of housing units in sampling stratum k within state;
- $H_{state}$ = number of housing units in the medium and large strata in state.

As an example, let's say that 402 total medium and large block clusters were allocated to a particular state. Assuming that there are an expected 9,000 housing units in all clusters in the medium sampling stratum and 12,060 housing units in both the medium and large sampling strata, the target number of clusters from the medium sampling stratum for the state is calculated as follows:

$$C_{state,\ medium} = 402 \times \frac{9{,}000}{12{,}060} = 300.$$

The target number of clusters from the large sampling stratum would then be 102.

**Small block clusters.** Because of cost considerations, small block clusters were generally sampled at a lower rate than either medium or large clusters. An overall allocation of 5,000 small block clusters was chosen because a total of 30,000 block clusters was deemed manageable for creating independent address lists. The high weights resulting from the lower sampling rates were not expected to have a serious impact on the estimates or variances for most clusters selected from the small block cluster sampling stratum. However, for clusters that were initially classified as small, but were observed to have a larger number of housing units, there was concern about high sampling weights disproportionately contributing to variance. In an attempt to avoid the problems associated with the high weights, a larger number of small clusters was initially selected, followed by an independent address list, followed by a subsample to remain in sample. Using updated measures of size for those 5,000 small block clusters in the small cluster reduction helped to target clusters that could have contributed disproportionately to the variance. These initial 5,000 small clusters were allocated to states proportionately to their estimated total number of housing units in small blocks.

Ideally, we would have allocated the 5,000 block clusters proportionally to states based on the number of small block clusters in the state. This was not possible because the first-phase sampling was done on a flow basis.

**American Indian Reservation block clusters.** To ensure sufficient sample for calculating reliable coverage estimates for American Indians living on reservations, we allocated 355 block clusters to American Indian Reservations nationwide. The 355 clusters were allocated to 26 states proportional to the 1990 population of American Indians living on reservations. Small block clusters on American Indian Reservations were not included in these 355 block clusters. These clusters were eligible for selection in the small cluster stratum. Block clusters within states containing little or no American Indian population on reservations were represented in the medium and large strata.

This sample allocation resulted in variable first-phase selection probabilities across the states despite our goal of having proportional allocation of the American Indian Reservation (AIR) sample. This occurred because the average number of housing units per American Indian Reservation block cluster varied across states. To get similar first-phase selection probabilities, we needed to have all of the block clustering completed before allocating the sample. However, the first-phase sampling was done on a flow basis.

### Selecting the Primary Sampling Units Within Each State

**Calculation of the sampling parameters.** The block cluster probability of selection (PS) for each of the four sampling strata in each state is the ratio of the target sample size to the number of clusters in the stratum. It takes the following form:

$$PS_{state,k} = \frac{c_{state,k} \times L_{state}}{C_{state,k}},$$

where,

$PS_{state,k}$ = probability of selection (sampling rate) in sampling stratum k within state;

$C_{state,k}$ = number of clusters in sampling stratum k within state;

$c_{state,k}$ = target number of clusters in sampling stratum k within state;

$L_{state}$ = the factor to reduce the number of clusters to select for the state, if the expected listing workload exceeded the planning estimate.

$$L_{state} = \begin{cases} 1 \text{ for small, medium and AIR sampling stratum} \\ 0 < L \leq 1 \text{ for large sampling stratum} \end{cases}$$

The large block cluster sampling rate was reduced if the expected number of housing units to list was greater than the planning estimate of the listing workload. A second step of sampling was necessary in Missouri and Indiana because the selected sample of clusters resulted in a greater number of housing units to list than was expected. To meet operational constraints, a subsample of the first-step selected block clusters was selected. The second step of sampling only occurred in the large sampling stratum, since that stratum disproportionately contributed to the listing workload. The second step occurred only if the estimated number of housing units in the medium and large strata was at least ten percent larger than the planning estimate of the number of housing units to be listed.

For states needing the second step of sampling, the sampling rate took the following form:

$$PS2_{state} = \frac{PW_{state}}{W_{state}}$$

where,

$PS2_{state}$ = second-step sampling rate for the large sampling stratum in state,

$W_{state}$ = resulting workload estimate from sample selection for the large sampling stratum in state,

$PW_{state}$ = planning workload estimate for the large sampling stratum in state.

**Sorting the PSUs.** The first-phase clusters were sorted within each sampling stratum as follows:

- American Indian Country Indicator
- Demographic/Tenure Group
- 1990 Urbanization
- County code
- Block cluster identification number

Although there was no differential sampling within the four first-phase sampling strata, the clusters were sorted by several variables in an attempt to improve the representativeness of the sample of block clusters. The first variable was the American Indian Country Indicator, which separated the block clusters into three American Indian categories:

1. American Indian Reservation or other trust land,
2. tribal jurisdiction statistical area, Alaska native village statistical area or tribal designated statistical area, and
3. all other areas.

The second sort variable was the demographic/tenure group. Block clusters containing similar demographic/tenure proportions, based on 1990 census data, were grouped. To aid in selecting a sample that was well represented by the six major race/origin groups, as well as owners and renters, block clusters were classified into 12 demographic/tenure groups. Although many block clusters tend to have a large proportion of one demographic/tenure group, rarely were they entirely composed of only one, thus many clusters fit well in two or more categories. To ensure that each cluster was assigned to only one group, a hierarchical assignment rule was developed so that when a cluster exceeded the first group threshold, it was assigned to that group. These thresholds were based on a multivariate clustering method applied to 1990 census blocks. Table 3-4 lists these threshold values. The hierarchy gives the smaller demographic groups priority over the larger ones and renters priority over owners. For example, if the approximate distribution of a block cluster population was 20 percent Asian Renter, 40 percent Asian Owner, and 40 percent White and other Renter, then the block cluster was assigned to the Asian Renter demographic/tenure group.

Table 3-4. **Demographic/Tenure Group Thresholds (50 States and the District of Columbia)**

| Order | Demographic/Tenure Group | Threshold |
|---|---|---|
| 1 | Hawaiian and Pacific Islander renters | 10% |
| 2 | Hawaiian and Pacific Islander owners | 10% |
| 3 | American Indian and Alaska Native renters | 10% |
| 4 | American Indian and Alaska Native owners | 10% |
| 5 | Asian renters | 20% |
| 6 | Asian owners | 20% |
| 7 | Hispanic renters | 20% |
| 8 | Hispanic owners | 20% |
| 9 | Black renters | 25% |
| 10 | Black owners | 25% |
| 11 | White and other renters | 30% |
| 12 | All others | all others |

A third sort variable was the estimated level of urbanization based on 1990 data for each block cluster. Each block cluster was categorized either as an urbanized area with 250,000 or more people, an urbanized area with less than 250,000 people, or a non-urban area. And finally, the clusters were sorted geographically using county and cluster number.

**General sampling procedure.** A systematic sample of block clusters was selected from each sampling stratum with each block cluster having the same probability of selection within a sampling stratum. The method used to select systematic samples follows:

1. Sampling units were sorted using the PSU sort criteria described at each sampling phase.
2. Each successive PSU was assigned an index number 1 through N within each sampling stratum where N is the number of PSUs in the stratum.
3. A random number (RN) between zero and one, $0 < RN \leq 1$, was generated.
4. A random start (RS) for the sampling stratum was calculated. The random start was the random number multiplied by the inverse of the probability of selection, $RS = RN \times 1/PS$, such that $0 < RS \leq 1/PS$.
5. Sampling sequence numbers were calculated. Given N PSUs, sequence numbers were: $RS, RS + 1 \times (1/PS), RS + 2 \times (1/PS),...,RS + n \times (1/PS)$ where n was the largest integer such that $[RS + (n-1) \times 1/PS] \leq N$. Sequence numbers were rounded up to the next integer. An integer number rounded to itself.
6. Sampling sequence numbers were compared to the index numbers assigned to PSUs. The PSU with the index number corresponding to the rounded sequence number was selected. All PSUs without corresponding index numbers were not in sample.

### First-Phase Sample Results

Table 3-5 lists the block cluster sample sizes and the number of housing units by sampling stratum for each state, the District of Columbia, and the nation.

Table 3-5. **State First-Phase Sample Results by First-Phase Stratum**

| State | First-phase housing units[1] | | | | | First-phase block clusters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Small | Medium | Large | AIR | Total | Small | Medium | Large | AIR | Total |
| Alabama | 60 | 7,900 | 19,000 | 0 | 26,960 | 116 | 286 | 109 | 0 | 511 |
| Alaska | 20 | 5,200 | 23,200 | 20 | 28,440 | 20 | 190 | 137 | 1 | 348 |
| Arizona | 20 | 7,800 | 44,700 | 2,600 | 55,120 | 86 | 269 | 180 | 113 | 648 |
| Arkansas | 40 | 9,600 | 15,900 | 0 | 25,540 | 90 | 353 | 101 | 0 | 544 |
| California | 50 | 45,000 | 227,600 | 230 | 272,880 | 184 | 1,442 | 1,311 | 11 | 2,948 |
| Colorado | 20 | 8,000 | 25,600 | 60 | 33,680 | 83 | 293 | 157 | 2 | 535 |
| Connecticut | 10 | 6,100 | 25,600 | 0 | 31,710 | 20 | 211 | 159 | 0 | 390 |
| Delaware | 20 | 7,200 | 28,700 | 0 | 35,920 | 20 | 243 | 156 | 0 | 419 |
| District of Columbia | 10 | 4,800 | 50,500 | 0 | 55,310 | 20 | 132 | 247 | 0 | 399 |
| Florida | 50 | 7,500 | 50,100 | 30 | 57,680 | 145 | 259 | 230 | 1 | 635 |
| Georgia | 70 | 6,100 | 30,300 | 0 | 36,470 | 154 | 220 | 162 | 0 | 536 |
| Hawaii | 10 | 3,000 | 42,400 | 0 | 45,410 | 20 | 103 | 161 | 0 | 284 |
| Idaho | 10 | 8,200 | 10,900 | 140 | 19,250 | 54 | 312 | 75 | 6 | 447 |
| Illinois | 100 | 8,600 | 22,300 | 0 | 31,000 | 185 | 281 | 140 | 0 | 606 |
| Indiana | 80 | 6,100 | 9,700 | 0 | 15,880 | 140 | 202 | 51 | 0 | 393 |
| Iowa | 120 | 6,800 | 9,500 | 0 | 16,420 | 147 | 242 | 53 | 0 | 442 |
| Kansas | 110 | 6,400 | 11,100 | 30 | 17,640 | 193 | 237 | 63 | 1 | 494 |
| Kentucky | 60 | 7,200 | 22,300 | 0 | 29,560 | 96 | 268 | 135 | 0 | 499 |
| Louisiana | 10 | 11,300 | 24,900 | 0 | 36,210 | 65 | 407 | 155 | 0 | 627 |
| Maine | 20 | 5,800 | 11,000 | 10 | 16,830 | 38 | 226 | 79 | 1 | 344 |
| Maryland | 20 | 5,300 | 38,000 | 0 | 43,320 | 36 | 177 | 175 | 0 | 388 |
| Massachusetts | 20 | 6,400 | 22,000 | 0 | 28,420 | 38 | 229 | 140 | 0 | 407 |
| Michigan | 50 | 7,900 | 15,100 | 150 | 23,200 | 122 | 268 | 104 | 5 | 499 |
| Minnesota | 70 | 6,000 | 14,000 | 270 | 20,340 | 141 | 208 | 83 | 10 | 442 |
| Mississippi | 40 | 8,400 | 11,700 | 120 | 20,260 | 81 | 303 | 77 | 3 | 464 |
| Missouri | 110 | 5,700 | 14,500 | 0 | 20,310 | 162 | 200 | 71 | 0 | 433 |
| Montana | 10 | 8,400 | 9,700 | 840 | 18,950 | 67 | 333 | 67 | 24 | 491 |
| Nebraska | 80 | 6,800 | 7,700 | 70 | 14,650 | 142 | 245 | 55 | 3 | 445 |
| Nevada | 10 | 6,400 | 57,800 | 190 | 64,400 | 46 | 225 | 230 | 5 | 506 |
| New Hampshire | 20 | 5,700 | 15,400 | 0 | 21,120 | 25 | 201 | 106 | 0 | 332 |
| New Jersey | 10 | 8,700 | 30,100 | 0 | 38,810 | 39 | 282 | 178 | 0 | 499 |
| New Mexico | 10 | 9,300 | 24,800 | 1,640 | 35,750 | 108 | 335 | 136 | 70 | 649 |
| New York | 80 | 17,600 | 124,700 | 70 | 142,450 | 143 | 603 | 631 | 5 | 1,382 |
| North Carolina | 100 | 6,700 | 20,700 | 80 | 27,580 | 143 | 236 | 121 | 4 | 504 |
| North Dakota | 100 | 5,900 | 9,100 | 340 | 15,440 | 121 | 236 | 64 | 12 | 433 |
| Ohio | 110 | 7,800 | 24,000 | 0 | 31,910 | 132 | 268 | 133 | 0 | 533 |
| Oklahoma | 60 | 9,000 | 17,300 | 270 | 26,630 | 142 | 314 | 101 | 8 | 565 |
| Oregon | 10 | 5,200 | 15,400 | 70 | 20,680 | 86 | 195 | 90 | 3 | 374 |
| Pennsylvania | 110 | 12,900 | 22,600 | 0 | 35,610 | 180 | 427 | 146 | 0 | 753 |
| Rhode Island | 10 | 7,600 | 18,000 | 0 | 25,610 | 20 | 256 | 108 | 0 | 384 |
| South Carolina | 40 | 8,200 | 19,100 | 0 | 27,340 | 95 | 285 | 112 | 0 | 492 |
| South Dakota | 50 | 5,800 | 9,200 | 450 | 15,500 | 106 | 242 | 57 | 27 | 432 |
| Tennessee | 90 | 7,800 | 25,400 | 0 | 33,290 | 133 | 285 | 137 | 0 | 555 |
| Texas | 70 | 34,700 | 148,500 | 30 | 183,300 | 349 | 1,222 | 681 | 1 | 2,253 |
| Utah | 10 | 9,100 | 23,900 | 120 | 33,130 | 38 | 312 | 144 | 7 | 501 |
| Vermont | 20 | 5,600 | 12,000 | 0 | 17,620 | 21 | 201 | 88 | 0 | 310 |
| Virginia | 60 | 5,600 | 31,900 | 0 | 37,560 | 98 | 96 | 166 | 0 | 460 |
| Washington | 20 | 5,600 | 21,400 | 480 | 27,500 | 73 | 187 | 120 | 17 | 397 |
| West Virginia | 30 | 5,000 | 13,100 | 0 | 18,130 | 46 | 189 | 79 | 0 | 314 |
| Wisconsin | 80 | 6,200 | 8,200 | 220 | 14,700 | 119 | 211 | 58 | 10 | 398 |
| Wyoming | 10 | 8,700 | 9,200 | 90 | 18,000 | 72 | 346 | 69 | 5 | 492 |
| Total U.S. | 2,400 | 438,600 | 1,539,800 | 8,620 | 1,989,420 | 5,000 | 15,393 | 8,388 | 355 | 29,136 |

[1]Preliminary census address list housing unit counts from spring 1999.

## SECOND PHASE OF THE A.C.E. SAMPLE DESIGN

The second phase, often referred to as the A.C.E. reduction phase, linked the first-phase sample selection to the A.C.E. sampling plan. The A.C.E. reduction was the first of several operations that reduced the number of housing units from the nearly two million housing units in the independent listing to the approximately 300,000 housing units that were sent for interview. Since not all of the first-phase block clusters were required for A.C.E., the reduction subsampled those clusters, with the selected clusters retained for the A.C.E. operations.

Following the selection of the A.C.E. first-phase sample, field staff visited the block clusters and created an independent address list for A.C.E. These updated housing

unit counts were used in the cluster subsampling phase. The cluster subsampling was done separately for:

- medium and large cluster reduction, and
- small block cluster reduction.

### Medium and Large Cluster Reduction

The medium and large cluster reduction was the transition to the A.C.E. sampling plan. The resulting national sample allocation was roughly proportional to state population with some differential sampling within states. Only block clusters from the medium and large first-phase sampling strata in the 50 states and the District of Columbia were subsampled in this phase. As part of the sample reduction, two other objectives of the A.C.E. sample were implemented.

One objective of the medium and large cluster reduction design was to stratify the first-phase clusters based on the relationship of current housing unit counts from the A.C.E. independent listing and the updated census address list as of January, 2000. Clusters were sampled with different selection probabilities in order to reduce the variance contribution due to inconsistent housing unit counts between the updated census list and the independent list. Clusters with significant differences between the counts were expected to have high erroneous enumeration and high omission rates. The objective of differentially sampling these types of clusters was to reduce the sampling weights associated with clusters having relatively high numbers of missed persons or those enumerated in error, and, thus, having potentially high variance contributions.

A second objective of the medium and large cluster reduction design was to differentially sample clusters based on the estimated demographic composition of the cluster. Clusters with a high proportion of persons of Hispanic origin or persons belonging to a census race group other than White were classified into a minority stratum. These types of clusters were sampled at a higher rate than predominantly non-Hispanic White clusters, in order to increase the sample size and improve the reliability of the A.C.E. population estimates for these historically undercounted subgroups.

**Stratifying second-phase clusters.** Each block cluster was put into two categories for the medium and large cluster reduction: a demographic group and a consistency group. Block clusters were put into reduction strata based on the combination of these two groups.

Demographic groups were based on the demographic/tenure groups created in the first-phase sample selection. The demographic/tenure groups represented a classification of block clusters, using the information of race/Hispanic origin and tenure of each block reported in the 1990 census. The demographic/tenure groups were used as a sort variable in the selection of the first-phase sample. For this reduction, clusters were put into two demographic groups by combining the 12 demographic/tenure groups in Table 3-4. The two demographic groups are:

1. Minority: block clusters from one of the ten minority demographic/tenure groups
2. Non-minority: block clusters from one of the two other demographic/tenure groups

For this reduction, two updated cluster housing unit counts were used: the independent listing housing unit count and the housing unit count from the updated census address list as of January 2000. The two housing unit counts were compared, and clusters were placed into consistency groups based on the relationship of the housing unit counts. Large differences between the counts indicated that coverage problems might occur; thus, the sampling weights for such clusters were controlled to avoid serious variance effects.

Clusters were placed into three consistency groups as shown in Table 3-6.

Table 3-6. **Second-Phase Sampling Consistency Groups**

| Relationship | Consistency group |
|---|---|
| Independent list is at least 25 percent lower than census | Low inconsistent |
| Independent list is at least 25 percent greater than census | High inconsistent |
| Independent list is within ± 25 percent of census | Consistent |

For List/Enumerate clusters (see attachment), the census housing unit count was not known at the time of the reduction since this census operation had not started. Thus, all such clusters were classified as high inconsistent.

Based on the demographic group, the consistency group, and the independent listing housing unit count, block clusters were assigned to one of five reduction strata:

1. Minority (low inconsistent, high inconsistent, consistent)
2. Non-minority low inconsistent
3. Non-minority high inconsistent
4. Non-minority consistent
5. Medium stratum jumper

Medium stratum jumper clusters were selected from the medium sampling stratum for the first-phase sample, but had 80 or more independent listing housing units. Medium clusters were sampled at lower rates than large clusters in the first-phase sample since large clusters eventually were to undergo within-cluster housing unit subsampling, an operation that increases sampling weights. Medium stratum jumper clusters also went through within-cluster housing unit subsampling, meaning the already higher sampling weights of these clusters became even larger. Retaining all of the medium stratum jumper clusters in this reduction avoided introducing significant weight variation in the sample.

**Allocating sample to strata.** The first step was to allocate the national sample of 300,000 housing units to the 50 states and the District of Columbia, in most cases proportional to 1998 population estimates, with a minimum of 1,800 housing units in each state. Hawaii was allocated approximately 3,750 housing units due to its concentration of Hawaiian and Pacific Islanders for which separate population coverage estimates were planned.

Within each state, the second-phase selection probabilities varied somewhat among the strata. First, all clusters in the medium stratum jumper reduction stratum were retained. For the remaining four reduction strata, higher retention rates were used in the minority, non-minority low inconsistent and the non-minority high inconsistent reduction strata than the non-minority consistent stratum. The stratum differential sampling factor is the ratio of the probability of selection for the stratum to the probability of selection for the consistent stratum.

The following statements describe how the stratum differential sampling factors were set to yield the overall state sample size. These are not exact rules, but give a sense of how much differential sampling within states was done.

- The maximum expected sampling weight after all subsampling, the inverse of the overall probability of selection, was 650 for the non-minority consistent reduction stratum.
- The maximum differential sampling factor was 3 for the two inconsistent reduction strata.
- The differential sampling factor was around 2 for the minority reduction stratum, except in small states where all of the minority clusters were retained.

The differential sampling factors were assigned using guidelines designed to achieve the two objectives of the reduction, while also controlling the size of the sampling weights and the amount of differential sampling. This led to the design of the differential sampling factors summarized in Table 3-7.

Using the stratum differential sampling factors and the estimated number of housing units, the sample allocation for each reduction stratum was derived as follows:

$$T_g = T \times \frac{DSF_g \times \hat{H}_g}{\sum_{g=1}^{4} DSF_g \times \hat{H}_g}$$

where,

| | | |
|---|---|---|
| g | = | A.C.E. second-phase sampling stratum, |
| $T_g$ | = | Target number of sample housing units allocated to reduction stratum g, |
| T | = | State target number of sample housing units modified for medium stratum jumper clusters, |
| $\hat{H}_g$ | = | Estimated number of housing units in the reduction stratum based on the independent listing housing unit counts, and |
| $DSF_g$ | = | Differential Sampling Factor for reduction stratum g. |

Table 3-7. **A.C.E. Second-Phase Sample Design Parameters for Large and Medium Clusters**

| State | Differential Sampling Factors[1] | | | | Target sample size[6] | First-phase sample size [7] |
|---|---|---|---|---|---|---|
| | Minority[2] | Low inconsistent[3] | High inconsistent[4] | Consistent[5] | | |
| Alabama | 1.78 | 1.78 | 1.78 | 1.00 | 4,470 | 26,960 |
| Alaska | 6.20 | 3.00 | 3.00 | 1.00 | 1,800 | 28,440 |
| Arizona | 1.78 | 1.78 | 1.78 | 1.00 | 4,800 | 55,120 |
| Arkansas | 2.00 | 3.00 | 3.00 | 1.00 | 2,610 | 25,540 |
| California | 2.00 | 3.00 | 3.00 | 1.00 | 33,510 | 272,880 |
| Colorado | 1.99 | 2.93 | 2.93 | 1.00 | 4,080 | 33,680 |
| Connecticut | 2.00 | 3.00 | 3.00 | 1.00 | 3,360 | 31,710 |
| Delaware | 2.91 | 3.00 | 3.00 | 1.00 | 1,800 | 35,920 |
| District of Columbia | 2.00 | 3.00 | 3.00 | 1.00 | 1,800 | 55,310 |
| Florida | 1.00 | 1.00 | 1.00 | 1.00 | 15,300 | 57,680 |
| Georgia | 2.01 | 2.01 | 2.01 | 1.00 | 7,830 | 63,470 |
| Hawaii | 2.00 | 3.00 | 3.00 | 1.00 | 3,750 | 45,410 |
| Idaho | 2.71 | 3.00 | 3.00 | 1.00 | 1,800 | 19,250 |
| Illinois | 1.19 | 1.19 | 1.19 | 1.00 | 12,360 | 31,000 |
| Indiana | 1.68 | 1.68 | 1.68 | 1.00 | 6,060 | 15,880 |
| Iowa | 2.00 | 3.00 | 3.00 | 1.00 | 2,940 | 16,420 |
| Kansas | 2.00 | 3.00 | 3.00 | 1.00 | 2,700 | 17,640 |
| Kentucky | 2.00 | 3.00 | 3.00 | 1.00 | 4,050 | 29,560 |
| Louisiana | 1.89 | 3.00 | 3.00 | 1.00 | 4,470 | 36,210 |
| Maine | 6.55 | 3.00 | 3.00 | 1.00 | 1,800 | 16,830 |
| Maryland | 1.87 | 2.46 | 2.46 | 1.00 | 5,280 | 43,320 |
| Massachusetts | 2.33 | 2.33 | 2.33 | 1.00 | 6,300 | 28,420 |
| Michigan | 1.25 | 1.25 | 1.25 | 1.00 | 10,080 | 23,200 |
| Minnesota | 2.11 | 2.11 | 2.11 | 1.00 | 4,860 | 20,340 |
| Mississippi | 1.96 | 2.83 | 2.83 | 1.00 | 2,820 | 20,260 |
| Missouri | 2.25 | 2.25 | 2.25 | 1.00 | 5,580 | 20,310 |
| Montana | 1.57 | 3.00 | 3.00 | 1.00 | 1,800 | 18,950 |
| Nebraska | 2.44 | 3.00 | 3.00 | 1.00 | 1,800 | 14,650 |
| Nevada | 1.95 | 2.76 | 2.76 | 1.00 | 1,800 | 64,400 |
| New Hampshire | 6.84 | 3.00 | 3.00 | 1.00 | 1,800 | 21,120 |
| New Jersey | 2.24 | 2.24 | 2.24 | 1.00 | 8,340 | 38,810 |
| New Mexico | 1.73 | 1.73 | 1.73 | 1.00 | 1,800 | 35,750 |
| New York | 2.00 | 3.00 | 3.00 | 1.00 | 18,660 | 142,450 |
| North Carolina | 1.83 | 1.83 | 1.83 | 1.00 | 7,740 | 27,580 |
| North Dakota | 2.14 | 3.00 | 3.00 | 1.00 | 1,800 | 15,440 |
| Ohio | 1.22 | 1.22 | 1.22 | 1.00 | 11,490 | 31,910 |
| Oklahoma | 2.00 | 3.00 | 3.00 | 1.00 | 3,420 | 26,630 |
| Oregon | 1.94 | 2.76 | 2.76 | 1.00 | 3,360 | 20,680 |
| Pennsylvania | 1.70 | 1.70 | 1.70 | 1.00 | 12,300 | 35,610 |
| Rhode Island | 2.94 | 3.00 | 3.00 | 1.00 | 1,800 | 25,610 |
| South Carolina | 1.60 | 1.60 | 1.60 | 1.00 | 3,930 | 27,340 |
| South Dakota | 1.83 | 3.00 | 3.00 | 1.00 | 1,800 | 15,500 |
| Tennessee | 1.99 | 2.86 | 2.86 | 1.00 | 5,580 | 33,290 |
| Texas | 1.86 | 2.36 | 2.36 | 1.00 | 20,280 | 183,300 |
| Utah | 2.00 | 3.00 | 3.00 | 1.00 | 2,160 | 33,130 |
| Vermont | 6.91 | 3.00 | 3.00 | 1.00 | 1,800 | 17,620 |
| Virginia | 1.90 | 1.90 | 1.90 | 1.00 | 6,960 | 37,560 |
| Washington | 2.23 | 2.23 | 2.23 | 1.00 | 5,850 | 27,500 |
| West Virginia | 2.00 | 3.00 | 3.00 | 1.00 | 1,860 | 18,130 |
| Wisconsin | 1.75 | 1.75 | 1.75 | 1.00 | 5,370 | 14,700 |
| Wyoming | 1.99 | 3.00 | 3.00 | 1.00 | 1,800 | 18,000 |

[1]The observed or actual sampling factors differed from the design sample rates. See the section on ''Selecting a subsample.''
[2]Clusters with high concentrations of minorities.
[3]Clusters where the independent listing housing unit count is at least 25 percent lower than the updated census list count.
[4]Clusters where the independent listing count is at least 25 percent higher than the updated census list.
[5]Clusters where the independent listing count and the updated census list do not differ by more than 25 percent.
[6]Target state housing unit interview sample size, excluding American Indian Reservation sample.
[7]First-phase preliminary census address list housing unit counts from Spring, 1999.

**Sorting the PSUs.** The first-phase clusters within each second-phase stratum by first-phase sampling stratum were sorted as follows:

- Consistency group
- List/Enumerate indicator
- American Indian Country Indicator
- Demographic/Tenure Group
- 1990 Urbanization
- County code
- Block cluster identification number

**Selecting a subsample.** Since the first-phase sample utilized different sampling rates for the medium and large sampling strata, separate samples were drawn for each second-phase stratum within the first-phase sampling strata. Selecting the sample required calculating the sampling rates, sorting the clusters, and drawing a systematic sample of clusters.

All of the medium stratum jumpers were retained in the sample. The sampling rates for the remaining four reduction strata were computed so that an integer number of block clusters was selected. This required computing a sampling rate based on the ratio of housing units which resulted in a non-integer expected number of clusters, determining an integer number of clusters to select, and calculating the final sampling rate based on the ratio of clusters. The medium and large cluster reduction followed the sampling procedure discussed earlier.

This resulted in a total of 9,765 out of 24,136 medium and large clusters retained in the A.C.E. sample for the 50 states and the District of Columbia.

**Medium and large cluster reduction sample results.**

Table 3-8 lists the number of housing units and clusters in sample.

**Small Cluster Reduction**

The first-phase sample contained 5,000 small clusters in the United States. Small clusters were expected to have between zero and two housing units based on an early census address list. Conducting interviewing and follow-up operations in clusters of this size was not as cost effective as in larger clusters. Therefore, to allocate A.C.E. resources more efficiently, only a subsample of these small clusters was retained in the A.C.E. sample.

This subsampling operation attempted a balance among three goals. One goal was to prevent any small clusters from having sampling weights that were extremely high compared to other clusters in the sample. Second, sampling weights should be lower on clusters where the number of housing units was different than expected. These first two goals attempted to reduce the contribution of small clusters to the variance of the dual system estimates. The third goal was to improve operational efficiency by reducing the number of clusters and future field visits. To achieve these goals, differential sampling was used.

**Stratifying first-phase clusters.** The first-phase small clusters were classified into nine possible reduction strata within each state. These strata were defined using three cluster characteristics: Size, American Indian Country status, and List/Enumerate status.

The size of a cluster was based on the greater of the independent listing housing unit count or the updated census address list housing unit count as of January 2000. For List/Enumerate clusters the size was always based on the actual independent listing count since the List/Enumerate operation had not yet started by the time of this reduction. The American Indian Country status had three categories as described in the first-phase of sampling. Table 3-9 contains the reduction strata for small block clusters.

Table 3-8. **Second-Phase Results–Medium and Large Block Cluster and Housing Unit Counts**

| Number of.... | Minority | Low inconsistent | High inconsistent | Consistent | Stratum jumpers | American Indian reservations | Nation |
|---|---|---|---|---|---|---|---|
| Housing units[1] | 230,529 | 49,086 | 94,850 | 403,806 | 32,064 | 9,251 | 819,586 |
| Clusters | 2,553 | 971 | 842 | 4,801 | 243 | 355 | 9,765 |

[1]Independent Listing counts as of December, 1999.

Table 3-9. **Small Block Cluster Second-Phase Strata**

| Second-phase stratum | Housing units | American Indian country | List/Enumerate status |
|---|---|---|---|
| 1 | 0 to 2 | No | No |
| 2 | 3 to 5 | No | No |
| 3 | 6 to 9 | No | No |
| 4 | 10+ | - | - |
| 5 | 0 to 2 | No | Yes |
| 6 | 3 to 9 | No | Yes |
| 7 | 0 to 9 | Reservation/Trust land | - |
| 8 | 0 to 2 | TJSA/TDSA/ANVSA[1] | - |
| 9 | 3 to 9 | TJSA/TDSA/ANVSA | - |

[1]Tribal Jurisdiction Statistical Area/Tribal Designated Statistical Area/Alaska Native Village Statistical Area

**Determining target sampling rates.** Using independent listing housing unit counts, target sampling rates were determined. These rates attempted to satisfy the previously discussed statistical and operational goals.

Generally, the small clusters were stratified into four groups based on the number of housing units in the cluster. All clusters with ten or more housing units, on American Indian land, or classified as List/Enumerate were retained in sample. For the remaining three reduction strata, some differential sampling was introduced.

To determine the sampling rates for these strata, two conditions were imposed. One of these conditions was that, if possible, the number of weighted housing units in a cluster did not exceed 2,400 housing units. Through computer simulations, a number of different limits were tried until a cap of 2,400 yielded a sample of appropriate size. The second condition was a minimum sampling rate, which varied among the three strata. Table 3-10 contains a summary of the sampling conditions. Table 3-11 illustrates the process for determining the second-phase sampling rate for each stratum.

The overall target selection probability was based on the maximum number of housing units within a stratum and the previously mentioned cap of 2,400 housing units. For example, the maximum number of housing units in stratum group one was two. Hence, the overall target selection probability was 1 in (2,400/2) or 1 in 1,200. The sampling rate for each second-phase stratum was then set at the rate required to attain these overall target probabilities of selection.

**Sorting the PSUs.** The first-phase clusters were sorted in the following order in each second-phase stratum:

- 1990 urbanization
- county code
- A.C.E. cluster identification number

Table 3-10. **Small Cluster Reduction Sampling Conditions**

| Second-phase stratum | Cluster size (HUs) | Overall target selection probability | Minimum second-phase sampling rate |
|---|---|---|---|
| 1 | 0 to 2 | 1/1,200 | 1/10 |
| 2 | 3 to 5 | 1/480 | 1/4 |
| 3 | 6 to 9 | 1/267 | 1/2.22 |

Table 3-11. **Second-Phase Sampling Rate Criterion**

| If . . . | Then, the second-phase sampling rate equals... |
|---|---|
| $\frac{\text{Overall target selection probability}}{\text{First-phase sampling rate}} \geq$ Minimum second-phase sampling rate | $\frac{\text{Overall target selection probability}}{\text{First-phase sampling rate}}$ |
| $\frac{\text{Overall target selection probability}}{\text{First-phase sampling rate}} <$ Minimum second-phase sampling rate | Minimum second-phase sampling rate |

**Selecting a subsample.** Separate samples were selected from each second-phase stratum within each state and the District of Columbia. This required calculating the actual sampling rate for the stratum, sorting the clusters and drawing a systematic sample of clusters.

All clusters with 10 or more housing units that were classified as List/Enumerate, or were in American Indian Country, were retained in sample. The sampling rates for the remaining three strata were computed to achieve an integer number of block clusters drawn from each stratum, similar to procedures used for the medium and large cluster reduction. This required computing a sampling rate, which resulted in a noninteger expected number of clusters determining an integer number of clusters to select, and calculating the final sampling rate based on the ratio of clusters. The small cluster reduction followed the sampling procedure discussed earlier.

This resulted in a total of 1,538 out of 5,000 small clusters retained in the A.C.E. sample for the 50 states and the District of Columbia.

**Small cluster reduction results.** Table 3-12 gives the distribution of block clusters and housing units after small block cluster reduction. As mentioned earlier, the larger of the independent listing housing unit count and the housing unit count from the updated census address list as of January 2000 was used to stratify the clusters. In Table 3-12, only the independent listing housing unit count is used in these tallies. Hence, with 55 clusters, as seen in the "6-9" cluster size, the number of housing units does not achieve the minimum of 330.

### Second-Phase Sampling Results

Table 3-13 lists the block cluster sample sizes and the number of housing units in each state, the District of Columbia, and the nation after the second phase of A.C.E. sampling.

Table 3-12. **Second-Phase Results—Small Block Cluster and Housing Unit Counts**

| Cluster size (HUs)[1] | American Indian country | List/enumerate status | Number of housing units[2] | Number of clusters |
|---|---|---|---|---|
| 0-2 | No | No | 209 | 692 |
| 3-5 | No | No | 358 | 117 |
| 6-9 | No | No | 325 | 55 |
| 10+ | - | - | 4,532 | 112 |
| 0-2 | No | Yes | 59 | 290 |
| 3-9 | No | Yes | 76 | 16 |
| 0-9 | Reservation/Trust land | - | 43 | 128 |
| 0-2 | TJSA/TDSA/ANVSA[3] | - | 40 | 121 |
| 3-9 | TJSA/TDSA/ANVSA | - | 30 | 7 |
| Total | | | 5,672 | 1,538 |

[1]The size of a cluster was based on the higher of the independent listing housing unit count or the January, 2000 census address list. For List/Enumerate clusters the size was always based on the actual independent listing count since the List/Enumerate operation had not yet been started by the time of this reduction.
[2]Keyed independent listing housing unit counts as of January, 2000.
[3]Tribal Jurisdiction Statistical Area/Tribal Designated Statistical Area/Alaska Native Village Statistical Area.

Table 3-13. **State Second-Phase Sample Results by First-Phase Stratum**

| State | Second-phase housing units[1] | | | | | Second-phase block clusters | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Small | Medium | Large | AIR | Total | Small | Medium | Large | AIR | Total |
| Alabama | 54 | 3,599 | 7,531 | 0 | 11,184 | 14 | 104 | 43 | 0 | 161 |
| Alaska | 24 | 1,401 | 3,099 | 16 | 4,540 | 7 | 40 | 22 | 1 | 70 |
| Arizona | 140 | 3,082 | 17,185 | 2,826 | 23,233 | 69 | 79 | 61 | 113 | 322 |
| Arkansas | 16 | 2,077 | 3,566 | 0 | 5,659 | 13 | 71 | 24 | 0 | 108 |
| California | 401 | 19,124 | 77,913 | 204 | 97,642 | 93 | 528 | 469 | 11 | 1,101 |
| Colorado | 19 | 2,722 | 9,248 | 52 | 12,041 | 24 | 85 | 55 | 2 | 166 |
| Connecticut | 37 | 1,699 | 6,718 | 0 | 8,454 | 5 | 59 | 47 | 0 | 111 |
| Delaware | 7 | 1,572 | 2,979 | 0 | 4,558 | 3 | 40 | 23 | 0 | 66 |
| District of Columbia | 0 | 1,251 | 5,403 | 0 | 6,654 | 2 | 25 | 31 | 0 | 58 |
| Florida | 265 | 7,976 | 54,986 | 20 | 63,247 | 43 | 259 | 230 | 1 | 533 |
| Georgia | 211 | 4,095 | 21,195 | 0 | 25,501 | 27 | 138 | 111 | 0 | 276 |
| Hawaii | 11 | 1,200 | 22,252 | 0 | 23,463 | 6 | 40 | 75 | 0 | 121 |
| Idaho | 12 | 1,632 | 2,714 | 152 | 4,510 | 32 | 53 | 16 | 6 | 107 |
| Illinois | 51 | 7,527 | 20,041 | 0 | 27,619 | 25 | 247 | 131 | 0 | 403 |
| Indiana | 125 | 4,141 | 7,431 | 0 | 11,697 | 24 | 141 | 46 | 0 | 211 |
| Iowa | 161 | 2,338 | 3,705 | 0 | 6,204 | 21 | 79 | 22 | 0 | 122 |
| Kansas | 33 | 2,193 | 3,488 | 31 | 5,745 | 24 | 70 | 22 | 1 | 117 |
| Kentucky | 92 | 2,329 | 9,621 | 0 | 12,042 | 14 | 92 | 52 | 0 | 158 |
| Louisiana | 7 | 3,332 | 6,574 | 0 | 9,913 | 40 | 109 | 50 | 0 | 199 |
| Maine | 38 | 1,447 | 2,020 | 1 | 3,506 | 24 | 53 | 16 | 1 | 94 |
| Maryland | 22 | 3,288 | 17,041 | 0 | 20,351 | 6 | 77 | 82 | 0 | 165 |
| Massachusetts | 105 | 3,467 | 11,471 | 0 | 15,043 | 10 | 120 | 80 | 0 | 210 |
| Michigan | 64 | 6,612 | 13,581 | 148 | 20,405 | 19 | 227 | 92 | 5 | 343 |
| Minnesota | 79 | 3,210 | 7,275 | 286 | 10,850 | 28 | 116 | 49 | 10 | 203 |
| Mississippi | 84 | 2,499 | 2,957 | 96 | 5,636 | 20 | 76 | 25 | 3 | 124 |
| Missouri | 269 | 3,229 | 11,558 | 0 | 15,056 | 24 | 113 | 51 | 0 | 188 |
| Montana | 15 | 1,880 | 2,365 | 905 | 5,165 | 41 | 60 | 14 | 24 | 139 |
| Nebraska | 25 | 1,685 | 1,317 | 91 | 3,118 | 31 | 53 | 13 | 3 | 100 |
| Nevada | 1 | 1,361 | 8,506 | 204 | 10,072 | 38 | 28 | 30 | 5 | 101 |
| New Hampshire | 50 | 1,658 | 2,535 | 0 | 4,243 | 11 | 46 | 19 | 0 | 76 |
| New Jersey | 4 | 4,883 | 14,960 | 0 | 19,847 | 8 | 147 | 103 | 0 | 258 |
| New Mexico | 29 | 1,813 | 2,666 | 1,854 | 6,362 | 76 | 47 | 19 | 70 | 212 |
| New York | 582 | 8,256 | 62,616 | 93 | 71,547 | 34 | 271 | 317 | 5 | 627 |
| North Carolina | 300 | 5,149 | 18,901 | 136 | 24,486 | 28 | 151 | 93 | 4 | 276 |
| North Dakota | 35 | 1,332 | 2,076 | 394 | 3,837 | 34 | 58 | 17 | 12 | 121 |
| Ohio | 146 | 6,906 | 22,631 | 0 | 29,683 | 22 | 230 | 127 | 0 | 379 |
| Oklahoma | 96 | 2,557 | 5,142 | 267 | 8,062 | 104 | 89 | 31 | 8 | 232 |
| Oregon | 7 | 2,165 | 7,231 | 124 | 9,527 | 52 | 70 | 44 | 3 | 169 |
| Pennsylvania | 203 | 8,622 | 15,227 | 0 | 24,052 | 28 | 293 | 107 | 0 | 428 |
| Rhode Island | 6 | 1,517 | 2,517 | 0 | 4,040 | 4 | 47 | 18 | 0 | 69 |
| South Carolina | 113 | 3,540 | 9,094 | 0 | 12,747 | 15 | 88 | 39 | 0 | 142 |
| South Dakota | 22 | 1,307 | 2,613 | 453 | 4,395 | 40 | 55 | 14 | 27 | 136 |
| Tennessee | 381 | 4,000 | 10,436 | 0 | 14,817 | 24 | 125 | 58 | 0 | 207 |
| Texas | 714 | 13,473 | 47,011 | 30 | 61,228 | 149 | 405 | 238 | 1 | 793 |
| Utah | 112 | 2,583 | 4,061 | 134 | 6,890 | 29 | 48 | 23 | 7 | 107 |
| Vermont | 16 | 1,191 | 3,237 | 0 | 4,444 | 10 | 45 | 20 | 0 | 75 |
| Virginia | 62 | 3,443 | 20,872 | 0 | 24,377 | 15 | 131 | 112 | 0 | 258 |
| Washington | 225 | 3,320 | 12,976 | 438 | 16,959 | 33 | 106 | 76 | 17 | 232 |
| West Virginia | 24 | 1,263 | 4,666 | 0 | 5,953 | 10 | 46 | 23 | 0 | 79 |
| Wisconsin | 164 | 4,380 | 5,909 | 219 | 10,672 | 24 | 138 | 39 | 10 | 211 |
| Wyoming | 13 | 1,778 | 1,186 | 89 | 3,066 | 61 | 62 | 11 | 5 | 139 |
| Total U.S. | 5,672 | 187,104 | 642,303 | 9,263 | 844,342 | 1,538 | 5,880 | 3,530 | 355 | 11,303 |

[1]Keyed independent listing housing unit counts as of January 2000. ''Keyed'' implies these counts went through a quality control review. Consequently, small discrepancies may exist between these independent listing housing unit counts and those from Table 3-8.

## THIRD PHASE OF THE A.C.E. SAMPLE DESIGN

In very large block clusters, the housing units within the cluster were subsampled. This achieved manageable field workloads for A.C.E. interviewing and person follow-up without having a big impact on reliability. The strategy of the A.C.E. large block cluster sampling plan was to increase the number of clusters in sample, while still attaining the targeted number of housing units for interview. Because housing units in a block cluster are often similar, interviewing all of them is not the most efficient use of resources. Instead, interviewing a manageable fraction of several different clusters provides a more geographically diverse sample.

In the first-phase sampling, large block clusters had a higher selection probability than medium block clusters to take into account this anticipated, subsequent housing unit reduction. The A.C.E. second-phase reduction maintained the differential selection probabilities between the large and medium block clusters. After the reduction of housing units in large block clusters, the housing unit selection probabilities in medium and large block clusters in the same second-phase sampling stratum were similar.

Another important goal of this housing unit reduction was to geographically overlap the P and E samples to reduce the E-sample person follow-up workload. An overlapping P and E sample was not necessary, but improved the precision of dual system estimates, the cost-effectiveness of the succeeding operation, and the data processing efficiency.

### Identifying the P-Sample Housing Units

The source of the P-sample housing units, which were subject to person interviewing by the field staff, was the independently listed housing units that were confirmed to exist following the housing unit matching and follow-up operations. (See Chapter 4.) In block clusters that had fewer than 80 of these housing units, all of the housing units were designated to be in the P sample. In addition, all housing units in a block cluster selected from the American Indian Reservation stratum were in the P sample, regardless of how many housing units were in the block cluster. Most block clusters from this stratum were expected to have fewer than 80 housing units and it was desirable to avoid introducing weight variation to the sample cases for this stratum. For block clusters with 80 or more housing units, the housing units were subsampled and the selected housing units were in the P sample.

The reduction of housing units within a large block cluster was done by forming groups of adjacent housing units called segments and selecting one or more segments of housing units to participate in the P sample. The segments had approximately equal numbers of housing units within a block cluster. Segments of housing units were used as the sampling units in order to obtain compact interviewing workloads and to facilitate overlapping P and E samples to reduce E-sample person follow-up workloads.

**Flow of operations.** A complication of this project was that large block clusters were ready for the housing unit subsampling on a flow basis as the preceding operations, housing unit matching and follow-up, were completed. To remain on schedule, it was essential that the P-sample housing units were selected and prepared for interview as quickly as possible. This meant that sampling parameters were computed based on the housing unit counts from the independent listing. If scheduling had not been an issue, the housing unit counts from the housing unit matching and follow-up would have been used. The time schedule constraints did not permit the entire country to be processed prior to subsampling. Further, there was no prespecified order in which block clusters were ready for housing unit subsampling. Thus, following the flow of block clusters from the preceding operations, the housing unit subsampling was performed daily.

**Stratifying third-phase clusters.** Before selecting the sample of segments, block clusters were divided into seven strata within each state. The first five strata were the same strata used for the second phase of sampling for the medium and large first-phase strata. The sixth stratum was the small to large stratum jumpers, block clusters from the small stratum observed to have more than 80 housing units during the independent listing. The seventh stratum was equivalent to the first-phase American Indian Reservation stratum, for which no housing unit reduction was done.

**Allocating the sample.** Nationally, the target distribution of the 300,000 P-sample housing unit sample was roughly proportional to population size, except for increases in sample size in the smaller states, which had roughly equal sizes. The second-phase introduced differential sampling within each state and generated overall target sample sizes for each reduction stratum in the state, the $T_g$ in the earlier section. Based on these targets and the observed second-phase sample block clusters, the sample was allocated to each stratum to provide approximately equal overall probabilities of selection for housing units from the same stratum.

**Determining sampling parameters.** Separate sampling parameters were computed for each stratum within a state. For each stratum, the selection probability was the ratio of the target number of housing units from large block clusters over the number of housing units from the independent listing in large block clusters.

$$\text{Within-cluster sampling rate} = \frac{\text{Target housing unit sample size in large block clusters}}{\text{Number of listed housing units in large block clusters}}$$

The target housing unit sample size was derived by subtracting the number of housing units in medium block clusters based on the independent list from the target stratum sample size. When tallying the housing unit counts from the independent list, any housing units classified as future construction were omitted from the count. Although some of this future construction was probably going to be built by Census Day, it was expected to be a rare occurrence.

Within a particular stratum in a state, a fixed number of segments was formed in each block cluster. This number was a function of the within-cluster sampling rate. This method yielded different size segments across block clusters within the same stratum. This method is a trade-off between having fewer segments to reduce nonsampling error and having more segments of a fixed size to reduce sample size variation. Nonsampling error was reduced by having fewer segment boundaries to identify. If the within-cluster sampling rate was less than or equal to 0.5, then

$$\text{Number of segments} = \frac{1}{\text{within-cluster sampling rate}}$$

rounded up to the nearest integer. When the within-cluster sampling rate was greater than 0.5, the above formula results in only two segments resulting in increased sample size variation with the larger segment size. To better control sample size variation when the sampling rate was greater than 0.5, the number of segments was calculated as

$$\text{Number of segments} = \frac{1}{(1 - \text{within-cluster sampling rate})}$$

**Forming the segments.** Within each block cluster the housing units were sorted by census block and geographic location within the block. Then based on the number of segments, approximately equal numbers of housing units were assigned to each segment.

**Selecting a subsample.** Within-cluster subsampling was done daily as the clusters completed the housing unit matching and follow-up operations. Despite the daily processing, the subsampling was equivalent to a one-time sample, since the results of the previous day were carried over to the next and continued. The one difference with the daily operation was the inability to control the block cluster sort across all block clusters in the stratum due to the flow of the block clusters. So, each day the block clusters that were to be subsampled were sorted by block cluster number within each stratum.

A sample of segments was selected by taking one systematic sample across all large block clusters in each stratum within a state. Selecting one systematic sample per sampling stratum, rather than a separate sample from each large cluster, reduced sample size variability. This allowed an observed sample size close to the target housing unit sample size to be achieved.

### P-Sample Results

Following within-cluster subsampling, the sample for the 50 states and the District of Columbia was 11,303 block clusters containing about 301,000 housing units. Table 3-14 displays the results for each state.

Table 3-14. **State Third-Phase Sample Results for the P Sample**

| State | Housing unit counts[1] by cluster size[2] | | | | Block cluster counts by cluster size[2] | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 - 79 | 80+ | AIR | Total | 0 - 79 | 80+ | AIR | Total |
| Alabama | 2,947 | 1,503 | 0 | 4,450 | 115 | 46 | 0 | 161 |
| Alaska | 1,152 | 587 | 16 | 1,739 | 48 | 21 | 1 | 70 |
| Arizona | 5,193 | 2,474 | 2,661 | 7,667 | 154 | 55 | 113 | 322 |
| Arkansas | 1,795 | 921 | 0 | 2,716 | 86 | 22 | 0 | 108 |
| California | 18,608 | 14,919 | 192 | 33,527 | 675 | 415 | 11 | 1,101 |
| Colorado | 2,662 | 1,491 | 50 | 4,153 | 113 | 51 | 2 | 166 |
| Connecticut | 1,971 | 1,272 | 0 | 3,243 | 72 | 39 | 0 | 111 |
| Delaware | 1,077 | 693 | 0 | 1,770 | 42 | 24 | 0 | 66 |
| District of Columbia | 1,106 | 1,084 | 0 | 2,190 | 29 | 29 | 0 | 58 |
| Florida | 8,736 | 6,518 | 20 | 15,254 | 329 | 203 | 1 | 533 |
| Georgia | 4,690 | 3,072 | 0 | 7,762 | 183 | 93 | 0 | 276 |
| Hawaii | 1,156 | 2,447 | 0 | 3,603 | 47 | 74 | 0 | 121 |
| Idaho | 1,653 | 342 | 146 | 1,995 | 86 | 15 | 6 | 107 |
| Illinois | 8,510 | 3,855 | 0 | 12,365 | 292 | 111 | 0 | 403 |
| Indiana | 4,172 | 1,773 | 0 | 5,945 | 169 | 42 | 0 | 211 |
| Iowa | 2,162 | 829 | 0 | 2,991 | 101 | 21 | 0 | 122 |
| Kansas | 2,114 | 552 | 29 | 2,666 | 101 | 15 | 1 | 117 |
| Kentucky | 2,607 | 1,372 | 0 | 3,979 | 111 | 47 | 0 | 158 |
| Louisiana | 3,031 | 1,386 | 0 | 4,417 | 153 | 46 | 0 | 199 |
| Maine | 1,571 | 361 | 1 | 1,932 | 80 | 13 | 1 | 94 |
| Maryland | 2,574 | 2,713 | 0 | 5,287 | 91 | 74 | 0 | 165 |
| Massachusetts | 4,500 | 1,893 | 0 | 6,393 | 151 | 59 | 0 | 210 |
| Michigan | 7,224 | 2,756 | 147 | 9,980 | 259 | 79 | 5 | 343 |
| Minnesota | 3,734 | 1,420 | 261 | 5,154 | 151 | 42 | 10 | 203 |
| Mississippi | 2,332 | 602 | 96 | 2,934 | 97 | 24 | 3 | 124 |
| Missouri | 3,389 | 2,120 | 0 | 5,509 | 141 | 47 | 0 | 188 |
| Montana | 2,146 | 654 | 863 | 2,800 | 100 | 15 | 24 | 139 |
| Nebraska | 1,736 | 225 | 79 | 1,961 | 86 | 11 | 3 | 100 |
| Nevada | 1,141 | 973 | 189 | 2,114 | 70 | 26 | 5 | 101 |
| New Hampshire | 1,156 | 609 | 0 | 1,765 | 53 | 23 | 0 | 76 |
| New Jersey | 5,369 | 2,902 | 0 | 8,271 | 175 | 83 | 0 | 258 |
| New Mexico | 2,600 | 988 | 1,736 | 3,588 | 119 | 23 | 70 | 212 |
| New York | 9,301 | 9,390 | 88 | 18,691 | 332 | 290 | 5 | 627 |
| North Carolina | 4,405 | 3,438 | 93 | 7,843 | 177 | 95 | 4 | 276 |
| North Dakota | 1,780 | 404 | 381 | 2,184 | 95 | 14 | 12 | 121 |
| Ohio | 7,369 | 3,973 | 0 | 11,342 | 262 | 117 | 0 | 379 |
| Oklahoma | 2,696 | 970 | 260 | 3,666 | 193 | 31 | 8 | 232 |
| Oregon | 1,866 | 1,606 | 124 | 3,472 | 127 | 39 | 3 | 169 |
| Pennsylvania | 9,463 | 2,801 | 0 | 12,264 | 344 | 84 | 0 | 428 |
| Rhode Island | 1,200 | 574 | 0 | 1,774 | 49 | 20 | 0 | 69 |
| South Carolina | 2,505 | 1,994 | 0 | 4,499 | 103 | 39 | 0 | 142 |
| South Dakota | 1,657 | 520 | 439 | 2,177 | 95 | 14 | 27 | 136 |
| Tennessee | 4,071 | 1,748 | 0 | 5,819 | 156 | 51 | 0 | 207 |
| Texas | 13,031 | 7,331 | 29 | 20,362 | 588 | 204 | 1 | 793 |
| Utah | 1,640 | 846 | 122 | 2,486 | 73 | 27 | 7 | 107 |
| Vermont | 1,345 | 571 | 0 | 1,916 | 57 | 18 | 0 | 75 |
| Virginia | 3,765 | 3,122 | 0 | 6,887 | 156 | 102 | 0 | 258 |
| Washington | 4,064 | 2,043 | 416 | 6,107 | 147 | 68 | 17 | 232 |
| West Virginia | 1,108 | 769 | 0 | 1,877 | 56 | 23 | 0 | 79 |
| Wisconsin | 4,323 | 1,186 | 209 | 5,509 | 164 | 27 | 10 | 211 |
| Wyoming | 1,391 | 527 | 83 | 1,918 | 121 | 13 | 5 | 139 |
| Total U.S. | 184,155 | 108,028 | 8,730 | 300,913 | 7,774 | 3,174 | 355 | 11,303 |

[1]The source of the P-sample housing unit counts was the independent list that was confirmed to exist following the housing unit matching and follow-up operations.

[2]Cluster size was based on number of confirmed A.C.E. housing units after housing unit matching and follow-up.

### Identifying the E-Sample Housing Units

The E sample consisted of the census enumerations in the same sample areas as the P sample. The source of the E-sample housing units was the unedited census files. Like the P sample, all housing units in block clusters that had fewer than 80 census housing units or in block clusters selected from the American Indian Reservation stratum were designated to be in the E sample. For block clusters with 80 or more housing units, the housing units were reduced and the selected housing units were in the E sample.

The reduction of housing units within a large block cluster was done by mapping the P-sample segments onto the census housing units. This was possible because when there was a match between an A.C.E. independently listed address and a census address during the housing unit matching, the census identification number was linked to the A.C.E. unit. Then the same segment selected for the P sample was selected for the E sample.

The census inventory of housing units changed between the housing unit matching operation and the identification of the E sample. Therefore, some census housing units did not have a link with an A.C.E. unit. These cases were assigned to a segment using pre-specified rules. Sometimes there were a large number of these cases in the segment selected to be in sample. If there were more than 80 of these, then an additional subsample was drawn from these census housing units without a link to an A.C.E. unit.

The data-defined census person enumerations in the E-sample housing units were in the E sample. To be a census data-defined person, the person record had two 100-percent data items filled. Name was not required for the person record to be considered data-defined, but could be one of the two items required to be data-defined.

**Census housing units not available for the E sample.** Not all housing units on the unedited census file were eligible to be in the E sample. As the census enumerations were being processed, the Census Bureau suspected that there was a significant number of duplicate addresses in the census files. As such, a new census operation, the Housing Unit Duplication Operation, was introduced in the fall of 2000. The primary goal of this operation was to improve the quality of the census; however its design allowed the A.C.E. operations to proceed. Essentially, suspected duplicate housing units were set aside and analyzed further. These housing units and the corresponding census person enumerations were not eligible for the E-sample component of the A.C.E. nor available for person matching and were excluded from the dual-system estimation calculation. Some of these set-aside housing units and the corresponding census enumerations were later put back into the final census counts.

**Subsampling criteria.** If a block cluster contained 80 or fewer available census housing units, then all available census housing units were in the E sample. If the block cluster was from the American Indian Reservation stratum, all available housing units were in the E sample. If the block cluster had 80 or more available census housing units, the housing units were subsampled.

**Assigning housing units to segments.** Within a block cluster, the census housing units were assigned to a segment based on the link to an A.C.E. housing unit address. If there was a link with an A.C.E. unit, then the census housing unit was assigned to the same segment as the A.C.E. unit. This helped to create overlapping P and E samples. Sometimes a census housing unit did not have a link with an A.C.E. housing unit. When this happened, all the available census housing units were sorted and then each census housing unit without a link was assigned to the same segment as the preceding census housing unit. When the block cluster contained city-style addresses, the census housing units were sorted by census block number, street name, house number, and unit designation. When the block cluster contained non-city-style addresses, the census housing units were sorted by census block number and geographic location within the block. For city-style census addresses, geographic location was not available.

**Selecting the E-sample housing units.** Once all the census housing units within a block cluster were assigned to a segment, then the census housing units in the segment or segments selected for the P sample were in the E sample. Occasionally, the selected segment or segments within the block cluster contained more than 80 census housing units that did not link to an A.C.E. housing unit. When this occurred, an additional step of subsampling was done to reduce the E sample follow-up workload, since the census housing units without this link were more likely to contribute to the follow-up workload than census housing units with this link.

A systematic subsample of census housing units without a link to an A.C.E. housing unit was drawn. Using the same sort used for assigning housing units to a segment, a subsample of 40 housing units was selected if the resulting subsampling rate was greater than 0.25. However, to avoid excessive sampling weight variation, the minimum subsampling rate was set to 0.25, resulting in more than 40 census housing units without a link to an A.C.E. housing unit being in the E sample from the particular block cluster.

**Special case block clusters.** There were special case block clusters when none of the census housing units in a block cluster linked to an A.C.E. housing unit address at the time of the housing unit matching. One example of a special case was a List/Enumerate cluster, since the List/Enumerate operation had not been conducted by the time that the housing unit matching was done. None of the housing units in a List/Enumerate cluster could be assigned to a segment. Instead of selecting a compact segment of housing units to be in the E sample, a systematic subsample of the housing units was drawn using the same method as discussed above. This prevented overlapping the P and E samples when these block clusters were large. This did not happen often.

### E-Sample Results

Following E-sample identification and subsampling, the E sample for the 50 states and the District of Columbia was 11,303 block clusters containing about 311,000 housing units. Table 3-15 displays the results for each state.

Table 3-15. **State Third-Phase Sample Results for the E Sample**

| State | Housing unit[1] counts by cluster size[2] | | | | Block cluster counts by cluster size[2] | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 - 79 | 80+ | AIR | Total | 0 - 79 | 80+ | AIR | Total |
| Alabama | 2,776 | 1,793 | 0 | 4,569 | 113 | 48 | 0 | 161 |
| Alaska | 925 | 926 | 16 | 1,867 | 44 | 25 | 1 | 70 |
| Arizona | 2,819 | 2,521 | 2,521 | 7,861 | 152 | 57 | 113 | 322 |
| Arkansas | 1,838 | 1,118 | 0 | 2,956 | 85 | 23 | 0 | 108 |
| California | 17,906 | 16,228 | 271 | 34,405 | 658 | 432 | 11 | 1,101 |
| Colorado | 2,623 | 1,587 | 49 | 4,259 | 114 | 50 | 2 | 166 |
| Connecticut | 2,074 | 1,241 | 0 | 3,315 | 73 | 38 | 0 | 111 |
| Delaware | 1,270 | 659 | 0 | 1,929 | 45 | 21 | 0 | 66 |
| District of Columbia | 1,144 | 1,216 | 0 | 2,360 | 29 | 29 | 0 | 58 |
| Florida | 8,108 | 7,037 | 26 | 15,171 | 320 | 212 | 1 | 533 |
| Georgia | 4,373 | 3,346 | 0 | 7,719 | 179 | 97 | 0 | 276 |
| Hawaii | 1,323 | 2,653 | 0 | 3,976 | 49 | 72 | 0 | 121 |
| Idaho | 1,243 | 850 | 155 | 2,248 | 72 | 19 | 6 | 107 |
| Illinois | 8,190 | 4,302 | 0 | 12,492 | 288 | 115 | 0 | 403 |
| Indiana | 4,082 | 1,870 | 0 | 5,952 | 170 | 41 | 0 | 211 |
| Iowa | 2,237 | 907 | 0 | 3,144 | 102 | 20 | 0 | 122 |
| Kansas | 2,097 | 734 | 27 | 2,858 | 100 | 16 | 1 | 117 |
| Kentucky | 2,360 | 1,692 | 0 | 4,052 | 107 | 51 | 0 | 158 |
| Louisiana | 3,078 | 1,809 | 0 | 4,887 | 152 | 47 | 0 | 199 |
| Maine | 1,595 | 429 | 1 | 2,025 | 80 | 13 | 1 | 94 |
| Maryland | 2,651 | 2,786 | 0 | 5,437 | 92 | 73 | 0 | 165 |
| Massachusetts | 4,249 | 2,736 | 0 | 6,985 | 146 | 64 | 0 | 210 |
| Michigan | 6,682 | 3,311 | 146 | 10,139 | 253 | 85 | 5 | 343 |
| Minnesota | 3,183 | 1,720 | 260 | 5,163 | 148 | 45 | 10 | 203 |
| Mississippi | 2,374 | 647 | 114 | 3,135 | 99 | 22 | 3 | 124 |
| Missouri | 3,231 | 2,098 | 0 | 5,329 | 141 | 47 | 0 | 188 |
| Montana | 1,480 | 521 | 866 | 2,867 | 98 | 17 | 24 | 139 |
| Nebraska | 1,637 | 258 | 68 | 1,963 | 86 | 11 | 3 | 100 |
| Nevada | 907 | 1,175 | 133 | 2,215 | 67 | 29 | 5 | 101 |
| New Hampshire | 1,426 | 467 | 0 | 1,893 | 57 | 19 | 0 | 76 |
| New Jersey | 4,952 | 3,666 | 0 | 8,618 | 170 | 88 | 0 | 258 |
| New Mexico | 1,136 | 754 | 1,536 | 3,426 | 121 | 21 | 70 | 212 |
| New York | 9,114 | 11,071 | 84 | 20,269 | 326 | 296 | 5 | 627 |
| North Carolina | 4,510 | 3,253 | 101 | 7,864 | 182 | 90 | 4 | 276 |
| North Dakota | 1,401 | 482 | 358 | 2,241 | 95 | 14 | 12 | 121 |
| Ohio | 7,223 | 4,016 | 0 | 11,239 | 263 | 116 | 0 | 379 |
| Oklahoma | 2,366 | 1,038 | 265 | 3,669 | 193 | 31 | 8 | 232 |
| Oregon | 1,644 | 2,378 | 125 | 4,147 | 122 | 44 | 3 | 169 |
| Pennsylvania | 9,143 | 3,449 | 0 | 12,592 | 336 | 92 | 0 | 428 |
| Rhode Island | 1,194 | 556 | 0 | 1,750 | 50 | 19 | 0 | 69 |
| South Carolina | 2,502 | 1,968 | 0 | 4,470 | 105 | 37 | 0 | 142 |
| South Dakota | 1,278 | 495 | 433 | 2,206 | 95 | 14 | 27 | 136 |
| Tennessee | 4,022 | 2,429 | 0 | 6,451 | 157 | 50 | 0 | 207 |
| Texas | 12,412 | 9,213 | 27 | 21,652 | 574 | 218 | 1 | 793 |
| Utah | 1,434 | 818 | 123 | 2,375 | 75 | 25 | 7 | 107 |
| Vermont | 1,263 | 640 | 0 | 1,903 | 56 | 19 | 0 | 75 |
| Virginia | 3,555 | 3,731 | 0 | 7,286 | 152 | 106 | 0 | 258 |
| Washington | 3,371 | 2,609 | 411 | 6,391 | 144 | 71 | 17 | 232 |
| West Virginia | 1,173 | 724 | 0 | 1,897 | 57 | 22 | 0 | 79 |
| Wisconsin | 4,067 | 1,159 | 211 | 5,437 | 167 | 34 | 10 | 211 |
| Wyoming | 1,337 | 554 | 84 | 1,975 | 121 | 13 | 5 | 139 |
| Total U.S. | 178,978 | 123,640 | 8,411 | 311,029 | 7,690 | 3,258 | 355 | 11,303 |

[1]Available housing unit counts from the unedited census file.
[2]Cluster size was based on available census housing unit tallies.

### Third-Phase Sampling Results

Table 3-16 gives the state weighted and unweighted P-sample and E-sample housing units. Also displayed are the average P-sample and E-sample weights, prior to weight trimming, TES adjustment, and nonresponse adjustments. The average weights ranged from approximately 100 to 500.

In Table 3-16, for most of the states, the average E-sample weight is smaller than the average P-sample weight. Nationally, despite the P- and E-sample sizes differing by about 10,000 housing units, after applying the weight, the weighted number of P-sample housing units is less than one percent larger than the weighted number of E-sample housing units.

Table 3-16. **P-Sample and E-Sample Housing Unit Sampling Results**

| State | Weighted housing unit estimates | | | Housing unit sample sizes | | | Average weight | |
|---|---|---|---|---|---|---|---|---|
| | P sample | E sample | P/E | P sample | E sample | P/E | P sample | E sample |
| Alabama | 1,967,703 | 1,953,559 | 1.007 | 4,450 | 4,569 | 0.974 | 442 | 428 |
| Alaska | 186,971 | 187,657 | 0.996 | 1,739 | 1,867 | 0.931 | 108 | 101 |
| Arizona | 2,291,735 | 2,419,098 | 0.947 | 7,667 | 7,861 | 0.975 | 299 | 308 |
| Arkansas | 1,204,014 | 1,214,878 | 0.991 | 2,716 | 2,956 | 0.919 | 443 | 411 |
| California | 12,255,066 | 12,129,849 | 1.010 | 33,527 | 34,405 | 0.974 | 366 | 353 |
| Colorado | 1,633,980 | 1,579,070 | 1.035 | 4,153 | 4,259 | 0.975 | 393 | 371 |
| Connecticut | 1,262,197 | 1,249,792 | 1.010 | 3,243 | 3,315 | 0.978 | 389 | 377 |
| Delaware | 282,962 | 285,557 | 0.991 | 1,770 | 1,929 | 0.918 | 160 | 148 |
| District of Columbia | 295,972 | 295,099 | 1.003 | 2,190 | 2,360 | 0.928 | 135 | 125 |
| Florida | 7,350,667 | 6,958,799 | 1.056 | 15,254 | 15,171 | 1.005 | 482 | 459 |
| Georgia | 3,178,003 | 3,101,337 | 1.025 | 7,762 | 7,719 | 1.006 | 409 | 402 |
| Hawaii | 446,780 | 467,582 | 0.956 | 3,603 | 3,976 | 0.906 | 124 | 118 |
| Idaho | 475,978 | 494,377 | 0.963 | 1,995 | 2,248 | 0.887 | 239 | 220 |
| Illinois | 4,752,616 | 4,723,175 | 1.006 | 12,365 | 12,492 | 0.990 | 384 | 378 |
| Indiana | 2,565,559 | 2,611,248 | 0.983 | 5,945 | 5,952 | 0.999 | 432 | 439 |
| Iowa | 1,286,159 | 1,303,393 | 0.987 | 2,991 | 3,144 | 0.951 | 430 | 415 |
| Kansas | 1,054,277 | 1,085,066 | 0.972 | 2,666 | 2,858 | 0.933 | 395 | 380 |
| Kentucky | 1,738,637 | 1,688,359 | 1.030 | 3,979 | 4,052 | 0.982 | 437 | 417 |
| Louisiana | 1,690,093 | 1,767,498 | 0.956 | 4,417 | 4,887 | 0.904 | 383 | 362 |
| Maine | 606,684 | 580,671 | 1.045 | 1,932 | 2,025 | 0.954 | 314 | 287 |
| Maryland | 2,240,463 | 2,237,811 | 1.001 | 5,287 | 5,437 | 0.972 | 424 | 412 |
| Massachusetts | 2,637,732 | 2,652,699 | 0.994 | 6,393 | 6,985 | 0.915 | 413 | 380 |
| Michigan | 3,945,568 | 3,948,348 | 0.999 | 9,980 | 10,139 | 0.984 | 395 | 389 |
| Minnesota | 1,976,410 | 1,940,302 | 1.019 | 5,154 | 5,163 | 0.998 | 383 | 376 |
| Mississippi | 1,067,393 | 1,065,495 | 1.002 | 2,934 | 3,135 | 0.936 | 364 | 340 |
| Missouri | 2,678,909 | 2,576,545 | 1.040 | 5,509 | 5,329 | 1.034 | 486 | 483 |
| Montana | 463,607 | 459,884 | 1.008 | 2,800 | 2,867 | 0.977 | 166 | 160 |
| Nebraska | 684,874 | 667,586 | 1.026 | 1,961 | 1,963 | 0.999 | 349 | 340 |
| Nevada | 895,050 | 862,509 | 1.038 | 2,114 | 2,215 | 0.954 | 423 | 389 |
| New Hampshire | 558,641 | 523,562 | 1.067 | 1,765 | 1,893 | 0.932 | 317 | 277 |
| New Jersey | 3,377,908 | 3,338,768 | 1.012 | 8,271 | 8,618 | 0.960 | 408 | 387 |
| New Mexico | 708,714 | 667,620 | 1.062 | 3,588 | 3,426 | 1.047 | 198 | 195 |
| New York | 7,573,292 | 7,706,526 | 0.983 | 18,691 | 20,269 | 0.922 | 405 | 380 |
| North Carolina | 3,857,166 | 3,748,539 | 1.029 | 7,843 | 7,864 | 0.997 | 492 | 477 |
| North Dakota | 294,040 | 288,677 | 1.019 | 2,184 | 2,241 | 0.975 | 135 | 129 |
| Ohio | 4,785,461 | 4,687,680 | 1.021 | 11,342 | 11,239 | 1.009 | 422 | 417 |
| Oklahoma | 1,461,163 | 1,465,046 | 0.997 | 3,666 | 3,669 | 0.999 | 399 | 399 |
| Oregon | 1,411,681 | 1,431,030 | 0.986 | 3,472 | 4,147 | 0.837 | 407 | 345 |
| Pennsylvania | 5130,010 | 5,179,175 | 0.991 | 12,264 | 12,592 | 0.974 | 418 | 411 |
| Rhode Island | 408,426 | 401,022 | 1.018 | 1,774 | 1,750 | 1.014 | 230 | 229 |
| South Carolina | 2,274,389 | 2,332,485 | 0.975 | 4,499 | 4,470 | 1.006 | 506 | 522 |
| South Dakota | 300,952 | 297,492 | 1.012 | 2,177 | 2,206 | 0.987 | 138 | 135 |
| Tennessee | 2,489,607 | 2,609,919 | 0.954 | 5,819 | 6,451 | 0.902 | 428 | 405 |
| Texas | 8,116,215 | 8,098,923 | 1.002 | 20,362 | 21,652 | 0.940 | 399 | 374 |
| Utah | 885,164 | 823,255 | 1.075 | 2,486 | 2,375 | 1.047 | 356 | 347 |
| Vermont | 307,822 | 296,414 | 1.038 | 1,916 | 1,903 | 1.007 | 161 | 156 |
| Virginia | 2,714,879 | 2,797,836 | 0.970 | 6,887 | 7,286 | 0.945 | 394 | 384 |
| Washington | 2,496,269 | 2,435,145 | 1.025 | 6,107 | 6,391 | 0.956 | 409 | 381 |
| West Virginia | 917,901 | 916,552 | 1.001 | 1,877 | 1,897 | 0.989 | 489 | 483 |
| Wisconsin | 2,274,773 | 2,268,976 | 1.003 | 5,509 | 5,437 | 1.013 | 413 | 417 |
| Wyoming | 190,271 | 194,844 | 0.977 | 1,918 | 1,975 | 0.971 | 99 | 99 |
| United States | 115,650,803 | 115,016,729 | 1.006 | 300,913 | 311,029 | 0.967 | 384 | 370 |

# Attachment.
# Census 2000 Type of Enumeration Areas (TEAs)[1]

The term "TEA" has been used for several decennial censuses. For Census 2000, it reflects not only the type of enumeration, but also the method of compiling the census address list that controls the enumeration process.

The Census Bureau defines TEA codes at the census collection block level. Each block must have a TEA code, and no block may have more than one TEA code.

**TEA 1 - Block Canvassing and Mailout/Mailback**

- Contains areas with predominantly city-style (house number/street name) addresses used for mail delivery.
- Census address list is created from USPS, 1990 census, local/tribal, and other potential supplementary address sources.
- Blocks are included in both Block Canvassing and the Postal Validation Check.
- Blocks are included in local/tribal program to identify "new construction."

Mailout/mailback is the most efficient, cost-effective enumeration method in heavily populated areas in which mail is delivered to city-style addresses in virtually all cases (there may be scattered non-city-style mailing addresses in use in these areas). In most instances, a census enumerator visits a residence once—during Block Canvassing. A subsequent visit is sometimes necessary during Nonresponse Follow-up.

The mailing list used for this operation is derived initially from automated address files (the USPS Delivery Sequence File and the 1990 Census Address Control File), and updated through various operations, including Address List Review (LUCA1998), ongoing DSF updates, Block Canvassing, the Postal Validation Check, and the New Construction Program.

**TEA 2 - Address Listing and Update/Leave**

- Contains areas with some number of non-city-style (e.g., P.O. Box or Rural Route) mailing addresses.
- Census address list is created from Address Listing, and updated from Address List Review (LUCA) 1999 Recanvassing (in selected areas) and Update/Leave
- Blocks are NOT included in Block Canvassing, the Postal Validation Check, or the New Construction Program
- Puerto Rico, including its military bases, is completely in TEA 2

Address Listing and Update/Leave are implemented in areas where mail often is delivered to non-city-style addresses. In these areas, it is difficult to obtain an up-to-date mailing address list and then "geocode" each address (that is, assign it to a collection block code), because of the constantly changing residential location/mailing address relationship (especially for P.O. Box addresses). The census address list therefore is compiled through a door-to-door independent listing operation (Address Listing) that is implemented in all TEA 2 blocks.

During Address Listing, enumerators knock on each residence door to obtain the occupant's name, phone number, residential address (or location description), and mailing address. (Enumerators do NOT revisit residences whose occupants are not present. This is why the census address list frequently does NOT contain a mailing address, and why the location description is the ONLY "address" in the census address list for many residences.) Enumerators identify the location of each building (containing living quarters) they encounter with a uniquely numbered map spot that they enter on their map and record in their address register; this number is linked to all residential units in the building, and stored in both the census address list and the TIGER data base. These areas will be included in Address List Review (LUCA) 1999.

At census time, enumerators deliver census questionnaires to all housing units compiled during Address Listing and that remain in TEA 2. In the course of delivering these questionnaires, the enumerators also update the census address list and map spotted map to reflect housing units that were not previously listed, and to eliminate residences that they cannot locate. (This operation is called Update/Leave, because the enumerators UPDATE the census address list and maps and LEAVE questionnaires.) Update/Leave enumerators use the residential address/location description in conjunction with the map spot location to determine the correct delivery point for all questionnaires.

Most housing units in TEA 2 areas are visited at least twice by enumerators — once during Address Listing, and again during Update/Listing. Respondents must mail their completed census questionnaires to the Census Bureau, and so some residences also will be visited a third time, during Nonresponse Follow-up.

[1]This documentation is reproduced from the Geography Division, U.S. Census Bureau, Web site located at http://www.geo.census.gov/mob/homep/teas.html.

**TEA 3 - List/Enumerate**

- Contains areas that are remote, sparsely populated, or not easily accessible
- Census address list is created and enumeration conducted concurrently
- Blocks are not included in Block Canvassing, the Postal Validation Check, the New Construction Program, or Address Listing
- Includes all military bases in TEA 3 areas
- All island areas (except Puerto Rico), including their military bases, are TEA 3

Some areas are remote, sparsely populated, and/or not easily visited. Many of the residences in these areas do not have city-style mail delivery. It is inefficient and expensive to implement Address Listing, Update/Listing, and Nonresponse Follow-up operations involving multiple visits. Instead, the creation of the address list and the delivery/completion of the census questionnaire are accomplished during a single operation, List/Enumerate. Enumerators visit residences in TEA 3 blocks, LIST them for inclusion in the census address list, mark their location on their map with a map spot and number, enter that map spot number in their address register, and ENUMERATE the residents on-site. They collect the same address information as in Address Listing, and include a map spot to reflect each building that contains one or more living quarters. These areas will NOT be included in any Address List Review (LUCA) program, because there is no address list for them in advance of the census.

**TEA 4 - Remote Alaska**

- Similar to List/Enumerate, but conducted earlier, before ice breakup/snow melt
- These areas will NOT be included in any Address List Review (LUCA) program, because there is no address list for them in advance of the census

**TEA 5 - "Rural" Update/Enumerate**

- Contains blocks initially in TEA 2, with map spots for all structures containing at least one housing unit
- In some instances, blocks initially in TEA 3 will be converted to TEA 5. These blocks were not included in Address Listing and LUCA 1999, and therefore lack structures and map spots in the MAF and TIGER at the times that LUCA 1999 and "Rural" Update/Enumerate are conducted
- Self-enumeration (through Update/Leave) is thought to be unlikely or problematic
- Census address list is updated, and enumeration is conducted, concurrently
- Blocks are NOT included in the Postal Validation Check or the New Construction Program
- The term "rural" reflects Address Listing as the initial source of the census address list, and does NOT reflect the official census definition of the term "rural"
- These areas will be included in Address List Review (LUCA) 1999 materials, as the MAF was compiled initially from Address Listing

In some areas that otherwise meet the criteria for inclusion in TEA 2, the Census Bureau has decided that having respondents enumerate themselves and return their questionnaires via the mail is not the best way to conduct the enumeration. Some targeted populations may be less likely to return their questionnaires in the mail, and more likely to respond to an enumerator. In other areas, housing units may be vacant because they are occupied seasonally.

In these and comparable situations, enumerators visit all residences on the census address list and complete the enumeration on-site. In the course of delivering these questionnaires, they also update the census address list to 1) reflect housing units that were not previously listed (including a map spot to reflect each building that contains one or more living quarters), and 2) eliminate housing units that they cannot locate. (This operation is called "Rural" Update/Enumerate, because the enumerators work in areas that were Address Listed, UPDATE the census address list [and assign map spots as well], and ENUMERATE the residents.)

**TEA 6 - Military**

- Contains blocks within TEA 2 that are on military bases
- Mailout/Mailback for family housing
- Separate enumeration procedures for barracks, hospitals, etc.
- Blocks are included in both Block Canvassing and the Postal Validation Check
- These blocks are included in Address List Review (LUCA) 1998 materials, as the MAF was compiled initially in the same manner as TEA 1 areas

The Department of Defense has advised the Census Bureau that virtually all family housing (that is, individual residences as opposed to barracks, hospitals, and jails) are assigned city-style addresses to which the Postal Service delivers mail. The Census Bureau therefore implements Mailout/Mailback methods to enumerate the population of these individual residences. Within TEA 1 areas, blocks on military bases are assigned a TEA code of 1. Within TEA 2 areas, blocks on military bases are assigned a TEA code of 6. There is no difference between TEA 1 blocks on military

bases and TEA 6 blocks in terms of either compiling the census address list or enumerating the population. Blocks within military bases in List/Enumerate areas (TEA 3) also are TEA 3.

### TEA 7 - "Urban" Update/Leave

- Contains blocks initially in TEA 1
- Census address list is updated, and questionnaires are delivered concurrently, by Census Bureau staff (following procedures employed in TEA 2 areas, but without assigning map spots)
- Blocks ARE included in the Postal Validation Check and the New Construction Program
- The term "urban" reflects the predominance of city-style addresses, and does NOT reflect official census definition of the term "urban"
- These blocks are included in Address List Review (LUCA) 1998 materials, as the MAF was compiled initially in the same manner as TEA 1 areas

In many areas where mail is delivered mostly to city-style addresses, older apartment buildings are common. In many of these buildings, unit designators (that is, apartment numbers), often do not exist. Further, the subdivision of existing units into multiple units, and the conversion of non-residential space to living quarters, may be frequent. Mail, therefore, often is not delivered to individual apartments (or individual mail boxes), but instead left at common drop points.

In some other areas with mostly city-style addresses, many residents have elected to receive their mail at post office boxes. The Census Bureau is concerned that the city-style addresses of these residents may not appear in the census address list.

To ensure questionnaire delivery to the largest number of residences, Update/Leave procedures are employed. As these residences have city-style addresses, there is no need for enumerators to assign map spots to assist enumerators in identifying these residences in subsequent operations.

### TEA 8 - "Urban" Update/Enumerate

- Contains blocks initially in TEA 1, without map spots for any addresses; maps generated for TEA 8 areas will not include map spots
- Contains mostly blocks on those American Indian reservations that initially were included in both TEA 1 and either TEA 2 or 3
- Same enumeration procedures as TEA 5
- The term "urban" reflects the initial inclusion of the block in TEA 1 due to the predominance of city-style mailing addresses
- These areas are included in Block Canvassing and the Postal Validation Check

Most American Indian Reservations will be enumerated using a single enumeration procedure (Mailout/Mailback, Update/Leave, or Update/Enumerate). Some of these initially contained blocks with a mixture of TEA codes. In these instances, the reservations will be enumerated using Update/Enumerate methods (see TEA 5). However, for affected blocks initially in TEA 1, the MAF and TIGER do not include map spots for structures containing at least one housing unit. Instead of converting these blocks to TEA 5 ("Rural" Update/Enumerate) and determining map spot locations, the blocks are being distinguished by a separate TEA.

### TEA 9 - Additions to Address Listing Universe of Blocks

- Contains groups of blocks (assignment areas) initially assigned to TEA 1
- Converted to Address Listing before Block Canvassing is conducted
- Blocks are NOT included in Block Canvassing, the Postal Validation Check, or the New Construction Program

Some blocks that are in TEA 1 contain a significant number of living quarters with non-city-style addresses. These blocks should not be included in Block Canvassing, which is an operation that is designed to confirm and correct the existence and/or location of city-style addresses. The Geography and Field Divisions are identifying Block Canvassing assignment areas (AAs) that likely contain blocks with significant numbers of non-city-style addresses. Some of these AAs will be removed from Block Canvassing, and included in Address Listing. The blocks in these AAs will be assigned a TEA code of 9, and the census address list compilation and census enumeration activities in TEA 9 blocks will be virtually identical to those in TEA 2 blocks (for instance, they will be included in Update/Leave and Nonresponse Follow-up).

Because most of these blocks had few, if any, addresses in the MAF from the USPS, the entities the blocks are in mostly had nothing to review during Address List Review (LUCA) 1998. For this reason, most of these blocks will have their Address List Reviewed during a new phase of LUCA, often called "LUCA 99 1/2 ."

# Chapter 4.
# A.C.E. Field and Processing Activities

## INTRODUCTION

This chapter describes the operational aspects of the A.C.E. which consisted of four major activities: housing unit listing, housing unit matching, person interviewing, and person matching. Housing unit listing and person interviews were conducted as field activities, whereas housing unit matching and person matching were processing activities carried out in the National Processing Center (NPC) in Jeffersonville, Indiana. As described earlier, all of these activities were completed prior to estimation. Once the sample clusters were selected, interviewers visited the clusters and independently listed all housing units. The A.C.E. and census housing units were then matched and, for those for which a match was not found, a follow-up interview was conducted to determine the status of the housing unit at the time of the census.

Following the resolution of the housing unit nonmatches, interviews were conducted with residents of the A.C.E. sample household (P sample) to obtain the roster of household residents and the detail required for matching. The P-sample persons were then matched to the list of persons enumerated in the census in the sample clusters. The search area was expanded to include one ring of surrounding blocks for those clusters identified as containing potential census geocoding errors. This operation was called the targeted extended search (TES) because it targeted clusters with high rates of A.C.E. housing unit nonmatches and census housing unit geocoding error. A further follow-up interview was conducted for selected mismatched people for whom additional information was required. Based on these activities, each person in the sample clusters, whether interviewed in the A.C.E. sample (P sample) or found in the census (E sample) was assigned a final match status code.

It is important to point out some key improvements of the A.C.E. 2000 operations over the 1990 Post-Enumeration Survey (PES). The 2000 A.C.E. improved on 1990 PES in several ways for interviewing and clerical matching.

- One problem in 1990 was the misreporting of Census Day addresses, with an estimated 0.7 percent of the P sample being erroneously reported as nonmovers (West 1991). The Computer Assisted Personal Interview (CAPI) instrument improved the quality of the reporting of mover status because it was a more automated process. In 2000, the Census Day household consisted of nonmovers and outmovers. The nonmovers lived in the housing unit at the time of the interview and on Census Day. The outmovers lived in the housing unit on Census Day, but moved before the A.C.E. interview. Nonmovers and outmovers in the P sample were matched to census people in their block cluster. In 1990, each inmover household (those that moved into PES block clusters after Census Day) had to be matched to a Census Day address, which was usually outside the cluster. In 2000, the reconstructed Census Day household was matched to the census enumerations in the sample block cluster.
- A study of clerical error in the 1990 PES found error in coding matches (Davis 1991) and erroneous enumerations (Davis 1991b). In 1990, codes were entered into a computer system, but the actual matching and duplicate searches were done using paper. In the 2000 A.C.E., the matching was better controlled and more efficient than 1990 because the clerical matching and quality assurance were automated and coded directly into the automated system. The automated interactive system did not prevent all matching error, but reduced the chances for error significantly. Software allowed searching for matches in the census based on first names, last names, characteristics, and addresses. For example, the system allowed searching for all people named George, all people whose last name begins with an H, all people on Elm Street, or everyone in the age 30 to 40 range. The software controlled the match codes that were relevant to the situation. For example, only P-sample nonmatch codes could be assigned to a P-sample nonmatch.
- The electronic searches for duplicates reduced the tedious searching through paper lists of census people. The searching in 1990 was limited to printouts in two sorts: last name and household by address. In 2000, the clerks had the capability to filter on name, characteristics, and address to help identify duplicates. The system monitored whether the matcher had completed all the necessary searches, such as looking for duplicates.
- There were built-in edits to ensure consistency of coding. For example, codes that applied to a household, such as geographic codes, were assigned to all people in the household. The system automatically assigned certain codes, reducing coding error.
- Clerical matchers could use a code indicating the case needed review at the next level of matching. This code allowed them to flag unusual cases to be examined by a person with more experience.

- All quality assurance for the clerical matching was automated.
- Clerical matching was centralized at the NPC instead of having separate groups of matchers in the seven processing offices, as was done in 1990. Forty-six technicians were hired in August, 1999 and were thoroughly trained in the design of the A.C.E. and methods of matching people and housing units. These technicians were responsible for quality assurance of the clerical matchers. Additionally, ten analysts who were among the most experienced matchers conducted quality assurance for the technicians and handled the most difficult cases.
- The computer matcher identified matches and possible matches within a block cluster. Additional computer programs were used to check the matching on cases after the before follow-up clerical matching to identify matches and duplicates in the expanded search area that were not identified by the clerical matchers. Consistency checks were also performed between housing unit and person match codes.
- Keying error in the data capture of the 1990 PES was reduced because the 2000 interview used a CAPI instrument. A more accurate capture of the data increased the efficiency of the computer matching.

## HOUSING UNIT LISTING

The first stage of sampling was the selection of A.C.E. block clusters. Then, in September through December of 1999, a listing of the addresses of all the housing units in the A.C.E. sample clusters was conducted. The listing was independent of the census. Training in how to list both city-style and non-city-style areas lasted 3 days and included a review of the first completed cluster assigned to each lister. There were 29,136 sampled block clusters in the 50 states and the District of Columbia. This list of housing units recorded in the Independent Listing Books (ILB) became the frame of A.C.E. housing units from which the P sample was later selected. Besides listing each housing unit in the cluster, the listers inquired about housing units present at each special place and commercial structure.

The housing unit listing was by basic street address. Each basic street address was assigned a map spot number and the map spot number was recorded on the A.C.E. map to identify the location of the basic street address. The address and coverage questions about the structure were asked for each basic street address. The number of housing units at the basic street address was obtained from a household member at the address, by proxy, from the apartment manager, or by observation. This contact helped to improve the coverage of housing units in the A.C.E. A page in the listing book for single and multiunit structures is shown in Figure 4-1. The individual housing units within a basic street address were listed on the pages of the listing book reserved for multiunits. Also, the A.C.E. lister recorded the number of units within a basic street address on the map in parentheses to conform with census methodology.

Mobile homes that were not in mobile home parks were listed like single units. Each mobile home was assigned a unique map spot number and each mobile home was listed on a separate line in the listing book. If the mobile homes were in a park, the park was listed in the housing unit section of the listing book, and each individual mobile home and vacant site was listed in the mobile home park section of the listing book. Each individual mobile home was assigned a unique map spot number, whether the mobile home was in a park or not. The location of the mobile home was identified by placing the map spot number for the mobile home on the map. This was the same procedure that was used in the census.

The following items were collected and recorded in the listing book for each basic street address:

- City-style addresses (house number and street names)
- Non-city-style addresses (route numbers, route and box numbers, or any other type of address that was not a city-style address)
- Householder names (rural areas only)
- Description of addresses (for only nonhouse number addresses in both urban and rural areas)
- Number of housing units in a basic street address
- Type of basic street address (single unit, multiunit, mobile home not in a mobile home park, mobile home in a mobile home park, housing unit in special place, multiunit in a special place, or other)
- Unit status for single units (occupied or intended for occupancy, under construction, future construction, unfit for habitation, boarded up, storage of household goods, and other)

The following items were also collected and recorded in the listing book for each unit within a multiunit basic street address:

- Unit designation
- Unit status for multi units (occupied or intended for occupancy, under construction, future construction, unfit for habitation, boarded up, storage of household goods, and other)

The following items were also collected and recorded in the listing book for each mobile home in a mobile home park:

- House number, lot number, or physical description
- Street name

- Rural address
- Unit status (intended for occupancy, unfit for habitation, boarded up, storage of household goods, vacant trailer site in a mobile home park, and other)

After the listing books were received in NPC, they were checked in and the data keyed into a computer file. The keying quality assurance was 100 percent. Keying rejects were reviewed clerically to correct errors before the matching began. A data file of the A.C.E. housing units was created to be used as input to the housing unit matching.

## HOUSING UNIT MATCH

The housing unit matching consisted of four steps: computer matching, clerical matching, housing unit follow-up, and after follow-up coding. The A.C.E. housing units were compared to the census housing units within cluster by computer, and then, clerically. Housing units that did not match, possible matches, and possible duplicates were followed up by field inspection and interview. The results of the follow-up interview were recorded during the after follow-up coding.

The purpose of housing unit matching was to create a list of addresses that existed as housing units in the block cluster on Census Day to use in the P-sample interviewing. The housing unit listing was conducted in the Fall of 1999. Addresses that had a chance to be housing units on Census Day, such as under construction, future construction, and vacant trailer sites, were listed. After the housing unit matching and follow-up, only the housing units originally listed and confirmed to exist as housing units were included in the P sample for CAPI interviewing. Housing units with unresolved status were also included in the interviewing.

Computer matching was conducted after the second phase of sampling, which consisted of sample reduction and small block subsampling. The results of the computer matching were reviewed clerically. All matching was conducted within the sample block clusters. The census addresses were the ones contained in the January, 2000 version of the Decennial Master Address File (DMAF). This was not the final version of the inventory of census addresses, because of later operations. The inventory of census housing units was final after the Hundred Percent Census Unedited File (HCUF) was completed.

As noted earlier, the P and E samples were located in the same block clusters. The advantages of linking the A.C.E. and census housing units were:

- The link of A.C.E. and census addresses allowed an overlapping P sample and E sample, (i.e., the housing units selected for the P sample were mostly the same as those in the E sample) eliminating the error prone clerical E-sample identification required to achieve the overlapping samples in the 1990 Post Enumeration Survey.
- The linking of addresses also allowed the person interviewing to begin earlier on the telephone using the census telephone number for the census questionnaire returned by mail. The telephone number from the census questionnaire was not available without the link between the A.C.E. housing unit and the census questionnaire for that housing unit.

After sample reduction of clusters, there were 11,303 clusters in the 50 states and the District of Columbia. See Chapter 3 for a discussion on the sample reduction. The 420 clusters in list/enumerate areas were not matched, because their census addresses were not available in the Spring of 2000. Therefore, 10,883 clusters were matched in the housing unit phase. Table 1 contains the number of housing units and clusters in housing unit matching for the A.C.E. The census numbers were preliminary; these were the addresses prior to mailing the census questionnaires. Subsequent census operations added and removed addresses from this list. Even though this census list contains more housing units than the A.C.E., this was not indicative of coverage differences due to the preliminary nature of the census numbers. See Chapter 3 for a discussion of the final P-sample and E-sample housing units and how they compare.

Table 4-1. **Sample Sizes for the A.C.E. Housing Unit Matching**

| | Clusters | Housing units |
|---|---|---|
| Clusters with housing units | 10,157 | |
| A.C.E. housing units | | 838,427 |
| Census housing units | | 859,296 |
| Clusters without housing units | 726 | |
| Total clusters in housing unit matching | 10,883 | |

### Computer Match

The census housing units included on the DMAF in January, 2000, in the block clusters retained in the A.C.E. after sample reduction and small block subsampling, were used in the housing unit matching. The housing unit data from the independent listing book file and the DMAF extract went through a series of data preparation steps, including address standardization. Addresses from either file that were blank or could not be standardized were matched clerically. The results of the computer matching and images of the A.C.E. and census maps with map spots in rural areas were inputs into an automated review and coding software for clerical matching.

### Clerical Match

The clerical matchers used the results of the computer matching to aid in their matching of addresses from the

A.C.E. and the census. There were 115 clerks, 46 technicians, and ten analysts involved in the matching operation. The clerks carried out the matching. The technicians applied quality control to the matching performed by the clerks. The analysts carried out quality control on the work of the technicians. The clerks and technicians used a "review" code when they saw something unusual or something that should have been looked at by the next level of matcher. The technicians and analysts examined the cases coded for review in the previous stage of matching, in addition to cases selected for quality control. The clerks used in the housing unit matching were given 4 weeks of training. The technicians were hired in August, 1999 and given extensive training on the background of coverage measurement and the design of the A.C.E. allowing them to make more informed decisions. The analysts were our most experienced people. The analysts have worked on coverage measurement for many years and were quite knowledgeable about the A.C.E. The three levels of staff produced a high quality of matching with a cost-efficient operation.

The clerical matching was conducted in the housing unit matching phase of the A.C.E. only for clusters expected to benefit from further examination. Since clerical matching was labor intensive, the amount of clerical work performed for the 2000 A.C.E. was reduced by an automated identification of clusters for follow-up interviewing without clerical review. These clusters had only a few nonmatches or nonmatches on only one side. For example, there could be 25 A.C.E. nonmatches and no census nonmatches, so there was nothing the clerical matchers could do. The clerical matchers were thus able to concentrate on the more difficult clusters where the review was beneficial. In 2000, 3,267 clusters were sent to the field for the follow-up phase without clerical review.

Supplemental materials were provided to facilitate the clerical matching, such as the maps with spots to identify the location of A.C.E. and census addresses in rural areas. The A.C.E. and census addresses that could not be matched by the computer were identified for the clerical matching. The matched addresses were not targeted for review, because experience in studies preparatory to the 2000 Census indicated a very high quality of the matches assigned by the computer. However, clerks were allowed to correct any errors in the computer matching that they noticed, while they were attempting to match the housing units that were not computer matched.

The clerical matchers used all housing unit information available to match housing units. The urban areas were almost totally city-style addresses. In rural areas, the addresses were more difficult to match, mainly because of the non-city-style addresses. The matchers had householder names and location descriptions to help in matching the A.C.E. and census addresses in rural areas. The spotted maps for the A.C.E. and the census were also used in the final determination of which housing units matched in rural areas. Computer images of the A.C.E. and census spotted maps that were used in the housing unit matching were accessed via the matching software and viewed on the screen.

There was also a clerical search, limited to the block cluster, for duplicate housing units during this phase of the matching. The possible duplicates were linked in the database for both the A.C.E. and the census. A follow-up interview was conducted to determine if the two addresses referred to the same housing unit.

One goal for the 2000 A.C.E. was not to use any paper in the clerical matching. Almost all materials needed for clerical matching were available on the computer. Paperless matching reduced the time needed for clerical matching, because the time spent waiting for an assignment and associated material was eliminated. There was thus no need for a large staff to maintain an A.C.E. library. Paper maps were available to use for cases where the image of the map was not available or was not easy to view in the software.

The quality assurance was applied as follows: all of the work done by each clerical matcher was reviewed initially until the matcher was determined to be performing at an acceptable level of quality. The number of records to be reviewed before a clerical matcher was classified as acceptable was 200, after which an "acceptable clerk" had a systematic sample of clusters reviewed for quality assurance. There was a computer record of the level of quality of each clerk's work. If the work in the sample of reviewed clusters fell below the acceptable level of quality, all of the subsequent work of that clerk was reviewed by technicians, until the clerk achieved an acceptable level of quality, then sampling was resumed. The analysts performed the same type of quality assurance on the technicians.

Table 4-2 contains the results of before follow-up clerical matching. These numbers include only the housing units in clusters that were processed in the housing unit matching. The list/enumerate clusters are therefore not included. The relisted clusters described at the end of this section are also not included in Table 4-2. The census had more possible duplicates and housing units not matching than the A.C.E. The follow-up interview resolved the housing unit status and determined if the possible duplicates were in fact duplicated.

Table 4-2. **Housing Unit Matching Results Before Follow-Up Interviewing**

| | A.C.E. | | Census | |
|---|---|---|---|---|
| | Housing units | Percent | Housing units | Percent |
| Matched | 681,385 | 81.6 | 681,385 | 79.7 |
| Possible match | 29,231 | 3.5 | 29,231 | 3.4 |
| Possible duplicate | 735 | 0.1 | 5,775 | 0.7 |
| Not matched | 123,469 | 14.8 | 138,657 | 16.2 |
| Remove from A.C.E. | 10 | 0.0 | | |
| Total | 834,830 | 100.0 | 855,048 | 100.0 |

### Housing Unit Follow-Up

All of the cases coded as not matched, possibly matched, or possibly duplicated were sent for a follow-up interview, regardless of the type of basic street address code. Selected matched cases were also sent to follow-up to collect additional information. Specifically, the cases identified for field follow-up were:

- **A.C.E. addresses with a before follow-up code of not matched.** Information was obtained to determine whether the addresses were housing units within the sample cluster.
- **Census addresses with a before follow-up code of not matched.** Information was obtained to determine whether the addresses were housing units within the sample cluster.
- **Possible matches.** The possible matches were sent to the field to determine if the A.C.E. and census addresses referred to the same housing unit. If they did not, they were identified as an A.C.E. nonmatch and a census nonmatch during the housing unit follow-up and information was obtained to determine whether the addresses were housing units within the sample cluster.
- **Possible census duplicates.** Census housing units that were identified as possible duplicates were followed up to determine if the two census addresses referred to the same housing unit.
- **Possible A.C.E. duplicates.** A.C.E. housing units that were identified as possible duplicates were followed up to determine if the two A.C.E. addresses referred to the same housing unit.
- **Matched housing units with a code of under construction, future construction, unfit for habitation, vacant trailer site in a mobile home park, other.** These matches were followed up to determine if they fit the definition of a housing unit at the time of the follow-up interview.

An A.C.E. housing unit with unit status indicating something other than an occupied or vacant housing unit that was intended for occupancy needed a follow-up interview to determine its status at the time of the follow-up interview. The address was either classified as a housing unit or removed from further processing. For example, a unit that was under construction or future construction at the time of listing may have fit the definition of a housing unit at the time of the follow-up interview. If the unit fit the definition of a housing unit, it was included in the A.C.E. housing unit processing. If construction had not progressed enough for it to fit the definition of a housing unit, it was coded as removed from the A.C.E. housing unit inventory.

The housing unit follow-up forms were computer generated. The questions for housing units requiring a follow-up interview were printed. In addition, all housing units in the block cluster were printed for reference. The questions for the A.C.E. nonmatches are in Figure 4-2. The same questions were asked for the census nonmatches.

The questions on the follow-up form were not designed to be read to respondents, but were intended to be used as a guide for an interviewer. Indeed, many questions were answered by observation. The answer to one question may have been the result of asking several other questions. The follow-up interviewer appropriately modified the questions, when necessary, to the situation that was encountered in the field and recorded the appropriate answers on the follow-up form. This approach was adopted because there were many situations that could occur and a form to cover every possible situation would be cumbersome to handle. It was necessary to find out if the housing unit satisfied the census housing unit definition at the time of the follow-up interview. There was no attempt to gather information about reasons for being something other than a housing unit.

For example, the follow-up interviewer determined if the address for an A.C.E. independent listing nonmatch or a census nonmatch existed as a housing unit. This was not a question meant for a respondent. There were several reasons why an address might not fit the definition of a housing unit, such as it burned, it was a mobile home that moved, it was converted to fewer housing units, it was group quarters, it was used for storage of farm machinery, it was the laundry room in an apartment complex, it was a

Table 4-3. **After Follow-Up Housing Unit Matching Results**

| | A.C.E. | | Census | |
|---|---|---|---|---|
| | Housing units | Percent | Housing units | Percent |
| Matched | 719,013 | 86.1 | 719,013 | 84.1 |
| Not matched, but existed in the block cluster | 76,418 | 9.2 | 28,874 | 3.4 |
| Did not exist as a housing unit | 30,770 | 3.7 | 48,684 | 5.7 |
| Geocoded outside the cluster | 6,316 | 0.8 | 45,053 | 5.3 |
| Duplicate | 1,157 | 0.1 | 12,296 | 1.4 |
| Unresolved | 1,156 | 0.1 | 1,128 | 0.1 |
| Total | 834,830 | 100.0 | 855,048 | 100.0 |

business, and so forth. The interviewer appropriately modified the questions, as necessary, to the situation that was encountered in the field. Furthermore, the interviewer could identify matches or duplicates in the field that had not been identified in the clerical matching.

Additional matches were also identified between the A.C.E. and census addresses during the follow-up interview, when the interviewer realized the two different addresses in the A.C.E. and census referred to the same unit. Corrections and updates to the addresses were also recorded on the follow-up form. The address updates were keyed into the database to accurately identify A.C.E. housing units for the person interviewing. The follow-up interviewers were instructed not to add housing units missed by both the A.C.E. and census for the 2000 A.C.E.

### After Follow-Up Coding

After the field follow-up, the completed forms were returned to the processing office. Using the information obtained during the field work, an after follow-up match code was assigned by the clerical matchers for cases sent to the field. The technicians and analysts reviewed the clusters containing housing units with a review code and carried out quality assurance for the clusters processed in the after follow-up housing unit matching.

The follow-up forms were reviewed clerically and codes were assigned to the A.C.E. and census housing units. Table 4-3 provides housing unit matching results for all A.C.E. and census housing units after the follow-up interview codes were assigned. A.C.E. housing units classified as existing in the block cluster and housing units with unresolved housing unit status were eligible for person interviewing. This included both matched and not matched units. A.C.E. addresses classified as not housing units, duplicates, and geocoding errors were removed from the A.C.E. universe, and therefore, were not eligible for person interviewing. The numbers in Table 4-4 are the A.C.E. housing units that were eligible for person interviewing before sample reduction. These numbers do not include the relisted clusters and clusters in list/enumerate areas. Census housing units with codes of not matched and unresolved statuses were not eligible to be included in the P sample for interviewing because they were not listed in the A.C.E. independent listing.

Table 4-4. **A.C.E. Housing Units Eligible for Person Interviewing**

| | A.C.E. | |
|---|---|---|
| | Housing units | Percent |
| Matched | 719,013 | 90.3 |
| Not matched, but existed in the block cluster | 76,418 | 9.6 |
| Unresolved | 1,156 | 0.1 |
| Total | 796,587 | 100.0 |

### Relisting for Clusters with A.C.E. Geocoding Errors

The follow-up operation also examined potential geocoding errors in the original A.C.E. housing unit listings. If a large proportion of the A.C.E. housing units in the cluster had wrong geocodes, the cluster was relisted. Clusters were identified for relisting when the after follow-up coding described in the previous section was completed. The decision to relist was automated. If 80 percent of the housing units in a cluster had geocoding error, the cluster was relisted. There were 62 relisted clusters in the 50 states and the District of Columbia. The field lister for relisted clusters had no previous contact with this cluster.

The relisting operation was carried out independently of the list of census housing units. To assure independence, the A.C.E. housing unit listings (both the original listing and the relisting) were done without the A.C.E. lister seeing the census inventory of housing units.

There was no housing unit matching in the relisted clusters during the housing unit matching phase of A.C.E. The addresses listed for A.C.E. during the relisting operation were the addresses used to conduct person interviewing. These clusters were treated in the same way as the list/enumerate clusters in 2000.

An unresolved code was assigned to all of the A.C.E. housing units in the relisted clusters and in the list/enumerate clusters. The census housing units in these clusters were assigned a blank housing unit code.

## PERSON INTERVIEW

Prior to person interviewing there was another stage of sampling, the within block subsampling of large block

clusters. See Chapter 3 for more details. The resulting housing units from A.C.E. comprised the P-sample housing units assigned for interviewing. There were 11,303 clusters selected for interviewing, and they contained 300,913 P-sample housing units. The person interview training lasted five days.

A.C.E. mover and residence status codes necessary to identify P-sample people from the person interview are assigned within the interview instrument. These codes are described in Figure 4-3.

The goal of the interview was to obtain a household roster for everyone living at the housing unit at the time of the interview and on Census Day, April 1, 2000. Procedure C was used for the 2000 A.C.E. With Procedure C, each A.C.E. person was assigned an A.C.E. mover code, an A.C.E. born since Census Day code, an A.C.E. group quarters code, and an A.C.E. other residence code. The A.C.E. status code combined all of the information from these codes to identify the people for whom matching was necessary. Attachment 1 contains the definitions for codes for the movers, those born since Census Day, members of group quarters, other residence code, and the A.C.E. status code. See the Chapter 7 attachment for more on Procedure C.

Group quarters were not listed in the A.C.E. and A.C.E. interviews were not conducted in group quarters. See Attachment 2 for a discussion of the treatment of group quarters in A.C.E.

### Mode of Interview

The A.C.E. person interview was conducted using a CAPI instrument on laptop computers. Attachment 3 contains a description of the procedures followed in the person interview. Some person interviews were conducted by telephone and some by personal visit.

To get an early start for the interviewing, a telephone interview was conducted at households where the census questionnaire included a telephone number and was received at a census processing office early enough for computer processing, before the start of person interviewing. The telephone number came from the census questionnaire of the matching census housing unit. The person interviews conducted by telephone were conducted from April 24, 2000 until June 13, 2000. See Byrne et al. (2001) for more details. A total of 88,573 interviews or 29.4 percent of the total workload were conducted by telephone.

The following cases were excluded from the A.C.E. telephone interviewing:

- Housing units in census large household and census coverage edit follow-up
- Questionnaires that were not returned by mail
- Housing units without house number and street name addresses
- Housing units in small multiunit structures (i.e., less than 20 units)

Large multiunits were able to be included in the telephone interviewing, because they tended to have unique unit designations. Many small multiunit structures and rural areas did not have addresses that allow the telephone interviewer to distinctly identify the address. Since there was no housing unit matching in relisted and list/enumerate clusters, all person interviewing in relisted and list/enumerate clusters was by personal visit.

All remaining interviews after the end of the telephone operation were conducted in person, except for some nonresponse conversion operation (NRCO) interviews and interviews in gated communities or secured buildings. The person interviews conducted by personal visit were conducted from June 18 until September 11, 2000. Crew leaders and supervisors conducted telephone interviews to give them experience in interviewing.

Table 4-5 contains the number of interviewers, crew leaders, and supervisors used during production interviewing and during the interviewing for person follow-up after the clerical matching.

Table 4-5. **Field Interview Personnel**

| | Telephone interview | Personal interview | Person follow-up |
|---|---|---|---|
| Interviewers | 450 | 4,502 | 4,470 |
| Crew leaders | 794 | 836 | 712 |
| Supervisors | 189 | 186 | 184 |

For the first 3 weeks of interviewing, the person interview was conducted only with a household member. If an interview with a household member could not be carried out within 3 weeks, an interview with a knowledgeable nonhousehold member was attempted, called a proxy interview. The proxy interviewing was allowed during the remainder of the interviewing period. During the last 2 weeks of interviewing for a cluster, a nonresponse conversion operation was conducted for the noninterviews using the best interviewers. This noninterview conversion attempted to obtain an interview with a household member or a knowledgeable proxy respondent, but not a last resort interview[1]. The nonresponse conversion operation converted 9,518 of the 9,735 total noninterviews to interviews.

[1]Last resort interviews were ones with minimal information, such as names like "White Female." The last resort interview is usually not from a knowledgeable proxy respondent. Last resort interviews were conducted in the census at the end of nonresponse follow-up, after all attempts to contact a knowledgeable respondent have not obtained an interview. Last resort interviews were not conducted for A.C.E.

### The Questionnaire

There were three paths or sections within the person interview. An interview was conducted using the first two paths, when at least one of the household members, for whom information was required, currently lived at the housing unit when the interview was conducted. One path collected data from a household member, and another path collected data from a nonhousehold member (i.e., proxy respondent) for these people. There were two paths, because the questions were worded differently for interviews with household members and with proxy respondents. The interviews from the first two paths were in housing units containing:

- Whole household nonmovers
- Whole household inmovers
- Households with a mixture of nonmovers, inmovers, and outmovers

The third path was for whole household outmovers. The data for outmovers was obtained by proxy with the current resident in the sample household or with other proxy respondents, when necessary. When there was an interview with whole household inmovers, there was also an interview using the third path for whole household outmovers.

When there were multiple interviews for the same housing unit, the CAPI data from the last interview was selected for processing. If there was also a quality assurance interview that replaced the original interview, the quality assurance interview was selected over any other interview.

After the interviewers obtained the names and characteristics of household members, they established the residence status on Census Day. For nonmovers and outmovers, mover status in addition to questions about group quarters and other residences on Census Day established the residence status.

College students living elsewhere in dormitories were not part of the A.C.E. universe. However, they were inadvertently included as inmovers in the A.C.E. instrument. To correct for this, an edit was performed for partial household inmovers who were in group quarters on Census Day. If the inmover was in group quarters on Census Day and was between the ages of 18 and 22, inclusive, the inmover was given an A.C.E. status code of removed.

### Quality Assurance of Person Interviewing

The quality assurance plan for the A.C.E. Person Interview operation consisted of a reinterview of a sample of the original A.C.E. interviews. The workload consisted of a preselected random sample of 5 percent of the total person interview caseload and another sample consisting of cases targeted by the supervisors in the regional offices using specially designed targeting reports. The targeting was based on various indicators likely to predict poor data quality or potential fabrication. The targeted sample was another 5 percent of the total workload.

A separate CAPI questionnaire was designed for the quality assurance interviews. The quality assurance questionnaire contained separate paths for telephone and personal visit quality assurance interviews. The questionnaire also included a complete version of the original interview to allow quality assurance interviewers to conduct the household interview on cases suspected of fabrication. Consequently, it was not necessary to assign another field representative at a later date to conduct the household interviews for such cases.

Quality assurance interviews were conducted either by telephone or personal visit. The interview determined whether or not the original respondent was contacted by an interviewer. If, after an initial set of questions, it appeared that the respondent had not been previously contacted, the quality assurance interview continued with a full household interview that replaced the original interview in all future processing.

The quality assurance plan centered on whether the original interviewer actually contacted the person who was reported to have been interviewed. When this was the case, the interview itself was assumed to be correct because, the person interview questionnaire was designed to ensure data quality using data edits and automated questionnaire skip patterns. When this was not the case (i.e., the proper household was not contacted), a full reinterview was conducted.

The quality assurance plan was designed to be most effective for the few interviewers who blatantly include data from fictitious interviews. This occurs in practice in similar surveys. Therefore, discrepant results were targeted by looking for inconsistent or conspicuous results identified using the targeting reports. Examples of inconsistent or conspicuous results include using the same name for respondents across cases, using famous names for household members, or completing cases too late in the day to really have been interviewing at someone's house.

Effectively identifying an interviewer with only one or two errors in a large workload of cases would require a prohibitively large random sample. Because, later A.C.E. operations such as the person follow-up interview were expected to identify such cases, the quality assurance plan did not attempt to identify these situations beyond what falls in the 5 percent random sample.

### Preliminary Estimation Outcome Codes

Preliminary P-sample estimation outcome codes were assigned to each P-sample housing unit before the computer and clerical matching. This outcome code was

assigned to the housing unit based on Census Day for nonmovers and outmovers. Only people with the following A.C.E. status codes were used in the matching operations:

- N = nonmover resident
- O = outmover resident
- U = unresolved residence status

The preliminary estimation outcome codes identified interviews and noninterviews in occupied housing units, vacant housing units, and housing units that were removed from the P sample. The interview outcomes described in this section were Census Day interview outcomes after data editing, which converts whole households of Census Day residents with insufficient information for matching to noninterviews and whole households of Census Day residents, who should not have been counted at the housing unit on Census Day to vacant housing units.

**Interviews**

- Complete interviews. Interviews conducted with a household member.
- Proxy interviews. Interviews conducted with someone outside the household.
- Sufficient partial interviews. Interviews with household members or proxies that did not collect all required data, but did collect enough information to be considered as interviews.

**Noninterviews**

- Field noninterview.
- Whole households of people with insufficient information to permit matching and follow-up.

**Vacant on Census Day**

- Housing units identified as vacant on Census Day by the interviewer.
- Whole households of people who should have been counted elsewhere on Census Day (i.e., whole household nonresidents).

**Not a Housing Unit on Census Day**

- The housing units identified during the person interview as not a housing unit on Census Day were removed from the P sample.

Table 4-6 contains the number of each category of preliminary outcome codes and the number and percentages of total occupied and vacant housing units for the preliminary outcome codes grouped into interview, noninterview, and vacant. The percentages of interview and noninterview for occupied housing units were also included. The noninterview rate for occupied housing units was 1.9 percent based on the preliminary outcome codes before clerical matching. The interviewers identified 10,206 addresses or 3.4 percent of the A.C.E. addresses as not being housing units on Census Day. The A.C.E. housing units identified as something other than housing units were not in the P sample. For more details see Childers et al. (2001).

Table 4-6. **Preliminary Census Day Estimation Outcome for A.C.E. Housing Units (Unweighted)**

| Outcome code | Total housing units | | Occupied housing units | |
|---|---|---|---|---|
| | Number | Percent | Number | Percent |
| Interview | 257,624 | 88.6 | 257,624 | 98.1 |
| Complete interview with a household member | 235,632 | | | |
| Complete interview with a proxy respondent | 19,380 | | | |
| Sufficient partial interview | 2,612 | | | |
| Noninterview | 4,988 | 1.7 | 4,988 | 1.9 |
| Field noninterview | 2,667 | | | |
| All people have insufficient information for matching and follow-up | 2,321 | | | |
| Total occupied housing units | | | 262,612 | 100.0 |
| Vacant | 28,095 | 9.7 | | |
| No Census Day residents | 4,184 | | | |
| Vacant on Census Day | 23,911 | | | |
| Total occupied and vacant housing units | 290,707 | 100.0 | | |
| Not a housing unit on Census Day | 10,206 | | | |
| Total interviewed housing units | 300,913 | | | |

The percent noninterview was calculated for the unweighted numbers of noninterviews divided by the occupied interviews, which was the interviews plus the noninterviews. Tables of preliminary noninterview rates are presented for respondent type and interview mode in Tables 4-7 and 4-8.

Table 4-7. **P-Sample Preliminary Percent Noninterview in Before Follow-Up by Respondent Type**

| Respondent type | P-sample preliminary percent noninterview |
|---|---|
| Household member | 0.9 |
| Proxy | 13.8 |
| Total | 1.9 |

Of all interviews at occupied housing units, 33.5 percent were completed by telephone, 66.1 percent were completed by personal visit, and 0.3 percent, which was 910 interviews, were completed by a quality assurance replacement interview. The percent noninterview of occupied housing units for each interview mode is shown in Table 4-8.

Table 4-8. **P-Sample Preliminary Percent Noninterview Before Follow-Up by Interview Mode**

| Interview mode | P-sample preliminary percent noninterview |
|---|---|
| Telephone | 0.9 |
| Personal visit | 2.2 |
| Quality assurance replacement | 36.0 |
| Total | 1.9 |

While telephone interviews were more likely than personal visit interviews to have insufficient information because neighbors could not be contacted, this was offset by the straightforward nature of the telephone interviews. These were cases where the respondent completed and returned the census form in a timely manner and provided a telephone number on the form. Conversely, personal visit cases tended to be the more difficult situations (such as movers or reluctant respondents), and were therefore, much more likely to result in noninterviews.

There were several reasons for a high noninterview rate for the quality assurance replacement interviews. These were difficult interviews, because they failed the quality assurance check and needed a reinterview. Many of the noninterviews were refusals. Additionally, because the instrument was monitoring both the quality assurance case and the replacement interview, it was difficult to obtain the Census Day residents in mover cases so that many of these were noninterviews. There was also a problem with the instrument in cases where the quality assurance interviewer could not find the address on the day of the QA interview. When this occurred, the case failed the quality assurance check, but no data were collected to replace the original interview since the QA interviewer could not find the address. However, unlike in personal visit cases, no attempt was made by the QA interviewer to determine if the sample address also did not exist on Census Day. Therefore, these cases were considered to be Census Day noninterviews. There were 108 such cases.

## PERSON MATCHING

After both the CAPI interviewing and the HCUF were completed, the E sample was identified from the HCUF and person matching began. People with incomplete names were identified by computer for both the P and E sample, because they did not contain sufficient information for matching and follow-up. See Attachment 4 for more information about census data-defined and insufficient information for matching and follow-up.

The P-sample people and those in the HCUF, within the sample clusters, were computer matched. The possible matches, P-sample nonmatches, and E-sample nonmatches were clerically reviewed using an automated matching and review system. Additional matches and possible matches were identified by the clerical staff. Duplicates on both lists were also identified clerically. After the matching was completed, field follow-up was conducted and the results of the field interview were coded in the matching database.

### Within Block Cluster Computer and Clerical Matching

With procedure C, the people in P-sample housing units, who were initially matched to the E-sample and non-E-sample census enumerations were:

- nonmovers and outmovers identified as residents (i.e., A.C.E. status equal to N and O), or
- people with unresolved residence status (i.e., A.C.E. status equal to U)

The matching within the sample clusters was done by the computer matcher followed by a computer assisted clerical review. The computer compared the nonmovers and outmovers to the E-sample census enumerations in sample clusters and when necessary to the non-E-sample enumerations. These non-E-sample enumerations were census people in housing units that were not included in the E sample after the subsampling of census housing units. The clerical matchers also searched among people enumerated in the census in group quarters. A match was assigned when the name and characteristics in the P sample for a person were found in the census data within the block cluster.

During computer matching, the P sample was matched to the census. However, this matching was prioritized; first

the P sample was matched to the E sample, then any leftover nonmatches from the P sample were matched to the non-E-sample people in housing units. The matching occurred in two steps:

- **Record Pair Ranking.** The standardized names from the P-sample person and the census person were compared along with the person characteristics using a string comparison (Winkler, 1994). A ranking score was assigned to each pair of people and the optimal pairs were identified.
- **Determination of Match Cutoffs.** The optimal pairs in the cluster were reviewed to determine the cutoffs for matches and nonmatches. All pairs above the match cutoff were identified as a match. All pairs between the match cutoff and nonmatch cutoff were identified as possible matches. All pairs below the nonmatch cutoff were classified as not matched. Match cutoffs were assigned conservatively to prevent false matches.

The goal of the matching and follow-up operation was to produce the correct ratio of cases classified as omitted from the census to those classified as correctly included in the census. After the computer matching, P-sample and E-sample people who did not match were reviewed clerically. The clerical matchers were able to match people the computer could not, because they had the whole household to aid in matching. The P-sample nonmatches were searched for in the census. A duplicate search was also conducted clerically. The matching and duplicate search was aided by the software in sorting and searching the census records. The computer assisted clerical matching software contained all A.C.E. and census information about P-sample and census people, including names, characteristics, outcome of the interview, and address.

The A.C.E. technicians carried out the quality assurance for the clerical matchers and resolved the cases flagged by the clerical matchers as needing further review. The A.C.E. analysts did the quality assurance for the technicians and resolved the cases flagged by the technicians as needing further review. There were 235 clerks, 46 technicians, and 10 analysts to do the clerical matching.

**Census Images.** Scanned images of census questionnaires were available for matching for the first time in Census 2000. The clerical matchers used these images as an aid in matching and when additional information (like names) was found, the new information was made available for the follow-up interview. An E-sample record could be updated by the clerks to provide sufficient information for matching and follow-up or to correct image capture errors. In addition, some information written outside the capture boxes was used to update the data.

For Census 2000, all census forms were scanned and the subsequent information was interpreted using Optical Mark Recognition and Optical Character Recognition or was keyed. For person matching, images were only available for housing units on the January, 2000 DMAF. Images were not available for census housing units added after January, 2000. An address identified by census identification number (ID) could return more than one form, including the following: original census form, Be Counted form, a foreign language form, and/or a Simplified Enumerator Questionnaire. Be Counted forms were not available to use for viewing images, since they did not have a census ID associated with the form when data captured.

The clerical matchers reviewed data for census people with insufficient information for matching and follow-up and searched for additional information that might allow them to be matched when the image was available. All review of census people with insufficient information for matching and follow-up was done before the clerical matching began and the census data in the matching software was updated. The software did not permit the assignment of a code until there were two characteristics and a complete name. After the software data were updated, the clerical matching process began and the matchers could match the P-sample person to the census people now containing sufficient information for matching. The matchers were also able to review data for nonmatches when they suspected data capture errors and to correct the records of name, relationship to person number one, sex, age, Hispanic origin, and race.

The corrected data were used on the follow-up form, but not sent to estimation. The updated data were not inserted into the HCUF. This updating was for matching in A.C.E. and for the follow-up form only. The matchers were **NOT** looking at people who were not data-defined to see if there was more information on the census form to make them data-defined. Therefore, people were **NOT** created in the census.

**Duplicate Search Within Cluster.** The search for duplicates was done clerically. A person was duplicated when the data collected for the person was repeated within the block cluster. The printouts used in 1990 for duplicate search were automated in 2000. Search routines in the 2000 clerical matching software made the searches quicker and more accurate. Duplicates were linked in the matching system for later analysis.

Duplicated People Were Identified:

- **Within the P sample.** A duplicated P-sample person was removed from the final P sample, because both people were not needed in that household in the P sample. When the whole households of P-sample people were duplicated, one of the housing units was converted to a noninterview because the interview was not a good one. The duplicated P-sample household was in a different housing unit and one of them was included instead of the people who actually lived at the address.

For example, the Smith family was collected in apartments A and B. Both apartments were housing units. The P-sample interview for the duplicated family is not a good interview and is converted to a noninterview after the P-sample people were removed.

- **Within the E sample.** An E-sample person duplicate was an erroneous enumeration in the census.
- **Between E-sample people and people not in the E sample.** The E-sample people were also compared to the census people in housing units within the same sample cluster who were not in sample in large block clusters after the E-sample identification. There was no duplicate search between E-sample people and people enumerated in group quarters. Also, there was no duplicate search within group quarters.

  When duplication between an E-sample person and a non-E-sample person was identified, it indicated that there was not a full erroneous enumeration. Therefore, the probability of erroneous enumeration caused by duplication was needed for the duplicated E-sample person. The formula for the probability of erroneous enumeration, was 100 times *d* divided by *c+d+1* percent or

$$\Pr(\text{EE}) = 100 \times d / (c + d + 1) \text{ percent}$$

where

*c* = number of times the E-sample person was duplicated with another E-sample person

*d* = number of times the E-sample person was duplicated with a non-E-sample person

In 1990, when there was duplication between a person in the E sample and a person in a household that was not in the large-cluster subsample, and therefore not in the E sample, the E-sample person was assigned a probability of erroneous enumeration of one half. This methodology was refined in the 2000 A.C.E. to accommodate triplicates. The 1990 estimate was biased when there was a triplicate enumeration in the census and this triplicate involved two E-sample duplicates and the triplicate was not in the E sample. However, there were only a few of these cases in 2000.

This assumes the E-sample person had been coded as correctly enumerated. If the E-sample person was coded unresolved, the final probability of erroneous enumeration included an imputation for unresolved enumeration status. If the E-sample person was assigned a match code that indicated erroneous enumeration, the number of times that the E-sample person was duplicated with non-E-sample people was irrelevant and ignored. A person could not have a probability of erroneous enumeration that was larger than 100 percent.

### Census Geocoding Errors

The clerical matchers reviewed people in census housing units identified in the housing unit matching as geocoding errors. The clerical matchers assigned a code indicating geocoding error to E-sample persons for whole household E-sample nonmatches. There was no need for a follow-up interview, since the housing unit follow-up operation identified these housing units with geocoding errors. These E-sample people were erroneously enumerated in this sample cluster because they were enumerated in a housing unit that was incorrectly geocoded to this sample cluster. In 1990, these people were followed up because it wasn't clear who was incorrectly geocoded until after the follow-up interview.

### Coding Nonmatches in Large Households

The mail return short form had a continuation roster to collect names for persons seven through twelve. The mail return long form had a roster for the names of persons one through twelve. Data were collected for the first six people in the household, for both long and short forms. If the large household follow-up was unsuccessful, there were only names for persons seven through twelve for the long and short mail return forms. Census records were not created for the people in households with only names, since they were not data-defined.

The names on the rosters were used to reduce the P-sample follow-up of nonmatches in large households. P-sample people in large households who were found on the large household roster were not followed up because they were residents of the housing unit on Census Day. They were still counted as not matched to a census enumeration, but a follow-up interview was not needed to establish their residence on Census Day.

### Targeted Extended Search

P-sample whole household nonmatches with no address match and E-sample whole households of nonmatched people in housing units coded as geocoding errors had their search area expanded into the first ring of surrounding blocks. The expanded search is referred to as targeted extended search (TES). See Chapter 5 for a full discussion. The targeted extended search for 2000 A.C.E. was a two-stage process. First, clusters were identified that would benefit most from expanding the search area to surrounding blocks. Second, blocks within the surrounding blocks were targeted for searching.

This extended search was targeted at the clusters most likely to benefit from expanding the search area. The clusters selected for targeted extended search for the 2000 Accuracy and Coverage Evaluation were:

- Clusters included with certainty
  - Relisted clusters in A.C.E.
  - The 5 percent of clusters having the most unweighted census geocoding errors and A.C.E. address nonmatches

- The 5 percent of clusters having the most weighted census geocoding errors and A.C.E. address nonmatches

- Clusters selected at random from the clusters with A.C.E. housing unit nonmatches (i.e., A.C.E. housing units coded CI or UI) or census housing units identified as geocoding errors (i.e., coded GE)

The clusters not selected for targeted extended search were:

- Clusters not selected from the clusters with A.C.E. housing unit nonmatches (i.e., A.C.E. housing units coded CI or UI) and census housing units identified as geocoding errors (i.e., coded GE), (i.e. TES eligible for sampling, but not selected)

- Clusters with no A.C.E. housing unit nonmatches or census geocoding errors identified in the housing unit matching.

- List/Enumerate clusters

Table 4-9 contains the number of clusters selected for TES and the number of P-sample and E-sample people in TES. The number of clusters includes the clusters included with certainty because they were relisted. P-sample people with a residence probability of zero have been excluded from the table.

Table 4-9. **The TES Sample**

| | Clusters | P-sample people | E-sample people |
|---|---|---|---|
| Included with certainty | 1,150 | 28,533 | 20,572 |
| Sampled for TES | 1,089 | 3,889 | 2,281 |
| Total | 2,239 | 32,422 | 22,853 |

Clusters with the most unweighted and weighted census geocoding errors and A.C.E. address nonmatches were included because some clusters with large weights contribute disproportionately to the estimates. Approximately 10 percent of the clusters of the remaining clusters with A.C.E. housing unit nonmatches and census geocoding errors (49 percent of all clusters) were selected at random. There were 2,239 clusters selected for targeted extended search.

In the second stage of targeting, the work was targeted to blocks within the search area where the geocoding error was located. In 1990, the effort required to search for matches and duplicates in large areas that had only a few possible matches or duplicates appeared to lead to errors. There was anecdotal evidence of clerks who did not bother to look in surrounding blocks because they rarely found anything. Targeting the expanded searching probably reduced clerical errors, as well as the cost of the operation.

### P-Sample Matching Extended Search

The search area was expanded to clerically search the ring of surrounding blocks for the P-sample whole household nonmatches, when a housing unit was not a match in housing unit matching, (i.e., the housing unit match code was a nonmatch or unresolved). There was no searching in surrounding blocks for partial household nonmatches or for whole household nonmatches with matching addresses.

How the search was done depended on whether the cluster and its surrounding blocks consisted solely of urban type addresses, or whether they consisted of some or all rural type addresses.

- In areas that are completely urban, if the clerk located the basic street address in the surrounding blocks or the clerk determined the range of addresses was in the surrounding blocks, person matching was conducted in that block where the basic street address or range was located. The matching was also conducted when there was a possible address match in a surrounding block.

- In rural or mixed urban and rural areas, because of the difficulties in matching rural type addresses, there was no attempt to match addresses in the surrounding blocks. Instead, people were searched for in all of the surrounding blocks.

### E-Sample Extended Search for Geocoding Errors

A census person in a housing unit that was coded as a geocoding error was an erroneous enumeration unless the housing unit was located inside the expanded search area. The census geocoding errors were identified in the housing unit phase of the A.C.E. Another interview identified the housing units that physically existed in the surrounding blocks, instead of within the cluster where they were enumerated. This field work was done for whole household E-sample nonmatches in housing units identified during the housing unit phase as geocoding errors. This field visit was conducted at about the same time as the A.C.E. person interview.

The people in these housing units were coded as follows:

- If the housing unit was found to exist in the surrounding blocks, the clerks coded the E-sample person as geocoded to the surrounding blocks during the before follow-up person matching.

- If the housing unit existed in the sample cluster, the E-sample person was coded as not a geocoding error, because that housing unit did exist in the sample cluster.

- If the housing unit did not exist in the surrounding blocks or could not be located on the map sent with the case, the E-sample person was coded as a geocoding error, indicating the person was erroneously enumerated because the housing unit was incorrectly geocoded in the block cluster.

- If the field work was not done or if it could not be determined if the block number entered on the form was in the block cluster or in the surrounding blocks, the unresolved code was used. There was no follow-up for the unresolved cases.

A person follow-up interview for the E-sample nonmatches coded in the sample cluster or in the surrounding blocks was needed to identify other reasons for erroneous enumeration, such as fictitious people and other residences where people should have been counted on Census Day.

### E-Sample Targeted Duplicate Search

A search for duplicated people was conducted clerically in the targeted extended search clusters, when the housing unit was identified during the field interview as physically existing in the surrounding blocks. Like the P-sample search for missed units, the duplicate search was created to identify people who were duplicated because of geocoding error. There was no searching for duplicates in the group quarters enumerations.

If an E-sample housing unit was identified as existing in the surrounding blocks, a housing unit duplicate search was conducted. How this was done depended on whether the cluster and its surrounding blocks consisted solely of urban style addresses or whether they were some or all rural style addresses.

- In urban areas, this duplicate search was done first on housing units and then on people. First, the clerks searched in the block where the housing unit should have been counted in the ring of surrounding blocks. If the housing unit was duplicated, a search was conducted to identify duplicated people. The duplicate search was conducted only in the block where the duplicated housing unit was located. These people were duplicated because the housing unit was enumerated correctly in a surrounding block and incorrectly in the sample cluster. If the housing unit was not duplicated, a search for person duplication was not conducted. The search concentrated on people who were duplicated and were in duplicated housing units caused by housing unit geocoding error in the surrounding blocks.
- The duplicate search in rural or mixed areas was a search throughout the entire search area for person duplicates.

### Added and Deleted Census Housing Units

Census coverage operations continued past the creation of the January, 2000 DMAF. As a result, an added census housing unit is one that was not in the initial housing unit matching, because it was added to the inventory of census housing units after the January, 2000 DMAF was created. A deleted census housing unit is one that was in the January, 2000 DMAF, but was removed from the cluster before the final inventory of housing units was created.

The targeted extended search was based on the A.C.E. housing unit matching to the January, 2000 DMAF and did not cover census housing units added to the block cluster since housing unit matching, thus excluding any geocoding errors that were not recognized in time to conduct the TES field follow-up. If a cluster was not identified for targeted extended search and a large building was added to the cluster, the first time it could have come to our attention was during person matching and any added housing units would be identified as geocoding errors during the person follow-up. If any of these cases should have been included in the targeted extended search and were incorrectly geocoded, another follow-up operation would have been needed to identify the ones that actually existed in the surrounding blocks and those that existed outside the expanded search area.

There was not sufficient time to conduct another interview to determine which added census housing units with geocoding error really existed in the first ring of surrounding blocks. These cases were handled in two ways:

- In TES clusters and clusters eligible for TES sampling, the people in added housing units where person follow-up identified geocoding error were treated as unresolved and the probability of correct enumeration was imputed. These new unresolved cases were treated the same as any other person coded with unresolved geography.
- When the housing unit was not in a TES cluster, the people remained coded as geocoding errors and were erroneous enumerations.

A similar limitation existed when a housing unit that was matched in the housing unit matching was later deleted. There was a concern that the deleted unit may have been moved to a surrounding block. Clusters, where matched housing units in the DMAF that were deleted from the HCUF, had no chance of being TES clusters, if the cluster had no A.C.E. housing unit nonmatches or census geocoding errors.

These deleted cases were also treated differently depending on whether they were in TES clusters:

- If in a TES cluster, they were identified as TES people and a surrounding block search was conducted for the housing units in the TES P-sample matching.
- If the housing unit was not in a TES cluster, there was no surrounding block matching. Surrounding block matching could not be done because there were no surrounding block people in non-TES clusters.

### Before Follow-Up Results

Tables 4-10 and 4-11 contain the results of before follow-up matching for the P sample and the E sample. For details of these codes, see Childers (2001). These before

follow-up matching results are from unweighted data from the fifty states and the District of Columbia.

**The P-sample codes are grouped into:**

- Matched
- Not matched
- Possible match
- Unresolved match status
- Removed from the P sample

**Matched.** The P-sample person was found in the census.

**Not matched.** The P-sample person was not found in the census. A follow-up interview was conducted for:

- Partial household nonmatches
- Whole households of conflicting household members (i.e., whole households of P-sample and census nonmatches)[2]
- Other whole household nonmatches where the P-sample interview was conducted with a nonhousehold member[3]

**Possible match.** The P-sample person may have been a match to the census person. A follow-up interview was needed to determine if the two names referred to the same person.

**Unresolved match status.** The only category of unresolved in the before follow-up matching was insufficient information for matching and follow-up.

**Removed from the P sample.** The only category of removed from the P sample in the before follow-up matching were the P-sample people coded as duplicates.

**The E-sample codes are grouped into:**

- Correctly enumerated
- Erroneously enumerated
- Nonmatch
- Possible match
- Unresolved

**Correctly enumerated.** The correctly enumerated people in before follow-up matching were the ones matching the P sample.

**Erroneously enumerated.** The categories during before follow-up were fictitious people, duplicates, insufficient information for matching and follow-up, and geocoding errors.

- The fictitious people were those where notes on the census image identified the person as one who died before or was born after Census Day, or as not a real person such as a dog or other pet.
- The E-sample people enumerated more than once were coded as duplicates.
- The E-sample people with insufficient information for matching and follow-up were those who were data-defined, but did not contain full name and at least two characteristics.[4]
- Census people in housing units identified as geocoding errors[5] during the initial housing unit follow-up were coded as erroneously enumerated because of geocoding error.

**Nonmatch.** All E-sample people who did not match to the P sample were sent for a follow-up interview.

**Possible match.** E-sample people who were coded as possible matches were followed up to determine whether they were, in fact, matches.

**Unresolved.** In before follow-up matching, the unresolved category only includes the census housing units that needed targeted extended search field work and that field work was not done.

Table 4-10. **P Sample Before Follow-Up Matching**

| P-sample match status | Unweighted people | Percent |
|---|---|---|
| Matched | 573,506 | 85.7 |
| Not matched | 76,804 | 11.5 |
| Possible match | 5,070 | 0.8 |
| Unresolved | 7,524 | 1.1 |
| Removed | 5,923 | 0.9 |
| Total | 668,827 | 100.0 |

[2]These cases have been called the Smith/Jones cases in the past.

[3]No follow-up interview was conducted when there were whole households of P-sample nonmatches from interviews with household members in a housing unit that did not match in the housing unit operation or matched to a housing unit containing no data-defined people.

[4]This is the same rule that was used in the 1990 PES. There must have been enough information about the person to have a chance at locating the person for a follow-up interview before the person was allowed into the matching process. See Childers (2001).

[5]A geocoding error is an error in assigning the housing unit to the correct location.

Table 4-11. **E-sample Before Follow-Up Matching**

| E-sample enumeration status | Unweighted people | Percent |
|---|---|---|
| Correctly enumerated | 544,995 | 76.4 |
| Erroneously enumerated | 27,934 | 3.9 |
| Not matched | 134,916 | 18.9 |
| Possible match | 4,751 | 0.7 |
| Unresolved | 304 | 0.0 |
| Total | 712,900 | 100.0 |

Note: Percentages in table may not add to total due to rounding.

### A.C.E. Person Follow-Up

The person follow-up was conducted to gather additional information to accurately code the residence status of the nonmatched P-sample people and the enumeration status of the E-sample people. In addition, the match status of the possible matches was resolved during the follow-up interview. The following cases were sent to person follow-up:

- P-sample partial household nonmatches
- P-sample whole household nonmatches where the census enumerated different E-sample people (i.e., conflicting households or Smith/Jones cases)
- P-sample whole household nonmatches where the A.C.E. person interview was with a proxy respondent
- E-sample nonmatches
- Possible matches between the P sample and the census
- P-sample matches and nonmatches with unresolved residence status
- P-sample nonmatches needing additional geographic work[6]

The results of the follow-up interview were recorded in the matching software by the matching clerks. Table 4-12 contains the results of the follow-up coding for the P-sample people who were followed up. The P-sample people who were followed up were clerically classified as:

- Matched
- Nonmatched resident of the cluster on Census Day
- Unresolved residence or match status
- Nonresident of the cluster on Census Day

[6]Housing units in relist and list/enumerate clusters did not have housing unit matching. Therefore, P-sample geocoding errors in such clusters needed to be identified during person matching. In addition, when the interviewer changed the address in the CAPI instrument, the P-sample geography was checked to make sure the interviewer did not interview outside the sample cluster.

**Matched.** The P-sample person was found in the census in the block cluster or in a surrounding block after the follow-up interview.

**Nonmatched resident of the cluster on Census Day.** The P-sample nonmatch was not found in the census, and the follow-up interview determined he or she should have been counted in the search area for this cluster.

**Unresolved residence or match status.** The person had unresolved residence status, because the follow-up interview did not successfully collect the information required to accurately identify this person as a resident of the cluster on Census Day. In the case of possible matches, the follow-up interview was not able to ascertain the match status of the people.

**Nonresident of the cluster on Census Day.** The P-sample person was not a resident of the housing unit on Census Day and was removed from the P sample. These people were duplicates, fictitious, living in a P-sample housing unit that was listed in the cluster in error (i.e., P-sample geocoding error), or the P-sample person should have been counted at another residence on Census Day.

The results of the follow-up interview in Table 4-12 indicate 14.7 percent unresolved and 12.5 percent removed from the P sample.

Table 4-12. **Results of P-sample Follow-Up Interview**

| After follow-up match code | Unweighted people | Percent |
|---|---|---|
| Matched | 9,793 | 19.4 |
| Nonmatched resident | 26,961 | 53.4 |
| Unresolved | 7,451 | 14.7 |
| Nonresident | 6,296 | 12.5 |
| Total | 50,501 | 100.0 |

Table 4-13 contains the results of the E-sample follow-up interviews. The followed-up E-sample people were classified as:

- Matched
- Correctly enumerated
- Erroneously enumerated
- Unresolved

**Matched.** The P-sample and E-sample enumerations refer to the same person. The match was made after the follow-up interview.

**Correctly enumerated.** The E-sample nonmatch was identified during the follow-up interview as correctly enumerated in the census.

**Erroneously enumerated.** The E-sample nonmatch was identified during the follow-up interview as erroneously enumerated in the census, because the person should have been counted at another residence on Census Day, was fictitious, had insufficient information for matching and follow-up, was duplicated, or lived in a household that was a geocoding error.

**Unresolved.** The follow-up interview for the census nonmatch was not successful.

The results of the E-sample follow-up in Table 4-13 indicate 7.4 percent of the E-sample people followed up were erroneously enumerated and 14.1 percent were unresolved.

Table 4-13. **Results of E-sample Follow-Up for Nonmatches and Possible Matches**

| After follow-up match code | Unweighted people | Percent |
|---|---|---|
| Matched | 9,088 | 6.3 |
| Correctly enumerated | 103,589 | 72.2 |
| Erroneously enumerated | 10,618 | 7.4 |
| Unresolved | 20,185 | 14.1 |
| Total | 143,480 | 100.0 |

### After Follow-Up Coding

After the follow-up was completed, the results of the interviews were reviewed and codes entered into the system by the matching clerks. See Attachments 5, 6, and 7 for definitions of the individual match, enumeration, and residence status codes assigned by the matching clerks.

The final P-sample results are shown in Tables 4-14 and 4-15. The P-sample people have been classified as matched, not matched, unresolved match status, and removed in Table 4-14 and also tabulated as resident, nonresident, and unresolved residence status in Table 4-15. The data are unweighted, but the people sampled out of the targeted extended search are removed from tabulations for this section.

The P-sample match status is defined as:

- Matched
- Not matched
- Unresolved match status
- Removed from the P sample

**Matched.** The P-sample person was found in the cluster or in the surrounding block in either a housing unit or in group quarters.

**Not matched.** The P-sample person was not found in the search area. If the nonmatch was sent to follow-up, the person was confirmed to be a resident of the cluster on Census Day. If the nonmatch was not sent for a follow-up interview, a household member identified the person as a resident of the housing unit during the original A.C.E. interview.

**Unresolved match status.** The match status was unresolved for possible matches with unsuccessful follow-up interviews and for P-sample people with insufficient information for matching and follow-up.

**Removed from the P sample.** People were removed from the P sample when they were fictitious, duplicates, geocoding errors, or not residents of the housing unit on Census Day.

Table 4-14. **P-sample Match Status After Follow-Up**

| P-sample after follow-up match status | Unweighted people | Percent |
|---|---|---|
| Matched | 578,695 | 88.6 |
| Not matched | 54,424 | 8.3 |
| Unresolved | 7,826 | 1.2 |
| Removed | 12,393 | 1.9 |
| Total | 653,338 | 100.0 |

The P-sample residence status was defined as:

- Resident
- Nonresident
- Unresolved residence status

**Resident.** The P-sample matched or not matched person was a resident of the housing unit on Census Day.

**Nonresident.** P-sample people were nonresidents of the cluster when they were fictitious, duplicates, geocoding errors, or should not have been included as a resident of the housing unit on Census Day. Nonresidents were removed from the P sample.

**Unresolved residence status.** A matched or not matched P-sample person had unresolved residence status when the follow-up interview did not successfully determine the person's residence on Census Day. The residence status of the possible match was unresolved when the follow-up interview was not successful. The residence status was also unresolved when the P-sample person had insufficient information for matching.

Table 4-15. **P-sample Residence Status After Follow-Up**

| P-sample after follow-up residence status | Unweighted people | Percent |
|---|---|---|
| Resident | 625,863 | 95.8 |
| Nonresident | 12,393 | 1.9 |
| Unresolved | 15,082 | 2.3 |
| Total | 653,338 | 100.0 |

The final E-sample results are in Table 4-16. The E-sample people were classified as correctly or erroneously enumerated or having an enumeration status of unresolved. These were the unweighted match results that go to imputation and estimation with the people sampled out of the targeted extended search removed.

The E-sample enumeration status was defined as:

- Correctly enumerated
- Erroneously enumerated
- Unresolved enumeration status

**Correctly enumerated.** E-sample people were correctly enumerated when they were matched to the P sample, or when they have been followed up and they should have been enumerated in this cluster.

**Erroneously enumerated.** E-sample people were erroneously enumerated when they have another residence where they should have been counted on Census Day, were fictitious, were duplicated, lived in a housing unit that was a geocoding error, or had insufficient information for matching and follow-up.

**Unresolved enumeration status.** E-sample people had unresolved enumeration status when the follow-up interview was unsuccessful. The E-sample person may have been followed up to obtain information about the E-sample nonmatch, possible match, matched person with unresolved residence status, or geographic work to obtain the location of the housing unit.

Table 4-16. **E-sample Matching After Follow-Up**

| E-sample enumeration status | Unweighted people | Percent |
|---|---|---|
| Correctly enumerated | 652,390 | 92.6 |
| Erroneously enumerated | 31,064 | 4.4 |
| Unresolved | 21,148 | 3.0 |
| Total | 704,602 | 100.0 |

There were unresolved codes assigned to P-sample and E-sample people. A probability of being matched was imputed for a P-sample person with unresolved match status. A probability that the P-sample person was a resident was imputed when the follow-up did not give enough information to resolve the person's residence status. The probability that a P-sample person was a resident was the probability that the person should have been included in the P-sample. The probability that the E-sample person was correctly enumerated was also imputed for the E-sample people with unresolved enumeration status. A P-sample person could be matched, but have unresolved residence status or have both match and residence status unresolved. Therefore, tabulations for match status and residence status are shown separately for the P-sample.

### Estimation Outcome Codes

Two sets of outcome codes were prepared, one for the Census Day household and one for the Interview Day household. The final P-sample estimation outcome code identified the status of the interview for estimation on Census Day and on the day of the interview. For example, there were cases that were complete interviews for the current residents, but were reported as noninterview or vacant for the Census Day residents.

The final Census Day outcome codes are in Table 4-17. Outcome codes were changed as a result of the follow-up interview in the following types of situations:

- **No Census Day residents noninterview.** Whole households of P-sample people who said they lived elsewhere on Census Day were converted to noninterviews.
- **No Census Day residents vacant.** Whole households who lived in group quarters on Census Day or should have been enumerated at another residence were converted to vacant.

The outcome codes for these two situations were changed because new information from the follow-up interview indicated the original interview was incorrect. The housing unit outcome code for people identified as residents of the housing unit from the person interview who said in the follow-up interview that they lived elsewhere was changed to noninterview. The original person interview listed this household as residents of the housing unit when they did not live at this address. The interview is incorrect and is converted to a noninterview.

The housing unit outcome codes for people identified as residents of the housing unit, from the person interview who said in the follow-up interview that they lived in group quarters or should have been enumerated at another residence, were changed to vacant. The original person interview should have classified the housing unit as vacant, because the people should have been enumerated at another address.

The table also contains numbers of housing units identified as interviews, noninterviews, and vacant and percentages of total housing units and numbers and percentages of occupied housing units. The noninterview rate for occupied housing units for Census Day was 3.0 percent. Addresses that were not housing units on Census Day were removed from the P sample.

Table 4-17. **Final Census Day Estimation Outcome Codes for A.C.E. Housing Units (Unweighted)**

| Census Day outcome code | Total housing units | | Occupied housing units | |
|---|---|---|---|---|
| | Number | Percent | Number | Percent |
| Census Day interview | 254,175 | 87.5 | 254,175 | 97.0 |
| Complete Census Day interview with a household member | 233,327 | | | |
| Complete Census Day interview with a proxy respondent | 18,335 | | | |
| Sufficient partial interview | 2,513 | | | |
| Census Day noninterview | 7,794 | 2.7 | 7,794 | 3.0 |
| No Census Day residents | 2,709 | | | |
| Field Census Day noninterview | 2,667 | | | |
| All people have insufficient information for matching and follow-up | 2,418 | | | |
| Total occupied Census Day housing units | | | 261,969 | 100.0 |
| Vacant | 28,472 | 9.8 | | |
| No Census Day residents | 4,561 | | | |
| Vacant on Census Day | 23,911 | | | |
| Total occupied and vacant housing units on Census Day | 290,441 | 100.0 | | |
| Not a housing unit on Census Day | 10,472 | | | |
| Total housing units | 300,913 | | | |

The Census Day noninterview rates in Tables 4-18 and 4-19 are for occupied housing units. The percent noninterview was calculated for the unweighted numbers of Census Day noninterviews divided by the occupied Census Day interviews, which was the interviews plus the noninterviews on Census Day. The Census Day noninterview rates were recalculated to reflect changes due to coding in after follow-up matching.

Table 4-18. **P-sample Noninterview Rates for Census Day in Occupied Housing Units by Interview Mode**

| Interview mode | Percent noninterview |
|---|---|
| Telephone | 1.1 |
| Personal | 3.7 |
| Quality assurance | 37.4 |
| Total | 3.0 |

Table 4-19. **P-sample Noninterview Rates for Census Day in Occupied Housing Units by Type of Interview**

| Type of interview | Percent noninterview |
|---|---|
| Interview with a household member | 1.8 |
| Proxy interview | 17.4 |
| Total | 3.0 |

### Comparison of Initial and Final P-Sample Estimation Outcome Codes for Census Day

Table 4-20 compares the preliminary and final Census Day interview outcome codes. The preliminary Census Day outcome codes were changed, when the follow-up interviews for the P-sample classified people as nonresidents because they did not live at the sample address at the time of the census, or they were considered as living at the sample address but should have been counted at another residence such as group quarters or another home. The housing unit could also be identified as not being a housing unit on Census Day.

Table 4-20. **Comparison of the Preliminary and Final Census Day Outcome Codes**

| Preliminary Census Day outcome codes | Final Census Day outcome codes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Interview with household member | Interview with proxy | Partial interview | No Census Day residents-noninterview | Field noninterview | Whole household insufficient information | No Census Day residents-vacant | Vacant | Not a housing unit |
| Interview with Household member | 233,327 | 0 | 0 | 2,033 | 0 | 0 | 125 | 0 | 147 |
| Interview with proxy | 0 | 18,335 | 0 | 676 | 0 | 0 | 252 | 0 | 117 |
| Partial interview | 0 | 0 | 2,513 | 0 | 0 | 97 | 0 | 0 | 2 |
| Field noninterview | 0 | 0 | 0 | 0 | 2,667 | 0 | 0 | 0 | 0 |
| Whole household insufficient information | 0 | 0 | 0 | 0 | 0 | 2,321 | 0 | 0 | 0 |
| No Census Day residents-vacant | 0 | 0 | 0 | 0 | 0 | 0 | 4,184 | 0 | 0 |
| Vacant | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23,911 | 0 |
| Not a housing unit | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10,206 |

Table 4-21. **Final Interview Day Estimation Outcome Codes for A.C.E. Housing Units (Unweighted)**

| Interview Day outcome code | Total housing units | | Occupied housing units | |
|---|---|---|---|---|
| | Number | Percent | Number | Percent |
| Interview Day interview | 264,103 | 89.0 | 264,103 | 98.9 |
| Complete interview on Interview Day with a household member | 249,854 | | | |
| Complete interview on Interview Day with a proxy respondent | 12,317 | | | |
| Sufficient partial interview | 1,932 | | | |
| Interview Day noninterview | 3,052 | 1.0 | 3,052 | 1.1 |
| No Interview Day residents-household converted to noninterview | 483 | | | |
| Field noninterview on Interview Day | 373 | | | |
| All people have insufficient information for matching and follow-up | 2,196 | | | |
| Total occupied housing units on Interview Day | | | 267,155 | 100.0 |
| Vacant on Interview Day | 29,662 | 10.0 | | |
| Total occupied and vacant housing units on Interview Day | 296,817 | 100.0 | | |
| Not a housing unit on Interview Day | 4,096 | | | |
| Total housing units | 300,913 | | | |

### Final P-Sample Estimation Outcome Codes for Interview Day

The final Interview Day outcome codes are in Table 4-21. The interview outcome, as of Interview Day, was for cases originally classified as nonmovers and inmovers. Changes as a result of the follow-up interview were from whole households of nonmovers who said they:

- Never lived at this residence
- Lived in group quarters on Census Day
- Lived at another residence on Census Day

The outcome codes for these cases were converted to noninterviews.

The Interview Day noninterview rates were recalculated to reflect changes due to coding in after follow-up matching. The final noninterview rates for Interview Day by interview mode and type of interview are in Tables 4-22 and 4-23.

Table 4-22. **P-sample Noninterview Rates for Interview Day in Occupied Housing Units by Interview Mode**

| Interview mode | Percent noninterview |
|---|---|
| Telephone | 0.7 |
| Personal | 1.0 |
| Quality assurance | 15.4 |
| Total | 1.1 |

Table 4-23. **P-sample Noninterview Rates for Interview Day in Occupied Housing Units by Type of Interview**

| Type of interview | Percent noninterview |
|---|---|
| Interview with a household member | 0.5 |
| Proxy interview | 8.6 |
| Total | 1.1 |

# Attachment 1.
# A.C.E. Mover and Residence Status Code

- **A.C.E. Mover Code**

  1 = Nonmover
  2 = Inmover
  3 = Outmover

- **A.C.E. Born Since Census Day Code**

  0 or blank = Default for inmovers
  1 = Born on or before Census Day
  2 = Born since Census Day
  D = Don't know
  R = Refused

- **A.C.E. Group Quarters Code**

  0 or blank = Default for whole household inmovers[7]
  1 = In group quarters on Census Day
  2 = Not in group quarters on Census Day
  D = Don't know
  R = Refused

- **A.C.E. Other Residence Code**

  0 or blank = Default for whole household inmovers
  1 = In other residence on Census Day
  2 = Not in other residence on Census Day
  D = Don't know
  R = Refused

- **A.C.E. Status**

  N = Nonmover, resident on Census Day
  O = Outmover, resident on Census Day
  I = Inmover, nonresident on Census Day
  R = Removed, nonresident on Census Day
  U = Unresolved residence status
  B = Born since Census Day, nonresident on Census Day

[7]Partial household inmovers were assigned the codes of 1, 2, D, or R during the edit for CAPI data review.

# Attachment 2.
# The Treatment of Group Quarters in A.C.E.

The A.C.E. was designed to provide estimates of person coverage in housing units. There was no sample of, and no estimates for, persons in group quarters. The P-sample housing units were selected for the A.C.E. and the people in the P-sample housing units were matched to the people enumerated in census housing units.

Classifying a structure as group quarters was difficult at times. For example, homes for the elderly have made it more common for a single structure to contain apartments for retired people, assisted living, and full care. Another example was college dormitories. A dormitory was group quarters when it was occupied by unmarried students. The dormitory contained housing units if it was occupied by married students. If the dormitory was mixed with married, unmarried, faculty, and staff, it contained housing units. As a result, housing units or group quarters could be misclassified, when they were not easily classified as housing units or group quarters. This misclassification could be found in both the A.C.E. and the census.

When the P-sample people in A.C.E. housing units did not match to people enumerated in housing units in the census, they were matched to people enumerated in the census in group quarters. That is, group quarters were searched for P-sample nonmatches. If the P-sample people were found in the group quarters enumerations, they were treated as matched. However, no attempt was made to discover whether the misclassification was in the A.C.E. or the census.

Likewise, if a census person in the E-sample was enumerated in a housing unit, but the housing unit was misclassified and should have been group quarters, the follow-up of the census nonmatch obtained information about the residence of the person. If it found the person should have been counted in this block in group quarters or a housing unit, the person was coded as correctly enumerated in A.C.E. processing. The ideal was not to classify someone as erroneous when they really should have been counted in this cluster, but the type of residence was misclassified.

If a structure contained both housing units and group quarters, the people who were enumerated in the census in a housing unit were eligible to be in the E sample. The follow-up interview identified such E-sample people who were not matched as living in the cluster and having no other residence. They were coded as correctly enumerated. There was no duplicate search between people enumerated in group quarters and housing units.

In summary, then, the approach was balanced:

- Look for P-sample people in group quarters when they were not found in census housing units.
- Follow up E-sample people in both housing units and group quarters in the cluster.

The population in housing units was covered, but there was no estimate of coverage in group quarters. If the housing unit was duplicated in the group quarters, the group quarters people were not counted as duplicates. Likewise, if a group quarter was missed, there was no determination of undercounted inhabitants.

# Attachment 3.
# The A.C.E. Person Interview[8]

**Household Roster**

If the person lived at the sample address, the interviewer began the interview with a series of questions to obtain the names of everyone currently living at the sample housing unit. The first question was:

> "I need to get a list of everyone living here permanently or staying temporarily at this address. What is your name?"

After obtaining the name of the person with whom the interviewer was speaking, the interviewer asked, "Anyone else?" If there was a "yes" response, the interviewer asked, "What is his or her name?" and followed that with "Anyone else?" until the interviewer received a "no" response.

As a check for types of people who were frequently left off listings of household members, there were two additional questions. The first question asked about people who may have lived at the household sometimes, but not all the time, such as children in joint custody or people who traveled a great deal of the time. The question was:

> "Are there any additional people who currently live or stay here, like someone who's temporarily away or someone who stays here off and on?"

If the response was "yes" the interviewer asked, "What was his or her name?" and followed with "Anyone else?" until the interviewer received a "no" response.

Other persons who were frequently omitted from household listings were roommates or live-in employees. The interviewer asked, "Is there anyone else like a roommate or a live-in employee who lives here?"

If the response was "yes" the interviewer asked, "What is his or her name?" and followed with "Anyone else?" until the interviewer received a "no" response.

At this point in the interview, the interviewer had collected a list of household members that the respondent had voluntarily mentioned, and the interviewer had also checked for two types of persons that research had shown were frequently left off household listings.

The interviewer then reviewed a screen that contained a list of the household members the respondent reported. The interviewer read the list of names and asked if the list was correct. The interviewer said, "I have listed [READS NAMES ON SCREEN]. Is that correct?" After the respondent had reviewed the names, the interviewer could change the spelling, or add or delete a name.

**Movers**

When the respondent agreed that the list was correct, the interviewer handed the respondent a calendar containing the months of March, April, and May of 2000 that had Census Day clearly marked. At this point in the interview, the goal was to begin determining whether the people listed as current residents were also residents of the sample housing unit on Census Day and if anyone else should have been included as a Census Day resident. The interviewer asked if any of the listed persons (current residents) had moved into the sample housing unit after Census Day. The interviewer said to the respondent:

> "Please look at this calendar. Did any of the people I just listed move into <sample address> after Census Day, April 1, 2000?"

If the answer was "yes," the interviewer asked, "Who moved in after April 1?" Any person mentioned was considered a nonresident of the sample housing unit on Census Day. If everyone in the household was mentioned, then the whole household was considered nonresidents on Census Day.

The interviewer now had a list of current residents who also lived at the sample housing unit on Census Day. It was necessary to determine if there was anyone living at the sample housing unit on Census Day who did not live there currently. The interviewer asked, "Was there anyone else living or staying here on April 1, 2000 who has moved out?" If the response was "yes," the interviewer asked, "What is his or her name?" and "Anyone else?" until a "no" response was received.

The interviewer now had a list of the names of everyone the respondent had reported living at the sample housing unit currently and on Census Day. The interviewer then established a reference person (relationships will be relative to this person) by asking who owns or rents the house or apartment. The interviewer asked, "In whose name is this (house/apartment) owned or rented?" The interviewer also asked whether the housing unit was owned or rented by saying, "Do you own this (house/apartment), rent it, or live here without payment of rent?"

[8]See Keeley (2000) for details.

### Demographics

At this point in the interview, the interviewer began to collect demographic characteristics about all listed persons to facilitate matching the persons collected in this interview to persons listed on the census questionnaire for the sample housing unit. Demographic characteristics are also used to create post-strata in dual system estimation. See Chapter 7 for more details.

The demographic characteristics collected in the interview were:

1. **Sex.** The interviewer may have entered the sex of the person or asked the question when in doubt. The question was, "Is [NAME[9]] male or female?"

2. **Age.** Age was collected in a series of questions. The interviewer asked for date of birth ("What is [NAME'S] date of birth?"). When the date of birth was entered in the instrument, the age of the person was calculated and the interviewer verifies the age by saying, "So [NAME] was about [AGE] on April 1?" If the age was not correct, the interviewer changed the date of birth in the previous question and the age was then recalculated.

   If the respondent did not know the date of birth, then the interviewer asked the person's age. The interviewer asked, "What was [NAME'S] age on April 1, 2000?"

3. **Relationship.** Relationship was to the person in whose name the house or apartment was owned or rented (called the Reference Person). The interviewer handed the respondent a card containing relationship categories and asked, "How is [NAME] related to [THE REFERENCE PERSON]?" for each person.

4. **Hispanic Origin.** Hispanic origin was collected in a series of questions. The first question was, "Is anyone of Spanish, Hispanic, or Latino origin?" If the response was "yes," the interviewer asked, "Who is?" followed by "Is there anyone else of Spanish, Hispanic, or Latino origin?" until the response was "no."

   If anyone was mentioned as being of Hispanic origin, the interviewer asked, "Is [NAME] of Mexican, Puerto Rican, Cuban, or some other Spanish origin?" for each person mentioned.

5. **Race.** Race was also collected in a series of questions. The interviewer referred the respondent to the part of the card containing racial categories and said, "I'm going to read a list of race categories. Please choose one or more categories that best describe [NAME'S] race."

   If the respondent said, "American Indian or Alaska Native," the interviewer asked, "What is [NAME'S] enrolled or principal tribe(s)?" The interviewer recorded as many responses as given.

   If the respondent said, "Asian," the interviewer asked, "To what Asian group did [NAME] belong? Is [NAME] Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, or, some other Asian group?" The interviewer recorded as many responses as given.

   If the respondent said, "Pacific Islander," the interviewer said, "To what Pacific Islander group did [NAME] belong? Is [NAME] Guamanian or Chamorro, Samoan, or some other Pacific Islander group?" The interviewer recorded as many responses as given.

At this point, the interviewer had a list of all reported current and Census Day residents and their demographic characteristics for use in matching these residents to residents reported on the census questionnaire for this housing unit.

For households that reported moving into the sample housing unit after Census Day, this information was verified. The interviewer said to the respondent:

> "So, everyone you mentioned today moved into <sample address> after April 1, 2000. Is that correct?"

If the information was correct, the interview was continued by asking the respondent if he or she knew and had information about the residents of the sample housing unit who lived there on Census Day. (This part of the interview was discussed in the section on movers.)

### Residence Section

For all households in which at least one member lived at the sample housing unit on Census Day, the interviewer continued with a few questions that checked for two types of special living situations that were potential sources of duplicate enumerations. Respondents tended to forget that household members may have been living or staying at a place away from the sample housing unit. This may have caused some persons to be reported more than once, at the sample housing unit and again at other places where they may have lived or stayed.

The first situation that had the potential to cause duplicate enumerations was when a person may have lived at a place that was not a private household on Census Day. Since the Census Bureau did special enumerations at

[9]The brackets containing name, age, and the Reference Person's name were filled by the instrument. When speaking to the respondent, "Are you" or other appropriate fillers replaced "Is [NAME]."

places such as college dorms, nursing homes, prisons, and emergency shelters, the interviewer inquired if anyone was staying at any of these types of places by saying:

> "Your answers to the next few questions help us count everyone at the right place. The Census Bureau did a special count at all places where groups of people stay. Examples include college dorms, nursing homes, prisons, and emergency shelters. On April 1, 2000, were any of the people you mentioned today staying elsewhere at any of these types of places?"[10]
>
> If the response was "yes," the interviewer asked, "Who stayed at one of these types of places?"

The next situation that could result in a duplicate enumeration was when a person might have had another residence. The interviewer said:

> "Some people have more than one place to live. Examples include a second residence for work, a friend's or relative's home, or a vacation home. On April 1, 2000, did any of the people you mentioned today have a residence other than <sample address>?"
>
> If the response was "yes," the interviewer asked, "Who had another residence?"

For each person mentioned as having another residence, the interviewer asked, "As of April 1, did [NAME] spend most of the time at <sample address> or at the other residence?" If the response was, "I don't know," the interviewer asked:

> "Which of the following categories, most accurately describes the amount of time [NAME] stays at the other residence? A few days of each week; entire weeks of each month; months at a time; or some other period of time."

If the respondent still was not sure where the person spent most of the time, there was a series of questions designed to assign an amount of time spent at some other residence, such as, "During a typical week, did [NAME] spend more days at <sample address> or at the other residence?" or "During a typical month, did [NAME] spend more weeks at <sample address> or at the other residence?"

If these questions did not help the respondent decide where the person spent "most of the time," the person's residence was determined by asking:

> "Was [NAME] staying at <sample address> or the other residence on April 1, 2000?"

At this point, the interviewer had a reported list of current and Census Day residents of the sample housing unit developed through an extensive household listing procedure. The interviewer had obtained the demographic characteristics of the listed persons. Through questions on mobility and other possible residences it had been determined:

- whether everyone listed in the household currently should be considered a Census Day resident of the sample housing unit
- whether anyone currently absent from the household should be considered a Census Day resident.

### Conclusion of Interview

The interviewer now was ready to conclude the interview. Before concluding, there was one last check of the household listing. The first name, middle initial, last name, sex, and age of each person listed as a current and Census Day resident was shown on the screen. The interviewer, again, showed the respondent the computer screen and asked, "Do I have the spelling, sex, and age correct for everyone?" If not, corrections could be made at this screen and the respondent was asked to verify and/or change the information until the respondent said that everything was correct.

The interviewer asked the respondent for his/her telephone number by saying, "In case we need to contact you again, may I please have your telephone number?" then thanked the respondent and concluded the interview by saying, "This concludes our interview. The Census Bureau thanks you for your participation."

[10] An interviewer help screen was available with a complete list of special enumeration places.

# Attachment 4.
# Insufficient Information for Matching and Follow-Up

The census person records were reviewed both by computer and clerically to identify people with insufficient information for matching and follow-up. Only people with sufficient information for matching and follow-up were allowed to be processed in the matching and follow-up interviewing phases of the person matching. The three types of insufficient information were:

- The census people were not data-defined.
- The census people were data-defined, but computer coded as insufficient information for matching and follow-up.
- The census people were computer coded as sufficient information, but converted clerically to insufficient information for matching and follow-up.

The first type of census people who were not data-defined were not included in the E sample. Only data-defined people were included in the E sample. These data-defined people create person records in the census.

**Census Data-Defined**

The term "data-defined" was a term that has been used in the past at the Census Bureau to mean that a census person record has been created. The term "Total Persons" was the total number of people counted in the census at a census housing unit. The term "Selected Persons" referred to data-defined census people in a census household. The difference was people who were not data-defined. These people had no census person record. A whole person imputation procedure was employed to create characteristic data in the census for these people.

Two characteristics were required to be data-defined, where name counts as a characteristic. Name must have had at least three characters in the first and last name together. Other characteristics that could be used in the counting were relationship, sex, race, Hispanic origin, and either age or year of birth[11]. Census records were created on the HCUF for all data-defined people. Anyone who was not data-defined was a whole person imputation.

The count of census people who were whole person imputations were identified separately from the other census people with insufficient information for matching, because they were treated differently in the Dual System Estimator. The number of whole person imputations was subtracted from the census count within post-strata. The E-sample people who were data-defined but with insufficient information for matching were included in the count of erroneous enumerations, and were, thus, excluded from the count of whole person imputations in the Dual System Estimator.

The mail return census forms were designed to collect characteristics for six people. However, space was provided for the names of the additional residents in households with seven to twelve people. The large household follow-up operation attempted to obtain characteristics for these people by telephone.

The exception was the enumerator questionnaire used in nonresponse follow-up. There was space for five people, but a continuation form was used to record data for persons six and above in large households.

There was some consideration given to using the names in the long form roster for persons seven through twelve to create person records and having them data-defined. However, it seemed preferable not to do this, and the A.C.E. did not attempt to create additional census data-defined

[11] Person one did not automatically have a relationship of head of household like it did in 1990, and the telephone number in item 2, on the mail return questionnaire, did not count as a characteristic. The age and date of birth were examined together. If age was present, age/year of birth counted as a characteristic. If age was blank, but year of birth was present, then the age/year of birth counted as a characteristic. If age and year of birth were both blank, the age/date of birth did not count as a characteristic. The month and day of birth were used in Dress Rehearsal in the determination of counting the age/date of birth as a characteristic, but not in Census 2000.

people for these people with only names in large households. These people were whole person imputations. The number of whole person imputations used in the Dual System Estimator will correspond to the counts used in the census.

### Computer Coding of Insufficient Information for Matching and Follow-Up for the E Sample

The A.C.E. requires a minimum amount of information for matching and follow-up. The data-defined census people were reviewed to identify the ones with sufficient information for matching and follow-up for A.C.E. The minimum amount of data required for data-defined census people to have sufficient information for matching and follow-up was complete name and two characteristics.

Complete name was defined as:

- First name[12], middle initial, and last name
- First name and last name
- First initial, middle initial, and last name

The A.C.E. used the same criteria for classifying age as data-defined as the census, which is only age and year of birth were used to determine if age was present in counting characteristics to determine if the person had enough data to be data-defined in the census. In other words, when the age and year of birth were both blank, month and day of birth were not considered.

### Clerical Coding of Insufficient Information for Matching and Follow-Up for the E Sample

There were cases where the name was not blank, but was too incomplete or unlikely to be real to permit matching and follow-up. Census names like Mr. Doe, Donald Duck, and White Female were coded insufficient information by the clerical matchers. The computer could not recognize names that were not real or were really incomplete names.

The retrieval system contained an image of the census questionnaire. The image of the census questionnaire was reviewed for census people coded as insufficient information for matching and follow-up to see if there was additional data that could be used to convert them to sufficient information for matching and follow-up. The data capture system may have had problems reading the hand written entries, or there may be information outside the boxes on the census questionnaire. Names were obtained from the roster on the image of the questionnaire for the long forms. Children with first names and no last names were converted to sufficient information for matching and follow-up when the last name could be assumed from an adult with first and last name in the household. These updates to the names were captured into the matching software, which was programmed to decide if the person had sufficient information for matching and follow-up.

### P-Sample Insufficient Information for Matching and Follow-Up

The P-sample people were reviewed by computer to identify people with insufficient information for matching and follow-up. The P-sample rules for sufficient information for matching and follow-up were the same as the E-sample rules, which was complete name and two characteristics. Cases identified by the computer as missing sufficient information were suppressed from viewing by the clerical matchers to prevent errors in matching people with insufficient information for matching. There were fewer than 4,000 P-sample people computer coded with insufficient information for matching and follow-up.

This computer review was established to avoid certain types of clerical errors in matching. For example, names like D K or Don't Know (D or Don't is the first name and K or Know is the last name), R R (refused for the first name and refused for the last name), or M Smith, which could not be matched with certainty or, if treated as a nonmatch, followed up with a high rate of success. The census might have recorded a person with a complete name, which might be matched by a clerk. If matching were allowed, it would have been biased by what was enumerated in the census. A match would have resulted if the names were present at the address, and a nonmatch if the names were not in the census. Since names like DK could not be followed up, they would have been coded as insufficient information for matching and follow-up. Therefore, a match would have been assigned when the census obtained complete names, and unresolved when no match was found. The best way to avoid a bias was to suppress the P-sample cases computer coded as insufficient information for matching and treat them as unresolved.

The probability of a match was imputed for the P-sample people coded as insufficient information for matching and follow-up. They were treated in the same way as other P-sample people with unresolved match status. If the whole household had insufficient information for matching and follow-up, the people were removed and converted to noninterview status.

[12]The minimum number of characters to be a name was two. Two characters were required in the first name and two characters in the last name.

# Attachment 5.
# Final P Sample Person Match Codes

**Matched**

| | | |
|---|---|---|
| M | = | The P-sample and census people were matched. The P-sample person was a resident of the housing unit on Census Day. |
| MR | = | The follow-up interview determined that the matched person with unresolved residence status was a resident. |
| MU | = | The A.C.E. person was matched, but the follow-up interview obtained no useful information to resolve the residence status for the matched person who had a residence status of unresolved before follow-up. The P-sample person's residence status was unresolved. |

**Not Matched**

| | | |
|---|---|---|
| NP | = | The P-sample person was not matched to a census person. There was no follow-up for the whole household nonmatches from person interviews with household members and the whole household nonmatches were not conflicting household nonmatches. |
| NC | = | The P-sample nonmatch was found on the census roster. This person in a partial nonmatch household was not matched to the census because only name was collected in the census for this person in a large household and the census person was not data-defined. No follow-up interview was necessary. |
| NR | = | The P-sample person was not matched and was identified as a resident in the block cluster on Census Day during the A.C.E. person follow-up interview. |
| NU | = | The P-sample person was not matched. Not enough information was collected during the A.C.E. person follow-up interview to identify the P-sample person as a resident or nonresident in the block cluster. The residence status for the P-sample person was unresolved. This code was also used when the P-sample person was followed up to collect geographic information and that information was not collected. The NU code was also used when the person did not live at the sample address on Census Day and the Census Day address was not complete enough to determine if the Census Day address was in the sample cluster. |

**Unresolved**

| | | |
|---|---|---|
| P | = | There was not enough information collected during the follow-up interview to determine if the possible match was a match or not. The match status of the P-sample person was unresolved. |
| KI | = | Match not attempted for the P-sample person, because the person had insufficient information for matching and follow-up. The name was blank or incomplete or the name was complete, but the person had only one characteristic. This was a computer assigned code and these people were suppressed from view by the matchers. |
| KP | = | Match not attempted for the P-sample person, because (1) the name was incomplete, such as "Mr. Jones," or (2) the name was not a valid name, such as "White Female" or "Donald Duck." This was a clerically assigned code. |

**Removed from the P Sample**

| | | |
|---|---|---|
| FP | = | The P-sample person was fictitious in this block cluster. The person was interviewed in error during the person interview. This person was not included in the final P sample. |
| NL | = | The P-sample person did not live at the sample address or in the block cluster on Census Day and was listed as a nonmover or outmover in error. This person was removed from the list of P-sample people, since he or she was collected during the person interview in error. |
| NN | = | The P-sample person was identified as a nonresident in the block cluster on Census Day during the A.C.E. person follow-up interview, because the person lived in group quarters on Census Day, or had another residence where the person should have been counted on Census Day according to census residence rules. This person was removed from the list of P-sample people, since he or she was collected during the person interview in error. |
| DP | = | The P-sample person was a duplicate of another P-sample person. |
| MN | = | The A.C.E. person follow-up interview determined that the matched person with unresolved residence status was not a resident in this housing unit or in this block cluster. The person was no longer in the list of P-sample people. |
| GP | = | The P-sample person was removed, because the person interview was conducted at a housing unit that exists outside the sample cluster. The person follow-up identified this housing unit as a P-sample geocoding error. |

# Attachment 6.
# E-Sample Person Enumeration Codes

**Correctly Enumerated**

| | | |
|---|---|---|
| M | = | The P-sample and E-sample people were matched. The E-sample person was correctly enumerated. |
| CE | = | The E-sample nonmatch was identified as correctly enumerated during the A.C.E. person follow-up interview. |
| MR | = | The A.C.E. person follow-up interview determined that the matched person with unresolved residence status was a resident. |

**Erroneously Enumerated[13]**

| | | |
|---|---|---|
| GE | = | The E-sample person was erroneously enumerated in this block cluster, because the census housing unit was a geocoding error (i.e., counted in the wrong block cluster). The E-sample person should have been enumerated elsewhere in the census. |
| EE | = | The E-sample nonmatch was identified during the person follow-up interview as erroneously enumerated. |
| FE | = | The E-sample nonmatch was determined to be fictitious in this block cluster during the follow-up interview. The person may have existed, but should not have been enumerated in the census within this block cluster. The E-sample person was erroneously enumerated in the census in this block cluster. |
| DE | = | The E-sample person was a duplicate of another E-sample person. The code was also used when the E-sample person was a duplicate of a census person in a surrounding block. The people in the E-sample housing unit were erroneously enumerated, because they were counted accurately in the surrounding block and duplicated in the sample cluster. |
| MN | = | The A.C.E. person follow-up interview determined that the matched person with unresolved residence status was not a resident in this housing unit or in this block cluster. The E-sample person was an erroneous enumeration. |
| KE | = | Match not attempted for the E-sample person. The name was blank or incomplete or the name was complete, but the person had only one characteristic. The name was incomplete or not a valid name, such as “Child Jones,” or “Mickey Mouse.” |

[13]The E-sample people who were duplicated with non-E-sample people were not full erroneous enumerations. See the section on Duplicate Search Within Cluster in this chapter for a discussion of the probability of erroneous enumeration when there was duplication between a census person in the E sample and a non-E-sample person.

**Unresolved**

| | | |
|---|---|---|
| UE | = | Not enough information was collected during the A.C.E. person follow-up interview to identify the E-sample person as correctly or erroneously enumerated in the block cluster. The enumeration status for the E-sample person was unresolved. The UE code was also used when the person did not live at the sample address on Census Day and the Census Day address was not complete enough to determine if the Census Day address was in the sample cluster. This code was also used when the E-sample person was followed up to collect geographic information and that information was not collected. |
| MU | = | The E-sample person was matched, but the follow-up interview obtained no useful information to resolve the residence status for the matched person who had a residence status of unresolved before follow-up. The E-sample person's enumeration status was unresolved. |
| P | = | There was not enough information collected during the follow-up interview to determine if the possible match was a match or not. The match status of the P-sample person was unresolved. |
| GU | = | The geographic work for the targeted extended search was unresolved. The code had the same definition in both the before and after follow-up matching. The difference was in after follow-up, the code was only used in the list/enumerate clusters. The field work for the targeted extended search was not done or the block number on the form was not in the surrounding blocks, in the block cluster, or on the map. It was not clear where the housing unit was located. |

# Attachment 7.
# Final P-Sample Person Residence Status Codes

**Resident**

| | | |
|---|---|---|
| M | = | The P-sample and census people were matched. |
| MR | = | The follow-up interview determined that the matched person with unresolved residence status was a resident. |
| NR | = | The P-sample person was not matched and was identified as a resident in the block cluster on Census Day during the A.C.E. person follow-up interview. The P-sample person was missed in the census. |
| NP | = | The P-sample person was not matched to a census person. There was no follow-up for the whole household nonmatches from person interviews with household members and the whole household nonmatches were not conflicting household nonmatches. These people were considered residents of the housing unit on Census Day. |
| NC | = | The P-sample nonmatch was found on the census roster. This person in a partial nonmatch household was not matched to the census because only name was collected in the census for this person in a large household and the census person was not data-defined. No follow-up interview was necessary. |

**Nonresident**

| | | |
|---|---|---|
| FP | = | The P-sample person was fictitious in this block cluster. The person was interviewed in error during the person interview. This person was not included in the final P sample. |
| NL | = | The P-sample person did not live at the sample address or in the block cluster on Census Day and was listed as a nonmover or outmover in error. This person was removed from the list of P-sample people, since he or she was collected during the person interview in error. |
| NN | = | The P-sample person was identified as a nonresident in the block cluster on Census Day during the A.C.E. person follow-up interview, because the person lived in group quarters on Census Day or had another residence where the person should have been counted on Census Day according to census residence rules. This person was removed from the list of P-sample people, since he or she was collected during the person interview in error. |
| DP | = | The P-sample person was a duplicate of another P-sample person. |
| MN | = | The A.C.E. person follow-up interview determined that the matched person with unresolved residence status was not a resident in this housing unit or in this block cluster. The person was no longer in the list of P-sample people. |
| GP | = | The P-sample person was removed because the person interview was conducted at a housing unit that exists outside the sample cluster. The person follow-up identified this housing unit as a P-sample geocoding error. |

**Unresolved**

| | | |
|---|---|---|
| MU | = | The A.C.E. person was matched, but the follow-up interview obtained no useful information to resolve the residence status for the matched person who had a residence status of unresolved before follow-up. The P-sample person's residence status was unresolved. |
| NU | = | The P-sample person was not matched. Not enough information was collected during the A.C.E. person follow-up interview to identify the P-sample person as a resident or nonresident in the block cluster. The residence status for the P-sample person was unresolved. This code was also used when the P-sample person was followed up to collect geographic information and that information was not collected. The NU code was also used when the person did not live at the sample address on Census Day and the Census Day address was not complete enough to determine if the Census Day address was in the sample cluster. |
| P | = | There was not enough information collected during the follow-up interview to determine if the possible match was a match or not. The residence status of the P-sample person was unresolved. |
| KI | = | Match not attempted for the P-sample person, because the person had insufficient information for matching and follow-up. The name was blank or incomplete or the name was complete, but the person had only one characteristic. This was a computer assigned code and these people were suppressed from view by the matchers. |
| KP | = | Match not attempted for the P-sample person, because (1) the name was incomplete, such as "Mr. Jones," or (2) the name was not a valid name, such as "White Female" or "Donald Duck." This was a clerically assigned code. |

Figure 4-1.
**Address Listing Book Page for Single and Multiunit Structures**

Figure 4-2.
**The Housing Unit Follow Up Questions for an A.C.E. Nonmatch**

| A.C.E. ADDRESS | | | | | | CENSUS ADDRESS | | |
|---|---|---|---|---|---|---|---|---|
| BLOCK | MSN | Within MSN ID | Type of Address | Unit Status | # Units | BLOCK | CENSUS ID | MSN |
| | | | | | | | | |

*If necessary, put address corrections here:*

***A.C.E. Nonmatch Address***

Notes:

1. Is ***A.C.E. Nonmatch Address*** located within the **non-**shaded area shown on the A.C.E. Cluster Map?
   - ☐ Yes
   - ☐ No, address does not exist - *Skip to item 6.*
   - ☐ No, address is outside cluster - *Skip to item 6.*

2. Is there a housing unit at ***A.C.E. Nonmatch Address***?
   - ☐ Yes
   - ☐ No - *Explain in notes, then skip to item 6.*

3. Does ***A.C.E. Nonmatch Address*** represent the same housing unit as any of the addresses listed in the **Census column** of the Housing Unit Reference List?
   - ☐ Yes - *Enter Census ID here:* _____.
   - ☐ No

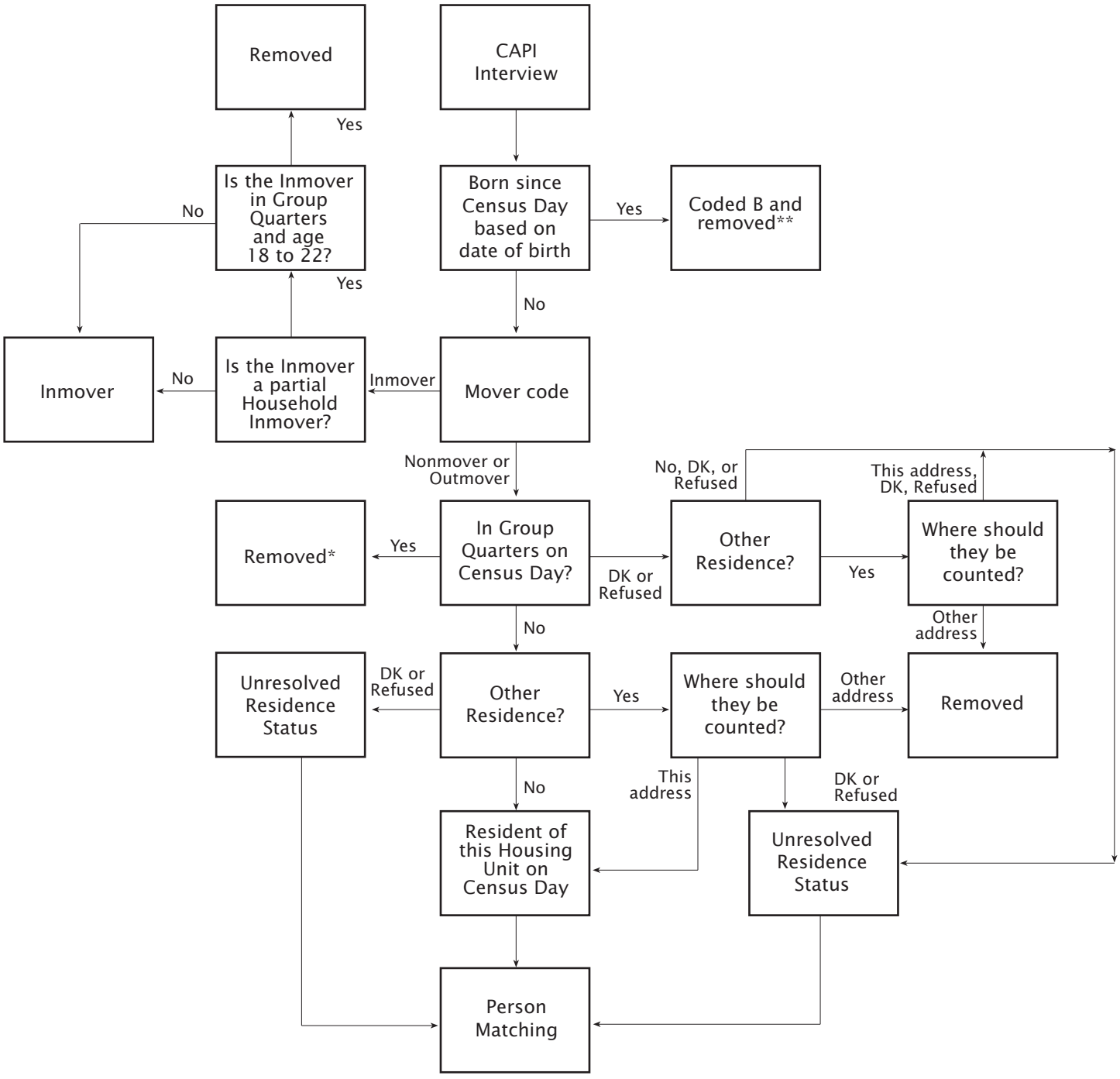
4. Does ***A.C.E. Nonmatch Address*** represent the same housing unit as any of the addresses listed in the **A.C.E. column** of the Housing Unit Reference List?
   - ☐ Yes - *Enter A.C.E. map spot number:* _____ *and the Within MSN ID* _____.
   - ☐ No

5. How many housing units are at this basic address? Enter number of HUs here: _____. *(Explain in Notes below.)*

6. Information from: *If more than one, mark the* ***main*** *source.*
   - ☐ Household member
   - ☐ Building manager or landlord
   - ☐ Proxy
   - ☐ Observation

Notes - Continue on back of page if needed:

Figure 4-3.
**A.C.E. Mover and Residence Status Flow**




* Other residence question is asked, but the person is removed regardless of the answer.

** Group quarters and Other residence questions are asked, but the person is removed regardless of the answer.

# Chapter 5.
# Targeted Extended Search

## INTRODUCTION

The concept behind the dual system estimate is to estimate the census omission rate using the P sample and the erroneous enumeration rate using the E sample. The complete definition of being omitted from or erroneously enumerated in the census includes the concept of "location," that is, a successful enumeration must have located the person in the right place. "Right location" in the census means anywhere in the block where the reported housing unit address was located, or in the "search area," defined as one ring of adjacent blocks. The operation concerned with locating and matching the persons in the surrounding areas is "Targeted Extended Search," or TES. The name was chosen because, unlike the similar procedure in the 1990 Post-Enumeration Survey (PES) where the surrounding area of every cluster was searched, the A.C.E. search was "targeted" in two ways:

1. Results from the initial housing unit matching operation were used to select the housing units that are candidates for TES.
2. In most cases, only clusters that include TES-eligible housing units were included in TES.

This chapter focuses on the statistical methods used in TES. A.C.E. field and processing activities, including TES, are described in Chapter 4.

### Overview

The 1990 PES included a search in all blocks surrounding each sample cluster. Every person in every house in every block adjoining every sample block cluster was included in the search. This was determined to be burdensome in terms of time, cost, and perhaps mental fatigue on the part of matchers performing low-payoff searches (Hogan, 1993). To improve efficiency, the Census 2000 A.C.E. took a more focused (i.e. "targeted") approach in selecting clusters, defining search areas, and determining which housing units and residents would be part of surrounding block operations.

The Census 2000 A.C.E. search operation differed from the 1990 PES in four primary ways:

1. Search area definition.
2. Amount of searching.
3. Persons eligible for search.
4. TES weighting.

### Search Area Definition

The search area for the 2000 A.C.E. was limited to either just the sample block cluster or one ring of adjacent blocks. An adjacent block is one that touches the cluster of sample blocks at one or more points. This definition includes the blocks that touch the corner of the block cluster. Results from empirical research, using Census 1998 Dress Rehearsal data, show that the additional benefits of using two rings of surrounding blocks are negligible (Wolfgang, 1999).

### Amount of Searching

There were two important differences between the extent of searching in the 1990 PES and the 2000 A.C.E.:

1. Only about 20 percent of A.C.E. block clusters had their surrounding areas searched, whereas in 1990 the surrounding area of every block cluster was searched.
2. The search was targeted (in most cases) to only housing units identified as being likely to exhibit geocoding error; in 1990, all persons in surrounding areas were eligible for search.

The clusters with a high number of potential geocoding errors were identified from the results of the initial housing unit matching operation and subsequent field follow-up (see Chapter 4). These were A.C.E. block clusters with a large number of Independent Listing housing units not found in the January 2000 Census Address List. These types of nonmatches are possibly census geocoding errors of exclusion (i.e. not included in the census within the sample area although they should have been). On the census side, A.C.E. block clusters with a large number of census geocoding errors are likely to be errors of inclusion (i.e. reported by the census in the block cluster, although the unit is physically outside the A.C.E. block cluster). These two types of housing units were eligible to be in the extended search as part of TES operations, and are thus ***TES-eligible housing units.***

Any cluster that included at least one potential census geocoding error, of either inclusion or exclusion, was eligible to have TES operations performed in it and is termed a ***TES-eligible cluster.*** Clusters with no such potential geocoding errors became ***non-TES-eligible.*** The clusters in which TES was actually done are ***TES clusters,*** and were selected from among the TES-eligible clusters either with certainty or by probability sampling.

Results from the 1990 PES show that geocoding errors are highly clustered. Slightly over 77 percent of the whole household nonmatches were concentrated in less than one-fourth of the PES sample block clusters. On the other hand, about 72 percent of the census geocoding errors were found in less than 3 percent of the PES sample block clusters. TES is a good example of Deming's "80-20" guideline—80 percent of the benefits are realized by solving 20 percent of the problems.

### Persons Eligible for Search

In order to be included in TES operations, a person must live in:

- a TES cluster; and
- a TES-eligible housing unit

A person in a housing unit that was not a TES-eligible housing unit, was a non-TES person, and thus, was not directly affected by TES operations. Any person in a TES-eligible housing unit was a TES person, unless someone in the housing unit matched (i.e. someone is confirmed to be not a TES person). TES persons in clusters that were not selected for TES operations were identified, but did not have TES operations applied. Instead, these cases were effectively removed from the sample by having an assigned weight of zero. They were represented by persons in other TES clusters selected by sampling.

### TES Weighting

Every selected TES cluster was assigned a sampling weight equal to the reciprocal of its selection probability. This TES cluster weight was assigned to all TES persons in that cluster and was multiplied by their A.C.E. sampling weights to produce their TES-adjusted weights. The TES-adjusted weight for TES persons in clusters not selected for TES is zero. In this way, the TES persons in the TES clusters represent the TES persons in non-TES clusters. All elements of the dual system estimate (DSE) calculation, except those involving inmovers, can be affected by the TES weighting because TES persons can be nonmovers or outmovers, matches or nonmatches, and correct or erroneous enumerations.

## CLUSTER SAMPLING

The decision to select 20 percent of the A.C.E. block clusters for TES was based on the assumption that most of the TES-eligible housing units and persons would be concentrated in a small fraction of the block clusters. Hence, most of the benefits of a complete surrounding area search could be realized at a substantial reduction in cost, if a disproportionate share of the effort was concentrated in the clusters with the greatest likely payoff—the ones with the most TES-eligible housing units and persons. Targeting these clusters would achieve one of the principal goals of the surrounding areas search—variance reduction. However, it is of at least equal importance that the surrounding area search be "balanced." There are two ways TES could have been out of balance: 1) the geographical area included in the search could have differed between the P and E samples; 2) the TES block cluster sampling could have selected clusters containing errors of inclusion with greater or less likelihood than clusters with errors of exclusion. To achieve the balancing in sample selection, it was necessary for each cluster with TES-eligible housing units and persons to have some probability of being selected for TES and be weighted by the inverse of the selection probability.

The information available for TES selection included the results of the initial housing unit matching, which included the results from the housing unit follow-up. Housing unit follow-up indicated, among other things, the count of potential geocoding errors of inclusion and exclusion. The geocoding errors of inclusion were census units found outside the cluster. Potential geocoding errors of exclusion were coded as address nonmatches in the independent listing. The combined number of census geocoding errors and independent listing address nonmatches were considered to be the number of potential geocoding errors in each cluster. The probability that any cluster would be selected for TES depended on the count of its potential geocoding errors for most clusters. Exceptions are relisted clusters and clusters that were enumerated in the census using the List/Enumerate methodology. Those clusters did not go through housing unit matching and follow-up.

A housing unit that represented a potential geocoding error could have been discovered by TES operations to be a geocoding error or an actual coverage error. Putting a particular housing unit in the category of "potential" made it, and the persons living in it, eligible for TES. This search was intended to determine whether the housing unit and persons were geocoded incorrectly into a neighboring block, in which case they would be counted as correctly enumerated, or were truly enumeration errors.

Hence, the following TES selection strategy was implemented:

- Clusters that did not have counts of potential geocoding errors available at the time of the TES sampling operation were assigned to a separate TES procedure. Clusters that were relisted (which were later included in TES with certainty) or enumerated using the List/Enumerate methodology (which were ultimately excluded from TES) fall into this group.
- The 5 percent of clusters that included the largest number of housing units that were potential geocoding errors were included in TES with certainty.
- The 5 percent of clusters that had the most housing units that were potential geocoding errors, when weighted by their A.C.E. cluster weight, were also

included in TES with certainty. The 5 percent of clusters included in the above bullet for having the most unweighted cases were excluded before this step was performed, so that a total of 10 percent of the A.C.E. clusters were selected based on the two certainty criteria.

- All clusters with at least one potential geocoding error housing unit were assigned to a noncertainty stratum to be sampled at a uniform national rate to be included in TES. The sampling rate was set so that the overall size of the TES sample, including those selected by certainty and by sampling, totaled 20 percent of A.C.E. clusters (excluding the first group).

Clusters with no potential geocoding errors were excluded from TES selection since there were no housing units or persons that were candidates for TES operations. This creates a potential for a small bias in TES, because housing units added to or deleted from the address lists after the selection of TES clusters were not eligible for TES operations.

### Sampling Methodology

For the United States as a whole, there were 11,303 A.C.E. clusters. Of these, 420 were excluded from TES selection because they used the List/Enumerate census method. Of the remaining 10,883 clusters, 20 percent, or 2,177 were selected for TES. Of the eligible clusters, 62 were relist clusters and were not part of the normal TES selection. (These clusters did not count as part of the 2,177 TES target sample size.) Five percent of the sampling universe, or 544 clusters, with the most potential geocoding errors were selected for TES with certainty and assigned a TES weight of 1. Of the remaining clusters, an additional 544 with the most potential geocoding errors, when weighted by the A.C.E. cluster weight, were also selected with certainty and assigned a TES weight of 1.

Of the remaining clusters that included at least one potential geocoding error, 1,089 were selected using systematic random sampling with equal probability. There were 5,326 clusters in the noncertainty stratum (i.e. all those that were not already selected by one of the other means and that contained at least one potential geocoding errors), so the selected clusters were assigned a TES weight of 5,326 divided by 1,089 or 4.8907. The remaining 4,407 clusters were out of scope for TES because they had no identified potential geocoding errors.

For purposes of drawing the systematic sample, clusters were sorted in the order:

- State
- First-phase Sampling Stratum
- Second-phase Sampling Stratum
- Small Block Cluster Sampling Stratum
- Cluster Number

The first four characteristics are the same ones used to select the A.C.E. sample. Sorting clusters in this order for TES improved the representativeness of TES with respect to the national A.C.E. sample. After sorting in this order, the clusters were systematically sampled with equal probability using a take-every of 4.8907 and were assigned a TES weight equal to that figure.

### Results of Cluster Sampling

The TES sample included 2,239 block clusters out of 11,303, or 19.8 percent. (Originally it had been intended to include a small number of List/Enumerate clusters in TES, and some sample was set aside for them but never used.) The clusters included 45,000 E-sample and 77,000 P-sample housing units, representing 80 and 73 percent of TES-eligible units in their respective samples before subsampling within large block clusters was performed. Because of differences in procedures, more E-sample units got into TES by certainty (76 percent versus 66 percent), while more P-sample units were selected by sampling, 7 percent to 5 percent. TES units represent about 7 percent of the housing units resulting from initial housing unit matching. (See Table 5-1.) This was not the final number of housing units included in TES field operations because:

- Subsampling within large block clusters reduced the number of A.C.E. housing units in clusters with 80 or more housing units; and
- Housing unit counts for Relist clusters were not available at the time the sample was selected

Subsampling within large block clusters reduced the final TES workload to 12,000 E-sample and 18,000 P-sample housing units.

Table 5-1. **TES Sampling Frame and Selection Results**

| | Clusters | Potential geocoding errors | | | | | |
|---|---|---|---|---|---|---|---|
| | | Total potential errors | | Errors of inclusion | | Errors of exclusion | |
| | | Number | Percent | Number | Percent | Number | Percent |
| **Total** | **11,303** | **122,440** | **100** | **45,053** | **100** | **77,387** | **100** |
| Out-of-scope | 4,827 | 0 | ... | 0 | ... | 0 | ... |
| List/Enumerate | 420 | 0 | ... | 0 | ... | 0 | ... |
| No TES HUs | 4,407 | 0 | ... | 0 | ... | 0 | ... |
| Eligible for TES | 6,476 | 122,440 | 100 | 45,053 | 100 | 77,387 | 100 |
| Certainty | 1,150 | 85,309 | 70 | 34,089 | 76 | 51,220 | 66 |
| Top weighted | 544 | 11,858 | 10 | 4,037 | 9 | 7,821 | 10 |
| Top unweighted | 544 | 73,451 | 60 | 30,052 | 67 | 43,399 | 56 |
| Relist | 62 | 0* | ... | 0* | ... | 0* | ... |
| Noncertainty | 5,326 | 37,131 | 30 | 10,964 | 24 | 26,167 | 34 |
| Selected into sample | 1,089 | 7,642 | 6 | 2,106 | 5 | 5,536 | 7 |
| Not selected | 4,237 | 29,489 | 24 | 8,858 | 20 | 20,631 | 27 |
| TES clusters | 2,239 | 92,951 | 76 | 36,195 | 80 | 56,756 | 73 |

*TES units in Relist clusters had not been determined at the time the sample was selected.

Note: Percentages in table may not add to total due to rounding.

## TES FIELD AND PROCESSING ACTIVITIES

Details on the operations involved in TES are described in Chapter 4. In summary, the main activities are:

- **Cluster selection (Spring 2000).** This operation selects the clusters for TES. Because of the need to select the cluster sample at a particular time, the final E and P samples had not been selected at the time of this operation.
- **Search for census units in surrounding blocks (Summer 2000).** Determines if census units erroneously included in the sample block cluster are located within the surrounding ring of blocks. This field operation is described more fully in Chapter 4.
- **Identify TES Persons (Fall 2000).** An automated activity performed at the National Processing Center in Jeffersonville, Indiana. See Chapter 4 for more information.
- **Extend the search area to surrounding blocks for TES persons (Fall 2000).** The P-sample TES persons were allowed to match to census records in the surrounding block. The E-sample TES persons were treated as correct enumerations if the census unit was located in a surrounding block. This was a clerical operation.
- **Assign TES weights (Winter 2000/2001).** TES persons identified in TES-eligible clusters were assigned the TES weight associated with that cluster, either 1.0 for a cluster selected with certainty or 4.8907 for a cluster selected by sampling. TES persons in TES clusters not selected into the sample were assigned a zero weight.

### Adds and Deletes

The preliminary census address list of housing units as of January 2000 was the source for the initial housing unit matching on which TES is based. Since some housing units on the January 2000 list were later deleted and others added, the final list of census housing units did not exactly match the initial housing unit matching counts of potential geocode errors. Therefore, procedures were necessary to update the TES identifications for adds and deletes.

In the vast majority of cases, where adds and deletes were not involved, P-sample housing units are TES-eligible if they did not match to a census address. However, if a P-sample unit was matched to an address during initial housing unit matching, but that address was deleted, then the unit was considered nonmatched. To adjust for deletions, P-sample persons in housing units that were matched to deleted census housing units were flagged as TES persons, as long as the unit did not contain any persons matched within the sample block (i.e. non-TES persons). This adjustment was performed only on persons in TES clusters.

E-sample housing units that were added to the final census list after January 2000 could represent geocoding errors, but they were not part of TES field operations. Without field operations, persons in such units would never be identified as surrounding block correct enumerations. Therefore, a correct enumeration probability was imputed for such persons in TES clusters. The imputed probability is the overall correct enumeration probability of all resolved persons in geocoding error housing units in the TES sample. See Chapter 6 for a description of the procedure.

Table 5-2. **Effect of Census Address List Changes after January 2000**

| | Count | Weighted | Matches/ correct enumerations |
|---|---|---|---|
| P sample - persons in housing units matched to deletes . . | 2,319 | 2,036,564 | 675,892 |
| E sample - geocode error adds . . . . . | 53 | 15,307 | 14,915 |

## TES IN DUAL SYSTEM ESTIMATION

Accounting for TES in the DSE calculation is primarily a matter of applying weights properly. Every person in the A.C.E. is either a TES person or a non-TES person, and every A.C.E. cluster is either a TES cluster or a non-TES cluster. Every TES person is assigned the TES weight of his A.C.E. cluster. The calculation of the DSE requires the use of seven distinct components, all but one of which represents the sum of the A.C.E. weights for some group of persons in the A.C.E., including both TES and non-TES persons. Hence, six of the seven components represents a weighted sum of TES and non-TES persons, the former with their TES cluster weights applied.

### Applying TES Weights

Every A.C.E. cluster including TES persons has a TES weight, although that weight is zero if the cluster is not selected for TES. A TES person must be weighted by the associated TES weight. The A.C.E. weight is multiplied by the TES weight to produce a person weight. TES weighting does not affect the weight of non-TES persons. Their individual weights are the same as the A.C.E. weights.

Table 5-3. **TES Weights by TES Status of the Person and Cluster**

| | TES cluster | Non-TES cluster |
|---|---|---|
| TES persons . . . . . | 1, if cluster in TES with certainty | 0 |
| | 4.8907, if cluster selected for TES by sampling | 0 |
| Non-TES persons . . . . . | 1 | 1 |

The issues related to inmovers, outmovers and noninterviews are the same for TES persons as for all other persons. From a calculation standpoint, the only effect that TES status has on the dual system estimates is in applying the cluster's TES weight.

### DSE Calculation

The DSE for Census 2000 is:

$$\hat{\mathrm{DSE}} = (\mathrm{DD})\left(\frac{\mathrm{CE}}{\mathrm{N}_e}\right)\left(\frac{\mathrm{N}_n + \mathrm{N}_i}{\mathrm{M}_n + \left(\frac{\mathrm{M}_o}{\mathrm{N}_o}\right)\mathrm{N}_i}\right)$$

where

DD = census data-defined persons
CE = estimated number of A.C.E. E-Sample correct enumerations
$\mathrm{N}_e$ = number of A.C.E. E-Sample persons
$\mathrm{N}_n$ = estimated number of A.C.E. P-Sample nonmovers
$\mathrm{N}_i$ = estimated number of A.C.E. P-Sample inmovers
$\mathrm{N}_o$ = estimated number of A.C.E. P-Sample outmovers
$\mathrm{M}_n$ = estimated number of A.C.E. P-Sample nonmover matches
$\mathrm{M}_o$ = estimated number of A.C.E. P-Sample outmover matches

The estimator has seven A.C.E. distinct components (plus DD from the census enumeration). Six of the seven components represent a weighted sum of persons, including both TES- and non-TES persons.

Other than inmovers, who cannot be TES persons, each of the DSE components is expressed as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n_p} w_{ij}^{*} m_{ij} x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n_p} w_{ij}^{*} m_{ij} y_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{n_p} w_{ij}^{*} t_{ij} m_{ij} z_{ij}$$

where

$i$ = cluster index
$j$ = person index
$n$ = number of block clusters in the A.C.E. sample
$n_p$ = number of persons in block cluster i
$x_{ij}$ = 1 if the person is not a TES person, 0 otherwise
$y_{ij}$ = 1 if the person is a TES person and is in the TES sample with certainty, 0 otherwise
$z_{ij}$ = 1 if the person is a TES person and is in the TES systematic sample, 0 otherwise
$m_{ij}$ = characteristic of interest, match, correct enumeration, E-sample person, or P-sample person
$w_{ij}^{*}$ = weight used for estimation (includes inverse of the probability of selection for A.C.E., adjustment for household noninterview and weight trimming)
$t_{ij}$ = TES sampling weight, the TES systematic sample take-every

## EFFECTS OF TES ON DUAL SYSTEM ESTIMATION

The principal effect of TES in Census 2000 is approximately what was expected–the overall correct enumeration rate was 2.9 percent higher with TES, than it would

Table 5-4. **Effect of TES at the National Level**

| | With TES | Without TES | Difference* | Effect of TES |
|---|---|---|---|---|
| | (1) | (2) | (1)-(2) | (1)/(2) |
| E sample | | | | |
| Persons ($N_e$) | 264,578,862 | 264,634,794 | (55,932)** | 1.000 |
| Correct Enumerations (CE) | 252,096,238 | 244,387,951 | 7,708,288 | 1.032 |
| CE Rate (%) | 95.3 | 92.3 | 2.9 | 1.032 |
| P sample | | | | |
| Persons ($N_p$) | 263,037,259 | 262,906,916 | 130,343** | 1.000 |
| Matches | 240,878,622 | 230,681,205 | 10,197,418 | 1.044 |
| Match Rate (%) | 91.6 | 87.7 | 3.8 | 1.044 |
| Ratio of CE to Match Rate | 1.040 | 1.053 | (0.012) | 0.989 |
| Coefficient of Variation for Ratio | 0.129 | 0.314 | (0.197) | 0.405 |

*Percentages were calculated on unrounded values.
**The weighted E- and P-sample sizes differed slightly because of variance in TES sampling.

Note: Table above reflects national totals without regard to post-stratification and differs from other totals in which post-stratum totals were aggregated.

have been without, and the overall match rate was 3.8 percent higher (see Table 5-4). The larger increase in the match rate, as compared to the correct enumeration rate, occurred because there were more identified potential geocoding errors in the P sample than in the E sample.

The difference in the number of matches versus correct enumerations (10.2 million and 7.7 million, respectively) from TES had been a source of concern, since it suggested the possibility of "balancing error." Balancing error would have occurred if the geographic boundaries included in the P and E samples had not been consistent. For instance, suppose the P sample was allowed to match to census persons in housing units beyond the first ring, while a census unit could only be classified correct if it was within the first ring. Adams and Liu (2001) performed an evaluation study of the P-sample housing units in A.C.E. and concluded that the main source of the measured imbalance was geocoding error in the P sample.

Table 5-4 shows that TES increased the number of correct enumerations from 244.4 million to 252.1 million and matches from 230.7 million to 240.9 million. Before TES, there had been 20.2 million erroneous enumerations, of which 7.7 million were geocoding errors that were classified as correct enumerations by TES. TES also allowed 10.2 million additional P-sample matches to occur out of 32.2 million original nonmatches. Improving both the match and correct enumeration rate this much significantly improves the variance of the DSE, since over 90 percent of people match or are correctly enumerated.

Table 5-5 shows the significant contribution that TES makes to variance reduction. For the A.C.E. considered as a whole (i.e. a direct DSE of the entire population without post-stratification), the coefficient of variation is 0.129 percent with TES and 0.314 percent without TES.

Table 5-5. **Effect of TES on Coefficient of Variation (CV)**

| | Standard error | CV (percent) |
|---|---|---|
| With TES | 355,451 | 0.129 |
| Without TES | 877,664 | 0.314 |

Note: Table above reflects national totals without regard to post-stratification and differs from other totals in which post-stratum totals were aggregated.

At the post-stratum level, the average improvement in the DSE standard error is about 33 percent. The gains in precision as measured by variance show that TES makes dual system estimates more precise, and that TES improves the quality of the A.C.E., so long as it does not make the DSEs less accurate by introducing bias. The coefficient of variation was reduced for a majority of the collapsed post-strata (448 original post-strata were collapsed into 416 post-strata for DSE calculation purposes).

Table 5-6. **Effect of TES on Post-Stratum CVs**
[Percent]

| | With TES | Without TES |
|---|---|---|
| Average CV | 2.07 | 2.66 |
| Median CV | 1.81 | 2.32 |
| Average CV weighted by census count | 1.30 | 1.93 |

# Chapter 6.
# Missing Data Procedures

## INTRODUCTION

This chapter gives an overview of missing data procedures for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.). General background information is presented first, while the following sections describe three types of procedures used to account for data missing in the A.C.E. The noninterview adjustment accounts for whole-household nonresponse. The next section describes the characteristic imputation used to assign values for specific missing demographic variables. Finally, for persons with unresolved match, residence, or enumeration status, a probability of matching, residence, or correct enumeration was assigned according to procedures.

As missing data in the A.C.E. were addressed after the completion of the field operations that produced the A.C.E. data files, a knowledge of the field activities and the circumstances that led to specific outcomes is necessary to understand the motivation for these procedures. For this information, the reader is referred to Chapter 4 for details on the field operations.

The missing data procedures used in the Census 2000 A.C.E. were similar to those used on the Integrated Coverage Measurement (ICM) sample in the Census 2000 Dress Rehearsal. An outline of the ICM procedures and a summary of related research are given in Ikeda, Kearney, and Petroni (1998). Kearney and Ikeda (1999) provide an overview of the results from the Dress Rehearsal. For detailed missing data procedures for the 2000 A.C.E., see Cantwell (2000) and Ikeda and McGrath (2001). A few basic results on missing data from 2000 are found in this chapter; many more results can be found in Cantwell et al. (2001).

## BACKGROUND

Before dual system estimates were calculated, it was necessary to account for missing information from the interviews of P-sample people and from the matching operations. It should be noted that the term "missing data" applies after all follow-up attempts have been made. Chapter 4 describes some of the extensive field procedures conducted to minimize the resulting level of such missing data. These activities – all specified in advance – included multiple attempts at interviews, the use of highly trained clerks and technicians to resolve cases, and the follow up of cases where a second interview could provide additional required information.

There were two main types of missing data in the A.C.E. and three processes used to correct for them. The first type was ***unit missing data.*** These were households that were not interviewed in the A.C.E. either because they could not be contacted or because the interview was refused. The noninterview adjustment process spread the weights of these households among households that were interviewed in the same noninterview cell.

The other type of missing data was ***item missing data.*** This situation occurred when some information for a household or person was available but portions of the data were missing. Two groups of missing data items had to be addressed: demographic items and items relating to a specific operational status. Missing age, sex, tenure, race, and Hispanic origin were imputed to allow the production of estimates of the census undercount by these characteristics, and because they were necessary to assign people to post-strata.

For a small number of people in the P sample, there was not enough information available to determine the match status (whether or not the person matched to someone in the census in the appropriate search area) or the residence status (whether or not the person was living in the block cluster on Census Day). Determining residence status was important for the P sample because Census Day residents of the block clusters in the sample were used to estimate the proportion of the population who were not counted in the census. Similarly, some people in the E sample lacked information to determine whether the person was correctly enumerated. Such cases where status could not be determined were said to be "unresolved." Generally for cases with missing status a probability of residence, match, or correct enumeration was assigned based on information available about the specific case and about cases with similar characteristics.

In the 1990 Post-Enumeration Survey, a hierarchical logistic regression program was used to calculate probabilities of match and correct enumeration for cases with missing information. (Due to the procedure used to treat movers in 1990, residence status played a different role then.) The model and some results are discussed in Belin et al. (1993). During census tests in 1995 and 1996, certain components of missing data were addressed using logistic regression, while for other components a simpler procedure called imputation cell estimation was used. The latter procedure was used exclusively in the Census 2000 Dress Rehearsal in 1998. Data from these tests indicate that the exact method of calculating probabilities for unresolved status (match, residence, or correct enumeration) has a

minor effect on the dual system estimates. More details of this research can be found in Ikeda (1997, 1998, 1998b, and 1998c) and Cantwell (1999). Based on these findings and concerns about implementing logistic regression in a production environment, the simpler procedure (that is, imputation cell estimation) was used to estimate missing data items in the A.C.E.

**Noninterview Adjustment (Household Level)**

At the time of the Computer Assisted Personal Interview (CAPI), questions were asked to determine who lived in the household on Interview Day and who lived there on Census Day, and a mover status was assigned based on the replies. Thus two rosters were created for each household–the Census Day roster and the Interview Day roster. The A.C.E. used inmovers to estimate the number of P-sample movers in the post-stratum, while using outmovers to estimate the match rate of the movers. This method is referred to as Mover Procedure C or PES-C in the research studies. See Chapter 4 for descriptions of the terms nonmover, inmover, and outmover.

All inmovers and all nonmovers were generally assumed to be A.C.E. Interview Day residents, with the exception of infants born after Census Day. People living in group quarters, such as college students in dormitories, were not eligible for the P sample. Therefore, for the purpose of estimating the number of inmovers, person inmovers aged 18 to 22 who were living in group quarters on Census Day were not considered to be Interview Day residents.

Noninterview adjustment was performed only on the P sample. The procedure was similar to that used in the Census Dress Rehearsal. Due to the mover procedure described above, there were two noninterview adjustments – one based on housing unit status as of Census Day (i.e., the Census Day roster), and the other based on housing unit status as of the day of the A.C.E. interview (i.e., the Interview Day roster). An occupied housing unit was defined as an interview (for the given reference day – Census Day or Interview Day) if there was at least one person (with a name and at least two demographic characteristics) who possibly or definitely was a resident of the housing unit on the given reference day. An occupied housing unit (as of the given reference day) that was not an interview was a noninterview. Thus a unit that was vacant, removed from the list of eligible housing units (because, for example, it was demolished or used only as a business), or in certain special places was not considered an interview or a noninterview. In the latter two situations, the unit was "deleted" from the list of A.C.E. sample housing units.

If a housing unit was found to be vacant on Census Day or deleted from the sample, then that household did not factor into the Census Day noninterview adjustment. The same concept applies to Interview Day. Thus, vacant and deleted units did not contribute toward dual system estimation. An example of an illustrative block cluster, provided in Figure 6-1, page 6-10, shows how the status of a housing unit on Census Day and Interview Day would be determined. Results of the A.C.E. interviewing operation are shown in Table 6-1.

Table 6-1. **Status of Household Interviews in the A.C.E.** [Unweighted]

| | Census Day | | A.C.E. Interview Day | |
|---|---|---|---|---|
| | Number | Percent | Number | Percent |
| Total housing units | 300,913 | 100.0 | 300,913 | 100.0 |
| Interviews | 254,175 | 84.5 | 264,103 | 87.8 |
| Noninterviews | 7,794 | 2.6 | 3,052 | 1.0 |
| Vacant units | 28,472 | 9.5 | 29,662 | 9.9 |
| Deleted units | 10,472 | 3.5 | 4,096 | 1.4 |
| Noninterview rate | 3.0% | | 1.1% | |

Note: Percentages in table may not add to total due to rounding.

Of the 261,969 housing units occupied on Census Day, 7,794 (3.0 percent) were noninterviews. The corresponding numbers for Interview Day were 267,155 and 3,052 (1.1 percent). The noninterview rate was higher for Census Day than Interview Day, because interview status was determined by results obtained on Interview Day. On that date, information was sought for both Census Day and Interview Day. Any time a household member or knowledgeable proxy could be reached, an interview for Interview Day was generally obtained. Census Day data was not always obtainable from the same respondent, usually in cases when the housing unit's occupants had moved in after Census Day. Each of the two noninterview adjustments generally spread the weights of noninterviewed units over interviewed units in the same noninterview cell, defined as the sample block cluster crossed with the type of basic address. For purposes of this adjustment, the types of basic address were single-family, multiunit (such as apartments), and all others. The Census Day noninterview adjustment, determined according to the status of housing units as of Census Day, was used to adjust the person weights of nonmovers and outmovers. Similarly, the Interview Day noninterview adjustment, determined according to the status of housing units as of Interview Day, was used to adjust the person weights of inmovers. The formulae are described as follows:

For a given block cluster and type of basic address, the Census Day noninterview adjustment factor was computed as

$$f^*_c = \frac{\sum_{\substack{\text{Census Day}\\ \text{interviews}}} w_i + \sum_{\substack{\text{Census Day}\\ \text{noninterviews}}} w_i}{\sum_{\substack{\text{Census Day}\\ \text{interviews}}} w_i}$$

where $w_i$ represents the weight of housing unit i, that is, the inverse of its probability of selection into the A.C.E. sample. When computing the noninterview adjustment factor, the weight $w_i$ incorporated the trimming that occurred in some block clusters. (See Appendix C.) However, the weights did not reflect the sampling for targeted extended search (TES, Chapter 5) for two reasons. First, the noninterview adjustment was done at the housing unit level, but a housing unit could contain some people with TES status and others without it. Second, TES status was not determined until after the matching operation, but information was usually not collected about people in noninterviewed units, and these people were generally not sent to be matched. Therefore, there was not a reasonable way of systematically classifying noninterviews into those with and without TES status.

Similarly, for a given cluster and type of basic address, the Interview Day noninterview adjustment factor was computed as

$$f^*_i = \frac{\sum_{\substack{\text{Interview Day} \\ \text{interviews}}} w_i + \sum_{\substack{\text{Interview Day} \\ \text{noninterviews}}} w_i}{\sum_{\substack{\text{Interview Day} \\ \text{interviews}}} w_i}$$

The example in Figure 6-1 on page 6-10, demonstrates the calculation of the noninterview adjustment. When the unweighted number of noninterviewed units in a given noninterview cell (sample block cluster by type of basic address category) was more than twice the unweighted number of interviewed units, then the weights of the noninterviewed units were spread over the interviewed units in a broader cell. This cell was formed by combining the sample block clusters in the same A.C.E. sampling stratum within the same type of basic address. Because the noninterview rates were so small, the noninterview adjustment factors were close to 1 for most housing units in the sample. For Census Day, the factors were smaller than 1.10 for more than 92 percent of the units; for Interview Day, the factors were less than 1.10 for over 98 percent of the units.

### Characteristic (Item) Imputation (Person Level)

Production of A.C.E. undercount estimates required data on age, sex, tenure (owner versus nonowner), race, and Hispanic origin to classify respondents by these important demographic characteristics, so they had to be imputed whenever the data were not collected. Characteristic imputation was not carried out for other missing variables (with the exception of the items with unresolved status). Several variables also used to assign post-strata, such as the location or return rate of the census tract, were the same for everyone in the block. The extent of the missing characteristics is portrayed in Table 6-2.

The imputation rates in the E sample for the five characteristics listed above ranged from 0.3 percent for sex up to 3.8 percent for tenure (using unweighted frequencies). Since the A.C.E. record for each person in the E sample was matched to the Census 2000 edited file and the five characteristics were extracted and copied, the following imputation procedures apply only to the P sample.

P-sample characteristic imputation for the A.C.E. was similar to that for the 1990 PES and the various Census 2000 tests, including the Dress Rehearsal. Age and sex were imputed based on the available demographic distributions determined from the P sample. Tenure was imputed using a form of nearest-neighbor hot-deck procedure. To impute for race and Hispanic origin, the two approaches were combined.

For missing tenure, race, and Hispanic origin, a hot-deck procedure was used to take advantage of the correlations often found in these characteristics among people living in the same block cluster (or, generally, in geographic proximity). The characteristics age and sex are geographically less clustered than tenure, race, and Hispanic origin. Further, the value of age or sex is often considerably affected by specific conditions, such as the person's relationship to the reference person, or whether information is available on the person's spouse. Thus, national distributions conditioned on relevant covariates were used to impute for age and sex. These distributions were constructed before the imputation began, without regard to the imputation for other missing characteristics.

Table 6-2. **Percent of Characteristic Imputation in the P and E Samples** [Unweighted]

| | Total people | Percent of people with imputed characteristic | | | | | Percent of people with one or more imputed characteristics |
|---|---|---|---|---|---|---|---|
| | | Age | Sex | Tenure | Race | Hispanic origin | |
| P sample | 706,245 | 2.5 | 1.7 | 1.9 | 1.4 | 2.4 | 5.5 |
| E sample | 704,602 | 3.1 | 0.3 | 3.8 | 3.5 | 3.6 | 11.2 |

**Age.** The value of age was missing for 2.5 percent (unweighted) of the P sample. When age was missing, one of four age categories (0-17, 18-29, 30-49 and 50 or older) – rather than a number – was imputed, because only the category was used to assign people to a post-stratum for estimation. In one-person households, missing age was imputed from the distribution of ages reported in such households. In multiperson households, if the relationship to the reference person was missing, the distribution of ages (excluding those of reference persons) in all multiperson households was used. Otherwise, if the person was the spouse, child, sibling, or parent of the reference person, missing age was generally imputed from a distribution of reported ages using the relationship to the reference person and the age of the reference person. For reference persons, other relatives, and nonrelatives, age was imputed from the distribution of ages reported by persons with the same "relationship." See Figure 6-2, on page 6-11, for details.

**Sex.** The imputation rate for sex was 1.7 percent in the P sample. For one-person households, sex was imputed from the distribution of sex in all one-person households. To impute the sex of a reference person, if the household had more than one person but no spouse was present, the distribution of sex for reference persons of multiperson households with no spouse present was used. If a spouse was present, the missing sex of the reference person or the reference person's spouse was imputed as the sex opposite to that of the spouse. If sex was missing for the reference person and the spouse, then the sex of the reference person was imputed from the distribution of sex for reference persons with a spouse present. The spouse was then assigned the sex opposite to that of the reference person.

For other persons in multiperson households (that is, other than reference persons and spouses): 1) if the relationship to the reference person was missing, and if no one else in the household was recorded as a spouse of the reference person, sex was imputed from the distribution of sex for persons (excluding reference persons) from all multiperson households; 2) otherwise, sex was imputed from the distribution of sex for persons (excluding reference persons, spouses, and persons with missing relationship) from all multiperson households. Figure 6-3, on page 6-12, illustrates the procedure.

**Tenure.** Household tenure (owner versus nonowner) was missing for 1.9 percent of the people in the P sample. Tenure was imputed from the previous household that had the same type of basic address and had tenure recorded. As with the adjustment for noninterviews, three types of basic address were used: single-family, multiunit, and all other types of units. See Figure 6-4, on page 6-13, for further information.

**Race.** When race was missing – 1.4 percent of the P sample – the imputed race could be any of the 63 possible combinations of the six basic race categories: White, Black, American Indian or Alaskan Native, Asian, Native Hawaiian or Other Pacific Islander, and Some Other Race. All 63 categories were treated the same in the imputation. That is, there were no special procedures for any categories or groups of categories.

Whenever possible, missing race was imputed from the same household. Independently for each household member with missing race, one person was selected at random from those household members with reported race and the selected person's race was imputed to the given household member. If race was missing for all household members but someone had reported origin (Hispanic or non-Hispanic), then the race distribution of the nearest previous household with any reported race and the same origin was used. Note that the Hispanic origin of the household was that of the first person on the household roster with origin reported. When race and Hispanic origin were missing for the whole household, the race distribution of the nearest previous household with reported race was used—regardless of Hispanic origin. See Figure 6-5, on page 6-14, for details.

**Hispanic Origin.** A value of origin – Hispanic or non-Hispanic – was imputed for 2.4 percent (unweighted) of the P sample. The procedure was analogous to that for imputing missing race. That is, whenever possible, origin was imputed from within the same household. If everyone in the household was missing origin, then the nearest previous household with reported origin and the same race category was used. When both Hispanic origin and race were missing for the whole household, the Hispanic origin distribution of the nearest previous household with reported Hispanic origin was used – regardless of race. For the imputation procedure and the race categories used in it, see Figure 6-6, on page 6-15.

For each of the five characteristics discussed, the distribution of imputed values did not necessarily mirror the distribution of reported values – nor was this expected. However, because the imputation rates were low in the P and E samples, the distributions before and after imputation were very similar. See the distribution of characteristics on the following page.

**Distribution of Characteristics Before and After Imputation** [Weighted]

| | P sample | | | E sample | | |
|---|---|---|---|---|---|---|
| | Before imputes | Imputed | After imputes | Before imputes | Imputed | After imputes |
| **Race** | **1.4% Imputed** | | | **3.2% Imputed** | | |
| White Only | 73.5% | 67.5% | 73.4% | 76.9% | 57.2% | 76.2% |
| Black Only | 11.0% | 10.2% | 11.0% | 11.8% | 6.6% | 11.6% |
| AIAN Only | 0.6% | 0.7% | 0.6% | 0.8% | 0.8% | 0.8% |
| Asian Only | 3.5% | 3.4% | 3.5% | 3.7% | 2.9% | 3.7% |
| NHPI only | 0.1% | 0.3% | 0.1% | 0.1% | 0.3% | 0.1% |
| Some other race only | 8.3% | 14.4% | 8.4% | 4.5% | 28.5% | 5.3% |
| Multiple races | 3.0% | 3.5% | 3.0% | 2.3% | 3.7% | 2.3% |
| **Hispanic origin** | **2.3% Imputed** | | | **3.4% Imputed** | | |
| Hispanic | 12.4% | 11.5% | 12.4% | 12.5% | 9.0% | 12.4% |
| **Age** | **2.4% Imputed** | | | **2.9% Imputed** | | |
| 0-17 | 26.1% | 21.7% | 26.0% | 25.9% | 19.7% | 25.7% |
| 18-29 | 16.7% | 18.9% | 16.7% | 15.5% | 19.0% | 15.6% |
| 30-49 | 30.7% | 33.0% | 30.8% | 31.0% | 30.9% | 31.0% |
| 50+ | 26.5% | 26.4% | 26.5% | 27.6% | 30.5% | 27.6% |
| **Sex** | **1.7% Imputed** | | | **0.2% Imputed** | | |
| Male | 48.4% | 47.2% | 48.3% | 48.8% | 53.9% | 48.8% |
| Female | 51.6% | 52.8% | 51.7% | 51.2% | 46.1% | 51.2% |
| **Tenure** | **1.9% Imputed** | | | **3.6% Imputed** | | |
| Owner | 68.4% | 70.3% | 68.4% | 69.9% | 65.1% | 69.7% |
| Nonowner | 31.6% | 29.7% | 31.6% | 30.1% | 34.9% | 30.3% |

### Assigning Probabilities for Unresolved Cases (Person Level)

After all follow-up activities were completed, there remained a small fraction of the A.C.E. sample without enough information to compute the components of the dual system estimator given in Chapter 7. Their status was said to be "unresolved." A procedure called imputation cell estimation was used to assign probabilities for P-sample people with unresolved match or Census Day residence status, and for E-sample people with unresolved enumeration status.

All P- and E-sample persons – resolved and unresolved – were placed into groups called imputation cells based on operational and demographic characteristics. Different variables were used to define cells for P-sample match and residence status and in the E-sample for enumeration status. Within each imputation cell the weighted average of 1's and 0's (representing, e.g., match and nonmatch, respectively) among the resolved cases was calculated, and that average was imputed for all unresolved persons in the cell.

One should note that the noninterview adjustment factor was not incorporated into the person weights when these averages were calculated. This is because the noninterview adjustment was designed to spread the weight of noninterviewed housing units over interviewed housing units. However, all persons with resolved residence status in noninterviewed units were nonresidents (since, by definition, if one person in the household was a resident then the household was considered an interview). Therefore, using the noninterview factor to calculate the averages for unresolved cases would have produced a biased estimate of residence probability. The issue of which weights to use was moot when resolving E-sample cases with missing enumeration status, as a noninterview adjustment was not applied to E-sample persons.

Thus, the weights, $w_i$, used here incorporated all stages of sampling, including the selection of people for targeted extended search, but were not adjusted for household noninterviews. Any trimming of the weights was also performed before these weighted averages were calculated.

### Unresolved Residence Status in the P Sample

After follow-up was completed, all persons in the P sample who were eligible to be matched to the Census (see Chapter 4) were classified into three types, according to their status as a resident in their sampled block at the time of the census: Census Day residents, Census Day nonresidents, and unresolved persons – those for whom there was not enough information to determine the residence status. The results are displayed in Table 6-3.

Table 6-3. **Final Residence Status for the P Sample by Mover Status**
[Unweighted]

| | Total people | Final residence status | | | Residence rate for resolved cases |
|---|---|---|---|---|---|
| | | Confirmed resident | Confirmed nonresident | Unresolved resident | |
| **U.S. total** | **653,337** | **95.8%** | **1.9%** | **2.3%** | **98.1%** |
| Mover status | | | | | |
| Nonmover | 627,992 | 96.6% | 1.7% | 1.7% | 98.3% |
| Outmover | 25,345 | 75.2% | 7.5% | 17.4% | 91.0% |

Because of the uncertainty of the actual status of the 15,082 people (2.3 percent of 653,337) with unresolved residence status, a probability of being a Census Day resident was assigned (see equation (6.2)). Then, when computing the dual system estimate, all person nonmovers and outmovers were included with their estimation weight (see Chapter 7) and the following residence probability:

(6.1)

$$Pr_{res,j} = \begin{cases} 1, \text{ if person j is a resident on Census Day} \\ 0, \text{ if person j is NOT a resident on Census Day} \\ Pr^*_{res,j}, \text{ if person j is unresolved} \end{cases}$$

To assign $Pr^*_{res,\,j}$ for unresolved cases, the Census Day residence probability for inmovers was irrelevant for estimation and was not used. Only nonmovers and outmovers in the P sample who had a resolved final residence status and went through the person matching operation (formally, those with a final match-code status) were used. They were placed into a number of imputation cells as defined in Table 6-4. Within each cell, among the resolved cases (those with $Pr_{res,\,j} = 1$ or 0) the weighted proportion of Census Day residents, that is, the weighted average of 1's and 0's, was computed:

$$Pr^*_{res,\,j} = \frac{\sum_{\substack{resolved \\ persons}} w_i \, Pr_{res,j}}{\sum_{\substack{resolved \\ persons}} w_i} \qquad (6.2)$$

where $w_i$ was defined at the beginning of this section.

This proportion was then assigned as $Pr^*_{res,j}$ to each unresolved case in the cell. (The exception is for follow-up match code group 7; this is explained below.) The cells used to resolve residence status, along with the probabilities assigned to the unresolved cases, are given in Table 6-4.

Match code groups 1 through 7, which partition the population into mutually exclusive and exhaustive groups, were determined from the match codes and other variables derived before the follow-up operation as explained in Chapter 4. Group 8 was formed differently. Some information from the follow-up operation was coded in time for

Table 6-4. **Imputation Cells and Probabilities Assigned for Resolving Residence Status in the P Sample**

| Match code group | Owner | | | | Nonowner | | | |
|---|---|---|---|---|---|---|---|---|
| | Non-Hispanic White | | Others | | Non-Hispanic White | | Others | |
| 1 = Matches needing follow-up | 0.982 | | 0.986 | | 0.993 | | 0.991 | |
| 2 = Possible matches | 0.973 | | 0.968 | | 0.966 | | 0.972 | |
| 3 = Partial household nonmatches needing follow-up | V3a* 0.755 | V3b* 0.956 | V3a* 0.901 | V3b* 0.971 | V3a* 0.883 | V3b* 0.959 | V3a* 0.928 | V3b* 0.969 |
| 4 = Whole household nonmatches needing follow-up, not conflicting households | 0.920 | | 0.943 | | 0.911 | | 0.914 | |
| 5 = Nonmatches from conflicting household | 0.910 | | 0.927 | | 0.945 | | 0.954 | |
| 6 = Resolved before follow-up | 0.993 | | 0.990 | | 0.990 | | 0.988 | |
| 7 = Insufficient information for matching (Weighted column average of groups 1-5 and 8) | 0.813 | | 0.867 | | 0.844 | | 0.872 | |
| 8 = Potentially fictitious or said to be living elsewhere on Census Day | 0.119 | | 0.123 | | 0.177 | | 0.157 | |

* V3a = Group 3 Persons age 18-29 listed as child of reference person; V3b= All other group 3 persons.

the A.C.E. missing data procedures. (Under the original schedule, this information would have become available too late to be of use.) After the follow-up operation, a small number of people in the P sample were coded as being potentially fictitious or said to be living elsewhere on Census Day. Such people were placed in Group 8, even though they also qualified for one of the Groups l1 through 7.

The two tenure categories were owners and nonowners. Persons were placed into one of two race categories: non-Hispanic White and all others. People of multiple races (for example, a person responding as White and Asian) were placed in the latter group. V3 was a variable defined only for match code group 3, partial household nonmatches. V3a comprised persons in group 3 who were 18 to 29 years of age and were listed on the A.C.E. household roster as a child of the reference person. V3b included all other persons in group 3.

The residence probability for unresolved P-sample persons was computed as described above, except for those in match code group 7 – people with insufficient information for matching. Within this set of four cells (see Table 6-4), there were almost no resolved cases from which to extract a probability of being a Census Day resident. Because of the lack of information – most of these cases did not even have a valid name – these people did not go through the matching operation and were not sent to follow-up. To adjust for these cases, a weighted proportion of Census Day residents (1's and 0's) was computed among the resolved cases in each of the four columns of Table 6-4 using match code groups 1 through 5 and 8. Separately for each of the four tenure race/ethnicity classes, the overall weighted probability of being a resident among those sent to follow-up (groups 1 through 5 and 8) was assigned to those with insufficient information for matching (group 7). Left out of this computation were those people who were resolved before follow-up (group 6). Observations from the Census 2000 Dress Rehearsal indicated that, in terms of their demographic and operational characteristics, people in group 7 tend to be more like those in groups 1 through 5 and 8, than like those in group 6.

In the Dress Rehearsal, only three weighted ratios were calculated for residence probability: a ratio for persons sent to follow-up, a ratio for persons not needing follow-up, and an overall ratio used for persons with insufficient information for matching. Based on Dress Rehearsal results, Kearney and Ikeda (1999) suggest calculating separate ratios by match code group and splitting persons from conflicting households into a separate match code group. The larger Accuracy and Coverage Evaluation sample size in Census 2000 than in the Dress Rehearsal made it possible to separate matches needing follow-up from possible matches. Additional research and discussion suggested adding additional variables within match code group.

### Unresolved Match Status

Computing the dual system estimator required measuring the total number of P-sample people who were matched to persons included in the census. (Separate estimates were obtained for nonmovers and outmovers, but that does not affect what follows.) After follow-up activities were completed, each confirmed or possible (unresolved) Census Day resident in the P sample was determined to be a match, a nonmatch, or unresolved (that is, persons for whom match status could not be determined). Match status of confirmed Census Day nonresidents was not used in the estimation. As is seen in Table 6-5, unresolved matches were infrequent in the P sample.

The treatment of unresolved matches was similar to that for unresolved residence status. For each confirmed or possible Census Day resident j in the P sample, the value $Pr_{m,\,j}$ was assigned as 1, 0, or $Pr^*_{m,\,j}$, in a manner analogous to equation (6.1), according to whether the person was a match, a nonmatch, or had unresolved match status, respectively. Unresolved matches accounted for 7,826 of 640,945 people in the P sample, or 1.2 percent. $Pr^*_{m,\,j}$ was assigned using imputation cell estimation based on those with a resolved match status. The formula is the same as in equation (6.2), but pertains to match status, that is, uses the values of $Pr_{m,\,j}$.

Table 6-5. **Final Match Status for the P Sample by Mover Status** [Unweighted]

| P sample (confirmed or possible residents) | Number of persons | Final match status | | | Match rate for resolved cases |
|---|---|---|---|---|---|
| | | Match | Nonmatch | Unresolved match | |
| U.S. total | 640,945 | 90.3% | 8.5% | 1.2% | 91.4% |
| Mover status | | | | | |
| Nonmover | 617,490 | 91.1% | 8.0% | 0.9% | 91.9% |
| Outmover | 23,455 | 67.8% | 21.7% | 10.5% | 75.8% |

As with residence status, the cases were first classified according to several characteristics. Within cells, the weighted proportion of matches among the resolved cases – excluding all confirmed Census Day nonresidents – was computed and assigned to each of the unresolved cases in the same cell. Again, the weights, $w_i$, are defined earlier.

The characteristics used to define the imputation cells for match status – different from those used for residence status – are shown in Table 6-6. They were based on observations from the Census 2000 Dress Rehearsal and an analysis of the A.C.E. operations. Kearney and Ikeda (1999) showed that mover status (nonmover versus outmover) discriminated well between matches and nonmatches among the resolved cases. The housing unit address match code refers to the initial match between housing units on the independent (A.C.E.) listing and the census address list; conflicting housing units were determined during A.C.E. person matching activities.

People with at least one imputed demographic variable (i.e., age, sex, race, Hispanic origin, or tenure) were grouped together for imputation of match status. Unpublished studies indicate that, at least in the Dress Rehearsal, the presence of these imputed characteristics among resolved cases is negatively associated with the propensity to be a match. For outmovers from a unit that was a nonmatch or a conflicting household, people were not separated according to their imputed characteristics. The reason was to maintain a reasonable number of resolved cases in each cell from which to estimate the weighted proportion of matches. The probabilities assigned to people with unresolved match status are provided in Table 6-6. It is useful to note that most persons with unresolved match status (7,693 of the 7,826) had insufficient information for matching; most of them did not have a valid name, and their rate of missing characteristics was much higher than the average. Further, almost all of these people (7,506) were in match code group 7. As such, they did not go through the matching process, nor were they sent for follow-up. This information was considered when cells were selected for imputation of match status. Variables such as age and ethnicity – that had a high chance of being imputed and might be of questionable quality – were avoided.

In the Dress Rehearsal, within each of the four geographic sites, one overall weighted ratio for match probability was calculated and used. Kearney and Ikeda (1999) suggest that separate ratios for outmovers and nonmovers should be calculated.

### Unresolved Enumeration Status (E Sample)

The dual system estimator also required the total number of correct enumerations in the E sample. As with operations previously discussed, follow-up activities left each person in the E sample with one of three types of enumeration status: correct, erroneous, or unresolved. The person was assigned a number, $Pr_{ce,\,j}$, equal to 1, 0, or $Pr^*_{ce,\,j}$, respectively, according to that status, similar to equation (6.1). Table 6-7 shows the distribution of persons according to enumeration status. The values of $Pr^*_{ce,j}$ for the 21,148 unresolved E-sample people (3.0 percent of 704,602) were determined through imputation cell estimation.

Table 6-6. **Imputation Cells and Probabilities Assigned for Resolving Match Status in the P Sample**

| Mover status | Housing Unit Address Match Code | | | |
|---|---|---|---|---|
| | Housing unit was a match (code 1) | | Housing unit was a nonmatch or the household is conflicting (code 2 or 4) | |
| | No imputes | 1 or more imputes | No imputes | 1 or more imputes |
| Nonmover | 0.945 | 0.901 | 0.690 | 0.567 |
| Outmover | 0.798 | 0.791 | 0.516 | |

Table 6-7. **Final Enumeration Status for the E Sample** [Unweighted]

| E sample | Number of persons | Final enumeration status | | | Correct enumeration rate for resolved cases |
|---|---|---|---|---|---|
| | | Correct enumeration | Erroneous enumeration | Unresolved enumeration | |
| U.S. total ..... | 704,602 | 92.6% | 4.4% | 3.0% | 95.5% |

The resolved and unresolved cases were placed in the cells defined shown in Table 6-8. Within each cell, the weighted proportion of correct enumerations among resolved cases was computed before accounting for duplication with non-E-sample people, analogous to equation (6.2), and then assigned to each unresolved case in the cell.

As with residence status for P-sample people, a key factor in determining enumeration status was the E-sample person's match code. These codes can be found in Chapter 4. People were placed in match code groups accordingly in the following sequence: 1) People coded as potentially fictitious or said to be living elsewhere on Census Day (based on information collected during the follow-up operation) were placed in groups 11 and 12, respectively. 2) All other people included in the operation for targeted extended search were placed in group 10. See Chapter 5 for details. 3) People in the remainder of the E sample were then placed in the appropriate match code group, as defined in Table 6-8. Other characteristics used to define cells were the presence or absence of imputed characteristics (as was used to define cells for match status); whether the person was non-Hispanic White or any other race-ethnicity combination; and V3, as defined in the section on residence status.

There was an additional adjustment made to the enumeration probability of E-sample people as a result of duplication with persons subsampled out of the E-sample in large clusters. If the same identity was assigned to $u$ E-sample persons and $v$ persons who were subsampled out of the E sample, 1) one of the $u$ E-sample persons was selected during the person matching operation, and 2) the initial correct enumeration probability was multiplied by $u/(u + v)$ during the missing data activities, as it was not known which person was the "actual" E-sample person. The other $u-1$ E-sample persons were assigned a correct enumeration probability of 0.

Table 6-8. **Probabilities Assigned for Resolving Enumeration Status in the E Sample**

| Match code group | No imputed characteristics | | 1 or more imputed characteristics | |
|---|---|---|---|---|
| 1 = Matches needing follow-up | 0.977 | | 0.977 | |
| 2 = Possible matches | 0.968 | | 0.968 | |
| 3 = Partial household nonmatches | V3a* 0.871 | V3b* 0.974 | V3a* 0.908 | V3b* 0.960 |
| 4 = Whole household nonmatches where the housing unit matched; not conflicting households | Non-Hispanic White 0.965 | Others 0.974 | 0.958 | |
| 5 = Nonmatches from conflicting households; for housing units not in regular nonresponse follow-up | 0.975 | | 0.965 | |
| 6 = Nonmatches from conflicting households; housing units in regular nonresponse follow-up | 0.914 | | 0.926 | |
| 7 = Whole household nonmatches, where the housing unit did not match in housing unit matching | Non-Hispanic White 0.959 | Others 0.947 | 0.950 | |
| 8 = Resolved before follow-up | Non-Hispanic White 0.995 | Others 0.990 | 0.979 | |
| 9 = Insufficient information for matching | 0.000 | | | |
| 10 = Targeted extended search people | 0.928 | | 0.858 | |
| 11 = Potentially fictitious people | 0.058 | | 0.088 | |
| 12 = People said to be living elsewhere on Census Day | 0.229 | | 0.210 | |

* V3a = Group 3 Persons age 18-29 listed as child of reference person; V3b= All other group 3 persons

Figure 6-1.

## Adjustment for Noninterviews: An Example

Consider a block cluster with nine housing units, all having the same type of basic address, for example, all single-family homes, as depicted below.

| Housing unit | Weight | Actual situation | Status of (and information from) A.C.E. Interview | Census Day interview status | A.C.E. Interview Day interview status |
|---|---|---|---|---|---|
| 1 | 100 | Resident on 4/1/00 and at time of A.C.E. interview | Interviewed in A.C.E. | Interview | Interview |
| 2 | 100 | Resident on 4/1/00 and at time of A.C.E. interview | Neighbor (proxy) interviewed in A.C.E. | Interview | Interview |
| 3 | 100 | Resident on 4/1/00 and at time of A.C.E. interview | No one interviewed in A.C.E. | Noninterview | Noninterview |
| 4 | 100 | Vacant on 4/1/00, resident at time of A.C.E. interview | Interviewed in A.C.E., knows of 4/1/00 status | Vacant | Interview |
| 5 | 100 | Vacant on 4/1/00, resident at time of A.C.E. interview | Interviewed in A.C.E., no knowledge of 4/1/00 status | Noninterview | Interview |
| 6 | 100 | Vacant on 4/1/00, resident at time of A.C.E. interview | No one interviewed in A.C.E. | Noninterview | Noninterview |
| 7 | 100 | Resident on 4/1/00, vacant at time of A.C.E. interview | Information obtained from proxy | Interview | Vacant |
| 8 | 100 | Resident on 4/1/00, vacant at time of A.C.E. interview | No information on 4/1/00 status; Census staff determines vacant at time of A.C.E. | Noninterview | Vacant |
| 9 | 100 | Resident on 4/1/00, different resident at time of A.C.E. interview | Interviewed in A.C.E., knows of 4/1/00 status | Interview | Interview |

Note: In this noninterview cell (sample block cluster × type of basic address), people in interviewed housing units would have received the following noninterview adjustments:

a) To the person weights of nonmovers and outmovers, Census Day Noninterview adjustment = 800 / 400 = 2.
b) To the person weights of inmovers, A.C.E. Interview Day Noninterview adjustment = 700 / 500 = 1.4.

Figure 6-2.
**Imputation of Age in the P Sample**

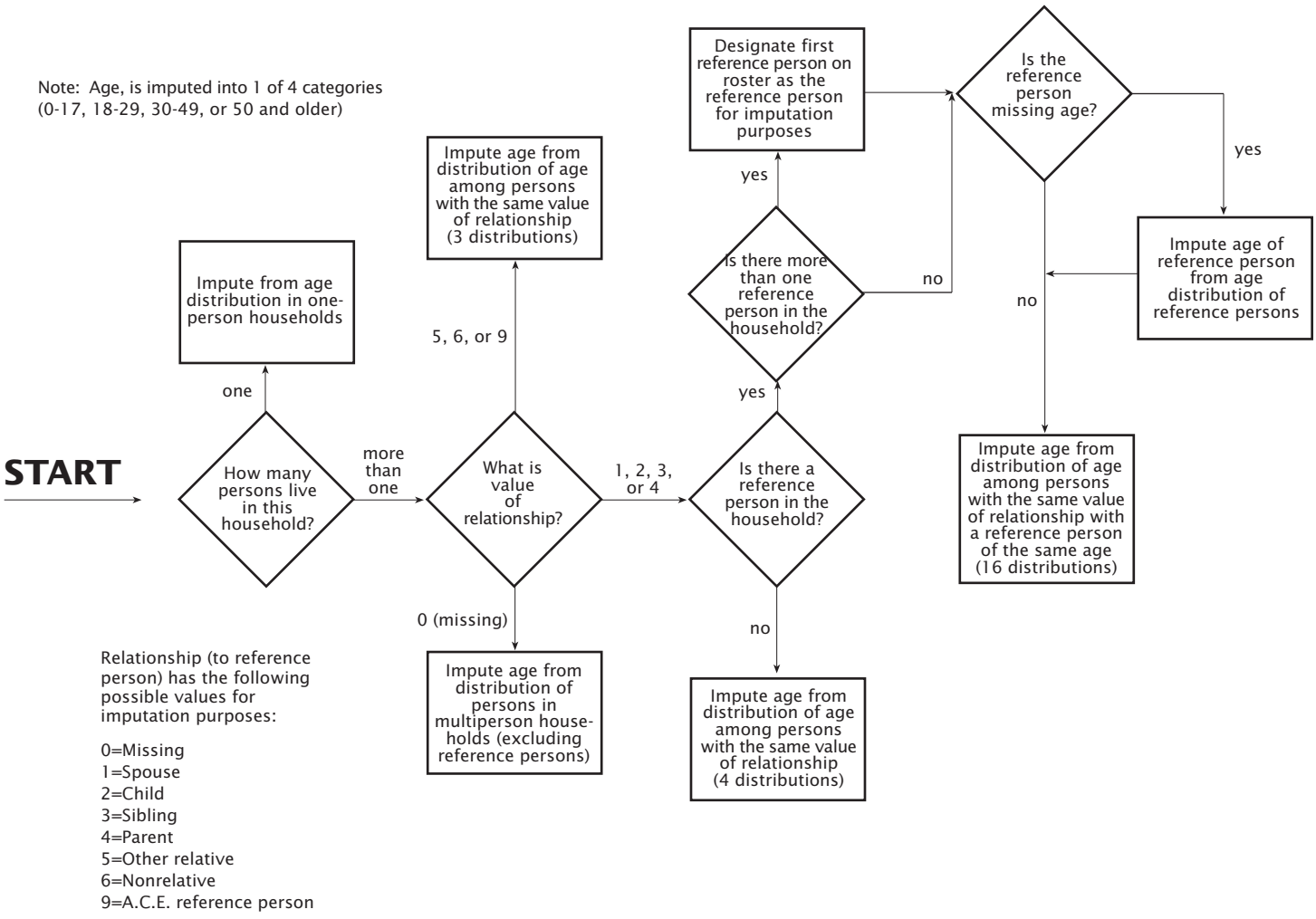
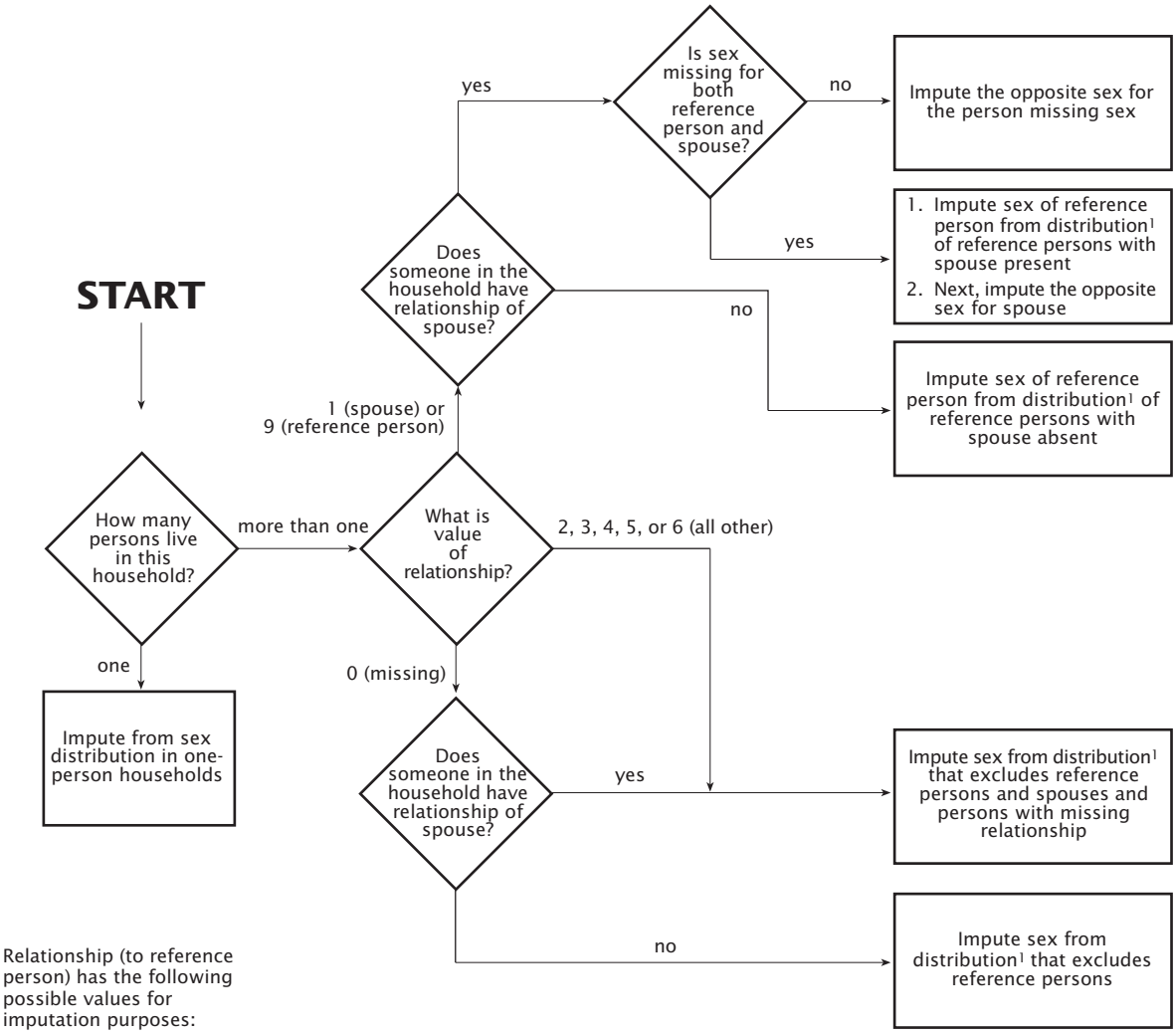


Figure 6-3.
**Imputation of Sex in the P Sample**




Relationship (to reference person) has the following possible values for imputation purposes:

0=Missing
1=Spouse
2=Child
3=Sibling
4=Parent
5=Other relative
6=Nonrelative
9=A.C.E. reference person

[1] This distribution only includes persons in multiperson households (i.e. exclude persons in one-person households)

Figure 6-4.
**Imputation of Tenure in the P Sample**

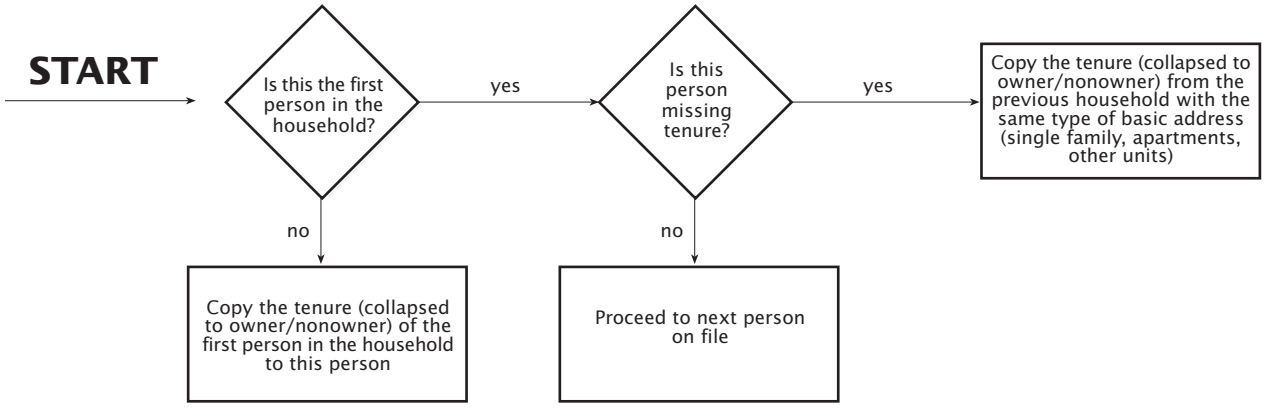


Figure 6-5.
**Imputation of Race in the P Sample**

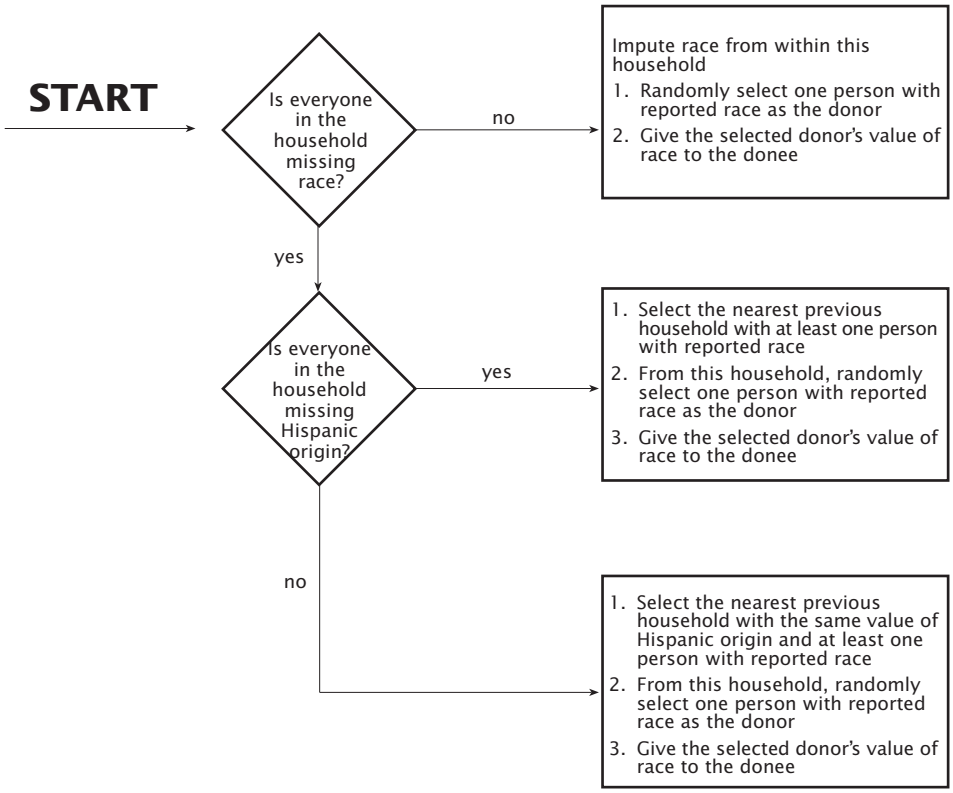
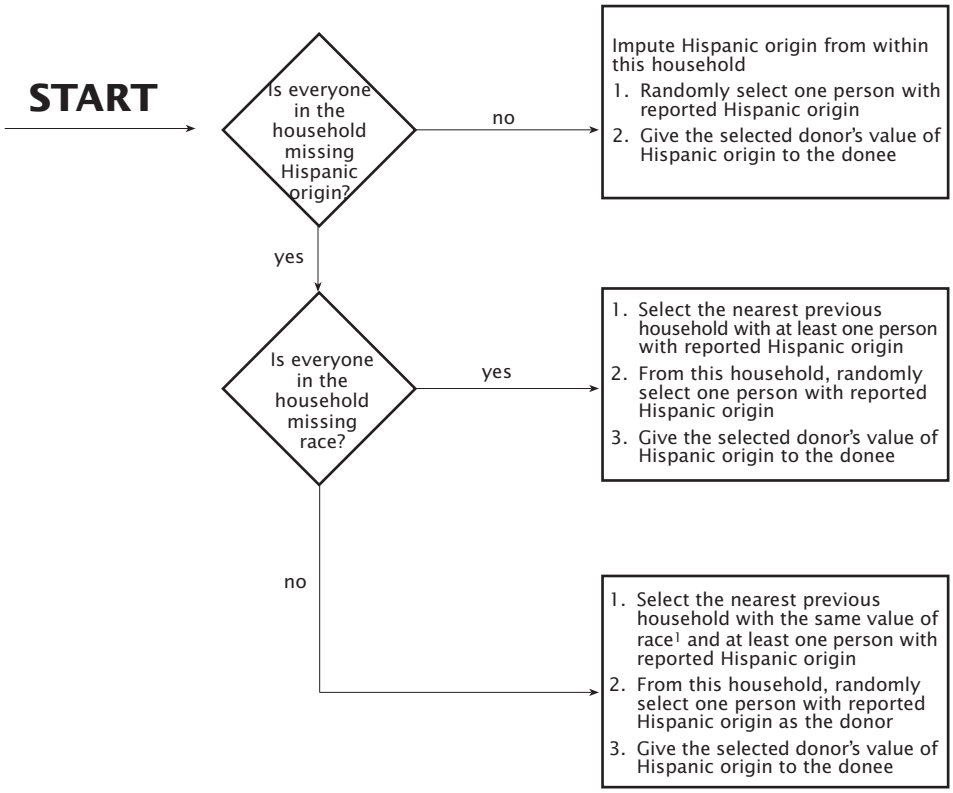


Figure 6-6.
**Imputation of Hispanic Origin in the P Sample**




[1] For imputing Hispanic origin, a household's race category, determined by the first person in the household with reported race, is coded as one of the following four values: 1) missing; 2) white; 3) 'other race,' or 'white' and 'other race'; or 4) any of the remaining race categories

# Chapter 7.
# Dual System Estimation

## INTRODUCTION

Dual System Estimation (DSE) was used to estimate coverage of Census 2000 using data from the Accuracy and Coverage Evaluation (A.C.E.) Survey. DSE was also used by the U.S. Census Bureau to estimate census coverage for the 1980 and 1990 censuses, and to evaluate coverage prior to 1980. The use of DSE for measurement of coverage in 1980 is described in Fay et al. (1988), while Hogan (1992,1993) describes the use of DSE in 1990. As described in Killion (1998), several alternatives to DSE were considered for Census 2000. These alternatives were either shown to produce results grossly inferior to DSE or research was not conclusive.

This chapter provides the details of DSE for the Census 2000 A.C.E. The DSE was calculated separately for a set of population groups referred to as post-strata. The post-stratification variables and the final post-stratification plan are discussed in detail. In addition, the variance estimation methodology used in each post-stratum is summarized and some basic results are given.

## DUAL SYSTEM ESTIMATION

This section contains the details of the DSE calculated within each final post-stratum. It describes the basic DSE model, including a discussion of the advantage of post-stratification. The details of the DSE computed within each final post-stratum for Census 2000 are presented. All components of the DSE are defined. The DSE accounted for special handling of missing data, search areas for matching, and movers. Missing data and search areas for matching are covered in detail in Chapters 6 and 5, respectively. The method used to handle special problems caused by movers in Census 2000 DSE is also discussed. The attachment provides detailed background on options for dealing with movers in census coverage measurement surveys. The section concludes with a short discussion of how the DSE results serve as input to synthetic estimation down to the block level. A detailed discussion of synthetic estimation is provided in Chapter 8 and Haines (2001).

### DSE Model

The DSE model is discussed in detail in Wolter (1986) and more generally in Hogan (1992). This chapter gives a general presentation. The DSE model (applied within each post-stratum) conceptualizes each person as having a probability of being either in or not in the census enumeration, as well as either in or not in the A.C.E.

Table 7-1. **DSE Model**

| | In census | Out of census | Total |
|---|---|---|---|
| In A.C.E. | $N_{11}$ | $N_{12}$ | $N_{1+}$ |
| Out of A.C.E. | $N_{21}$ | $N_{22}$ | $N_{2+}$ |
| Total | $N_{+1}$ | $N_{+2}$ | $N_{++}$ |

All cells are conceptually observable except $N_{22}$ and any of the marginal cells that include $N_{22}$ (i.e., $N_{2+}$, $N_{+2}$, and $N_{++}$). The model assumes independence between the census and the A.C.E. This means that the probability of being in the $ij^{th}$ cell, $p_{ij}$, is the product of the marginal probabilities, $p_{i+}p_{+j}$. The estimate of total population in a post-stratum with the independence assumption is

$$\mathrm{DSE} = N_{++} = \frac{(N_{+1})\,(N_{1+})}{N_{11}}.$$

The independence assumption can be in error, either due to causal dependence between the census enumeration and the A.C.E. enumeration, or due to heterogeneity in capture probabilities within a post-stratum. Causal dependence occurs when the event of an individual's inclusion or exclusion from one system affects his or her probability of inclusion in the other system. For example, some people who did answer the census may not have cooperated with the A.C.E., thinking they had helped enough. As another example, a person contacted during A.C.E. listing may not have responded to the census thinking that the A.C.E. lister already recorded them. However, even if causal independence is true for all individuals ($p_{ij} = p_{i+}p_{+j}$), the independence assumption can be violated by heterogeneity. Either the census inclusion probabilities $p_{+1}$ or the A.C.E. inclusion probabilities $p_{1+}$ must be the same for all individuals. This means that homogeneity in both systems is not required. For example, some people may try their best to avoid being counted in both the census and A.C.E., resulting in these people having much smaller inclusion probabilities than other people. Error in the independence assumption for either reason results in correlation bias.

Post-stratification, or grouping of individuals likely to have similar inclusion probabilities, and calculating DSEs within post-strata was done to decrease correlation bias. Research was carried out to determine effective variables

for the A.C.E. post-stratification design. All variables included in the 1990 PES post-stratification were considered as were several new ones. The specific variables considered were race/Hispanic origin, age/sex, tenure, household composition, relationship, urbanicity, percent owner, return rate, percent minority, type of enumeration area, household size, hard-to-count scores, census division, census region, and regional census center. From these variables, fifteen post-stratification options were developed for empirical research. For each post-stratification option, mean square errors of total population estimates and synthetic estimates were computed at the national, state, and congressional district levels, as well as for selected cities. The major conclusions were as follows:

- The demographic variables used in the 1990 PES were effective, but did not fully capture the geographic differences, especially those affected by the quality of the Master Address File. An urbanicity/type of enumeration variable appeared to capture much of the geographic differences.
- The tract-level return rate variable captured some of the socioeconomic differences for synthetic estimates at lower levels of aggregation.

Details of the Census 2000 post-stratification research methodology are given in Kostanich et al. (1999) and Griffin (1999). Results of this research are given in Griffin and Haines (2000) and Schindler (2000). The post-stratification design chosen for Census 2000 is provided in this chapter.

The DSE can be written as follows:

$$DSE = N_{+1}\left(\frac{N_{1+}}{N_{11}}\right)$$

That is, the total population is estimated by the number captured in the census times the ratio of those captured in the A.C.E. survey to those captured in both systems. In practice, the components of the DSE are estimated from a sample survey. $N_{+1}$ is not the census count; the census count (C) must be corrected for erroneous enumerations, as well as for persons enumerated in the census with insufficient information to match to the A.C.E. enumeration. To actually estimate the number of people correctly enumerated in the census, a sample of all data-defined persons is selected. This sample of data-defined census persons is called the enumeration or E sample. To estimate the ratio of those captured in both systems to those captured in A.C.E., the population or P sample is used. The P sample consists of persons interviewed during A.C.E. enumeration.

The form of the DSE used in census coverage measurement surveys such as A.C.E. is as follows:

$$DSE = DD \times \frac{CE}{N_e} \times \frac{N_p}{M}$$

where

DD = the number of census data-defined persons eligible and available for A.C.E. matching,
CE = the estimated number of correct enumerations from the E sample,
$N_e$ = the estimated number of people from the E sample,
$N_p$ = the estimated total population from the P sample,
M = the estimated number of persons from the P-sample population who match to the census.

Note: Persons in Group Quarters are excluded from all the above counts for A.C.E., as were persons in housing units who were added to the census after E sample Identification (late adds).

## Definitions

**Block Cluster.** A grouping of one or more census blocks. Block clusters are the primary sampling units for A.C.E. and average about 30 housing units each.

**Correct Enumeration (CE).** A correct enumeration is a person who is enumerated in a sample block cluster during the census who is also determined by A.C.E. operations to have lived in that block cluster (or if appropriate a surrounding block) on Census Day. Correct enumerations have a correct enumeration probability, $Pr_{ce,j,}$ equal to 1 for each person j.

**Correct Enumeration Probability ($Pr_{ce,j}$).** This is defined as the probability that person j in the E sample was correctly enumerated in the A.C.E. (or surrounding block) block cluster. The probability of correct enumeration is typically 0 or 1, but it can take on values within this range due to missing data imputation.

**Coverage Correction Factor (CCF).** The coverage correction factor for a post-stratum is calculated by dividing the DSE for that post-stratum by its census count. A.C.E. synthetic estimates for any data item for any geographic area are obtained by multiplying the coverage correction factor by the census count within each post-stratum, then summing over all post-strata (see Chapter 8 for details on synthetic estimation).

**Data-Defined Person.** This concept is defined for all census persons. A data-defined person is a person who has two or more of the 100-percent data items answered on the census form. Any items can be selected from the 100-percent data items, which include name, age, sex, race, and Hispanic origin. Relationship to person one is also a 100-percent data item for all persons besides person one. Persons not satisfying this criteria are referred to as non-data-defined.

**E Sample.** The E sample is the Enumeration sample. It consists of all data-defined persons in the A.C.E. block clusters who were enumerated in the census.

**Group Quarters (GQ) Persons.** Persons living in GQs, such as college dormitories, prisons, or military barracks. GQ persons were not covered in the A.C.E. and are excluded from the A.C.E. universe.

**Inmover.** A person who moved into a P-sample housing unit after Census Day.

**Insufficient Information in Census (II).** Those persons in the census for whom there is insufficient information for inclusion in the E sample. Very little data is available for these persons. This category includes non-data-defined persons and persons in whole household imputations. Note that insufficient information in census is different than insufficient information for matching. The former are excluded from the E sample and the latter are included in the E sample.

**Late Adds.** Late Adds are persons in housing units who were added to the census after E-sample Identification. These housing units had an unknown final status at the time of A.C.E. matching but were subsequently included in the census. Persons who are Late Adds were ineligible for matching and, therefore, not included in the census DSE component.

**Match Probability ($Pr_{m,j}$).** This is defined as the probability that person j in the P sample was matched to a census person in the search area (or in a TES block) . The match probability is typically 0 or 1, but it can take on values within this range due to missing data imputation.

**Mover Status.** Each person in the P sample was classified as a nonmover, outmover, or inmover.

**Nonmover.** An A.C.E. sample person whose housing unit on Census Day and A.C.E. Interview Day are identical.

**Outmover.** A person who moved out of an A.C.E. housing unit between Census Day and the date of the A.C.E. interview.

**P Sample.** Also known as the Population sample. The P sample consists of those persons confirmed to be residents of the housing units in the A.C.E. block clusters as of Census Day by the independent portion of the A.C.E. reinterview and subsequent operations.

**Residence Probability ($Pr_{res,j}$).** The probability that person j on the P-sample file is a resident of the sample household on Census Day. All inmovers are assumed to be A.C.E. Interview Day residents. Nonmovers and outmovers can be Census Day nonresidents, if information indicates they were not a resident of the sample household based on census residency rules. The residence probability is typically 0 or 1 but it can take on values within this range due to missing data imputation.

**Targeted Extended Search (TES).** A.C.E. operation in which block clusters are identified and selected for a search of the immediate surrounding area to find persons geographically mis-located in a block neighboring the A.C.E. block cluster. More generally, it is the methodology for targeting, sampling, and implementing the search operations in the field.

### DSE Formula

The DSE for any given post-stratum was calculated by:

$$\hat{\mathrm{DSE}} = \mathrm{DD}\left(\frac{\mathrm{CE}}{\mathrm{N}_e}\right)\left[\frac{\mathrm{N}_n + \mathrm{N}_i}{\left(\mathrm{M}_n + \left(\frac{\mathrm{M}_o}{\mathrm{N}_o}\right)\mathrm{N}_i\right)}\right]$$

All counts and estimates are for a specific post-stratum and the subscripts n, i, and o stand for nonmovers, inmovers, and outmovers, respectively. Adjustments to this DSE were occasionally made to avoid the unlikely event that the formula results in division by zero. For post-strata with less than ten (unweighted) outmover persons, the ratio inside the square brackets was changed to the following:

$$\frac{\mathrm{N}_n + \mathrm{N}_o}{\mathrm{M}_n + \mathrm{M}_o}.$$

### Coverage Correction Factor Formula

The coverage correction factor (CCF) is a measure of the net overcount or net undercount of the household population within the census. The CCF for a post-stratum is the ratio of the DSE to the census count:

$$\mathrm{CCF} = \frac{\mathrm{DSE}}{\mathrm{C}}$$

where

C = the final census household population count where C = DD + II + LA,

II = the number of census people with insufficient information,

LA = the number of people added (late) to the census and not available for A.C.E. matching. Late Adds include both data-defined and non-data-defined records.

Note: The numerator of the CCF is based on data-defined persons. The denominator includes data-defined and non-data-defined persons as well as late adds. Thus, we are implicitly assuming the coverage of late adds and non-data-defined persons is the same as that for data-defined persons. For example, a coverage correction factor of 1.05 would imply that for every 100 people within the given post-stratum, the net undercount is five persons.

### DSE Components

Each component of the DSE is described next.

DD is the census count (unweighted) of data-defined persons in the post-stratum.

The estimated number of E-sample persons is written as:

$$N_e = \sum_{j \in E\ sample} W_j^*$$

where $W_j^*$ = inverse of the probability of selection, including a factor for Targeted Extended Search sampling.

The estimated number of correct enumerations is calculated as:

$$CE = \sum_{j \in E\ sample} Pr_{ce,j} W_j^*$$

where $Pr_{ce,j}$ is:

1 if person j correctly enumerated,
0 if person j NOT correctly enumerated, or
$Pr^*_{ce,j}$ if person j is unresolved, where $Pr^*_{ce,j}$ is estimated through missing data imputation.

Note: Probabilities for persons with unresolved final correct enumeration status in the E sample or unresolved final residence or match status in the P sample are assigned using imputation cell estimation within groups. See Chapter 6 for details. Within each group, a probability equal to a simple proportion is imputed for unresolved persons. For example, E-sample (or P-sample) persons in a group with unresolved enumeration (match) status were assigned a correct enumeration (match) probability that is the proportion of correct enumerations (matches) among persons with resolved enumeration (match) status in the group. The probabilities are estimated in the DSE formulas as:

$Pr^*_{m,j}$ is the estimated match probability for unresolved match status
$Pr^*_{res,j}$ is the estimated residence probability for unresolved residence status
$Pr^*_{ce,j}$ is the estimated enumeration probability for unresolved enumeration status

Some persons moved between Census Day and A.C.E. Interview Day. A mover is a person whose location on the day of the A.C.E. interview differs from his or her location on Census Day. The treatment of movers has important ramifications for estimation. The attachment to this chapter titled "The Effect of Movers on Dual System Estimation" provides a discussion on alternative methodologies for handling movers. For Census 2000, movers were treated by a procedure known as Procedure C, unless a post-stratum had less than ten (unweighted) outmover persons. In this case, Procedure A was implemented. Procedure C identifies all current residents living or staying at the sample address at the time of the A.C.E. interview (nonmovers and inmovers), plus all other persons who lived at the sample address on Census Day who have since moved (outmovers). The P sample includes nonmovers and outmovers. For outmovers, the interviewers attempted a proxy interview to obtain data such as name, sex, and age that was used for matching. The match rate for inmovers was estimated by the match rate of outmovers. In contrast, the number of movers in the P sample for A.C.E. sample areas was estimated by the inmovers. Note that no matching was done for inmovers.

$N_n$ is the weighted total population for nonmovers for the post-stratum from the P sample. The weight for each person j is the product of three values:

1. the inverse of the P-sample selection probability including a factor for the Targeted Extended Search sampling ($W_j^*$),
2. a noninterview adjustment based on Census Day interview status ($f^*_{c,j}$), and
3. a Census Day residence probability ($Pr_{res,j}$).

The estimated number of P-sample nonmovers is calculated as:

$$N_n = \sum_{j \in Nonmovers} f^*_{c,j} Pr_{res,j} W_j^*$$

where, $Pr_{res,j}$ *is*:

1 if person j is a resident on Census Day,
0 if person j is NOT a resident on Census Day, or
$Pr^*_{res,j}$ if person j is unresolved, where $Pr^*_{res,j}$ is estimated through missing data imputation.

Note: Persons who were not residents on Census Day are not included in $N_n$ since $Pr_{res,j} = 0$ is a multiplicative factor in each person's contribution to $N_n$.

The estimated number of P-sample nonmover matches is written as:

$$M_n = \sum_{j \in Nonmovers} Pr_{m,j} f^*_{c,j} Pr_{res,j} W_j^*$$

where, $Pr_{m,j}$ is:

1 if person j is a match on Census Day,
0 if person j is NOT a match on Census Day, or
$Pr^*_{m,j}$ if person j is unresolved, where $Pr^*_{m,j}$ is estimated through missing data imputation

$N_i$ is the weighted total population for inmovers for the post-stratum from the P sample. The weight for each person j is the product of two values:

1. the inverse of the P-sample probability of selection ($W_j^*$ as defined above), and
2. a noninterview adjustment factor based on A.C.E. Interview Day status ($f^*_{a,j}$).

The estimated number of P-sample inmovers is denoted:

$$N_i = \sum_{j \in Inmovers} f^*_{a,j} W_j^*$$

Note that all inmovers are assumed to be A.C.E. Interview Day residents.

The estimated number of P-sample outmovers is written:

$$N_o = \sum_{j \in Outmovers} f^*_{c,j} Pr_{res,j} W_j^*$$

The estimated number of P-sample outmover matches is calculated as:

$$M_o = \sum_{j \in Outmovers} Pr_{m,j} f^*_{c,j} Pr_{res,j} W_j^*$$

### Synthetic Estimation

The estimated coverage correction factors for each post-stratum were used to form synthetic estimates. Synthetic estimation combines coverage error results with census counts at the block level to produce adjusted block-level population estimates. The synthetic methodology assumes coverage correction factors do not vary within a post-stratum. As a result, one coverage correction factor is assumed to be appropriate for all geographic areas within each post-stratum. To obtain block-level synthetic estimates, block-level census counts for post-strata are multiplied by post-stratum coverage correction factors and aggregated. There is one coverage correction factor for each post-stratum, and each person in a block is in one post-stratum. For example, suppose all persons in a block fall into one of six post-strata. A synthetic estimate for this block is formed by summing the product of census counts for that block and post-stratum with its corresponding coverage correction factor. A controlled rounding technique was implemented, resulting in the creation of person records at the block level. Subsequent tabulations, based on the original and replicated records, are corrected for coverage error. A detailed discussion of synthetic estimation is provided in Chapter 8 and Haines (2001).

## POST-STRATIFICATION

### Background

The goal of post-stratification for dual system estimation is to establish groups of persons who are expected to have similar coverage. A common assumption is that people who are subject to similar housing, language, education, and cultural attitudes would also share similar census coverage. Hogan (1993) indicated that tenure, race and ethnic origin, age/sex, and degree of urban development were reasonable markers for these similarities in the 1990 census. An earlier section noted, however, that the independence assumption of the DSE model can be in error due to heterogeneous capture probabilities within a post-stratum. Post-strata are formed to support DSE by grouping persons with similar census coverage, so as to reduce heterogeneity in capture probabilities for DSEs. In many surveys, post-stratification is done to reduce variances and partially correct for problems in sampling or undercoverage. For DSE, the primary reason for post-stratification is to reduce heterogeneity bias. Any variance reduction or sampling bias correction associated with post-stratification is a bonus. In fact, the usual trade-off is that forming many post-strata reduces heterogeneity at the expense of adding variance. As the number of post-strata increases, fewer people in the coverage measurement survey fall into each individual post-strata.

The post-stratification plan for Census 2000 A.C.E. is summarized in this section. Also, the detailed definitions of the post-stratification variables and the race/Hispanic origin domains are given. See Haines (2001b) for further details. The 2000 A.C.E. differs from the 1990 Post-Enumeration Survey (PES) in that it has approximately twice the sample size of the PES. This larger sample size permitted the formation of more post-strata that has the advantage of reducing correlation bias, as well as sampling variance. Additionally in 2000, multiple responses to the race question were permitted; in 1990 only one race could be selected.

The 1990 PES post-strata started with a cross-classification of seven variables: age, sex, race, Hispanic origin, tenure, urbanicity, and region. There were 840 cells in the cross-classification. Collapsing was necessary in order to produce post-strata with sufficient sample for reliable Dual System Estimation (DSE). The collapsing reduced the number of post-strata to 357.

Race and Hispanic origin were considered the most important variables to retain in 1990. After collapsing, five race/Hispanic origin post-strata were maintained: Non-Hispanic White or Other, Black, Hispanic White or Other, Asian and Pacific Islander, and Reservation Indians. Off-reservation American Indians were placed in either the Non-Hispanic White or Other group or the Hispanic White or Other group, depending on whether they were of Hispanic origin. Within each of these race/Hispanic origin post-strata, seven age/sex categories were maintained.

The other variables were collapsed in the following order: region, urbanicity, then tenure, if necessary. For American Indians residing on reservations, all these variables were collapsed. For Asian and Pacific Islanders, region and urbanicity were collapsed and tenure maintained. For the Black and Hispanic White or Other groups, region was collapsed for two levels of urbanicity. For Non-Hispanic White or Other, the full cross-classification of region, urbanicity and tenure were maintained. Griffin and Haines (2000b) provides a detailed table on the 1990 PES post-stratification.

### Post-Stratification Plan

The Census 2000 A.C.E. retained most of the 1990 PES post-stratification variables and included several additional ones. Nine variables were used in 2000: age, sex,

race, Hispanic origin, tenure, region, Metropolitan Statistical Area size/Type of Enumeration Area, and tract-level return rate. The Metropolitan Statistical Area size variable replaced the urbanicity variable that was not available until the summer of 2001. Type of Enumeration Area (TEA) and the tract return rate were two new features of the 2000 A.C.E. post-stratification. The mailout/mailback areas were differentiated from other types of enumeration areas. In addition, tracts were classified by high or low return rates. Multiple responses to the race question were reflected in the race and Hispanic origin groupings.

Table 7-2 shows the 64 post-stratum groups for the Census 2000 A.C.E. Within each post-stratum group, there are seven age/sex groups (shown in Table 7-3). Thus, there was a maximum of $64 \times 7 = 448$ post-strata. The P-sample size was too small or the sampling variance too high for eight of the 64 post-stratum groups. For each of these eight groups, the 7 age/sex post-strata were collapsed into 3 post-strata (under 18; males 18+ and females 18+). As a result, direct DSEs were calculated within each of 416 post-strata, which were expanded to 448 DSEs using synthetic estimation for the collapsed groups. The post-stratification plan was chosen to reduce correlation bias without having an adverse effect on the variance of the dual system estimator. Following is a detailed description of the post-stratification variables including an explanation of the race/Hispanic origin domain assignment

Table 7-2. **Census 2000 A.C.E. 64 Post-Stratum Groups (U.S.)**

| 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 | 01 | 02 | 03 | 04 | 05 | 06 | 07 | 08 |
| | | | Medium MSA MO/MB | 09 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| | | | Small MSA & Non-MSA MO/MB | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| | | | All other TEAs | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| | | Nonowner | Large MSA MO/MB | 33 | | | | 34 | | | |
| | | | Medium MSA MO/MB | 35 | | | | 36 | | | |
| | | | Small MSA & Non-MSA MO/MB | 37 | | | | 38 | | | |
| | | | All other TEAs | 39 | | | | 40 | | | |
| Domain 4 (Non-Hispanic Black) | | Owner | Large MSA MO/MB | 41 | | | | 42 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 43 | | | | 44 | | | |
| | | | All other TEAs | | | | | | | | |
| | | Nonowner | Large MSA MO/MB | 45 | | | | 46 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 47 | | | | 48 | | | |
| | | | All other TEAs | | | | | | | | |
| Domain 3 (Hispanic) | | Owner | Large MSA MO/MB | 49 | | | | 50 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 51 | | | | 52 | | | |
| | | | All other TEAs | | | | | | | | |
| | | Nonowner | Large MSA MO/MB | 53 | | | | 54 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 55 | | | | 56 | | | |
| | | | All other TEAs | | | | | | | | |
| Domain 5 (Native Hawaiian or Pacific Islander) | | Owner | | 57 | | | | | | | |
| | | Nonowner | | 58 | | | | | | | |
| Domain 6 (Non-Hispanic Asian) | | Owner | | 59 | | | | | | | |
| | | Nonowner | | 60 | | | | | | | |
| American Indian or Alaska Native | Domain 1 (On Reservation) | Owner | | 61 | | | | | | | |
| | | Nonowner | | 62 | | | | | | | |
| | Domain 2 (Off Reservation) | Owner | | 63 | | | | | | | |
| | | Nonowner | | 64 | | | | | | | |

*For Census 2000, persons can self-identify with more than one race group. For post-stratification purposes, persons are included in a single Race/Hispanic Origin Domain. This classification does not change a person's actual response. Further, all official tabulations are based on actual responses to the census.

Table 7-3. **Census 2000 A.C.E. Age/Sex Groups**

| | Male | Female |
|---|---|---|
| Under 18 | 1 | |
| 18 to 29 | 2 | 3 |
| 30 to 49 | 4 | 5 |
| 50+ | 6 | 7 |

### Post-stratification Variables

This section gives a detailed description of the post-stratification variables including the handling of multiple responses to the race question. A.C.E. post-stratification used the following variables:

- race/Hispanic origin - seven categories
- age/sex - seven categories
- tenure - two categories
- Metropolitan Statistical Area (MSA) by Type of Enumeration (TEA) - four categories
- return rate - two categories
- region - four categories

The seven race/Hispanic origin domains were:

- American Indian or Alaska Native on Reservations
- Off-Reservation American Indian or Alaska Native
- Hispanic
- Non-Hispanic Black
- Native Hawaiian or Pacific Islander
- Non-Hispanic Asian
- Non-Hispanic White or "Some other race"

Inclusion in a race/Hispanic origin domain is complicated, as it depends on several variables and whether there are multiple race responses. In addition, inclusion in a race/Hispanic origin domain does not change a person's race/Hispanic origin response. All Census 2000 tabulations are based on the actual responses. For example, a person who responded as American Indian on a reservation and Black was placed in the first race/Hispanic origin category (Domain 1) for post-stratification purposes, but was tabulated in the census as American Indian/Black.

The seven age/sex categories were:

1. Under 18
2. 18 - 29 male
3. 18 - 29 female
4. 30 - 49 male
5. 30 - 49 female
6. 50+ male
7. 50+ female

The two tenure categories were:

1. Owner
2. Nonowner

The four MSA/TEA categories were:

1. Large MSA Mailout/ Mailback (MO/MB)
2. Medium MSA MO/MB
3. Small MSA or Non-MSA MO/MB
4. All other TEAs

MSA/CMSA FIPS codes, as defined by the Office of Management and Budget, were used for post-stratification. For simplification, MSA/CMSA will herein be referred to as MSA. Large MSA consists of the ten largest MSAs based on unadjusted, Census 2000 total population counts including the population in Group Quarters. Medium MSAs are those (besides the largest 10) that have at least 500,000 total population. Small MSAs are those with a total population size less than 500,000. For post-stratification purposes, MO/MB areas were contrasted with the non-MO/MB areas.

The two return rate categories were:

1. High
2. Low

Return rate is a tract-level variable measuring the proportion of occupied housing units in the mailback universe that returned a census questionnaire. Low (high) return rate tracts are those tracts whose return rate is less than or equal to (greater than) the 25th percentile return rate. Separate 25th percentile cut-off values were formed for the six applicable race/Hispanic origin by tenure groups. Persons in List/Enumerate, Rural Update/Enumerate, and Urban Update/Enumerate TEAs were automatically placed in the High category.

The four region categories were:

1. Northeast
2. Midwest
3. South
4. West

### Pre-Collapsing

Pre-collapsing was done prior to data collection and knowledge of the exact sample size in each post-stratum. All race/Hispanic origin, age/sex, and tenure categories for the U.S. were initially maintained. The research for the determination of the important post-stratification variables

provided information on the expected sample size in each category which was then used to define a collapsing hierarchy. The pre-collapsing plan for the region, MSA/TEA and return rate variables was as follows:

- Non-Hispanic White or "Some other race" Owners: No collapsing.
- Non-Hispanic White or "Some other race" Non-owners: Region was eliminated.
- Non-Hispanic Black: Region was eliminated. In addition there was partial collapsing of the MSA/TEA variable within return rate and tenure categories.
- Hispanic: Region was eliminated. In addition there was partial collapsing of the MSA/TEA variable within return rate and tenure categories.
- Native Hawaiian or Pacific Islander: The region, return rate and MSA/TEA variables were eliminated. Only tenure and age/sex were retained.
- Non-Hispanic Asian: The region, return rate and MSA/TEA variables were eliminated. Only tenure and age/sex were retained.
- American Indian or Alaska Native on Reservations: The region, return rate and MSA/TEA variables were eliminated. Only tenure and age/sex were retained.
- Off-Reservation American Indian or Alaska Native: The region, return rate and MSA/TEA variables were eliminated. Only tenure and age/sex were retained.

### Post-Collapsing

A.C.E. post-stratification included a plan to collapse post-strata that contained less than 100 (unweighted) P-sample persons, called post-collapsing, considering such a post-stratum too small to produce reliable estimates. If a collapsed post-strata was still too small, it could have been further collapsed. The collapsing procedure was hierarchical and required a pre-defined collapsing order. Given the pre-collapsing plan that yielded 448 post-strata, not much post-collapsing was anticipated, but an extensive post-collapsing strategy was designed for completeness and to satisfy the requirement of pre-specification.

Note that collapsing does not necessarily imply elimination of a variable. Collapsing can refer to a reduction in the number of categories for a variable. The following general outline describes the post-collapsing hierarchy that was planned:

- If any of the 448 post-strata are too small, collapse age/sex first. This means that within any of the 64 U.S. post-stratum groups, if at least one of the seven age/sex categories defined in Table 7-3 has less than 100 P-sample persons, reduce age/sex to the following three categories: Under 18, 18+ male, and 18+ female.
- If some post-strata are still too small and require collapsing, collapse region next, if applicable. This collapsing applies only to the Non-Hispanic White or "Some other race" domain since the variable region is only included in their post-stratification definition. In this case, all levels of region (Northeast, Midwest, South, West) are combined to eliminate the variable.
- Next, collapse the four-level MSA/TEA variable, into the following two groups:
  - Large and medium MSA MO/MB
  - Small MSA and non-MSA MO/MB and all other TEAs
- If further collapsing is necessary, return rate is the next variable to collapse. High and Low return rate categories are combined to eliminate the variable.
- Next collapse the variable MSA/TEA. If necessary, the two groups defined above would be combined together to eliminate the variable MSA/TEA completely.
- The next variable to collapse is tenure. Owner and non-owner categories are combined to eliminate the variable entirely, if necessary.
- If collapsing is still needed, the three remaining age/sex post-strata are combined to eliminate the age/sex variable completely.
- In the event that there are less than 100 P-sample persons in a race/Hispanic origin domain, combine all persons in that domain with Domain 7, which includes non-Hispanic White and "Some other race."

In practice, only the first step of collapsing was necessary. Eight of the 64 post-stratum groups had their 7 age/sex post-strata collapsed to 3 age/sex groups, resulting in 32 fewer post-strata. Thus, there were 448 - 32 = 416 post-strata.

### Race and Hispanic Origin Classifications

The Census 2000 questionnaire has 15 possible race responses. The 15 responses are collapsed into six major race groups as shown below. Races that are included in the major groups are shown in parentheses. Persons self-identifying with a single race essentially place themselves into one of these six categories.

- White
- Black (Black, African American, Negro)
- American Indian or Alaska Native
- Asian (Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian)
- Native Hawaiian or Pacific Islander (Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander)

- "Some other race" (There was a box on the questionnaire labeled "Some other race - Print race" with a line to enter any race the respondent desired.)

For the first time in census history, persons were able to respond to more than one race category. Allowing persons to self-identify with multiple races results in many more than six race groups. In fact, after collapsing race to the six major groups, there are $2^6 - 1 = 63$ possible race combinations. It is necessary to subtract the 1 in this equation since each individual is assumed to have a race.

The race variable defined above is often cross-classified with the Hispanic origin variable to define post-strata. The Hispanic origin variable consists of two responses, No and Yes. Categories that are included in the Yes response are shown in parentheses.

1. No, not Spanish/Hispanic/Latino
2. Yes (Mexican, Mexican American, Chicano, Puerto Rican, Cuban, Other Spanish/Hispanic/Latino)

Combining the race and Hispanic origin variables yields $63 \times 2 = 126$ possible race/Hispanic origin groups. It is important to note that in a survey the size of A.C.E., no post-stratification plan of interest can support 126 race/Hispanic origin groups. Consequently, each of the 126 race/Hispanic origin response possibilities was assigned to one of seven race/Hispanic origin domains. The seven race/Hispanic origin domains are defined as follows:

1. American Indian or Alaska Native on Reservations
2. Off-Reservation American Indian or Alaska Native
3. Hispanic
4. Non-Hispanic Black
5. Native Hawaiian or Pacific Islander
6. Non-Hispanic Asian
7. Non-Hispanic White or "Some other race"

Note that missing race and Hispanic origin data are imputed. The rules used to classify the 126 race and Hispanic origin combinations into one of the seven race/Hispanic origin domains are now presented. Many of the decisions on how multiple race persons were classified are based on cultural, linguistic, and sociological factors, which are known to affect coverage and are not necessarily data-driven.

A hierarchy was used to assign persons to a race/Hispanic origin domain. The race/Hispanic origin designation occurs in the following order: American Indian or Alaska Native on Reservations, Off-Reservation American Indian or Alaska Native, Hispanic, Non-Hispanic Black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian, and Non-Hispanic White or "Some other race." This collapsing was only used for the post-stratification, all census data were tabulated in accordance with the race and Hispanic origin categories selected by census respondents.

For the following tables, Indian Country (IC) is a block-level variable that indicates whether a block is (wholly or partly) inside an American Indian reservation/trust land, Oklahoma Tribal Statistical Area (OTSA), Tribal Designated Statistical Area (TDSA), or Alaska Native Village Statistical Area (ANVSA).

Tables 7-4 and 7-5 display the assignment of race/Hispanic origin domains. Table 7-4 applies to Hispanic persons, while Table 7-5 applies to non-Hispanic persons. The first six rows of Tables 7-4 and 7-5 correspond to a single race response. The remaining portion of the tables address the assignment of multiple race responses to a single race/Hispanic origin domain. Although a person may be associated with multiple race responses, each person is included in only one of the seven race/Hispanic origin domains. All persons with a common number are assigned to the same race/Hispanic origin domain. The number for each race/Hispanic origin domain was assigned as follows:

**Domain 1 (Includes American Indian or Alaska Native on Reservations).** This domain includes any person living on a reservation marking American Indian or Alaska Native either as their single race or as one of many races, regardless of their Hispanic origin.

**Domain 2 (Includes Off-Reservation American Indian or Alaska Native).** This domain includes any person living in Indian Country, but not on a reservation who marks American Indian or Alaska Native either as a single race or as one of many races, regardless of their Hispanic origin. This domain also includes any Non-Hispanic person not living in Indian Country who marks American Indian or Alaska Native as a single race.

**Domain 3 (Includes Hispanic).** This domain includes all Hispanic persons who are not included in Domains 1 or 2. All Hispanic persons (excluding American Indian or Alaska Native in Indian Country) are included in Domain 3. The only exception to this rule occurs when a Hispanic person lives in the state of Hawaii and classifies himself or herself as Native Hawaiian or Pacific Islander, regardless of whether he or she identifies with a single or multiple race. All Hispanic persons satisfying this condition are re-classified into Domain 5.

**Domain 4 (Includes Non-Hispanic Black).** This domain includes any non-Hispanic person who marks Black as their only race. It also includes the combination of Black and American Indian or Alaska Native not in Indian Country. In addition, people who mark Black and another single race group (Native Hawaiian or Pacific Islander, Asian, White, or "Some other race") are included in Domain 4.

The only exception to this rule occurs when a NonHispanic Black person lives in the state of Hawaii and classifies himself or herself as Native Hawaiian or Pacific Islander. All Non-Hispanic Black persons satisfying this condition are reclassified into Domain 5.

**Domain 5 (Includes Native Hawaiian or Pacific Islander).** This domain includes any Non-Hispanic person marking the single race Native Hawaiian or Pacific Islander. For NonHispanic persons, it also includes the race combination of Native Hawaiian or Pacific Islander and American Indian or Alaska Native not in Indian Country. Also included is the race combination of Native Hawaiian or Pacific Islander with Asian for Non-Hispanic persons. All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race, are also included in Domain 5.

**Domain 6 (Includes Non-Hispanic Asian).** This domain includes any non-Hispanic person marking Asian as their single race. If a person self-identifies with Asian and American Indian or Alaska Native not in Indian Country, they are included in Domain 6.

**Domain 7 (Includes Non-Hispanic White or "Some other race").** Non-Hispanic White or Non-Hispanic "Some other race" persons are included in Domain 7. Non-Hispanic persons who self-identify with American Indian or Alaska Native not in Indian Country and are White or "Some other race" are classified into Domain 7. If a Native Hawaiian or Pacific Islander response is combined with a White or "Some other race" response, they also are included in Domain 7. A person who self-identifies with Asian and White or Asian and "Some other race" is also included in this domain. Finally, all Non-Hispanic persons who self-identify with three or more races (excluding American Indian or Alaska Native in Indian Country) are included in Domain 7. The only exception to this rule occurs when a Non-Hispanic White or Non-Hispanic "Some other race" person lives in Hawaii and classifies themselves as Native Hawaiian or Pacific Islander, regardless of whether they identify with other races. Persons who satisfy this criteria are re-classified into Domain 5.

Table 7-4. **Census 2000 A.C.E. Race/Origin Post-stratification Domains for Hispanic Indian country (IC)**

| | Not in IC | Indian country (IC) | |
|---|---|---|---|
| | | Not on reservation | On reservation |
| Single race: | | | |
| American Indian or Alaska Native | 3 | 2 | 1 |
| Black | 3 | 3 | 3 |
| Native Hawaiian or Pacific Islander | *3 | 3 | 3 |
| Asian | 3 | 3 | 3 |
| White | 3 | 3 | 3 |
| ''Some other race'' | 3 | 3 | 3 |
| American Indian or Alaska Native and: | | | |
| Black | 3 | 2 | 1 |
| Native Hawaiian or Pacific Islander | *3 | 2 | 1 |
| Asian | 3 | 2 | 1 |
| White | 3 | 2 | 1 |
| ''Some other race'' | 3 | 2 | 1 |
| Black and: | | | |
| Native Hawaiian or Pacific Islander | *3 | 3 | 3 |
| Asian | 3 | 3 | 3 |
| White | 3 | 3 | 3 |
| ''Some other race'' | 3 | 3 | 3 |
| Native Hawaiian or Pacific Islander and: | | | |
| Asian | *3 | 3 | 3 |
| White | *3 | 3 | 3 |
| ''Some other race'' | *3 | 3 | 3 |
| Asian and: | | | |
| White | 3 | 3 | 3 |
| ''Some other race'' | 3 | 3 | 3 |
| American Indian or Alaska Native and: | | | |
| Two or More Races | *3 | 2 | 1 |
| All Else** | *3 | 3 | 3 |

*All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race, are included in Domain 5, which includes Native Hawaiian or Pacific Islander.

**All Else encompasses all remaining combinations that exclude American Indian or Alaska Native.

Table 7-5. **Census 2000 A.C.E. Race/Origin Post-stratification Domains for Non-Hispanic**

| | Not in IC | Indian country (IC) | |
|---|---|---|---|
| | | Not on reservation | On reservation |
| Single race: | | | |
| American Indian or Alaska Native | 2 | 2 | 1 |
| Black | 4 | 4 | 4 |
| Native Hawaiian or Pacific Islander | 5 | 5 | 5 |
| Asian | 6 | 6 | 6 |
| White | 7 | 7 | 7 |
| ''Some other race'' | 7 | 7 | 7 |
| American Indian or Alaska Native and: | | | |
| Black | 4 | 2 | 1 |
| Native Hawaiian or Pacific Islander | 5 | 2 | 1 |
| Asian | 6 | 2 | 1 |
| White | 7 | 2 | 1 |
| ''Some other race'' | 7 | 2 | 1 |
| Black and: | | | |
| Native Hawaiian or Pacific Islander | *4 | 4 | 4 |
| Asian | 4 | 4 | 4 |
| White | 4 | 4 | 4 |
| ''Some other race'' | 4 | 4 | 4 |
| Native Hawaiian or Pacific Islander and: | | | |
| Asian | 5 | 5 | 5 |
| White | *7 | 7 | 7 |
| ''Some other race'' | *7 | 7 | 7 |
| Asian and: | | | |
| White | 7 | 7 | 7 |
| ''Some other race'' | 7 | 7 | 7 |
| American Indian or Alaska Native and: | | | |
| Two or More Races | *7 | 2 | 1 |
| All Else** | *7 | 7 | 7 |

*All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race, are included in Domain 5, which includes Native Hawaiian or Pacific Islander.

**All Else encompasses all remaining combinations which exclude American Indian or Alaska Native.

## VARIANCE ESTIMATION

The A.C.E. sample was considered a three-phase sample—the initial listing sample was the first phase; A.C.E. reduction and small block cluster subsampling was the second phase; and Targeted Extended Search (TES) was the third phase. Multiphase sampling differs from multistage in the following way. In a multistage design, the information needed to draw all stages of the sample is known before the sampling begins; in a multiphase design, the information needed to draw any phase of the sample is not available until the previous phase is completed. Because of the multiphase nature of the design (housing counts not available until after the first-phase listing), a new variance estimator needed to be developed. Full details are given in Starsinic and Kim (2001).

Our goal is to obtain a variance estimator for the Dual System Estimator (DSE), of the form:

$$\hat{\mathrm{DSE}} = \mathrm{DD}\left(\frac{\mathrm{CE}}{\mathrm{N}_e}\right)\left(\frac{\mathrm{N}_n + \mathrm{N}_i}{\mathrm{M}_n + \left(\frac{\mathrm{M}_o}{\mathrm{N}_o}\right)\mathrm{N}_i}\right) \quad (1)$$

where:

- DD = number of census data-defined persons
- CE = estimated number of A.C.E. E-sample correct enumerations
- $\mathrm{N}_e$ = estimated number of A.C.E. E-sample persons
- $\mathrm{N}_n$ = estimated number of A.C.E. P-sample nonmovers
- $\mathrm{N}_i$ = estimated number of A.C.E. P-sample inmovers
- $\mathrm{N}_o$ = estimated number of A.C.E. P-sample outmovers
- $\mathrm{M}_n$ = estimated number of A.C.E. P-sample nonmover matches
- $\mathrm{M}_o$ = estimated number of A.C.E. P-sample outmover matches

The DSE is computed separately for each post-stratum denoted by h. The national corrected population estimate is computed as:

$$\hat{\mathrm{T}}_{\mathrm{US}} = \sum_h \hat{\mathrm{DSE}}_h \quad (2)$$

There is no closed-form solution for the variance estimator, and the Taylor linearization variance estimator is very complex. That leaves replication methodology as the only practical variance estimator. Specifically, a stratified jackknife estimator was the type of replication method chosen for the implementation.

A jackknife estimator is calculated from a set of replicates where the number of replicates is equal to the number of observations (clusters in this case) in the sample. Each replicate represents what the DSE would have been had each particular cluster not been part of the sample. The overall variance is calculated by summing the squares of the differences between the replicate DSE and the whole-sample DSE.

The most important challenge for the Census 2000 A.C.E. variance estimation was the precise form for calculating the contribution of replicate DSEs to the variance estimator; in particular, new weights had to be calculated for replicates to represent the effect of removing the cluster whose replicate was being calculated. No previous results were directly applicable to the DSE, but a methodology was developed based on the work of Rao and Shao (1992). The remaining part of this section describes the precise formulas in detail. They require somewhat complex notation and mathematical steps.

### Detailed Methodology

A general estimator of a total is:

$$\mathrm{T}_y = \sum_i w_i\, y_i \quad (3)$$

The estimator for the *$j^{th}$ replicate is*

$$\mathrm{T}_y^{(i)} = \sum_i w_i^{(j)}\, y_i \quad (4)$$

where $y_i$ is the characteristic of interest, and $w_i^{(j)}$ is the replicate weight for the $i^{th}$ unit, which differs from the original weight in a prespecified subset of the observations. With these replicate estimators, a variance estimator can be constructed:

$$\hat{\mathrm{V}ar}(\mathrm{T}_y) = \sum_j c_j\,(\mathrm{T}_y^{(i)} - \mathrm{T}_y)^2 \quad (5)$$

Before continuing, we must set down some specific notation. Let $w_i$ be the first phase sampling weight, and let $y_i$ be the cluster-level total of any of the seven estimated components of the DSE (CE, $N_n$, etc.). Let A and $A_2$ indicate the first and second phase samples, respectively. Let $x_{ig}$=1 if unit i is in "group" (second phase stratum) g and zero otherwise. Let $n_h$ be the number of units selected in first-phase stratum h. Let $n_g$ be the number of units in stratum h that are also in group g, and let $r_g$ be the number of the $n_g$ units selected in the second phase. In all of the following equations, "j" will represent one cluster that is being dropped to calculate its associate replicate estimate $T^{(j)}$; "k" is one cluster other than the one being dropped.

For two-phase stratified sampling, there are two different point estimators, the Double Expansion Estimator (DEE)

$$\mathrm{DEE} = \sum_g \sum_{i\,\epsilon\,\mathrm{A}_2} \frac{n_g}{r_g}\, w_i\, x_{ig}\, y_i \quad (6)$$

and the Reweighted Expansion Estimator (REE)

$$\mathrm{REE} = \sum_g \sum_{i\,\epsilon\,\mathrm{A}_2} \left(\frac{\sum_{i\,\epsilon\,\mathrm{A}} w_i x_{ig}}{\sum_{i\,\epsilon\,\mathrm{A}_2} w_i x_{ig}}\right) w_i\, x_{ig}\, y_i \quad (7)$$

There is an established result by Rao and Shao (1992) which gives a replicate variance estimator for the REE under two-phase stratified sampling. Unfortunately, all the individual components of the DSE, such as $N_e$, the number of E-sample people, are DEE's. Taking a closer look at the DEE, however, suggested a procedure that could be applied.

$$n_g = \sum_{i \epsilon A} x_{ig},\ r_g = \sum_{i \epsilon A_2} x_{ig}$$

$$\text{DEE} = \sum_g \sum_{i \epsilon A_2} \frac{n_g}{r_g} w_i x_{ig} y_i = \sum_g \sum_{i \epsilon A_2} \left( \frac{\sum_{k \epsilon A} x_{kg}}{\sum_{k \epsilon A_2} x_{kg}} \right) w_i x_{ig} y_i$$

$$= \sum_g \sum_{i \epsilon A_2} \left( \frac{\sum_{\kappa \epsilon A} w_k x_{kg} w_k^{-1}}{\sum_{k \epsilon A_2} w_k x_{kg} w_k^{-1}} \right) w_i x_{ig} y_i \qquad (8)$$

The DEE has just been rewritten in a form that is quite similar to the REE. This suggests the following generalization:

$$T_{y2} = \sum_{i \epsilon A_2} \alpha_i y_i, \text{ where}$$

$$\alpha_i = \sum_g \left( \frac{\sum_{k \epsilon A} w_k x_{kg} q_k}{\sum_{k \epsilon A_2} w_k x_{kg} q_k} \right) w_i x_{ig} \qquad (9)$$

and where $q_j = 1$ for the REE and $w_i^{-1}$ for the DEE.

Replicates are then naturally written as:

$$T_{y_2}^{(j)} = \sum_{i \epsilon A_2} \alpha_i^{(j)} y_i, \text{ where}$$

$$\alpha_i^{(j)} = \sum_g \left( \frac{\sum_{k \epsilon A} w_k^{(j)} x_{kg} q_k}{\sum_{k \epsilon A_2} w_k^{(j)} x_{kg} q_k} \right) w_i^{(j)} x_{ig} \qquad (10)$$

When $q_j=1$ (i.e. the REE case), the replicate variance estimator of this generalized estimator, based on equation (5), is the same as the REE replicate variance estimator of Rao and Shao (1992).

## Application To a Three-Phase Dual System Estimator

Within any of the seven components of the DSE that are subject to sampling error (CE, $N_e$, $N_n$, $N_o$, $N_i$, $M_n$, and $M_o$), the cluster sums ($y_i$) can be broken down into two components: the total prior to any adjustments made by TES ($u_i$), and the additional total from the TES sample ($v_i$). This second piece can be further subdivided into TES totals from clusters sampled with certainty, and TES totals from clusters sampled systematically. The estimator (a DEE) of one of the components is

$$\hat{T}_{y3} = \sum_{i \epsilon A_2} \alpha_i u_i + \sum_{k=1}^{2} \sum_{i \epsilon A_2} \alpha_i t_k s_{ik} a_i v_i \qquad (11)$$

where $s_{ik}$ is the third phase stratum indicator ($s_{ik}=1$ if the cluster is selected with certainty, 0 otherwise; $s_{i2}=1-s_{i1}$, an indicator that the cluster is eligible to be selected systematically), $a_i$ is the third phase sample indicator ($a_i=1$ if the cluster is in $A_3$, 0 otherwise), and $t_k$, the TES conditional weight, is equal to

$$t_k = \frac{\sum_{i \epsilon A_2} s_{ik}}{\sum_{i \epsilon A_2} s_{ik} a_i} = \frac{\sum_{i \epsilon A_2} s_{ik}}{\sum_{i \epsilon A_3} s_{ik}} = \frac{\text{number of clusers selected in phase 2}}{\text{number of clusters selected in phase 3}} \qquad (12)$$

For $s_{i1}$, the certainty stratum, all clusters within it have $a_i=1$, so $t_k=1$ for all clusters in the stratum.

To create the replicate estimator, simply apply what was learned above in equations (8) and (10).

$$\hat{T}_{y3}^{(j)} = \sum_{i \epsilon A_2} \alpha_i^{(j)} u_i + \sum_{k=1}^{2} \sum_{i \epsilon A_2} \alpha_i^{(j)} t_k^{(j)} s_{ik} a_i v_i \qquad (13)$$

$$= \sum_{i \epsilon A_2} \alpha_i^{(j)} u_i + \sum_{i \epsilon A_2} \alpha_i^{(j)} t_l^{(j)} s_{il} a_i v_i + \sum_{i \epsilon A_2} \alpha_i^{(j)} t_2^{(j)} s_{i2} a_i v_i$$

where,

$$t_1^{(j)} \equiv 1$$

$$t_2^{(j)} = \frac{\sum_{i \epsilon A_2} \alpha_i^{(j)} s_{i2} \alpha_i^{-1}}{\sum_{i \epsilon A_2} \alpha_i^{(j)} s_{i2} a_i \alpha_i^{-1}}$$

## Implementation of Variance Estimation for the A.C.E.

The first step in implementing this variance estimation methodology is calculating the replicate weights. To this point, the method of replication used to arrive at the variance is immaterial, but we will now state that the jackknife will be used. Let the replicate weights after the first stage of sampling be the standard jackknife replicate weights

$$w_i^{(j)} = \begin{cases} 0 & \text{if i= j} \\ \dfrac{n_h}{n_h - 1} w_{hi} & \text{if i and j are in the same first phase stratum} \\ w_{hi} & \text{otherwise} \end{cases} \qquad (14)$$

Then, the final weights are obtained by applying equation (10).

Note that this is an unusual form of the jackknife. Normally, the jackknife has as many replicates as observations. Here, there are 11,303 clusters remaining after the second phase of the sample, but the number of replicates is equal to the first phase's sample size of 29,136 clusters. The clusters sampled out in the second phase obviously do not contribute to the variance due to the second and third phases, but they must be included to accurately

account for the first phase of sampling. "Deleting" a cluster that was sampled out changes the weights of the other clusters that were in the same first phase sampling stratum.

The second step of the implementation is to adjust the imputation of certain probabilities to account for the replication. This is a component of the variance that can be accounted for by including the effect of the replicate weights in the imputation. For some persons, their match, residence, or correct enumeration status remains unresolved even after follow-up operations. In these cases, a probability for each unresolved status is imputed using an imputation cell technique, with each unresolved case in an imputation cell getting the same imputed probability. The general form for the "replicated" imputation of the probability for an unresolved person in imputation cell k is:

$$\Pr_k^{*(j)} = \frac{\sum_{resolved\ p \in k} w_p^{*(j)} t_p^{*(j)} \Pr_p}{\sum_{resolved\ p \in k} w_p^{*(j)} t_p^{*(j)}} \qquad (15)$$

where the summation is over all resolved persons in imputation cell k, and:

$w_p^*$ = person-level weight for replicate j, incorporating all sampling operations except TES, and not including the noninterview adjustment

$t_p^{*(j)} = \begin{cases} \text{conditional TES weight for replicate j, the inverse of the probability of selection in the TES sample, if the person is a TES person} \\ 1 \text{ if the person is NOT a TES person} \end{cases}$

$\Pr_p =$

$\begin{cases} 1 \text{ if a person is a \{match/resident/correct enumeration\}} \\ 0 \text{ if a person is NOT a \{match/resident/correct enumeration\}} \end{cases}$

To complete the estimation of the variances, the 29,136 replicate dual system estimates were computed for each of the 448 post-strata:

$$\hat{\mathrm{DSE}}_h^{(j)} = (\mathrm{C} - \mathrm{II}) \left( \frac{\mathrm{CE}^{(j)}}{\mathrm{N}_e^{(j)}} \right) \left( \frac{\mathrm{N}_n^{(j)} + \mathrm{N}_i^{(j)}}{\mathrm{M}_n^{(j)} + \left( \frac{\mathrm{M}_o^{(j)}}{\mathrm{N}_o^{(j)}} \right) \mathrm{N}_i^{(j)}} \right) \qquad (16)$$

Equation (13) was used for the separate computation of each of the seven replicated terms of the DSE: $CE^{(j)}$, $N_e^{(j)}$, $N_n^{(j)}$, $N_i^{(j)}$, $N_o^{(j)}$, $M_n^{(j)}$, and $M_o^{(j)}$.

The variance estimates for post-stratum h used formula (5):

$$Var(\hat{\mathrm{DSE}}_h) = \sum_j \frac{n_{1,i} - 1}{n_{1,i}} (\hat{\mathrm{DSE}}_h^{(j)} - \hat{\mathrm{DSE}}_h)^2 \qquad (17)$$

finally, the variance of the national adjusted population estimate is:

$$Var(\hat{\mathrm{T}}_{\mathrm{US}}) = \sum_{post-stratum\ h} \sum_{post-stratum\ h'} Cov(\hat{\mathrm{DSE}}_h, \hat{\mathrm{DSE}}_{h'}),$$

where $Cov(\hat{\mathrm{DSE}}_h, \hat{\mathrm{DSE}}_h) = Var(\hat{\mathrm{DSE}}_h)$, and

$$Cov(\hat{\mathrm{DSE}}_h, \hat{\mathrm{DSE}}_{h'}) = \sum_j \frac{n_{1,i} - 1}{n_{1,i}} (\hat{\mathrm{DSE}}_h^{(j)} - \hat{\mathrm{DSE}}_h)(\hat{\mathrm{DSE}}_{h'}^{(j)} - \hat{\mathrm{DSE}}_{h'}) \qquad (18)$$

Covariances exist between post-strata mostly because of correlations between members of the same household being in different post-strata but having the same probability of being included in the sample. For instance, within a given race/Hispanic origin/tenure/region group there exists some covariance among males 30-49, females 30-49 and children 0-17, because such persons are likely to live in the same household, and hence, show very similar census and A.C.E. inclusion probabilities.

## RESULTS

The percent net undercount (UC) is the estimated net undercount (or net overcount) divided by the dual system estimate for a post-stratum expressed as a percentage. A positive number implies undercoverage, while a negative number implies overcoverage. The percent net undercount for Census 2000 shown in this document is strictly for the household population and excludes group quarters persons.

$$\mathrm{UC} = \left( \frac{\mathrm{DSE} - \mathrm{C}}{\mathrm{DSE}} \right) \times 100$$

Table 7-6 presents the estimated percent net undercount for each of the 64 post-stratum groups. Table 7-7 presents the standard error of each of these estimates. Many more results are available in Davis (2001).

Table 7-6. **Census 2000 A.C.E. 64 Post-Stratum Groups - Percent Net Undercount**

| 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 | 0.81 | 0.01 | 0.36 | -0.38 | -3.62 | -2.61 | 2.19 | 1.14 |
| | | Medium MSA MO/MB | 0.30 | -0.12 | 0.46 | -0.28 | -4.39 | -0.33 | 0.66 | 1.81 |
| | | Small MSA & Non-MSA MO/MB | -0.25 | 0.14 | 0.44 | 0.30 | 2.29 | 2.61 | 2.09 | 2.71 |
| | | All other TEAs | 1.84 | -1.11 | 1.34 | 0.85 | 0.56 | -0.16 | 0.15 | 1.59 |
| | Nonowner | Large MSA MO/MB | 1.82 | | | | 1.02 | | | |
| | | Medium MSA MO/MB | 0.61 | | | | 2.83 | | | |
| | | Small MSA & Non-MSA MO/MB | 2.45 | | | | 3.61 | | | |
| | | All other TEAs | 1.64 | | | | 4.08 | | | |
| Domain 4 (Non-Hispanic Black) | Owner | Large MSA MO/MB | 1.63 | | | | -1.31 | | | |
| | | Medium MSA MO/MB | | | | | | | | |
| | | Small MSA & Non-MSA MO/MB | 0.07 | | | | 0.46 | | | |
| | | All other TEAs | | | | | | | | |
| | Nonowner | Large MSA MO/MB | 4.18 | | | | 3.42 | | | |
| | | Medium MSA MO/MB | | | | | | | | |
| | | Small MSA & Non-MSA MO/MB | 2.64 | | | | 0.12 | | | |
| | | All other TEAs | | | | | | | | |
| Domain 3 (Hispanic) | Owner | Large MSA MO/MB | 1.46 | | | | 0.04 | | | |
| | | Medium MSA MO/MB | | | | | | | | |
| | | Small MSA & Non-MSA MO/MB | 1.66 | | | | 1.08 | | | |
| | | All other TEAs | | | | | | | | |
| | Nonowner | Large MSA MO/MB | 3.52 | | | | 4.98 | | | |
| | | Medium MSA MO/MB | | | | | | | | |
| | | Small MSA & Non-MSA MO/MB | 4.88 | | | | 10.74 | | | |
| | | All other TEAs | | | | | | | | |
| Domain 5 (Native Hawaiian or Pacific Islander) | Owner | | 2.71 | | | | | | | |
| | Nonowner | | 6.58 | | | | | | | |
| Domain 6 (Non-Hispanic Asian) | Owner | | 0.55 | | | | | | | |
| | Nonowner | | 1.58 | | | | | | | |
| American Indian or Alaska Native – Domain 1 (On Reservation) | Owner | | 5.04 | | | | | | | |
| | Nonowner | | 4.10 | | | | | | | |
| American Indian or Alaska Native – Domain 2 (Off Reservation) | Owner | | 1.60 | | | | | | | |
| | Nonowner | | 5.57 | | | | | | | |

*For Census 2000, persons can self-identify with more than one race group. For post-stratification purposes, persons are included in a single Race/Hispanic Origin Domain. This classification does not change a person's actual response. Further, all official tabulations are based on actual responses to the census.

A negative net undercount denotes a net overcount.

Table 7-7. **Census 2000 A.C.E. 64 Post-Stratum Groups - Standard Error of the Net Undercount in Percent**

| 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 | 0.43 | 0.36 | 0.87 | -0.45 | 1.05 | 1.43 | 1.54 | 2.09 |
| | | | Medium MSA MO/MB | 0.85 | -0.28 | 0.42 | 0.38 | 1.52 | 0.84 | 1.10 | 2.79 |
| | | | Small MSA & Non-MSA MO/MB | 1.33 | 0.40 | 0.43 | 0.57 | 3.60 | 2.12 | 1.08 | 1.49 |
| | | | All other TEAs | 1.06 | 0.39 | 0.97 | 1.66 | 2.17 | 1.21 | 0.65 | 1.89 |
| | | Nonowner | Large MSA MO/MB | 0.63 | | | | 1.01 | | | |
| | | | Medium MSA MO/MB | 0.71 | | | | 1.24 | | | |
| | | | Small MSA & Non-MSA MO/MB | 0.51 | | | | 1.24 | | | |
| | | | All other TEAs | 0.94 | | | | 1.67 | | | |
| Domain 4 (Non-Hispanic Black) | | Owner | Large MSA MO/MB | 0.56 | | | | 1.24 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 1.07 | | | | 1.86 | | | |
| | | | All other TEAs | | | | | | | | |
| | | Nonowner | Large MSA MO/MB | 0.66 | | | | 1.05 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 0.96 | | | | 2.08 | | | |
| | | | All other TEAs | | | | | | | | |
| Domain 3 (Hispanic) | | Owner | Large MSA MO/MB | 0.52 | | | | 1.26 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 1.01 | | | | 2.09 | | | |
| | | | All other TEAs | | | | | | | | |
| | | Nonowner | Large MSA MO/MB | 0.67 | | | | 1.12 | | | |
| | | | Medium MSA MO/MB | | | | | | | | |
| | | | Small MSA & Non-MSA MO/MB | 1.55 | | | | 4.12 | | | |
| | | | All other TEAs | | | | | | | | |
| Domain 5 (Native Hawaiian or Pacific Islander) | | Owner | | 3.83 | | | | | | | |
| | | Nonowner | | 4.07 | | | | | | | |
| Domain 6 (Non-Hispanic Asian) | | Owner | | 0.87 | | | | | | | |
| | | Nonowner | | 0.98 | | | | | | | |
| American Indian or Alaska Native | Domain 1 (On Reservation) | Owner | | 1.45 | | | | | | | |
| | | Nonowner | | 1.42 | | | | | | | |
| | Domain 2 (Off Reservation) | Owner | | 1.95 | | | | | | | |
| | | Nonowner | | 2.02 | | | | | | | |

*For Census 2000, persons can self-identify with more than one race group. For post-stratification purposes, persons are included in a single Race/Hispanic Origin Domain. This classification does not change a person's actual response. Further, all official tabulations are based on actual responses to the census.

A negative net undercount denotes a net overcount.

# Attachment.
# The Effect of Movers on Dual System Estimation

This attachment discusses the effect of movers on Dual System Estimation (DSE). Three alternative methodologies for handling movers in DSE have been considered by the U.S. Census Bureau. Historically, they are referred to as PES-A, PES-B, and PES-C. However, the current terminology is to refer to them as Procedures A, B, and C. Following are the definitions of these methodologies as described in U.S. Bureau of the Census (1985).

**Procedure A.** This procedure reconstructs the households as they existed at the time of the census. A respondent is asked to identify all persons who were living or staying in the sample household on Census Day. These persons are then matched against names on the census questionnaire for the sample address (and surrounding area). From this information, estimates of the number and percent matched for nonmovers and outmovers can be made.

**Procedure B.** This procedure identifies all current residents living or staying in the sample household at the time of the interview. The respondent is asked to provide the address(es) where these persons were living or staying on Census Day. These persons are then matched against names on corresponding census questionnaire(s) at the nonmover's or inmover's census address. Estimates of the number and percent matched for nonmovers and inmovers can be made.

**Procedure C.** This procedure identifies all current residents living or staying at the sample address at the time of the interview plus all other persons who lived at the sample address on Census Day and have moved since Census Day. However, only the Census Day residents (nonmovers and outmovers) are matched with the census questionnaire(s) at the sample address. Estimates of the number of nonmovers, outmovers, inmovers, and the percent matched for nonmovers and outmovers, can then be made. Estimates of nonmovers and movers come from Procedure B and match rate estimates for the movers from Procedure A (using outmover matching). Thus, Procedure C is a combination of Procedures A and B.

In 1990, Procedure B was used. The unresolved match rate for inmovers in 1990 was high, around 13 percent. In addition, with sampling for nonresponse initially planned for Census 2000, inmover matching would have had an even higher level of difficulty. A decision was made that Procedure B would NOT be used for Census 2000. When the Supreme Court decided against sampling for apportionment (no sampling for nonresponse), it was too late to change the decision on Procedure B.

In the 1995 and 1996 Census tests, Procedure A was used. The U.S. Census Bureau reasoned that an outmover match rate would be more accurate than an inmover match rate, particularly with sampling for nonresponse. For outmovers, interviewers attempted to obtain the names, new addresses and other data that could be used for matching from the new occupants or neighbors. Then an attempt could be made to trace the people to obtain an interview with a household member. The best available data for outmovers was matched to their Census Day addresses in the same manner as for the nonmovers.

Outmover tracing had problems in 1995 and was tested in 1996 and in the Census 2000 Dress Rehearsal. The outmover tracing evaluation by Raglin and Bean (1999) showed that there is little gain in an outmover tracing operation. A decision was made to use the outmover proxy interview data for outmover matching for Census 2000.

Procedure C was tested in the Census 2000 Dress Rehearsal and it was used in Census 2000 (Schindler, 1999). The advantage of Procedure C is that the estimate of the number of movers uses inmover data, which is more reliable since it is collected from the inmovers themselves. The match rate of the movers is estimated using the outmover match rate so that the difficulties of inmover matching are avoided. Outmover tracing is a problem, however, and in many cases it is necessary to use proxy data for matching. There was no outmover tracing for Census 2000. Procedure C attempts to obtain a Procedure B estimate with no inmover matching. Procedure C and Procedure B estimates are different since outmovers do not have the same match rate as inmovers. However, the disadvantage of the Procedure B inmover match rate estimate is that it may yield a high percentage of unresolved cases.

# Chapter 8.
# Model-Based Estimation for Small Areas

## INTRODUCTION

This chapter documents the Accuracy and Coverage Evaluation (A.C.E.) methodology of synthetic estimation for small areas including the estimation of sampling variances of synthetic estimates and the generalization of the variances. Synthetic estimation is the particular model used for coverage adjustment for small areas for A.C.E. First, the synthetic estimation methodology and the implied model are described. Then, the methodology for estimating sampling variances of these synthetic estimates and for generalizing these variances are discussed.

## SYNTHETIC ESTIMATION METHODOLOGY FOR SMALL AREAS

### Background

As discussed in Chapter 7, dual system estimates (DSE) and coverage correction factors were calculated at the post-stratum level. These are direct A.C.E. Survey estimates, based only on data from sample units in the post-stratum. However, census counts adjusted for coverage error are desirable for small geographic areas much smaller than any post-stratum such as blocks, tracts, counties and congressional districts. The adjusted counts were expected to improve data used for congressional redistricting as well as states, most metropolitan areas, and larger counties and cities and to provide consistent totals when census data are aggregated over many small areas. Many of these areas do not include any A.C.E. sample units, making a direct estimate impossible (see Chapter 3 for details of A.C.E. sampling). The geographic areas that include A.C.E. sample units only have a small number of sample units. A direct estimate would result in unacceptably large standard errors. Synthetic estimation is discussed in Ghosh and Rao (1994), Gonzalez (1973), and Gonzalez and Waksberg (1973). Gonzalez (1973) describes synthetic estimation as follows: " An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates." Synthetic estimation was first used by the National Center for Health Statistics (1968) to calculate state estimates of long and short term physical disabilities from the National Health Interview Survey data (Ghosh and Rao, 1994). Synthetic estimation is a useful procedure for small area estimation, mainly due to its simplicity and potential to increase accuracy in estimation by borrowing information from similar small areas.

Synthetic estimation was used for Census 2000 to provide adjusted population estimates for small geographic areas such as blocks, tracts, counties, and congressional districts. These block-level estimates can then be aggregated to any geographic level. The synthetic estimates provide revised population counts for both all persons and persons 18 and over. Counts are also provided for Hispanic or Latino persons by race (63 categories) and Not Hispanic or Latino persons by race (63 categories) for both the total population and the population 18 years and over. For example, counts of single-race Asian persons who are Not Hispanic or Latino are given for both the total population and the population 18 years and over. Counts of single-race Asians who are Not Hispanic or Latino who are less than 18 years of age can be obtained by subtraction. Synthetic estimates are formed by combining coverage measurement results with census counts to produce population estimates for any geographic area of interest. For example, a block-level synthetic estimate is formed by distributing a post-stratum's coverage correction factor to blocks proportional to the size of the post-stratum's population within the block. Rounded, adjusted synthetic estimates at the tabulation block level constitute the adjusted redistricting[1] data file.

The synthetic estimation model assumes that coverage correction factors are uniform within a given post-stratum, meaning that the coverage error rate for a given post-stratum is the same within all blocks. To the extent that the synthetic assumption is incorrect, the estimates of coverage for individual areas are biased and, hence, so are the population size estimates based on the coverage correction factors. Synthetic estimation bias decreases as the size of the geographic area increases.

### Synthetic Estimation

This section describes the calculation of synthetic estimates. Synthetic estimation includes a controlled rounding procedure used to produce estimates that are integer-valued. The visual representation of the twelve steps in the controlled rounding process given in Haines (2001) is provided here.

[1] Since it was originally intended that the A.C.E. might be used to adjust census counts for redistricting, such data is called "redistricting data," although it was not ultimately used for that purpose.

## Calculation

Consider forming synthetic estimates for geographic level g for a given post-stratum. Let $C_{i,g}$ denote the census count for post-stratum i in geographic level g and define $CCF_i$ to be the coverage correction factor for post-stratum i. The general form for a synthetic estimate for post-stratum i at geographic level g is calculated as

$$\hat{N}^{S}_{i,g} = C_{i,g} \times CCF_{i}.$$

Aggregating synthetic estimates over all the post-strata in geographic level g yields a synthetic estimate for the total population of geographic level g. This is denoted as

$$\hat{N}^{s}_{g} = \sum_{i} C_{i,g} \times CCF_{i}.$$

One purpose of synthetic estimation and the controlled rounding procedure is to produce integer-valued adjusted synthetic estimates at the tabulation block level. Then, summing over different geographies within a larger area yields the same estimate as that for the larger geographic area. These estimates comprise the adjusted redistricting data file.

## Geography

Components of synthetic estimates use two slightly different organizations of geography. Both collection and tabulation blocks are used in the synthetic estimation process. A collection block is a geographic area used during census data-collection activities. The Hundred-Percent Census Edited File (HCEF) is based on collection block geography. Tabulation blocks, on the other hand, are geographic areas used for tabulating census data. The Hundred-Percent Detail File (HDF) is based on tabulation block geography. Synthetic estimation census counts are based on tabulation block geography while the coverage correction factors associated with post-strata are based on collection block geography. This could have ramifications on variables with a geographic component, although any such effects are probably small.

For example, consider the post-stratification variable "return rate." Return rate was calculated at the tract level and based on collection-tract definitions. People were assigned to post-strata based on the return rate of tracts defined using collection blocks. Now consider the case where people are assigned to post-strata based on the return rate of tracts defined using tabulation blocks. It could be the case that the change in geography causes an individual's post-stratum assignment to change. For example, suppose the return rate of a collection-tract is 80 percent and that the collection tract is split into two pieces by a tabulation-tract. A person who belonged to the collection-tract (with an 80 percent return rate) may now belong to a tabulation-tract with a different return rate.

Changes in an individual's post-stratum would also cause changes in the dual system estimates, coverage correction factors, and synthetic estimates. To avoid potential inconsistencies in the assignment of people to post-strata, there was only one assignment of people to post-strata. The assignment was based on collection-block geography, which was consistent with the geography used in the A.C.E. Further, this post-stratification assignment was maintained for all estimation purposes.

## Controlled Rounding

Synthetic estimates at any geographic level are not typically integer-valued. A controlled rounding program, developed by the Statistical Research Division (SRD) of the U.S. Census Bureau, was utilized that produces integer-valued estimates. The theory of controlled rounding is given in Cox and Ernst (1982). The problem is represented as a transportation theory problem to minimize an objective function that measures the change due to controlled rounding. In essence, the controlled rounding program takes a two-dimensional matrix of numbers and rounds each to an adjacent integer value based on an efficiency algorithm. An optimal solution that minimizes the change due to controlled rounding is guaranteed; there can, however, be more than one optimal solution. The two dimensions of the matrix are: 1) the post-strata for one level of geography; and 2) totals for a lower level of geography. The controlled rounding procedure ensures that the sum of the synthetic estimates within a geographic level are rounded up or down by an amount strictly less than one person.

The overall goal of controlled rounding was to obtain an integer number of persons for each post-stratum i within each tabulation block b, reflecting the estimates of overcount and undercount. The controlled rounding program could not be implemented in one step due to the size of the post-strata by tabulation block matrix. As a result, controlled rounding was implemented in steps such that the rounded, adjusted synthetic estimates for blocks sum to:

- the rounded, adjusted synthetic estimates for tracts,
- the rounded, adjusted synthetic estimates for counties, and
- the rounded, adjusted synthetic estimates for states.

In other words, the block, tract and county rounded, adjusted synthetic estimates would all be consistent with each other. Also, the state-level synthetic estimates are adjusted in order to guarantee that total population estimates at the state level sum to the national total population estimate.

A controlled rounding procedure for the U.S. can be implemented as follows:

1. Form the ratio of the control-rounded dual system estimate ($DSE^R$) to the unrounded DSE for post-stratum i. It is written as

$$\frac{DSE_i^R}{DSE_i}$$

2. For each post-stratum i within state s, multiply the state-level synthetic estimate by the ratio formed in step 1. The superscript AS denotes an adjusted synthetic estimate. The resulting product is the adjusted synthetic estimate for post-stratum i, within state s written as

$$\hat{N}_{i,s}^{AS} = \hat{N}_{i,s}^{S}\,[\,DSE_i^R / DSE_i\,] \text{ where } \hat{N}_{i,s}^{S} = C_{i,s} \times CCF_i\,.$$

3. Apply the controlled rounding procedure to the adjusted state-level synthetic estimates to produce rounded, state-level synthetic estimates, denoted $\hat{N}_{i,s}^{RS}$. The superscript RS denotes a rounded, synthetic estimate. The two dimensions of this matrix are state s by post-stratum i.

| Post-stratum i | | | | | |
|---|---|---|---|---|---|
| State | 1 | 2 | .. | i | .. |
| 1 | | | | | |
| 2 | | | | | |
| : | | $\hat{N}_{i,s}^{AS}$ | | | |
| s | | | | | |
| : | | | | | |

→

| Post-stratum i | | | | | |
|---|---|---|---|---|---|
| State | 1 | 2 | .. | i | .. |
| 1 | | | | | |
| 2 | | | | | |
| : | | $\hat{N}_{i,s}^{RS}$ | | | |
| s | | | | | |
| : | | | | | |

4. Calculate the ratio of the rounded state-level synthetic estimate to the state-level synthetic estimate for post-stratum i in state s.

5. For each post-stratum i within county c for state s, multiply the county-level synthetic estimate by the ratio formed in step 4. The resulting product is the adjusted county-level synthetic estimate for post-stratum i, written as

$$\hat{N}_{i,c}^{AS} = \hat{N}_{i,c}^{S}\,[\hat{N}_{i,s}^{RS} / \hat{N}_{i,s}^{S}] \text{ where } \hat{N}_{i,c}^{S} = C_{i,c} \times CCF_i\,.$$

6. Apply the controlled rounding procedure to the adjusted county-level synthetic estimates to produce rounded, adjusted, county-level synthetic estimates, denoted $\hat{N}_{i,c}^{RS}$. The two dimensions of this matrix are county c (in state s) by post-stratum i (in state s).

| Post-stratum i in state s | | | | | |
|---|---|---|---|---|---|
| County | 1 | 2 | .. | i | .. |
| 1 | | | | | |
| 2 | | | | | |
| : | | $\hat{N}_{i,c}^{AS}$ | | | |
| c | | | | | |
| : | | | | | |

→

| Post-stratum i in state s | | | | | |
|---|---|---|---|---|---|
| County | 1 | 2 | .. | i | .. |
| 1 | | | | | |
| 2 | | | | | |
| : | | $\hat{N}_{i,c}^{RS}$ | | | |
| c | | | | | |
| : | | | | | |

7. Form the ratio of the rounded, adjusted, county-level synthetic estimate to the county-level synthetic estimate for post-stratum i in county c in state s.

8. For each post-stratum i within tract t in county c for state s, form the product of the tract-level synthetic estimate and the ratio formed in step 7. This results in the adjusted tract-level synthetic estimate for post-stratum i, written as

$$\hat{N}_{i,t}^{AS} = \hat{N}_{i,t}^{S}\,[\hat{N}_{i,c}^{RS} / \hat{N}_{i,c}^{S}] \text{ where } \hat{N}_{i,t}^{S} = C_{i,t} \times CCF_i\,.$$

9. Apply the controlled rounding procedure to the adjusted tract-level synthetic estimates to produce rounded, adjusted tract-level synthetic estimates, denoted $\hat{N}_{i,t}^{RS}$. The two dimensions of this matrix are tract t (in county c in state s) by post-stratum i (in county c in state s).

| Post-stratum i in county c in state s | | | | | |
|---|---|---|---|---|---|
| Tract | 1 | 2 | .. | i | .. |
| 1 | | | | | |
| 2 | | | | | |
| : | | $\hat{N}_{i,t}^{AS}$ | | | |
| t | | | | | |
| : | | | | | |

→

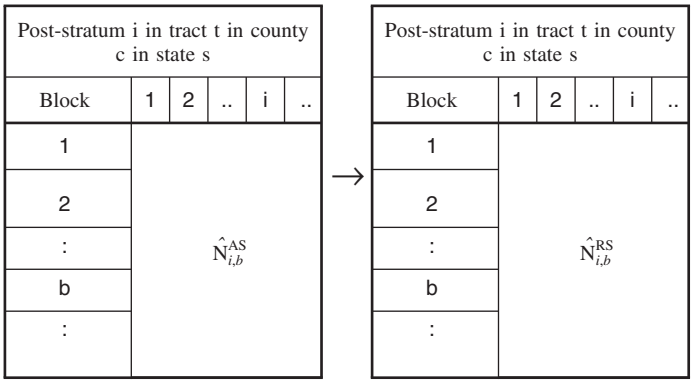
| Post-stratum i in county c in state s | | | | | |
|---|---|---|---|---|---|
| Tract | 1 | 2 | .. | i | .. |
| 1 | | | | | |
| 2 | | | | | |
| : | | $\hat{N}_{i,t}^{RS}$ | | | |
| t | | | | | |
| : | | | | | |

10. Calculate the ratio of the rounded, adjusted tract-level synthetic estimate to the tract-level synthetic estimate for post-stratum i in tract t in county c in state s.

11. For each post-stratum i within block b in tract t in county c for state s, multiply the block-level synthetic estimate by the ratio formed in step 10. The resulting product is the adjusted block-level synthetic estimate for post-stratum i, written as

$$\hat{N}^{AS}_{i,b} = \hat{N}^{S}_{i,b}\,[\hat{N}^{RS}_{i,t} / \hat{N}^{S}_{i,t}] \text{ where } \hat{N}^{S}_{i,b} = C_{i,b} \times CCF_i\,.$$

12. Again, apply the controlled rounding procedure to the adjusted block-level synthetic estimates to produce rounded, adjusted block-level synthetic estimates, denoted $\hat{N}^{RS}_{i,b}$. The two dimensions of this matrix are block b in tract t in county c in state s by post-stratum i in tract t in county c in state s.




### Record Replication for Coverage Correction

Once the rounded, adjusted block-level synthetic estimates were formed, they were compared with the census counts for post-stratum i in tabulation block b. Person records were then replicated at the post-stratum level to reflect the coverage correction for the census blocks. No attempt was made to place these persons in households. Thus, for example, the number of persons per household does not change due to coverage correction. The number of records replicated depends on the value of the coverage correction factor that is reflected in the rounded synthetic estimate for post-stratum i and tabulation block b.

### Coverage Correction Factors ≥1

When the coverage correction factor for post-stratum i was greater than one, undercount person records were replicated to reflect the undercount in post-stratum i and block b as follows:

$$U_{i,b} = \hat{N}^{RS}_{i,b} - C_{i,b}$$

If $U_{i,b} = 0$, then no additional records were necessary. If $U_{i,b} > 0$, then we replicated $U_{i,b}$ undercount person records for post-stratum i in tabulation block b.

Undercount person records were replicated by randomly selecting without replacement $U_{i,b}$ records from the $C_{i,b}$ available person records in post-stratum i and tabulation block b. The selected records were replicated and appended to the file of person records. The undercount person record for each of the replicated records was given an effective weight of +1 for tabulations. This resulted in an upward adjustment of people in post-stratum i in tabulation block b.

### Coverage Correction Factors <1

When the coverage correction factor for post-stratum i was less than one, overcount person records were replicated to reflect the overcount in post-stratum i and block b as follows:

$$O_{i,b} = C_{i,b} - \hat{N}^{RS}_{i,b}$$

If $O_{i,b} > 0$, then $O_{i,b}$ overcount person records were replicated for post-stratum i in tabulation block b.

Overcount person records were replicated by randomly selecting without replacement $O_{i,b}$ records from the $C_{i,b}$ available person records in post-stratum i and tabulation block b. The selected records were replicated and appended to the file of person records. The overcount person record for each of the replicated records was given an effective weight of -1 for tabulations, resulting in a downward adjustment of people in post-stratum i in tabulation block b.

## VARIANCE ESTIMATION FOR SMALL AREAS

Estimating the error due to sampling for any published estimate is a policy of the Census Bureau. This policy applies to synthetic estimates as well as the more traditional estimates. Due to the large number of estimates at lower levels of geography, it is not feasible to provide tables listing the standard error of each published estimate. Instead, a parameter, the generalized coefficient of variation (GCV), is provided, that allows users to approximate the standard error for any desired estimate. The coefficient of variation of an estimate is simply the ratio of the estimate's standard error to the estimate itself.

Small area variance estimation is a two-step process. The first step consists of producing direct variance estimates for the synthetic count estimates for small areas such as census tracts. This process is explained under the heading Direct Variance Estimates. The second step is to model the direct variance estimates using the generalized coefficient of variation, or GCV. This method is explained under the heading Generalized Variance Estimates, along with an example.

Variances calculated for small areas do not account for all sources of synthetic error; they only reflect variations due to sampling. Synthetic population bias can exist since the same coverage correction factors are applied to areas with different net census coverage. See Griffin and Malec (2001) for details on estimating synthetic bias. In most very small geographic areas such as blocks and tracts, the

biases are likely to be the principal source of errors. Sampling errors dominate the total error for larger areas such as states, metro area, etc. Bias in the post-stratum-level dual system estimates can stem from matching bias, data collection errors, and correlation bias, among other sources. Bell (2001) investigates and estimates correlation bias in the A.C.E. dual system estimates by comparing them to results from Demographic Analysis.

### Direct Variance Estimates

During the post-stratum-level A.C.E. variance estimation operation, a variance-covariance matrix of the A.C.E. coverage correction factors (CCFs) was produced. The estimated variance of any synthetic population estimate can be computed using this matrix and the unadjusted census counts, broken down by post-stratum and excluding out-of-scope persons in the A.C.E. See Starsinic (2001) for details. A synthetic household population estimate (Group Quarters persons are not included) for tract t is written as

$$\hat{X}_t = \sum_{\text{post-strata } h} \hat{X}_{th} = \sum_{h=1}^{416} C_{th} \times CCF_{h'}$$

where

$C_{th}$ is the final, unadjusted census count for post-stratum h in tract t.

There were 416 post-strata used to estimate coverage. The variance for the synthetic household population estimate $\hat{X}_t$ is

$$\begin{aligned} Var(\hat{X}_t) &= Var\left(\sum_{h=1}^{416} \hat{X}_{th}\right) \\ &= \sum_{h=1}^{416}\sum_{h'=1}^{416} Cov(\hat{X}_{th}, \hat{X}_{th'}) \\ &= \sum_{h=1}^{416}\sum_{h'=1}^{416} Cov(C_{th} \times CCF_h, C_{th'} \times CCF_{h'}) \\ &= \sum_{h=1}^{416}\sum_{h'=1}^{416} C_{th} \times C_{th'} \times Cov(CCF_h, CCF_{h'}). \end{aligned}$$

For a given data item j in tract t, the synthetic variance for the synthetic household population estimate $\hat{X}_{jt}$ is expressed as

$$Var(\hat{X}_{jt}) = \sum_{h=1}^{416}\sum_{h'=1}^{416} C_{jth} \times C_{jth'} \times Cov(CCF_h, CCF_{h'}), \qquad (1)$$

where

$C_{jth}$ is the final, unadjusted census count for data item j in post-stratum h in tract t.

Here h and h′ refer to particular post-strata and j refers to a data item.

### Generalized Variance Estimates

The generalized coefficient of variation (GCV) is the variance estimation methodology used for estimating variances of adjusted redistricting data and for estimates of adjusted population counts for the thousands of geographic areas that can be tabulated using synthetic estimation. For a given count in a particular state, the coefficient of variation (CV) was calculated for all tracts in that state that had population in the particular demographic category. The CV of an estimate is estimated as the ratio of the standard error of the estimate to the estimate itself, i.e.

$$CV(\hat{X}) = \frac{SE(\hat{X})}{\hat{X}}.$$

The standard error in the numerator is the square root of the variance estimate from (1). Tracts composed entirely of persons out-of-scope for the A.C.E. sample had no sampling variance (and therefore a CV of 0) and were removed from the processing. Also removed were tracts with a very small population in the demographic category, as these were shown in the Census 2000 Dress Rehearsal analysis to have a disproportionate downward effect on the parameters. The process of removing tracts was controlled to prevent removing an overly large fraction of "small" tracts for any adjusted demographic data item. In addition, outliers were identified using the relative absolute deviation (RAD) statistic for each data item j. Tracts with a RAD value above the cutoff value were removed and a new GCV was computed using CVs of remaining tracts. There were four iterations of identifying and removing outliers. Of the 286 unique demographic categories, GCVs were calculated for the 50 states and the District of Columbia for each of the 56 largest categories and 4 additional "catch-all" groups.

The average of the direct CVs for data items in a state is a GCV parameter. The state-level GCV parameters can then be used to estimate the standard error of a data item for all geographic areas within that state. Consider the following table of GCV parameters for a given state.

## State Parameters for Calculating the Standard Error of A.C.E.–Adjusted Data

| Demographic category | All persons | | Not Hispanic or Latino | |
|---|---|---|---|---|
| | All ages | 18 and over | All ages | 18 and over |
| | GCV | GCV | GCV | GCV |
| All persons | 0.0063 | 0.0067 | 0.0066 | 0.0069 |
| Hispanic or Latino | 0.0106 | 0.0115 | X | X |
| Population of one race | 0.0064 | 0.0067 | 0.0066 | 0.0069 |
| White alone | 0.0073 | 0.0077 | 0.0081 | 0.0083 |
| Black or African American alone | 0.0073 | 0.0083 | 0.0073 | 0.0083 |
| American Indian and Alaska Native alone | 0.0143 | 0.0147 | 0.0188 | 0.0190 |
| Asian alone | 0.0080 | 0.0085 | 0.0081 | 0.0086 |
| Native Hawaiian and Other Pacific Islander alone | 0.0391 | 0.0495 | 0.0507 | 0.0545 |
| Some Other Race alone | 0.0109 | 0.0119 | 0.0126 | 0.0139 |
| Population of two or more races | 0.0070 | 0.0077 | 0.0071 | 0.0082 |
| Population of two races | 0.0071 | 0.0078 | 0.0071 | 0.0082 |
| White; Black or African American | 0.0103 | 0.0156 | 0.0103 | 0.0157 |
| White; American Indian and Alaska Native | 0.0088 | 0.0092 | 0.0096 | 0.0100 |
| White; Asian | 0.0116 | 0.0131 | 0.0120 | 0.0133 |
| Black or African American; American Indian and Alaska Native | 0.0129 | 0.0140 | 0.0128 | 0.0140 |
| Asian; Native Hawaiian and Other Pacific Islander | 0.0524 | 0.0560 | 0.0530 | 0.0566 |
| All other combinations of two or more races | 0.0088 | 0.0095 | 0.0088 | 0.0099 |

Suppose a data user is interested in calculating the standard error of the population estimate of all Asians in a given county. The data user would locate the GCV parameter that corresponds to the "Asian alone" demographic category and the "All persons, All ages" classification in the appropriate state table. For the table above, the GCV parameter is 0.0080. Now assume that the population estimate for all Asians in this county is 370 people. Users are instructed to use the formula

$$\mathrm{SE}(\hat{X}) = \mathrm{GCV} \times \hat{X},$$

to calculate the estimated synthetic standard error, yielding $0.0080 \times 370 = 2.96$, or about 3 people in this example. Similar calculations can be done for any geographic level and demographic category.

# Appendix A.
# Census 2000 Missing Data

## INTRODUCTION

The Census Bureau used imputation in the 2000 Decennial Census, as it has in prior censuses, to address the problem of missing, incomplete, or contradictory data, an inevitable aspect of censuses and surveys. It is impossible not to have missing data in an endeavor as massive and complex as a decennial census. In Census 2000, the Census Bureau processed data for over 120 million households, including over 147 million paper questionnaires and 1.5 billion pages of printed material. In the 2000 Census, the various situations that resulted in missing data included incomplete or unavailable responses from housing units with previously confirmed addresses, conflicting data about the same housing unit, and failures in the data-capture process. The various types of missing data included characteristic data (information about an enumerated person, such as sex, race, age), population count data (information about the number of occupants in an identified housing unit), and housing unit status data (whether the unit is vacant, occupied, or nonexistent). The 2000 Census used two primary types of imputation.

1. The first type, called "count imputation," is imputation of the number of occupants of a housing unit. Count imputation applies when the Census Bureau is unable to secure any information regarding a given address, or when the Census Bureau has limited information about the address and does not have definitive information on the number of occupants.
2. The second type, called "characteristic imputation," is imputation that supplies missing characteristic data for a housing unit's response, but does not involve the number of occupants for a housing unit. For example, if a given housing unit did not provide ages for the individuals living in the housing unit, but supplied all other information, age would be imputed for the individuals in that housing unit. Sometimes the household size is known for the housing unit; however, none of the characteristics about the people are known. In this case all of the person's characteristics are imputed.[1]

This appendix summarizes the methods used to impute these types of missing data in the census. Some summary statistics showing the degree of imputation for these categories is given in the last section of this appendix.

[1]This does not include geographic characteristics such as location, urban or rural residency etc., which are generally known for all households.

## BACKGROUND

The census data collection activities started around mid-March of 2000, through the mail or directly using census-enumerators. From June to September, census staff conducted nonresponse follow-up (NRFU) and coverage improvement follow-up (CIFU) operations to revisit addresses for which census reports were not completed, i.e. did not respond to mailout/mailback or early enumeration operations. Based on the results of these operations, the Census Bureau was able to designate more than 99.5 percent of housing unit records as occupied, vacant, or nonexistent housing unit addresses. To designate an address as vacant or nonexistent required at least two independent census operations. This was to ensure complete census coverage. The nonexistent housing units were addresses of places used only for nonresidential purposes, or places that were uninhabitable and were not included in the census counts.

To permit the production of census population counts, it was necessary for each census address to have a status of occupied/vacant/nonexistent and a household size if occupied. To permit the production of the redistricting file and other more detailed census products, it was also necessary to have information about each person such as age, race, and sex. The count imputation covered status for housing units with undetermined status and household size for occupied units with an unknown number of occupants. The characteristic imputation was used to fill in the missing person data.

Census housing units identified in the Accuracy and Coverage Evaluation ( A.C.E.) block clusters were defined as the E-sample housing units. Persons residing in these housing units were E-sample persons. It took several different census operations to establish a list of census housing unit records and a list of census person records. One of these operations was the creation of a Hundred Percent Census Unedited File (HCUF). At the housing unit level, all housing units designated as occupied or vacant through data collection or through imputation were included in the HCUF. The file was used as a source file to identify the E-sample housing units for the A.C.E. operations. At the person level, the HCUF was used as a source file for person matching between the census and the A.C.E. (However, this does not include imputed persons, since they were not sent to A.C.E. matching). Chapter 3 provides detailed information on E-sample identification, while Chapter 4 provides information on person matching.

Persons imputed to an occupied unit with an unknown number of occupants or persons with all their characteristics imputed were considered as non-data-defined persons in the person Dual System Estimation (DSE). For data-defined persons, characteristic imputation filled in census missing data, such as sex, age, ethnicity, and owner /renter status for person DSE poststratification purposes.

### COUNT IMPUTATION

The Census Bureau used count imputation for three categories of cases in Census 2000.

1. **Household size imputation.** The Census Bureau imputed the number of occupants for a housing unit when Census Bureau records indicated that the housing unit was occupied, but did not show the number of individuals residing in the unit.
2. **Occupancy imputation.** When Census Bureau records indicated that a housing unit existed, but not whether it was occupied or vacant, the Census Bureau imputed occupancy status (occupied or vacant). If the unit was imputed to be occupied, the household size was also imputed.
3. **Status imputation.** When the Census Bureau's records had conflicting or insufficient information about whether an address represented a valid, nonduplicated housing unit, the Census Bureau first imputed for the status of the unit (occupied, vacant, nonexistent), then, if occupied, the household size was imputed.

#### Methodology

The Census Bureau used the nearest-neighbor hot deck imputation methodology to perform the count imputation. "Nearest" was defined by the geographical closeness of housing units. Group quarters addresses were included in the measure of distance, although not otherwise involved in count imputation. Census geographical identifiers, such as tract number, block number, or map spot number, along with street name, house number or apartment number were used to describe geographical proximity of housing records. To properly assign status and number of occupants to the housing units requiring imputation, limited donor pools and expanded donor pools were developed for each imputation category, which were further subdivided by type of structure.

All cases with missing status, occupancy or household size went through intensive follow-up operations to reduce the amount of imputation as much as possible. This was the main purpose of the NRFU and CIFU for mailout/mailback areas and enumerator visit in list/enumerate or update/enumerate areas. To properly represent these cases (donees), the primary donor pools were also housing units from NRFU, CIFU or from other enumerator visited cases. In the design phase, the Census Bureau did develop a standby procedure to include all enumerations in an expanded donor pool. With 99.5 percent of housing units having status and household size information available from data collection activities, the expanded donor pools were never used. The chart below characterizes the relationship between donees and the primary donor pool by imputation category.

### Donors and Donees by Imputation Category

| Imputation categories | Donees | Donor pool |
|---|---|---|
| Household size imputation:<br>a. Single units<br>b. Multiunits | Occupied with unknown household population | Occupied units with known population (in NRFU, or CIFU, or from list/enumerate or update/enumerate areas) |
| Occupancy imputation:<br>a. Single units<br>b. Multiunits | Units known to be either occupied or vacant | Occupied units with known population or vacant units (in NRFU, or CIFU, or from list/enumerate or update/enumerate areas) |
| Status Imputation:<br>a. Single units<br>b. Multiunits | Units with no status information | Occupied units with known population, vacant units, or nonexistent units (in NRFU, or CIFU, or from list/enumerate or update/enumerate areas) |

In general, type of structure (multi or single), type of enumeration (mail or list/enumerate), and final stage of the data collection for a housing unit (initial collection, NRFU, or CIFU) determined whether a housing record could be used as a primary donor. Each available donor could only be used once. Most of the time, the nearest potential donor was selected as the donor. Occasionally, a second nearest neighbor was designated as the donor, because the nearest donor had been taken by some previously processed donee. Whenever possible, the donor and donee were to be in the same tract, or in the same multiunit if the donee was located in a multiunit building.

To identify the nearest donor, a search was conducted in both directions: forward and backward. Using the donee as a reference point, potential donors surrounding the donee record were searched, and the donor housing unit geographically closest to the donee housing unit was determined. The search was done separately for single units and multiunits.

## CHARACTERISTIC IMPUTATION

Characteristic imputation was the process of filling in missing person characteristics, which include sex, age/date of birth, relationship, Hispanic origin and race. The Census Bureau used characteristic imputation for three categories of cases in Census 2000.

1. **Whole household imputation.** The Census Bureau imputed all of the characteristics for all of the persons in the household when the household record did not contain any data defined persons. To be data-defined, a person record must contain two or more of the 100-percent population data items, or a name.
2. **Within household imputation.** The Census Bureau imputed all the 100-percent characteristics for any non-data-defined persons in the household when the household record contained at least one data-defined person.
3. **Within Person Imputation.** Sometimes some of the 100-percent characteristic data for data-defined persons were missing and were imputed.

### Methodology

The categories of characteristic imputation employ different methodologies. For whole household imputations, the process replicates all of the 100-percent person data items by substituting data from a hot deck nearest neighbor donor pool record of the same household size. This process is sometimes referred to as substitution, since it assigns all the characteristics for all of the persons in the selected donor household to the household requiring imputation. This substitution process is also used to obtain the person characteristics for those housing units that were imputed as occupied or had their household size imputed during the count imputation process. By definition these households do not contain any data-defined persons. However, the majority of whole household imputations occur for cases where a census response on household size was obtained.

For within household imputations as well as within person imputations, the process allocates missing values for individual person characteristic data items on the basis of other reported information for other persons in the household, or from other persons in households with similar characteristics.

## RESULTS

This section briefly summarizes the overall level of imputation for people whose 100-percent characteristics were totally imputed in Census 2000 (within person imputations are excluded) for the U.S. population residing in housing units.

### Census 2000 Housing Unit Persons by Imputation Category

(Excludes within person imputations)

| | Number of persons | Percent of total persons |
|---|---|---|
| Total housing unit population | 273,643,272 | 100.00 |
| 100-percent characteristic imputation not required | 267,869,007 | 97.89 |
| 100-percent characteristic imputation required | 5,774,266 | 2.11 |
| Count imputations: | 1,172,144 | 0.43 |
| Household size | 495,600 | 0.18 |
| Occupancy | 260,652 | 0.10 |
| Status | 415,892 | 0.15 |
| Characteristic imputations | 4,602,122 | 1.68 |
| Whole household[1] | 2,269,010 | 0.83 |
| Within household | 2,333,112 | 0.85 |

[1]The count imputation cases (also requiring characteristic imputation) are not included in this figure to avoid duplication.

About 2 percent of persons residing in housing units required imputations of all 100-person characteristics. The majority of these cases , about 1.7 percent, occurred in situations where a census response on household size was obtained. Less than a half of a percent were situations where household size or the status of the housing unit was unknown.

# Appendix B. Demographic Analysis

## INTRODUCTION

The Census Bureau has used Demographic Analysis (DA) to measure population coverage, trends between censuses, and differences in coverage by age, sex, and race (Black, non-Black) at the national level in every census since 1960 (Siegel and Zelnik (1966), Siegel (1974), Fay et al. (1988), and Robinson et al. (1993)). DA produces estimates of the U.S. population through the use of data from administrative records and other noncensus sources. It has documented both the long-term reduction in the census net undercount rate and the persistent and disproportionate undercount of certain demographic groups, such as Black men. One goal of Census 2000 was to reduce these differential undercounts, which has been a continuing effort for the last several censuses.

The independence from the census and internal consistency of the DA estimation process allows us to compare the results with the survey-based Accuracy and Coverage Evaluation (A.C.E.) coverage estimates; in particular, the consistency of the age-sex results can be assessed. DA and A.C.E. use entirely different methodologies. Because the sources and patterns of errors in the two estimates are sufficiently different, any disagreement in the results can shed light on both the quality of the census and potential problems in methodology in the A.C.E. or the DA. Because of data limitations, DA estimates and comparisons are only possible at the national level and for certain large demographic groups. A further discussion of DA limitations is found in the section "Limitations of DA Estimates" of this appendix.

The U.S. Census Bureau released two sets of DA results as part of its evaluation of Census 2000 and the A.C.E. All DA results in this section are from the revised values released in October 2001. See Robinson (2001) for details.

## DESCRIPTION OF THE DEMOGRAPHIC ANALYSIS METHOD

Demographic Analysis represents a macro-level approach for measuring coverage. Estimates of net undercount are obtained by comparing census counts to independent estimates of the population derived from other measures (mostly administrative data). In general, DA population estimates are developed for the census date by combining various types of demographic data that are independent of the census and are highly reliable, such as administrative statistics on births, deaths, and Medicare data and estimates of immigration and emigration. The difference between the DA estimated population (P) and the census count (C) provides an estimate of the net census undercount (u). Dividing the net undercount by the DA benchmark provides an estimate of the net undercount rate (r):

$$u = \mathrm{P} - \mathrm{C}$$

$$r = (u/\mathrm{P}) \times 100$$

The particular analytic procedure used to estimate coverage nationally for the various demographic subgroups depends primarily on the nature and availability of the required demographic data. Two different demographic techniques were used to produce the demographic analysis estimates for 2000, one for the population under age 65 and another for the population 65 and over.

**Ages under 65.** The Demographic Analysis estimates for the population below age 65 are based on the compilation of historical estimates of the components of population change: births since 1935 (B), deaths to persons born since 1935 (D), immigrants born since 1935 (I), and emigrants born since 1935 (E). Presuming that the components are accurately measured, the population estimates ($P_{0\text{-}64}$) are derived by the basic demographic accounting equation applied to each birth cohort:

$$\mathrm{P}_{0-64} = \mathrm{B} - \mathrm{D} + \mathrm{I} - \mathrm{E}$$

The size of the component estimates used to develop the DA population under age 65 for 2000 is shown in Table B-1:

Table B-1. **DA Estimates of the Components of Change for the U.S. Resident Population: April 1, 2000**

| Component | Estimate |
|---|---|
| Total population | 281,759,858 |
| Under age 65 in 2000 | |
| + Births since 1935 (B) | 234,860,298 |
| - Deaths to persons born since 1935 (D) | 14,766,736 |
| + Immigration of persons born since 1935 (I) | 32,563,971 |
| - Emigration of persons born since 1935 (E) | 5,485,117 |
| Ages 65 and over in 2000 | |
| Medicare-based population | 34,587,440 |

Clearly, births (234.9 million) represent by far the largest component. The immigration component (32.6 million) is second largest, followed by deaths (14.8 million) and emigrants (5.5 million).

The actual calculations are carried out for single-year birth cohorts. For example, the estimate of the population age 40 on April 1, 2000 is based on births from April 1959 to March 1960 (adjusted for under-registration), reduced by deaths to the cohort in each year between 1959 and 2000, and incremented by estimated immigration and emigration of the cohort over the 40-year period.

The components for births and deaths are compiled principally from vital statistics records augmented by correction factors. The immigration component is estimated from its subcomponents:

Table B-2. **DA Estimates of the Components of Immigration for the U.S. Resident Population Under 65 Years of Age: April 1, 2000**

| Component | Estimate |
|---|---|
| Legally admitted permanent residents | 20,332,038 |
| Other measured migration | 2,249,001 |
| Migrants from Puerto Rico | 905,698 |
| Temporary migrants | 776,002 |
| Civilian citizen migration | 891,940 |
| Armed Forces overseas | -324,639 |
| Residual foreign-born migration (includes unauthorized migrants) | 9,982,932 |

**Age 65 and over.** Administrative data on aggregate Medicare enrollments are used to estimate the population age 65 and over ($P_{65+}$):

$$P_{65+} = M + m$$

where M is the aggregate Medicare enrollment and m is the estimate of underenrollment in Medicare. The DA population 65 and over is based on 2000 Medicare enrollments. Medicare is an administrative data set from the Health Care Financing Administration. Although Medicare enrollment is generally presumed to be quite complete, adjustments are made to the basic data to account for individuals who are omitted. An allowance is made for the estimated 1.3 million not enrolled (3.9 percent). Underenrollment factors are based on survey estimates of Medicare coverage and data on age at enrollment in the Medicare file. The DA population aged 65 and over (34.6 million) represents 12.3 percent of the total population in 2000.

The demographic component estimates for the population under 65 are combined with the Medicare-based estimate for the population 65 and over to produce the total DA population estimate of 281.8 million as of April 1, 2000.

## LIMITATIONS OF DA ESTIMATES

DA estimates for the total population are available only at the national level and only for the broad categories Black and non-Black. DA cannot provide estimates for subnational geographic areas like states or metropolitan areas; or for other demographic groups, such as Hispanics. DA also cannot provide separate estimates for census overcoverage and undercoverage, but is limited to estimating net undercount.

There are also certain inherent limitations on DA estimates because of data quality. The race categories reflect the race as assigned at the time of the event (e.g. birth or Medicare enrollment), which for some persons will differ from the race reported in the census. There is also considerable uncertainty in the quality of the data for some of the components related to immigration, most importantly the components which capture those who entered illegally or temporarily, or whose legal status had not yet been determined.

## DA ESTIMATES

Compared to the Census 2000 count of 281.4 million, the DA estimate of 281.8 million implies a net census undercount of 0.12 percent (see Table B-3). The net census undercount in 2000 was dramatically different from that in 1990, which was 4.2 million, or 1.65 percent. However, the fact that DA provides only a net undercount estimate, not separate measures of gross undercount and overcount, is a limitation on its ability to shed light on specific undercoverage or overcoverage problems in the census.

Table B-3. **Demographic Analysis Estimate and Net Census Undercount for the Total Population: 1990 and 2000**

| Category | 1990 Census | 2000 Census |
|---|---|---|
| DA (millions) | 252.9 | 281.8 |
| Difference from Census | 4.2 | 0.3 |
| Percent difference | 1.65 | 0.12 |

The DA estimates indicate that the substantial reduction in net census undercount from 1990 to 2000 was shared by almost all demographic groups. The net census undercount of males and females each fell by about 1.5 percentage points (to an estimated net census undercount of 0.86 percent for males and estimated net census overcount of 0.60 percent for females in 2000). The estimated net undercount rate dropped more for Blacks (estimated net census undercount of 2.78 percent in 2000) than non-Blacks (estimated net census overcount of 0.29 percent in 2000), reducing the differential undercount of Blacks relative to non-Blacks from 4.4 percentage points in 1990 to 3.1 points in 2000.

Table B-4. **Demographic Analysis Estimates of Percent Net Census Undercount for the Total Population and Selected Demographic Groups: 1990 and 2000**

| Category | 1990 DA | 2000 DA |
|---|---|---|
| Total | 1.65 | 0.12 |
| Male | 2.39 | 0.86 |
| Female | 0.93 | -0.60 |
| Black | 5.52 | 2.78 |
| Non-Black | 1.08 | -0.29 |
| Black male, ages 20-64 | 11.31 | 8.44 |
| Children, ages 0-4 | 3.72 | 3.84 |

(a minus sign denotes a net overcount)

# Appendix C. Weight Trimming

## INTRODUCTION

This appendix contains a general overview of the Accuracy and Coverage Evaluation (A.C.E.) weight trimming plan. The procedure was designed to protect against undue influence from a small fraction of the sample. The weight trimming criteria were established prior to the completion of data processing operations to ensure that there was no manipulation of the dual system estimates. This procedure was implemented according to the pre-specified criteria. Since only one cluster was trimmed, the impact on the dual system estimates was very minimal.

The A.C.E. weight trimming procedure was designed to reduce the sampling weights for clusters that potentially could have had an extreme influence on the dual system estimates and variances. The measure of cluster influence was the net cluster error, the absolute difference between the weighted estimate of nonmatches and the weighted estimate of erroneous enumerations. When the net error exceeded a pre-set maximum value, the sampling weights were reduced. This approach reduced variance and may have introduced some bias, but it is highly likely to have reduced the mean square error for most items. If the net error of the cluster did not exceed the pre-set maximum value, the sampling weights were unchanged.

The net error criteria was examined after the A.C.E. person matching operation was completed. If the criteria for weight trimming was met, it was done for all sample cases in a cluster even though a cluster contributed sample to multiple post-strata. This was done prior to the missing data process.

## BACKGROUND

Weight trimming guards against the possibility of a certain small number of clusters exerting an undue influence on post-stratum estimates and variances. In the A.C.E., these are expected to be due to a disproportionate number of census nonmatches or census erroneous enumerations within a few block clusters. Although extreme sampling weights can also be a source of influence in surveys, the A.C.E. sampling weights, the inverse of the probability of selection, were reasonably controlled in the sample design. These were not expected to be an important source of variance in the A.C.E.

While the A.C.E. sample design helped minimize the occurrence of highly influential clusters, a weight trimming plan was developed to reduce the effect of these potential extreme clusters. The A.C.E. weight trimming plan was a modification of the method used for the 1990 Post-Enumeration Survey (PES). As in 1990, the weights for extremely influential clusters were trimmed to yield a pre-specified net error. The intention of the plan was to lessen the impact of such clusters on the dual system estimates and variances.

The plan did not redistribute the weights across the remaining clusters to preserve totals. This would imply treating the E and P samples differently to preserve these separate totals, and contradicts the preference for consistent treatments of both samples. Since the primary interest was in the dual system estimation ratios, and not E- and P-sample totals, the weights were not redistributed.

## A.C.E. WEIGHT TRIMMING METHODOLOGY

Each cluster was evaluated to determine if it contributed disproportionately to the dual system estimates and variances. If the cluster was an outlier, the cluster sampling weight was multiplied by a factor to decrease the influence of the cluster on the dual system estimates and variances.

### Identify Outlier Clusters

A measure of the cluster influence was calculated for each cluster. Then, based on pre-set criteria, a decision was made whether the cluster should be identified as an outlier.

**Cluster Influence.** The measure of cluster influence was the net error. For purposes of weight trimming, the net error was the absolute difference between the weighted number of nonmatches and the weighted number of erroneous enumerations. The form of the weighted net error was

$$Z_i = \left| (P_i - M_i) - (E_i - CE_i) \right| \qquad (1)$$

where

- $Z_i$ = the net error estimate for cluster i,
- $P_i$ = the weighted P-sample population estimate for cluster i,
- $M_i$ = the weighted P-sample match estimate for cluster i,
- $E_i$ = the weighted E-sample population estimate for cluster i, and
- $CE_i$ = the weighted E-sample correct enumeration estimate for cluster i.

The first term of equation (1) was the weighted number of nonmatches in the $i^{th}$ cluster, while the second term was the weighted number of erroneous enumerations in the $i^{th}$ cluster.

**Outlier Criteria.** The outlier criterion was the maximum allowable net error for a single cluster. There were two different criteria based on the cluster geography. The nation was classified into two levels of geography: American Indian Reservations and the balance of the nation. The American Indian Reservation clusters were sampled at disproportionately higher rates relative to the balance of the country. In addition, separate American Indian on American Indian Reservation post-stratum estimates were planned. If the American Indian Reservation clusters were included with the rest of the nation, it is unlikely that an influential cluster would be detected. The two outlier criteria are defined in Table C-1.

Table C-1. **Outlier Cluster Criteria**

| Cluster geography | Maximum net error |
|---|---|
| American Indian Reservations | 6,250 |
| Balance of the United States | 75,000 |

All clusters with net error greater than the maximum allowable net error were considered influential clusters. They were expected to disproportionately influence the dual system estimates and variances. The sampling weights of these clusters were decreased.

The maximum net error for the balance of the country was based on experience in the 1990 PES. Since the A.C.E. was roughly double the PES sample size, the maximum allowable net error was set to be half the 1990 value. For the American Indian Reservation clusters, the maximum allowable value was a function of the average sampling rates. The American Indian Reservation average P-sample cluster sampling weight was approximately one-twelfth the balance of the U.S. average P-sample cluster sampling weight. Because of this, the American Indian Reservation maximum allowable net error was one-twelfth the balance of the U.S. criteria.

**Implementation Strategy.** The outlier clusters were identified after the person matching operation (Chapter 4) was completed, but before the missing data process (Chapter 6). The person matching results were the major input into this process. The weight trimming estimate used the best estimate of cluster net error at that time that was operationally feasible. This timing had several implications:

- Only nonmovers and outmovers were used for deriving the estimate of nonmatches above. For dual system estimation, if the number of outmovers in a post-stratum was less than 10 then only the non-movers and outmovers were used. Because of the small number of movers expected in most clusters, this process only used nonmovers and outmovers.
- Some nonmovers and outmovers had unresolved match status and residence status. Some E-sample cases had unresolved enumeration status. This meant the status of unresolved cases had to be estimated to identify outlier clusters. Information available at the time of the weight trimming process was used to approximately estimate the unresolved status cases. Since the weight trimming process was done before the missing data process, there was some information that the missing data process used to estimate unresolved status that was not yet available.
- A P-sample noninterview adjustment was approximated in the estimate of nonmatches. Information available during the weight trimming process was used to approximately estimate the noninterview adjustment for each cluster. As with the unresolved cases, since the weight trimming process was done before the missing data process, there was some information that the missing data process used to do the noninterview adjustment that was not yet available.
- The targeted extended search results and sampling rates were reflected in the estimate of nonmatches and erroneous enumerations (Chapter 5).

### Down-Weighting Outlier Cluster

All outlier clusters were down-weighted, so that no cluster contributed more than the maximum allowable number of net errors for the appropriate geography. A separate down-weighting factor was computed for each outlier cluster. The down-weighting factor was the ratio of the outlier cluster criteria to the cluster net error computed above.

$$D_i = \frac{C}{Z_i} \qquad (2)$$

where

$D_i$ = the down-weighting factor for cluster i, and
$C$ = the maximum net error from Table C-1 for the appropriate level of geography, and
$Z_i$ = the net error estimate for cluster i from (1).

The cluster down-weighting factor was applied to the P-sample and the E-sample weights of the outlier cluster. The P-sample and E-sample weights for the remaining clusters were unchanged.

## A.C.E. WEIGHT TRIMMING RESULTS

Table C-2 shows the one cluster down-weighted by the weight trimming process in the balance of the United States. No clusters were trimmed on American Indian Reservations.

Figures C-1 and C-2 show the distributions of net error before weight trimming for the balance of the United States and the American Indian Reservation areas.

Table C-2. **A.C.E. Weighted Net Errors for Down-Weighted Cluster**

| Geographic area | Before trimming | | | After trimming |
|---|---|---|---|---|
| | Estimated erroneous enumerations | Estimated weighted nonmatches | Estimated weighted net error | Estimated weighted net error |
| Balance of the United States ..... | 79,371 | 1,396 | 77,975 | 75,000 |

Figure C-1.
**Distribution of Net Error for the Balance of the United States**

(Number of clusters)

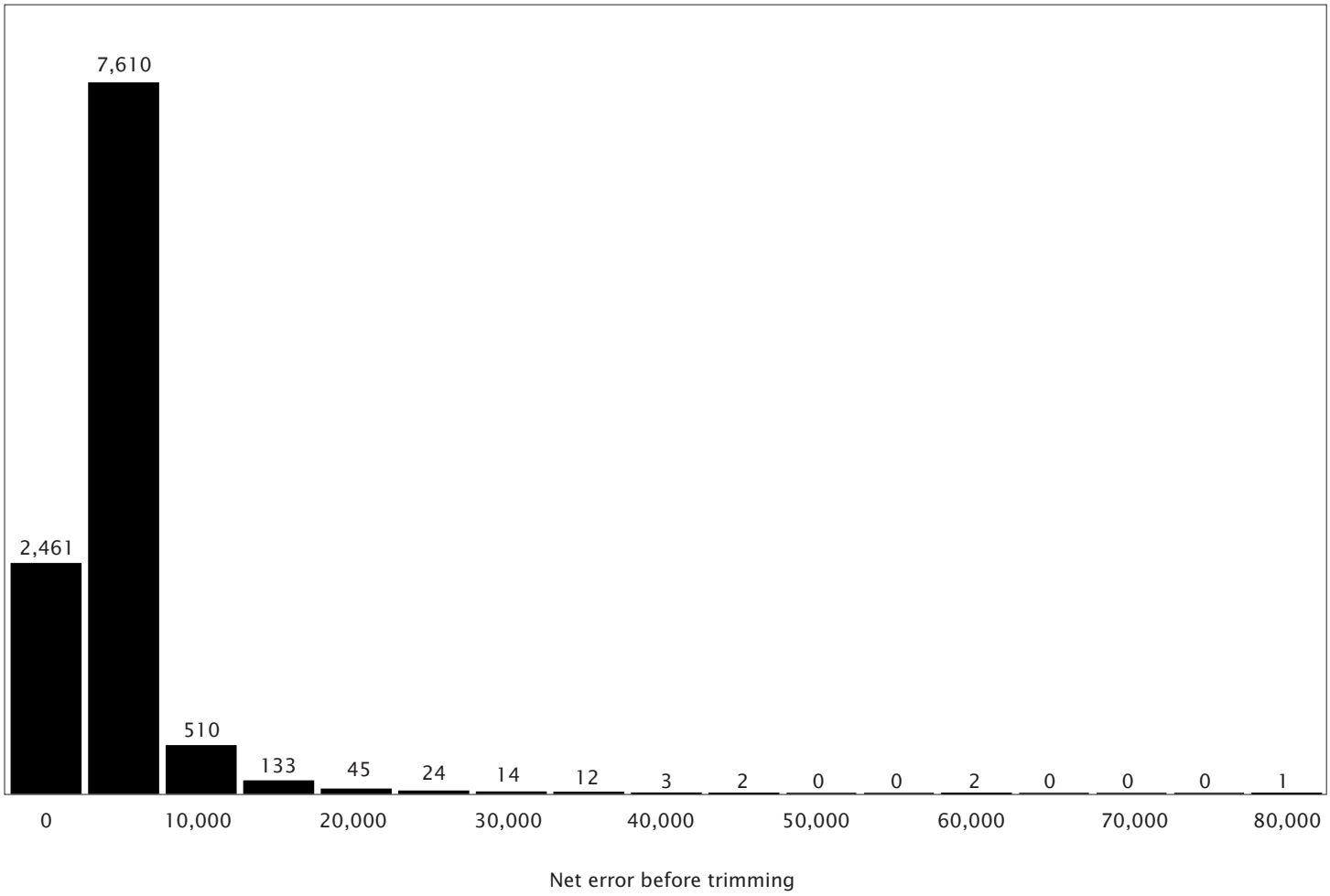
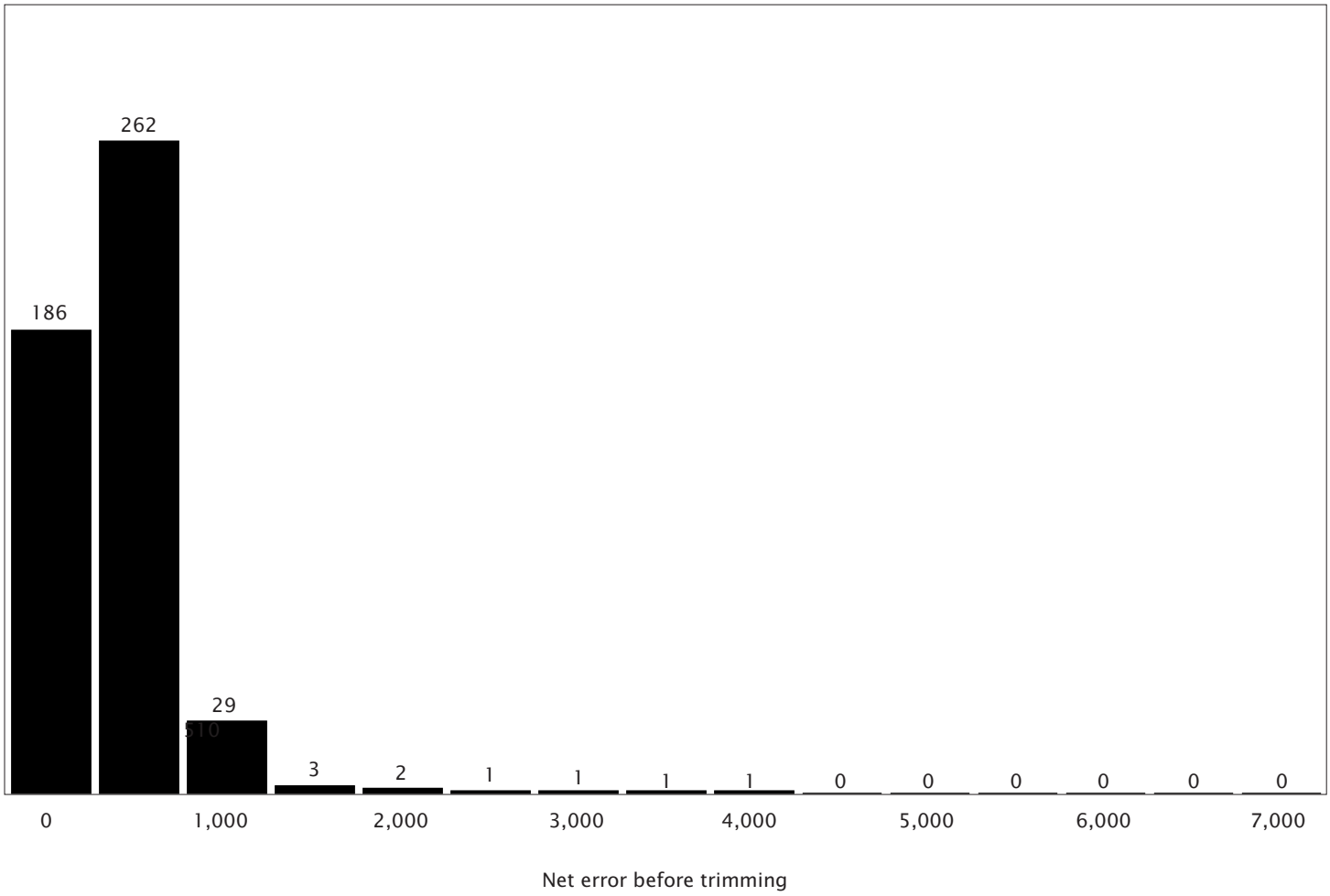


Figure C-2.
**Distribution of Net Error for American Indian Reservations**

(Number of clusters)



# Appendix D.
# Error Profile for A.C.E. Estimates

## INTRODUCTION

The Accuracy and Coverage Evaluation (A.C.E.) survey provided estimates of census coverage error that have been considered for adjusting Census 2000. The estimation used the PES-C version of dual system estimation with the data collected by the A.C.E. The adjusted estimates are subject to nonsampling error, as well as sampling error. This appendix discusses the types of errors found in the use of PES-C and the measurement of these errors.

## OVERVIEW OF ADJUSTED ESTIMATES

Define the following notation for each post-stratum h.

$C_h$ = census "count" for post-stratum h
$II_h$ = number of persons imputed into the original enumeration for post-stratum h
$\hat{I}_{E,h}$ = estimated number of enumerations in post-stratum h with insufficient information for matching[1]
$\hat{E}_{E,h}$ = estimated number of erroneous enumerations in post-stratum h
$\hat{N}_{E,h}$ = estimated population size for post-stratum h from the E sample
$\widehat{CE_h}$ = estimated population size for post-stratum h who could possibly be matched
$\widehat{CE_h}$ = $\hat{N}_{E,h} - \hat{I}_{E,h} - \hat{E}_{E,h}$
$\hat{N}_{P,h}$ = estimated size of the P-sample population
$\hat{M}_h$ = estimated number of the P-sample population enumerated in the census

The dual system estimator for the population size $N_h$ in post-stratum h is defined by

$$\hat{N}_h = (C_h - II_h)(\widehat{CE_h}/\hat{N}_{E,h})(\hat{N}_{P,h}/\hat{M}_h).$$

The 2000 A.C.E. used the PES-C formulation of the dual system estimator which uses the number of inmovers to estimate the number of outmovers, but uses the match rate for the outmovers to obtain the estimate of the number of outmovers that match the census. The post-stratum index h is suppressed in the following formula.

$$(\hat{N}_P/\hat{M}) = (\hat{N}_n + \hat{N}_i)/(\hat{M}_n + (\hat{M}_o/\hat{N}_o)\hat{N}_i).$$

where

$\hat{N}_n$ = estimated number of nonmovers
$\hat{N}_o$ = estimated number of outmovers
$\hat{N}_i$ = estimated number of inmovers
$\hat{M}_n$ = estimated number nonmovers enumerated in the census
$\hat{M}_o$ = estimated number outmovers enumerated in the census

When a post-stratum had fewer than 10 outmovers, the PES-A version of the dual system estimator that does not use inmovers was employed as follows:

$$(\hat{N}_P/\hat{M}) = (\hat{N}_n + \hat{N}_o)/(\hat{M}_n + \hat{M}_o)$$

The adjustment factor for post-stratum h is defined as $\hat{A}_h = \hat{N}_h/C_h$. The unadjusted estimate for area j is $N_{unadj,\,j} = \sum_h C_{h,\,j}$ and the adjusted estimate is $\hat{N}_{adj,\,j} = \sum_h A_h C_{h,\,j}$. The estimates of undercount in the population size of area j is $\hat{N}_{adj,\,j} - N_{unadj,\,j}$ and the estimate of the corresponding undercount rate is $(\hat{N}_{adj,\,j} - N_{unadj,\,j})/\hat{N}_{adj,\,j}$.

## SOURCES OF ERROR IN ADJUSTMENTS

The adjusted estimates are subject to a variety of possible sources of error: sampling error, data collection and survey operations error, missing data, error from exclusion of late census data and data with insufficient information for matching, contamination error, correlation bias, synthetic estimation bias, inconsistent post-stratification, and balancing error.

### P-Sample Matching Error and E-Sample Processing Error

**Source.** The term 'P-sample matching' has been used to describe the search of the census records for enumerations for P-sample respondents. The P-sample respondents are designated as matching an enumeration in the census or as not enumerated. The counterpart for the E sample is called "E-sample processing" where census enumerations are designated as correctly enumerated or erroneously enumerated. When the status of a P-sample or E-sample case can not be determined, it is designated as unresolved.

"P-sample matching error" refers to the net effect of errors that occur during the processing that affect the determination whether a P-sample person matches a census enumeration. Likewise, the net effect of errors in assigning enumeration status to E-sample enumerations during the office processing is called "E-sample processing error."

[1] Late enumerations are included with imputations in the original enumeration.

Errors may occur in either direction. P-sample people may be designated as matching a census enumeration although they are not in the census, called a "false match," or people may be designated as not enumerated although they are, called a "false nonmatch." E-sample enumerations may be falsely assigned a correct enumeration status, called a "false correct enumeration," or enumerations may be incorrectly designated as an erroneous enumeration, a "false erroneous enumeration."

Matching error also encompasses errors in the size of the P-sample population that may happen during the processing of the P-sample. These errors also may occur in either direction. A person included as a member of a household may really reside at another location or not be in the population of interest. For example, the census residency rules consider family members away at college to reside at their college address. A family member in a nursing center is considered to be in the group quarters population, which is not part of the population of interest. Vice versa, a person with two homes, may be designated as living at the other home, but really live at the one in the sample.

In the application of PES-C, respondents have the potential of many more statuses than was possible in the processing of the P-sample than in 1990. The reason is that a P-sample respondent may be a nonmover, an outmover, an inmover, or an out-of-scope person. The nonmovers and outmovers have another characteristic, which is resident or nonresident. A person who is living at the sample address on Census Day is called a resident.

Errors in mover status may go in all directions. A person designated as a nonmover may be an inmover or an outmover. All combinations of errors may happen and affect the DSE in different ways.

**Definition.** P-sample matching error affects both the estimates of nonmovers and inmovers in the estimate of the size of the P-sample population. In addition, matching error affects the estimates of the number of nonmover matches, the number of outmovers and outmover matches, and the number of inmovers in the estimate of the number of matches. E-sample data collection error affects the estimate of the number of erroneous enumerations. The post-stratum index h is suppressed in the following definitions.

$m_{nms}$ = net P-sample matching error in the nonmover component of $\hat{M}$

$m_{omims}$ = net P-sample matching error in the outmover and inmover component of $\hat{M}$

$n_{nms}$ = net P-sample matching error in the nonmover component of $\hat{N}_P$

$n_{ims}$ = net P-sample matching error in the inmover component of $\hat{N}_P$

$ee_s$ = net E-sample office processing error in $\widehat{CE}$

Under the assumption that all other errors are zero, the bias in the adjustment factors caused by P-sample matching error and E-sample processing error is defined as

$$\hat{B}_{process,h} = \hat{A}_h - \frac{C_h - II_h}{C_h} \times \frac{\widehat{CE}_h - \hat{B}_{CE-process,h}}{\hat{N}_{E,h}} \times \frac{[\hat{N}_{P,h} - \hat{B}_{P-process,h}]}{\hat{M}_h - \hat{B}_{M-process,h}}.$$

The error component definitions include a ratio adjustment because they are estimated using the Evaluation Sample. The ratio adjustment for components from the P-sample is the ratio of the P-sample population total from the A.C.E. to the P-sample population total based on the Evaluation Sample $\hat{N}_{PF}$. The ratio adjustment for the components from the E-Sample is ratio of the two E-sample totals defined comparably.

$$\hat{B}_{M-process} = [m_{nms} + m_{omims}] \times [\hat{N}_P/\hat{N}_{PF}]$$
$$\hat{B}_{P-process} = [n_{nms}] \times [\hat{N}_P/\hat{N}_{PF}]$$
$$\hat{B}_{CE\text{-}process} = -[ee_s] \times [\hat{N}_E/\hat{N}_{EF}]$$

### P-Sample and E-Sample Data Collection Error

**Source.** Errors may occur during the data collection. While an interview is in progress, the respondent may make an error in answering a question, or the interviewer may make an error in asking a question or recording the answer. Errors also occur when an interviewer goes to the wrong address. Regardless of whether the error is caused by the respondent, the interviewer, or a combination of the two, such errors may cause the matching operation to assign mover status, residency status, or match status incorrectly to a person on the household roster. The A.C.E. interviewer collects both a Census Day roster and an Interview Day roster. A person who resides at the household on both days is classified as a nonmover. A person who lived there only on Census Day is an outmover, while a person who lived there only on Interview Day is an inmover. Persons classified as outmovers and nonmovers may or may not have been a resident at the address on Census Day. Errors in the mover status, residency status, or other errors may cause the matching operation to fail to determine that a person was enumerated and to classify the person as a nonmatch incorrectly.

Sometimes people listed on household rosters do not exist. A more likely scenario is an interviewer who is having trouble contacting the residents of a housing units may copy the name from a mail box and fill in the characteristics. This type of error is called "P-sample fabrication." Usually fabricated households cause an underestimate of the match rate, because they are smaller than the average household size and do not match,

A special type of E-sample data collection error is the failure to identify duplicate enumerations. The processing includes a search for duplicate enumerations within the block cluster and the surrounding blocks. Duplicate enumerations outside the block cluster and surrounding

blocks are more difficult to find. Identifying these duplications requires the respondent to provide information concerning another address where a household member may also be enumerated. Errors may occur when the respondent does not understand the residency rules or is unaware that a household member may be enumerated at another address. The situations most prone to causing duplicate enumerations are college students enumerated at their family home and their college address, children in joint custody agreements enumerated at both parents' addresses, and people with two residences.

Another type of field error occurs during the listing of the housing units for the census or for the P-sample. The housing units listed as being in the sample block may be in another block or vice versa. These types of errors are called "geocoding error." To account for minor geocoding errors in 2000, the search for matches occurred within all block-clusters and also in surrounding blocks for a sample of the cases with geocoding errors recorded in the E sample a design called "Targeted Extended Search (TES)." The variance estimates for the A. C. E. account for the TES design. Flaws in the execution of the TES may result in biases.

**Definition.** P-sample fabrication and data collection error affect both the estimates of nonmovers and inmovers in the estimate of the size of the P-sample population. In addition, fabrication and data collection error affect the estimates of the number of nonmover matches, the number of outmovers and outmover matches, and the number of inmovers in the estimate of the number of matches. E-sample data collection error affects the estimate of the number of erroneous enumerations. Again, the post-stratum index h is suppressed in the following definitions.

$m_{nmr}$ = net P-sample data collection error in the nonmover component of $\hat{M}$

$m_{omimr}$ = net P-sample data collection error in the outmover and inmover component of $\hat{M}$

$n_{nmr}$ = net P-sample data collection error in the nonmover component of $\hat{N}_P$

$n_{imr}$ = net P-sample data collection error in the inmover component of $\hat{N}_P$

$ee_r$ = net E-sample data collection error in $\widehat{CE}$

$m_{nmfp}$ = net P-sample fabrication error in the nonmover component of $\hat{M}$

$m_{omimfp}$ = net P-sample fabrication error in the outmover and inmover component of $\hat{M}$

$n_{nmfp}$ = net P-sample data collection error in the nonmover component of $\hat{N}_P$

$n_{imfp}$ = net P-sample data collection error in the inmover component of $\hat{N}_P$

Under the assumption that all other errors are zero, the bias in the adjustment factors caused by P-sample data collection error and E-sample data collection error is defined as

$$\hat{B}_{collect,h} = \hat{A}_h - \frac{C_h - II_h}{C_h} \times \frac{\widehat{CE}_h - \hat{B}_{CE-collect,h}}{\hat{N}_{E,h}} \times \frac{[\hat{N}_{P,h} - \hat{B}_{P-collect,h}]}{\hat{M}_h - \hat{B}_{m-collect,h}}.$$

The error component definitions include a ratio adjustment because they are estimated using the Evaluation Sample. The ratio adjustment for components from the P-sample is the ratio of the P-sample population total from the A.C.E. to the P-sample population total based on the Evaluation Sample. The ratio adjustment for the components from the E-Sample is ratio of the two E-sample totals defined comparably.

$\hat{B}_{M-collect} = [m_{nmr} + m_{nmfp} + m_{omimr} + m_{omimfp}] \times [\hat{N}_P/_{PF}]$

$\hat{B}_{P-collect} = [n_{nmr} + n_{nmfp} + n_{im}] \times [N_P/\hat{N}_{PF}]$

$\hat{B}_{CE-collect} = -[ee_r] \times [N_E/\hat{N}_{EF}]$

## Missing Data

**Source.** A.C.E. data may be missing for a variety of reasons—some A.C.E. interviews fail to take place, some households provide incomplete data on questionnaire items, and in some cases the information for classification as a match or nonmatch is ambiguous. The methods used to compensate for missing data effectively assume that the match status for the case with missing data is equal on average to the status for cases that are similar, except that they have complete data. Missing data on characteristics are imputed from otherwise similar cases with complete data. Nonresponse weighting adjustments are used to account for sampled but noninterviewed households. The P-sample matching and E-sample processing operation assigns "unresolved" enumeration status to a case when the available data is inadequate to determine whether the person is enumerated in the census and a probability of being correctly enumerated is imputed for such cases.

Also, error in the resolved cases causes error in the imputations, because the resolved cases are used to form the imputations. Even if the imputation model were perfect, the imputations will have error if the data used to fit the model has error. This type of error is called "imputation error due to data error."

Although one can consider the range of effects on the DSE by considering extreme alternatives—e.g., all unresolved matches truly are matches or truly are nonmatches—the range is too wide to be informative about the likely bias. The bias from the method used to compensate for missing data can in principle be estimated from intensive follow-up of cases with missing data, but in practice the fraction completed by follow-up is too low. The Census Bureau analyzed the missing-data bias by looking at the changes in the DSE when alternative methods were used to compensate for missing data.

Results from the Analysis of Reasonable Alternative Imputation Models are used to estimate the variance component. See Keathley et al. (2001) for details. The results of

Reasonable Alternative Imputation Models provided the data for the calculation of the variance-covariance matrix for adjustment factors the missing data component. The missing data variance-covariance matrix was added to the sampling variance-covariance matrix to obtain a variance-covariance matrix for the adjustment factors that contained the random error due to sampling and imputation for missing data.

**Definition.** The Census Bureau models the error due to imputation as a random effect and estimates its variance-covariance matrix. Modeling imputation error as a random effect is motivated by practicalities. In principle, the bias from the method used to compensate for missing data can be estimated from intensive follow-up of cases with missing data, but in practice the fraction completed by follow-up is too low. The variance component due to imputation for missing data has three components.

$V_M$ = variance due to imputation = $V_{RA} + V_B + V_I$

where

$V_{RA}$ = variance due to the imputation model selection
$V_B$ = variance due to the model parameter estimation
$V_I$ = within-person imputation variance.

The imputation variance components due to parameter estimation and within person estimation are included in the sampling error estimates, leaving the variance due to model selection to be estimated separately. The missing data variance-covariance matrix is added to the sampling variance-covariance matrix to obtain a variance-covariance matrix for the adjustment factors that contained the random error due to sampling and imputation for missing data.

The components of imputation error due to data error affect estimate of the number of nonmovers, the estimate of the number of nonmovers enumerated, the estimate of the match rate for the outmovers, and the estimate of the number of erroneous enumerations. The post-stratum index h is suppressed in the following definitions.

$m_{nmi}$ = net imputation error due to data error in the nonmover component of $\hat{M}$
$m_{omi}$ = net imputation error due to data error in the outmover match rate component of $\hat{M}$
$n_{nmi}$ = net imputation error due to data error in the nonmover component of $\hat{N}_P$
$ee_i$ = net imputation error due to data error in $\widehat{CE}$.

Under the assumption that all other errors are zero, the bias in the adjustment factor $\hat{A}_h$ caused by imputation error due to data error is defined as

$$\hat{B}_{impdata,h} = \hat{A}_h - \frac{C_h - II_h}{C_h} \times \frac{\widehat{CE}_h - \hat{B}_{CE-impdata,h}}{\hat{N}_{E,h}} \times \frac{[\hat{N}_{P,h} - \hat{B}_{P-impdata,h}]}{\hat{M}_h - \hat{B}_{M-impdata,h}}$$

The error component definitions include a ratio adjustment because they are estimated using the Evaluation Sample. The ratio adjustment for components from the P-sample is the ratio of the P-sample population total from the A.C.E. to the P-sample population total based on the Evaluation Sample. The ratio adjustment for the components from the E Sample is the ratio of the two E-sample totals defined comparably.

$\hat{B}_{M-impdata} = [m_{nmi} + m_{omi}] \times [\hat{N}_P/\hat{N}_{PF}]$
$\hat{B}_{P-impdata} = [n_{nmi}] \times [\hat{N}_P/\hat{N}_{PF}]$
$\hat{B}_{CE-impdata} = -[ee_i] \times [\hat{N}_E/\hat{N}_{EF}]$

### Sampling Error

**Source.** Sampling error gives rise to random error, quantified by sampling variance, and to a systematic error known as ratio-estimator bias. The sampling variance is present in any estimate based on a sample instead of the whole population. Ratio-estimator bias arises because even if X and Y are unbiased estimators, X/Y typically is biased.

**Definition.** The sampling variance and ratio-estimator bias for the adjustment factors are

$S^2$ = sampling variance-covariance matrix for the adjustment factors
$\hat{B}_{ratio,h}$ = ratio-estimator bias in the adjustment factor $\hat{A}_h$ for post-stratum h.

Random sampling error is reflected in the estimated variance-covariance matrix of the $\hat{A}_h$'s. The covariance matrix is estimated by the Census Bureau's sampling-error software applied to the A.C.E. data. The software also can be used to produce estimates of ratio-estimator bias.

### Correlation Bias

**Source.** If there is variability of the enumeration probabilities for persons in the same post-stratum, or if there is a dependence between enumeration in the census and in the A.C.E. e.g., people less likely to be enumerated in the census may also be less likely to be found in the A.C.E., then correlation bias may arise. Correlation bias is most likely a source of downward bias in the DSE. Evidence of correlation bias in national estimates is provided by sex ratios (males to females) for adjusted numbers that are low relative to ratios derived from demographic analysis of data on births, deaths, and migration.

The information from demographic analysis is insufficient to estimate correlation bias at the post-stratum level, however, and alternative parametric models have been used to allocate correlation bias estimates for national age-race-sex groups down to post-strata. Estimates of correlation bias at the national level provided by demographic analysis information also account for possible error from groups whose probabilities of enumeration are so low that the DSE will fail to account for them. The estimates of correlation bias based on sex ratios are affected by error in the demographic-analysis sex ratios and by possible other biases in the sex ratios in the DSE.

The assumptions and model underlying the measurement of correlation bias are discussed in detail in Bell (2001,2001b), but are described briefly here. Although there are several models for how correlation bias is distributed, the 'two-group' model was selected. The two-group model relies on the basic assumptions listed below for the estimation of correlation bias. In addition, sensitivity analyses assess the impact of variations in these assumptions.

- The ratio of males to females measured in demographic analysis is more reliable for the two racial groups, Black and non-Black, than the A.C.E.
- There is no correlation bias present in the A.C.E. estimates for females.
- The relative correlation bias is equal across all A.C.E. post-strata within an age-race category.
- The relative impact of other nonsampling errors is equal for males and females at the national level.

The assumption with the two-group model of the relative correlation bias being equal across post-strata within an age-sex category has the advantage of permitting the estimation of correlation bias through a multiplicative factor applied to the corrected DSE. Even more important, an unbiased estimate of the factor is available under the assumption that the relative impact of the other nonsampling errors is equal for males and females without actually having to estimate the nonsampling errors.

**Definition.** Correlation bias usually causes the DSE to underestimate the population size.

$\hat{B}_{correl,h}$ = correlation bias in the adjustment factor $\hat{A}_h$ for post-stratum h.

### Excluded-data Error from Reinstated Census Enumerations

**Source.** The DSE treats late census data as nonenumerations. Thus, duplicate enumerations among the late data do not contribute to census data, but valid enumerations among the late data are treated as census misses and are estimated by the DSE. If the late census data were excluded from the entire adjustment process and estimation, no new source of error would be present. The adjusted estimates do partially incorporate late census data by including them in $C_h$, but excluding them from the computation of $\widetilde{N}_h$. This use of late data affects the estimates for areas with disproportionately many or few late adds, with an effect that is similar to synthetic estimation error. In addition, the exclusion of late census data from the E sample could bias the estimates at the post-stratum level.

There are two conditions that have to be met for the exclusion of the late adds from the processing of the A.C.E. not to bias the dual system estimates at the post-stratum level:

- The P sample covers the correct enumerations among the late adds at the same rate as other correct enumerations.
- The late adds occur in the E sample at the same rate as they occur in the census (excluding the imputations).

**Definition.** Error due to excluding the reinstated census enumerations in the calculation of the DSE affects the estimate of the DSE and therefore the adjustment factor directly.

$\hat{B}_{reinstate,h}$ = bias in the adjustment factor $\hat{A}_h$ for post-stratum h due to excluding reinstated census enumerations.

### Contamination Error

**Source.** Contamination occurs when the A.C.E. selection of a given block cluster alters the implementation of the census there and affects enumeration results, e.g, by increasing or decreasing erroneous enumerations or by increasing or decreasing coverage rates. Contact with residents of the sample blocks during the listing for the P-sample may cause them to not respond to the census, because they think that the listing contact is a response to the census.

**Definition.** The bias in the adjustment factor for post-stratum h from contamination is defined as follows.

$\hat{B}_{contam,h}$ = bias in the adjustment factor $\hat{A}_h$ for post-stratum h due to contamination error.

### Synthetic Estimation Bias

**Source.** The adjustment methodology relies on a method called synthetic estimation to provide the same adjustment factor $\hat{A}_h$ for all enumerations in a given post-stratum, regardless of whether the enumerations are from the same geographic area. Synthetic estimation bias arises when the census from different areas but in the same post-stratum should have different adjustment factors.

To assess synthetic estimation bias for a given area one needs to develop an estimate based on data from the area alone, which is rarely possible. Attempts to estimate synthetic estimation bias in undercount estimates from analysis of "artificial populations" or "surrogate" variables, whose geographic distributions are known, are unconvincing. Therefore, sensitivity analyses have been conducted to assess the impact of synthetic estimation bias. These studies show that assuming synthetic estimation has a minor effect on uses of the data is reasonable.

**Definition.** The synthetic estimates may cause a bias in the adjusted estimates for area j. Error from synthetic estimation does not affect the dual system estimate for a post-stratum, only areas within a post-stratum.

$\hat{B}_{syn,j}$ = bias in the adjusted estimate $\hat{N}_{adj,j}$ for area j due to synthetic estimation error.

### Inconsistent Post-stratification

**Source.** The computation of $\widehat{CE}_h/\hat{N}_{E,h}$ requires census enumerations to be assigned to post-strata, and the computation of $\hat{N}_{P,h}/\hat{M}_h$ requires P-sample enumerations to be assigned to post-strata. When the assignments are not made consistently for the two samples, error arises in the ratio $\hat{N}_{P,h}/\hat{M}_h$. Inconsistent assignments to post-strata may be caused by mis-reporting of characteristics used in post-stratification.

Cases that are most prone to inconsistent classification are those where there is a different respondent for the household in the census and the A.C.E. For example, a household member's age or race may be reported differently in a self-response than when another household members responds for the person. Such inconsistencies also may be due to computer processing errors, as well as inconsistent reporting.

The matches in the A.C.E. sample provide a source of data for estimating the error due to inconsistent post-stratification. An estimate of the error for a post-stratum may be formed by assuming the inconsistency rate observed in the matches also holds for those not matched.

**Definition.** Error due to inconsistent post-stratification affects the estimate of the DSE and therefore the adjustment factor directly.

$\hat{B}_{inconsist,h}$ = bias in the adjustment factor $\hat{A}_h$ for post-stratum h due to inconsistent post-stratification.

### Error from Estimating Outmovers with Inmovers

**Source.** This error is unique to the PES-C model used in the A.C.E. For the PES-C model, the members of the P-sample are the residents of the housing units on Census Day. There is some difficulty in identifying all the residents of all the housing units on Census Day because some move prior to the A.C.E. interview. The A.C.E. interview relies on the respondents to identify those who have moved out, the outmovers. Since the outmovers are identified by proxies, many of the outmovers are not recorded. Therefore, the estimate of outmovers is too low. PES-C uses the number of inmovers to estimate the number of outmovers. The inmovers are those who did not live in the sample blocks on Census Day, but moved in prior to the A.C.E. interview. Theoretically the number of inmovers in the whole country should equal the number of outmovers. However, the number of inmovers may not equal the number of outmovers in a post-stratum because of circumstances such as economic conditions causing more people to move out of an area than to move into an area.

**Definition.** The error due to using the inmovers to estimate the outmovers affects the estimates of the size of the P-sample population and the number of matches.

$m_{io,h}$ = net P-sample data collection error in the mover component of $\hat{M}_h$ in post-stratum h

$n_{io,h}$ = net P-sample data collection error in the mover component of $\hat{N}_{P,h}$ in post-stratum h.

Under the assumption that all other errors are zero, the bias in the adjustment factors caused by P-sample matching error and E-sample processing error is defined as

$$\hat{B}_{inout,h} = \hat{A}_h - \frac{C_h - II_h}{C_h} \times \frac{\widehat{CE}_h}{\hat{N}_{E,h}} \times \frac{[\hat{N}_{P,h} - \hat{B}_{P-inout,h}]}{\hat{M}_h - \hat{B}_{M-inout,h}}.$$

The error component definitions include a ratio adjustment because they are estimated using the Evaluation Sample. The ratio adjustment for components from the P-sample is the ratio of the P-sample population total from the A.C.E. to the P-sample population total based on the Evaluation Sample. The post-stratum index h is suppressed in the following definitions.

$\hat{B}_{M-inout} = [m_{io}] \times [\hat{N}_P/\hat{N}_{PF}]$

$\hat{B}_{P-inout} = [n_{io}] \times [\hat{N}_P/\hat{N}_{PF}]$

### Balancing Error

**Source.** Balancing error must be addressed in the design of the search areas used to search for E-sample correct enumerations and P-sample matches. Limiting the search for correct enumerations and matches is necessary because the matching operation cannot search the entire census. By limiting the search area, a small percentage of correct enumerations will not be found and a small percentage of matches will not be found. This causes an underestimate of the correct enumerations and an underestimate of the matches. However, the estimate of the net error is not biased as long as the percentage error in the correct enumerations equals the percentage error in the matches. The A.C.E. design avoids balancing error by choosing the same block clusters for the E-sample and the P-sample and drawing the search areas consistently.

**Definition.** There is not a separate measurement of balancing error. Any balancing error that may arise during the implementation of the A.C.E. will be included in the measurement of data collection error.

# Section I. References


Adams, T. and Liu, X. (2001). "ESCAP II: Evaluation of Lack of Balance and Geographic Errors Affecting Person Estimates," Executive Steering Committee for A.C.E. Policy II, Report 2.

Baker v. Carr 369 U.S. 186 (1962).

Belin, T., Diffendal, G., Mack, S., Rubin, D., Schafer, J., and Zaslavsky, A. (1993). "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," Journal of the American Statistical Association, 88, 1149-1166.

Bell, W. (2001). "Accuracy and Coverage Evaluation: Correlation Bias," DSSD Census 2000 Procedures and Operations Memorandum Series #B-12*.

Bell, W. (2001b). "ESCAP II: Estimation of Correlation Bias in 2000 A.C.E. Estimates Using Revised Demographic Analysis Results," Executive Steering Committee for A.C.E. Policy II, Report 10.

Byrne, R., Imel, L., Ramos, M., and Stallone, P. (2001). "Accuracy and Coverage Evaluation: Person Interviewing Results," Census 2000 Procedures and Operations Memorandum Series #B-5*.

Cantwell, P. (1999). "Accuracy and Coverage Evaluation Survey: Overview of Missing Data for P & E Samples," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-3.

Cantwell, P. (2000). "Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-25.

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

Childers, D. (2001). "Accuracy and Coverage Evaluation: The Design Document," Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised.

Childers, D., Byrne, R., Adams, T., and Feldpausch, R. (2001). "Accuracy and Coverage Evaluation: Person Matching and Followup Results," Census 2000 Procedures and Operations Memorandum Series #B-6*.

Coale, A. (1955). "The Population of the United States in 1950 Classified by Age, Sex, and Color A Revision of Census Figures," Journal of the American Statistical Association, 50, 16-54.

Coale, A. and Rives, N. (1973). "A Statistical Reconstruction of the Black Population of the United States, 1880-1970: Estimates of True Numbers by Age and Sex, Birth Rates, and Total Fertility," Population Index, 39(1), 3-36.

Coale, A. and Zelnick, M. (1963). "New Estimates of Fertility and Population in the United States," Princeton University Press.

Cox, L. and Ernst, L. (1982). "Controlled Rounding," INFOR, Vol. 20, No. 4.

Davis, M. (1991). "Preliminary Final Report for PES Evaluation Project P7: Estimates of P-sample Clerical Matching Error from a Rematching Evaluation," 1990 Coverage Studies and Evaluation Memorandum Series #H-2.

Davis, M. (1991b). "Preliminary Final Report for PES Evaluation Project P10: Measurement of the Census Erroneous Enumerations - Clerical Error Made in the Assignment of Enumeration Status," 1990 Coverage Studies and Evaluation Memorandum Series #L-2.

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

Fay, R., Passel, J., Robinson, J. G., and Cowan, C. (1988). "The Coverage of Population in the 1980 Census," 1980 Census of Population and Housing: Evaluation and Research Reports, PHC80-E4, U.S. Bureau of the Census, Washington, D.C.

Ghosh, M. and Rao, J. N. K. (1994). "Small Area Estimation: An Appraisal," Statistical Science, Vol. 9, No. 1, 55-93.

Gonzalez, M. (1973). "Use and Evaluation of Synthetic Estimators," Proceedings of the Social Statistics Section, American Statistical Association.

Gonzalez, M. and Waksberg, J. (1973). "Estimation of the Error of Synthetic Estimates," paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

Griffin, R. (1999). "Accuracy and Coverage Evaluation Survey: Post-stratification Research Methodology," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-5.

Griffin, R. and Haines, D. (2000). "Accuracy and Coverage Evaluation Survey: Post stratification for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-21.

Griffin, R. and Haines, D. (2000b). "Accuracy and Coverage Evaluation Survey: Final Post stratification Plan for Dual System Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-24 .

Griffin, R. and Malec, D. (2001). "Accuracy and Coverage Evaluation: Assessment of Synthetic Assumption," DSSD Census 2000 Procedures and Operations Memorandum Series #B-14*.

Haines, D. (2001). "Accuracy and Coverage Evaluation Survey: Synthetic Estimation," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-46.

Haines, D. (2001b). "Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (U.S.) - Re-issue of Q-37," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-48.

Himes, C. and Clogg, C. (1992). "An Overview of 'Demographic Analysis' as a Method for Evaluating Census Coverage in the United States," Population Index, 58(4), 587-607.

Hogan, H. (1992). " The 1990 Post-Enumeration Survey: An Overview," The American Statistician, Vol. 46(4), 261-269.

Hogan, H. (1993). "The 1990 Post-Enumeration Survey: Operations and Results," Journal of the American Statistical Association, 88, 1047-1060.

Hogan, H. (2000). "The Accuracy and Coverage Evaluation: Theory and Application," Proceedings of the Survey Research Methods Section, American Statistical Association.

Hogan, H. (2001). "Accuracy and Coverage Evaluation Survey: Effect of Excluding 'Late Census Adds,'" DSSD Census 2000 Procedures and Operations Memorandum Series #Q-43.

Ikeda, M. (1997). "Effect of Using the 1996 ICM Characteristic Imputation and Probability Modeling Methodology on the 1995 ICM P and E-Sample Data," DSSD Census 2000 Dress Rehearsal Memorandum Series #A-20.

Ikeda, M. (1998). "Effect of Different Methods for Calculating Match and Residence Probabilities for the 1995 P-Sample Data," DSSD Census 2000 Dress Rehearsal Memorandum Series #A-23.

Ikeda, M. (1998b). "Effect of Different Methods for Calculating Correct Enumeration Probabilities for the 1995 E-Sample Data," DSSD Census 2000 Dress Rehearsal Memorandum Series #A-28.

Ikeda, M. (1998c). "Effect of Using Simple Ratio Methods to Calculate P-Sample Residence Probabilities and E-Sample Correct Enumeration Probabilities for the 1995 Data," DSSD Census 2000 Dress Rehearsal Memorandum Series #A-30.

Ikeda, M., Kearney, A., and Petroni, R. (1998). "Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," Proceedings of the Survey Research Methods Section, American Statistical Association.

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

Kearney, A. and Ikeda, M. (1999). "Handling of Missing Data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," Proceedings of the Survey Research Methods Section, American Statistical Association.

Keathley, D., Kearney, A., and Bell, W. (2001). "ESCAP II: Analysis of Missing Data Alternatives for the Accuracy and Coverage Evaluation," Executive Steering Committee for A.C.E. Policy II, Report 12.

Keeley, C. (2000). "Census 2000 Accuracy and Coverage Evaluation Computer Assisted Interview," DSSD Census 2000 Procedures and Operations Memorandum Series #S-QD-02.

Killion, R.A. (1998). "Estimation Decisions for the Integrated Coverage Measurement Survey for Census 2000," Census 2000 Decision Memorandum No. 42.

Kostanich, D., Griffin, R., and Fenstermaker, D. (1999). "Census 2000 Accuracy and Coverage Evaluation Survey: Sample Allocation and Post-stratification Plans," DSSD Census 2000 Procedures and Operations Memorandum Series #R-2.

Marks, E. (1979). "The Role of Dual System Estimation in Census Evaluation," in K. Krotki (Ed.), Developments in Dual System Estimation of Population Size and Growth, Edmonton: University of Alberta Press, 156-188.

Nash, F. (2000). "Overview of the Duplicate Housing Unit Operations," Census 2000 Information Memorandum Number 78.

National Center for Health Statistics (1968). "Synthetic State Estimates of Disability," P. H. S. Publication 1759, U.S. Government Printing Office, Washington, D.C.

Raglin, D. and Bean, S. (1999). "Outmover Tracing and Interviewing," Census 2000 Dress Rehearsal Evaluation Results Memorandum Series #C-3.

Rao, J.N.K. and Shao, J. (1992). "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation," Biometrika, 79, 811-822.

Reynolds v. Simms, 377 U.S. 533 (1964).

Robinson, J. G. (2001). "ESCAP II: Demographic Analysis Results," Executive Steering Committee for A.C.E. Policy II, Report 1.

Robinson, J. G., Ahmed, B., Gupta, P., and Woodrow, K. (1993). "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," Journal of the American Statistical Association, 88, 1061-1071.

Schindler, E. (1998). "Allocation of the ICM Sample to the States for Census 2000," Proceedings of the Survey Research Methods Section, American Statistical Association.

Schindler, E. (1999). "Comparison of DSE C and DSE A," Census 2000 Dress Rehearsal Evaluation Memorandum # C-8a.

Schindler, E. (2000). "Accuracy and Coverage Evaluation Survey: Post-stratification Preliminary Research Results," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-23.

Sekar, C. C. and Deming, W. E. (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration," Journal of the American Statistical Association, 44, 101-115.

Siegel, J. (1974). "Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis," 1970 Census of Population and Housing: Evaluation and Research Program, PHC(E)-4, U.S. Bureau of the Census, Washington, D.C.

Siegel, J. and Zelnik, M. (1966). "An Evaluation of Coverage in the 1960 Census of Population by Techniques of Demographic Analysis and by Composite Methods," Proceedings of the Social Statistics Section, American Statistical Association.

Starsinic, M. (2001). "Accuracy and Coverage Evaluation Survey: Specifications for Covariance Matrix Output Files from Variance Estimation for Census 2000," DSSD Census 2000 Procedures and Operations Memorandum Series #V-4.

Starsinic, M. and Kim, J. (2001). "Accuracy and Coverage Evaluation: Computer Specifications for Variance Estimation for Census 2000 - Revision," DSSD Census 2000 Procedures and Operations Memorandum Series #V-5.

U.S. Bureau of the Census (1985). Evaluating Censuses of Population and Housing, Statistical Training Document, ISP-TR-5, Washington, D.C.

West, K. (1991). "Final Report for PES Evaluation Project P4: Quality of Reported Census Day Address - Evaluation Follow-up," 1990 Coverage Studies and Evaluation Memorandum Series #D-2.

Winkler, W. (1994). "Advanced Methods for Record Linkage," Proceedings of the Survey Research Methods Section, American Statistical Association.

Wolfgang, G. (1999). "Request for Dress Rehearsal Surrounding Block Files for A.C.E. Research," unpublished Census Bureau Memorandum.

Wolter, K. (1986). "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, 338-346.

Woltman, H., Alberti, N., and Moriarity, C. (1988). "Sample Design for the 1990 Census Post Enumeration Survey," Proceedings of the Survey Research Methods Section, American Statistical Association.

ZuWallack, R. (2000). "Sample Design for the Census 2000 Accuracy and Coverage Evaluation," Proceedings of the Survey Research Methods Section, American Statistical Association.

# Accuracy and Coverage Evaluation of Census 2000: Design and Methodology

## Section II

## A.C.E. Revision II March 2003

# Chapter 1.
# Introduction to A.C.E. Revision II

## INTRODUCTION

The Accuracy and Coverage Evaluation (A.C.E.) survey was designed to measure and possibly correct net coverage error in Census 2000. However, because A.C.E. failed to measure a significant number of erroneous enumerations, A.C.E. did not meet these objectives. The Census Bureau's Executive Steering Committee for A.C.E. Policy (ESCAP) recommended twice **NOT** to correct the census counts.[1] There are, however, concerns about differential coverage error in Census 2000 data. While the Census 2000 data products will not be corrected, it is possible that improvements could be made to the post-censal population estimates used for survey controls. This is the Census Bureau's motivation for correcting errors in the A.C.E. data and developing improved estimates of the net undercount. The improved estimates are called A.C.E. Revision II estimates. These revised estimates provide a better picture of Census 2000 coverage and will help us design a better coverage measurement program for 2010. This part of the document provides a description of the methodology used to produce the A.C.E. Revision II estimates. A comprehensive technical description of the methodology used to produce the original
estimates of net undercount released in March 2001 is presented in the first half of this publication.

This chapter summarizes the history of the two adjustment decisions and discusses key findings and limitations. It also introduces the key components of the revision and describes the major errors being corrected. The next chapter provides an overview of the revision process and subsequent chapters provide detailed methodology as follows:

Chapter 2: Summary of A.C.E. Revision II Methodology

Chapter 3: Correcting Data for Measurement Error

Chapter 4: A.C.E. Revision II Missing Data Methods

Chapter 5: Further Study of Person Duplication in Census 2000

Chapter 6: A.C.E. Revision II Estimation

Chapter 7: Assessing the Estimates

[1]The ESCAP recommendations, supporting analyses, technical assessments, and limitations can be found on the Census Bureau's Web site at www.census.gov/dmd/www/EscapRep.html.

## BACKGROUND

The original A.C.E. estimates were available in February of 2001, in time to allow for the possibility of correcting Census 2000 redistricting files. The Census Bureau's ESCAP recommended in March 2001 not to correct the Census 2000 counts for purposes of redistricting (ESCAP I, 2001). The Secretary of Commerce concurred. Given the information available at this time, this decision was not based on any clear evidence that the census counts were more accurate, but rather concern that there was some yet undiscovered error in the A.C.E. The A.C.E estimate of a 3.3 million net undercount was much larger than the Demographic Analysis (DA) estimate of only 340,000. Further evaluations were conducted over the next 6 months to examine the reasons for the discrepancy and to determine if other Census 2000 data products should be corrected. The Census 2000 redistricting files were the first of many Census 2000 data products scheduled for public release. See the Census Bureau's Web site, www.census.gov, for released data products. The question remained as to whether these other Census 2000 data releases should be corrected.

In October 2001, the ESCAP again decided not to correct the census counts for other Census 2000 data products. Analysis of A.C.E. evaluation data and a study of duplicates in the census revealed that the A.C.E. failed to measure large numbers of erroneous census enumerations (ESCAP II, 2001). This error called into question the quality of the A.C.E. survey results. Some of the key findings from the analyses are:

- An evaluation by Krejsa and Raglin (2001) was the first indication that A.C.E. seriously underestimated erroneous enumerations. This analysis revealed an additional net 1.9 million erroneously enumerated persons for those cases that could be resolved. These results are based on an independent reinterview and matching of about 70,000 E-sample persons. Because of the serious implications of this finding, a further Review Study was conducted.
- The findings from the Review Study by Adams and Krejsa (2001) showed that A.C.E. underestimated erroneous enumerations by a net of 1.45 million persons, which was smaller than the evaluation figure but still a significantly large amount. This figure does not include unresolved cases, so the estimated amount is probably somewhat higher. This study was based on a sample of

about 17,000 persons selected from the 70,000 person evaluation follow-up E sample. The most experienced analysts reviewed these cases using both the original A.C.E. person follow-up interviews as well as the reinterview results to determine their enumeration status.

- Mule (2001) showed that Census 2000 suffers from a large number of duplicate enumerations, i.e., persons who were double counted. Mule computer-matched census enumerations in A.C.E. block clusters to those across the entire country. The matching used by Mule was conservative in picking up census duplicates given his requirement for exact matching at the first stage. Within the A.C.E. block clusters, Mule found only 38 percent of the in-scope duplicates that A.C.E. found, leading us to believe that his matching algorithm was underestimating duplicates in the census. Note that A.C.E. was not designed to estimate duplicates outside the search area, and this, itself, was not a design flaw. A.C.E. was, however, expected to determine which census enumerations were erroneous because they were reported at the wrong residence. The design of A.C.E. Revision II accounts for duplicates outside the search area. Mule's study did not distinguish which of the duplicate pair was correct and which was erroneous, but one could easily speculate that half of these should be correct and half should be erroneous.
- Feldpausch (2001) examined the A.C.E. enumeration status for E-sample cases identified by Mule (2001) as duplicates outside the search area. Only 14 percent of the E-sample persons that were duplicates of a person in a housing unit were coded as erroneous by A.C.E. This was much lower than the expected 50 percent, indicating that A.C.E. underestimated erroneous enumerations due to not perceiving that these E-sample persons should have been counted at other residences. Note that these results suggest measurement error in the original A.C.E. figures released.
- Fay (2001, 2002) then compared the enumeration status for the E-sample Review Study cases to the duplicates identified by Mule (2001) outside the search area. Only 19 percent of the review cases that were duplicates of a person in a housing unit were coded as erroneous by the Review Study. Again, this was much less than the expected 50 percent, indicating that the evaluation data and the special review did not identify all the erroneous enumerations. Using these data, Fay then produced a lower bound on the level of unmeasured erroneous enumerations of 2.9 million.
- There was also evidence that similar problems may have affected the population sample (P sample) which is used to measure the omission rate. A.C.E. evaluation data from Raglin and Krejsa (2001) show that there are measurement errors in determining residency and mover status.

Using Fay's lower bound on the level of unmeasured erroneous enumerations, Thompson et al. (2001) produced a "Revised Early Approximation" of undercount for three race/Hispanic origin groups. These estimates were intended to be illustrative of net undercount and possible coverage differences. The same methodology and data were later used to expand the calculations to seven race/Hispanic origin groups. See Fay (2002) and Mule (2002) for details. These preliminary estimates show a very small net undercount. The data also indicate that the differential undercount has not been eliminated. These results are limited to the extent that they only provide information at the national level for broad population groups. Furthermore, these preliminary approximations were based on a small subset of A.C.E data and only partially correct for errors in measuring erroneous enumerations. Potential errors in measuring omissions were not accounted for.

In summary, the A.C.E. results were not acceptable because A.C.E. failed to measure large numbers of erroneous census enumerations. This was the reason for not using the A.C.E., but this does not mean that there were no other errors in the A.C.E. In particular, there was concern about P-sample cases that matched to enumerations suspected of being duplicates. If the E-sample case was erroneous, then that match cannot be valid. The extent of this problem was not quantified at the time of the ESCAP II decision. The level of other errors was small by comparison, and therefore, was not a major factor in this decision. See Hogan et al. (2002) and Mulry and Petroni (2002) for further information.

### Plans for Revising the 2000 A.C.E. Estimates

Even though the ESCAP recommended twice not to correct the census counts, they had concerns about differential coverage error in Census 2000 data. They thought it possible that further research resulting in revised estimates of coverage could potentially be used to improve the post-censal estimates. In addition, revised estimates would provide a better understanding of Census 2000 coverage error that could be used to improve census operations for 2010 and would help in developing better methodologies for the 2010 coverage measurement program.

The major objective was to produce improved estimates of the household population that could be used to measure net coverage error in Census 2000. This meant obtaining better estimates of erroneous census enumeration from the E sample and obtaining better estimates of census omissions from the P sample. Furthermore, since the national net undercount, as indicated by both DA and the

"Revised Early Approximations," was very close to zero and the census included large numbers of erroneous enumerations in the form of duplicates, it was imperative that the revised methodology carefully account for both overcounts and undercounts. Hogan (2002) summarized the major revision issues in the form of the following five challenges:

1. Improve estimates of erroneous census enumerations
2. Improve estimates of census omissions
3. Develop new models for missing data
4. Enhance the estimation post-stratification
5. Consider adjustment for correlation bias

There were no field operations associated with the A.C.E. Revision II process. Because of the late date, it was not feasible (or practical) to revisit households for additional data collection. Consequently, the revisions were based on data that had already been collected. One aspect of the strategy for revising the coverage estimates involves correcting measurement error using information from the A.C.E. evaluation data. This is referred to as the recoding operation. Another aspect of these corrections involves conducting a more extensive person duplicate study to correct for measurement error that was not detected by A.C.E. evaluations. This is referred to as the Further Study of Person Duplication (FSPD). The estimation method, discussed briefly in Chapter 2 and more fully in Chapter 6, is designed to handle overlap of errors detected by both of these studies to avoid overcorrecting for measurement error.

The recoding operation was designed to improve estimates of erroneous census enumerations and census omissions. It uses 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 Study[2] to correct for measurement error in enumeration status, residence status, mover status and matching status. This effort involved extensive recoding of about 60,000 P-sample cases and more than 70,000 E-sample cases.[3] An automated computer algorithm was used to recode most of the cases, but many required a clerical review by experienced analysts at the National Processing Center. The analysts had access to the questionnaire responses, as well as interviewer notes that put them in a better position to resolve apparent discrepancies. It was not possible to completely code all cases because of missing or conflicting information; however, the number of conflicting cases was relatively small.

The duplicate study was designed to further improve estimates of erroneous census enumerations and census omissions. This study used computer matching and modeling techniques to identify E- and P-sample cases that link to census enumerations across the entire country, including group quarters, reinstated, and deleted census cases. For the E-sample links, this study does not identify which enumeration is correct and which is the duplicate. For P-sample links, this study does not identify whether the correct Census Day residence is at the P-sample location or the census location. This information is used to model the probability that an E-sample linked case is a correct enumeration or that a P-sample case is a resident on Census Day.

New missing data models were developed to reflect the different types of missing data now possible as a result of the recoding operation. There were three new types of missing data to deal with:

1. P-sample households that were originally considered interviews, but the recoding determined that there were no valid Census Day residents,
2. cases with unresolved match, enumeration, or residency status because of incomplete or ambiguous interview data, and
3. cases with conflicting enumeration or residency status, because contradictory information was collected in the A.C.E. PFU and EFU interviews.

It was impossible to determine which data were valid for these cases. A household noninterview weighting adjustment using new cell definitions was used for type 1 above. Imputation cells and donor pools were developed for the second type of missing data based on detailed responses to the questionnaire. For the conflicting cases in type 3 above, there were no applicable donor pools, and probabilities of 0.5 were imputed for correct enumeration status and Census Day residency status. Fortunately, the recoding operation resulted in a relatively small number of these cases.

The revision effort incorporates separate post-strata for estimating census omissions and erroneous census enumerations because the factors related to each of these are likely to be different. The research effort focused on determining variables related to erroneous enumerations. This was because much of the previous work on developing post-strata focused only on the census omissions, and by default, the same post-strata were applied to the erroneous inclusions. For the E sample, some of the original post-stratification variables have been eliminated and additional variables have been included. Variables such as

[2]The PFU/EFU Review Study was not a planned evaluation. It was a special study conducted in a subsample of the evaluation data to resolve discrepancies between enumeration status in the PFU and EFU.

[3]These are probability subsamples of the original A.C.E. P and E samples and in the context of A.C.E. Revision II are called revision samples, but they are in fact equivalent to the evaluation follow-up samples.

region, Metropolitan Statistical Area/type of census enumeration area, and tract-level return rate were replaced by proxy status, type and date of census return, and household relationship and size. For the P sample, only the age variable was modified to define separate post-strata for children aged 0 to 9 and those 10 to 17. This was done because the DA estimates suggested different coverage for these groups. The estimated correct enumeration rates and estimated match rates are used to calculate Dual System Estimates (DSEs) for the cross-classification of the E and P post-strata.

The A.C.E. Revision II DSEs include an adjustment for correlation bias. Correlation bias exists whenever the probability that an individual is included in the census is dependent on 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. Since the intent of the A.C.E. Revision II is to estimate the net coverage error, it is important to carefully account for errors of omissions and errors of erroneous inclusions. In previous coverage measurement surveys, the erroneous inclusions were assumed to be much smaller than omissions. Consequently, not adjusting for correlation bias had the effect of understating the net undercount and relative to the census was a correction that was in the right direction, but just not big enough. In the presence of large numbers of overcounts, this assumption is no longer valid and it's possible that a correction might not even be in the right direction when the estimate is close to zero. For example, if there is a small true net undercount, it's possible to estimate an overcount because the DSE would underestimate population in the presence of correlation bias. Estimates of correlation bias were calculated using the "two-group model" and sex ratios from 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 less than 18 years of age, and that Black males have a relative correlation bias that is different from the relative correlation bias for non-Black males. The correlation bias adjustment is also done by three age categories: 18-29, 30-49, and 50 and over with the exception of non-Black males 18 to 29 years of age. This is because the A.C.E. Revision II sex ratios for non-Blacks 18-29 exceed the corresponding modified DA sex ratio and is likely a result of a data problem. This model further assumes that relative correlation bias is constant over male post-strata within age groups.

The DSEs, adjusted for correlation bias, are used to produce coverage correction factors for each of the cross-classified post-strata. These factors are applied or carried down within the post-strata to produce estimates for geographic areas such as counties or places. This process is referred to as synthetic estimation. The key assumption underlying this methodology is that the net census coverage, estimated by the coverage correction factor, is relatively uniform within the post-strata. Failure of this assumption leads to synthetic error.

# Chapter 2.
# Summary of A.C.E Revision II Methodology

## INTRODUCTION

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 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 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. The purpose of this chapter is to present a high-level overview of the process used to produce A.C.E. Revision II estimates of the population coverage of Census 2000. Further details concerning the methodology and procedures are included in subsequent chapters.

### Background

The chronology of events leading to the corrected A.C.E. Revision II 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 11,303 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. The MES measured errors introduced when the census and A.C.E. interviews were matched. The EFU, which was designed to study unusual living situations, entailed another interview. It evaluated the Census Day residency, enumeration status and mover status assigned during the A.C.E. interview and A.C.E. Person Follow-up (PFU) interview. The MES and EFU were conducted in a subsample of 2,259 block clusters selected from the original 11,303 block clusters. A further subsample of persons within these block clusters was selected for the EFU evaluation.

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. At this point the A.C.E. Revision II program commenced. The Revision E and P samples were developed for purposes of producing A.C.E. Revision II estimates. They are each comprised of about 70,000 sample persons. 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. 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.

The A.C.E. Revision II process is presented below. First, the corrections for measurement error (undetected erroneous enumerations and P-sample nonresidents) in the Revision Samples are explained. Then, a discussion is given of the missing data methods applied to cases whose match, residency or enumeration status had changed in the Revision Samples. Next, the process for identifying census duplicates across the entire nation is discussed. An applicable dual system estimation formula that incorporates these changes and accounts for correlation bias is presented. Finally, synthetic estimation was employed to produce A.C.E. Revision II results. See Kostanich (2003) for a summary of the methodology.

## CORRECTING MEASUREMENT ERROR IN THE REVISION SAMPLES

As previously stated, the original A.C.E. process (step 1. above) failed to detect significant numbers of erroneous census enumerations (EEs). These undetected EEs (one part of "measurement error" in the A.C.E.) were uncovered during the evaluations of the A.C.E. (step 2. above). In general, the original A.C.E. Person Interview (PI) and PFU, the EFU interview, the MES, and the PFU/EFU Review results were used to correct for measurement error in the enumeration, residency, mover, and match statuses for subsamples of the Full A.C.E., called the Revision E and P samples. No additional data were collected in this measurement error correction process.

The Revision 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 cases in the 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 that described the reason why the particular code was assigned.

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

1. **Validation.** Determine, for categories of 'Why' codes, if the automated coding was of high quality based on level of agreement with the Review data.
2. **Targeting.** Target only those 'Why' code categories that had codes produced by automated coding that had low levels of agreement with the Review data.
3. **Clerical coding.** Clerically recode only cases in the targeted 'Why' code categories. The clerical recoding took advantage of handwritten interviewer comments.

In general, cases did not go to clerical review if both the PFU and EFU automated codes agreed, the mover statuses also agreed, 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, residency, and mover status, the results of the MES were used to correct for false matches and false nonmatches. Some matching errors were a result of incorrect residency 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. See Krejsa and Adams (2002) for further details.

## 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. In addition, imputation methods were used to adjust for missing characteristics such as age or tenure, as well as enumeration, residency, and match status. For the A.C.E. Revision II work, 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 Revision E and P samples, however, there were three new types of missing data to deal with:

1. Noninterviewed households: Revision P-sample households that were considered interviews in the A.C.E. P sample, but were identified as noninterviews in the Revision coding because it was determined that there were no valid Census Day residents;
2. Revision E- or P-sample cases with unresolved match, enumeration, or residency status because of incomplete or ambiguous interview data;
3. Revision E- or P-sample cases with conflicting enumeration or residency status. This occurred when contradictory information was collected in the A.C.E. PFU and the EFU interviews and it could not be determined which was valid.

### Household Noninterview Adjustment for the Revision P Sample

For the original March 2001 A.C.E. estimates, the household noninterview adjustment generally spread the weights of the Full P-sample noninterviewed housing units over interviewed housing units in the same block cluster with the same housing unit structure type. The methodology for the Revision P-sample household noninterview adjustment for Interview Day was essentially unchanged from that used for the Full P sample. There was, however, an important change for the noninterview adjustment for Census Day residency. A separate cell was defined for new noninterviews 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.

### 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 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.

Revision 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 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.

For Revision P- and E-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 Revision 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.

### Imputation for Revision E- or P-Sample Conflicting Cases

When the A.C.E. PFU and 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 better and assign an appropriate 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.

## FURTHER STUDY OF PERSON DUPLICATION

Earlier work showed that correcting measurement error by recoding was not going to correct all the missed erroneous enumerations. 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 census erroneous enumerations was duplicate census enumerations; that is, 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 (the area where census and A.C.E. person matching was conducted). However, there was an expectation that the A.C.E. would detect that these E-sample enumerations had another residence and that roughly half the time this other residence was the usual residence. Feldpausch (2001) showed this expectation was not met.

For purposes of constructing A.C.E. Revision II estimates, the study of person duplication used matching and modeling techniques to identify duplicate links between the Full E and P samples to census enumerations. Links to group quarters, reinstated, deleted and E-sample eligible records throughout the entire nation were allowed. 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.

### 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.

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 household members in the sample housing unit to all household members in the census housing unit.

### 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. See Chapter 5 for a full discussion on the person duplication study.

## THE A.C.E. REVISION II DSE FORMULA

With the correction of measurement error in the Revision E and P samples, the adjustment for missing data in the Revision E and P samples, and the determination of census

duplicate links between the Full E and P samples and census enumerations, the dual system estimation formula can be applied. The following sections explain the formula and its adjustment for the A.C.E. Revision II work.

Using procedure C for movers and different post-strata for the E and P samples, the DSE formula can be written as:

$$\mathrm{DSE}^{\mathrm{C}}_{ij} = (C\mathit{en}'_{ij} - \mathrm{II}'_{ij}) \frac{\left[\frac{\mathrm{CE}_i}{\mathrm{E}_i}\right]}{\left[\frac{\mathrm{M}_{nm,j} + \left[\frac{\mathrm{M}_{om,j}}{\mathrm{P}_{om,j}}\right]\mathrm{P}_{im,j}}{\mathrm{P}_{nm,j} + \mathrm{P}_{im,j}}\right]}$$

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

$$\mathrm{ReDSE}^{\mathrm{C}}_{ij} = (C\mathit{en}'_{ij} - \mathrm{II}'_{ij}) \frac{\left[\frac{\mathrm{CE}_i^{\mathrm{ND}} f_{1,i'} + \mathrm{C}\widetilde{\mathrm{E}}_i^{\mathrm{D}}}{\mathrm{E}_i}\right]}{\left[\frac{\mathrm{M}^{\mathrm{ND}}_{nm,j} f_{2,j'} + \widetilde{\mathrm{M}}^{\mathrm{D}}_{nm,j} + \left[\frac{\mathrm{M}_{om,j} f_{3,j'}}{\mathrm{P}_{om,j} f_{4,j'}}\right](\mathrm{P}_{im,j} f_{5,j'} + g(\mathrm{P}^{\mathrm{D}}_{nm,j} - \widetilde{\mathrm{P}}^{\mathrm{D}}_{nm,j}))}{\mathrm{P}^{\mathrm{ND}}_{nm,j} f_{6,j'} + \widetilde{\mathrm{P}}^{\mathrm{D}}_{nm,j} + \mathrm{P}_{im,j} f_{5,j'} + g\,(\mathrm{P}^{\mathrm{D}}_{nm,j} - \widetilde{\mathrm{P}}^{\mathrm{D}}_{nm,j})}\right]}$$

Recall that the II´ term excludes the late census adds.

| Notation | | |
|---|---|---|
| Terms | CE | Correct enumerations |
| | E | E-sample total |
| | M | Matches |
| | P | P-sample total |
| | f | 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 Procedure 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 |

## Adjustment for Duplicates using the Duplicate Study

The first task was to adjust the usual dual system estimate formula for those cases that have a link to a census enumeration outside the A.C.E. search area. P- and E-sample cases with links to census enumerations 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.

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 was necessary to estimate two conditional probabilities, which are reflected for the E sample in $\mathrm{C}\widetilde{\mathrm{E}}_i^{\mathrm{D}}$. In the P sample, these probabilities are reflected in the nonmover terms $\widetilde{\mathrm{P}}^{\mathrm{D}}_{nm,j}$ and $\widetilde{\mathrm{M}}^{\mathrm{D}}_{nm,j}$.

## Adjustment for Measurement Error Using the Revision E and P Samples

Next, an adjustment is made for other measurement errors not accounted for by the duplicate study. This adjustment was applied only to nonduplicate terms to avoid over-correction due to any overlap between the duplicate study and correction of measurement error.

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 Samples. This recoded data from the Revision 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 E or P Revision Sample and were a ratio of an estimate using the Revision coding to the estimate using the original coding. These adjustments were done by measurement error correction post-strata $i'$ or $j'$ and are denoted by the $f$ terms 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 $\mathrm{P}^{\mathrm{D}}_{nm,j} - \widetilde{\mathrm{P}}^{\mathrm{D}}_{nm,j}$ is an estimate of nonresidents among nonmovers with duplicate links.

## 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 non-Black 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 non-Black.

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 have been made would have been 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. See U.S. Census Bureau (2003) for technical details.

## SYNTHETIC ESTIMATION

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

$$\mathrm{CCF}_{ij} = \frac{\widetilde{\mathrm{R}}e\mathrm{DSE}^{\mathrm{C}}{}_{ij}}{Cen_{ij}}$$

where:

$\widetilde{\mathrm{R}}e\mathrm{DSE}^{\mathrm{C}}_{ij}$'s are the correlation bias adjusted DSEs for post-strata $ij$.

$Cen_{ij}$'s are the census counts for post-strata $ij$, including late census adds.

A coverage correction factor was assigned to each post-stratum. The post-strata excluded persons in group quarters or in remote Alaska. Effectively, these persons have a coverage correction factor of 1.0. In dealing with duplicate links to group quarters persons, the person in the group quarters was treated as if (s)he was a correct enumeration or as if this was their correct residence on Census Day. A synthetic estimate for any area or population subgroup b is given by:

$$\widetilde{\mathrm{N}}_{\,b} = \sum_{ij \in b} Cen_{b,ij}\,\mathrm{CCF}_{\,ij}$$

# Chapter 3.
# Correcting Data for Measurement Error

## INTRODUCTION

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 involves correcting measurement errors that cannot be detected with the information available in the subsample. These additional errors, which are identified via a duplicate study, are discussed in Chapter 5.

To understand the measurement error correction process, it is important to be familiar with the various sources of available information. These are summarized in the following table.

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 11,303 block clusters. The Matching Error Study (MES) and the Evaluation Follow-up (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 11,303 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 conducted 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

Table 3-1. **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 11,303 block clusters | A.C.E. Person Interviewing (PI), Summer 2000<br>A.C.E. Person Follow-up (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 Follow-up (EFU) | EFU E and P samples[1] | E: About 77,000 persons in 2,259 block clusters<br>P: About 61,000 persons in 2,259 block clusters | Evaluation Person Follow-up (EFU), January - February, 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<br>P: About 61,000 persons in 2,259 block clusters | Recoding Operation, Summer 2002 |

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

error. This chapter discusses the measurement error corrections made to the E- and P- Revision samples. These corrected 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.

## GOALS AND BACKGROUND

The goal for A.C.E. Revision II was to correct as much measurement error as possible in the original A.C.E. estimates, given resource and timing constraints.[2] The primary sources of measurement error were determining residence and enumeration status, match status, and mover status.

**Residence and Enumeration Status.** The original A.C.E. did not detect all of the erroneous enumerations. See Adams and Krejsa (2001) and Fay (2002) for documentation. The Evaluation Follow-up (EFU) detected approximately 1.4 million additional erroneous enumerations in the E sample. Since the coding of enumeration status in the E sample was identical to the coding of residence status in the P sample, similar results for P-sample residence status coding were expected (i.e., additional nonresidents were expected to be found as a result of the EFU). To correct for the residence status errors, the A.C.E. Revision II utilized a recoding of the Evaluation Follow-up Interview in combination with the original A.C.E. to determine the best residence or enumeration status for each person in the Revision sample.

**Matching Error.** The Matching Error Study showed a net difference in match codes between the original March 2001 matching results and the evaluation matching results of 0.41 percent in the E sample and 0.20 percent in the P sample. Bean (2001) suggested this net difference translated into an increase in the dual system estimate of 483,938 people. To correct for matching error, results of the Matching Error Study and the A.C.E. Revision II recoding were used in conjunction to determine the appropriate match status for each person.

**Mover Status.** Raglin and Krejsa (2001) estimated a 2.6 percent gross difference rate in the mover status between the original A.C.E. and the Evaluation Follow-up. This translated into a negative bias of 465,000 in the DSE (assuming no other biases). Results of the Evaluation Follow-up were used to correct for mover status errors. The EFU questionnaire contained questions designed to probe for a person's mover status. This information was captured during the clerical recoding and during the initial coding of the Evaluation Follow-up form. These types of measurement errors were corrected either by computer or clerically.

Two other sources of error were not part of the measurement error recoding portion of the A.C.E. Revision II. These errors included geocoding errors and duplicates outside the search area. Certain geocoding errors detected by various geocoding evaluations were not included in the A.C.E. Revision II.[3] Within the P sample, 245,926 production nonmatched residents were found outside the search area[4] and 195,321 production correct enumerations in the E sample were found outside the search area. See Adams and Liu (2001). Some of the correct enumerations outside the search area were identified by the EFU interview and, hence, were reflected in the revised coding.[5] Duplicates found outside the search area as a result of computer matching (see Chapter 5) were not handled by clerical coding. They were accounted for in the dual system estimator using estimation techniques. See Chapter 6 for a full description of the estimation techniques.

## RESIDENCE STATUS AND ENUMERATION STATUS

As already noted, the original March 2001 A.C.E. underestimated the number of erroneous enumerations. To correct for this, the best residence status code was based on available field follow-up data. Duplicates were corrected using a separate process. The following data were available for measurement error correction:

- **Person Interview (PI).** The PI was the original A.C.E. enumeration of the P sample. It was a Computer-Assisted Personal Interview questionnaire designed to fully enumerate persons in the A.C.E. It was conducted by either phone or personal visit between April and September, 2000.
- **Person Follow-up (PFU).** The PFU was the follow up used to assign residence and enumeration status, whenever those items were not determined, after the before follow-up matching (Childers, 2001). It was conducted by personal visit in October and November, 2000, approximately 6-7 months after Census Day.
- **Evaluation Follow-up (EFU).** The EFU was an evaluation of the A.C.E. designed to detect unusual living situations using additional probes and additional interviewing techniques (e.g., flashcards). It was conducted by personal visit in January and February, 2001, approximately 9-10 months after Census Day.

[2] In order to complete the A.C.E. Revision II estimates on time, 12 weeks were allotted for coding. Analysts at the National Processing Center were expected to code approximately 25,000 cases in this time frame.

[3] As part of the A.C.E., several evaluations of geocoding error were conducted on various subsamples of the A.C.E., most notably Targeted Extended Search 2 (TES2) and Targeted Extended Search 3 (TES3). Results of these evaluations can be found in Adams and Liu (2001).

[4] For the 2000 A.C.E., the search area, or area in which a person can be considered a correct enumeration or match, was the cluster and any census block touching the cluster.

[5] Some of the cases in TES2 were evaluated using the Evaluation Follow-up questionnaire. For these cases, results of the geocoding evaluation were included in the Evaluation Follow-up; however, if a case was in TES2, but not in the Evaluation Follow-up, no geocoding evaluation results were included.

Results of the Person Interview were used to assign A.C.E. residence status by computer to all people in A.C.E. who did not need follow-up. In contrast, the PFU was used to assign residence status for anyone who was eligible for follow-up (Childers, 2001). The PFU is similar to the PI. The PFU process interviewed both P-sample and E-sample people. The EFU followed up a sample of people sent to PFU and a sample of those not sent to PFU. This allowed the residence/enumeration status of a representative sample of people eligible for field follow up to be evaluated.

There were measurement errors in both the A.C.E. PFU and EFU resulting from limitations of their respective interviews. These errors are documented in Bean (2001) and Adams and Krejsa (2001), respectively. Also, the EFU was not strictly coded according to census residence rules. To evaluate the E sample for ESCAP II, the Census Bureau conducted the PFU/EFU Review in the summer of 2001. Expert matchers reviewed a subsample of the EFU E sample and applied consistent census residency rules. These analysts were assumed to make negligible errors; therefore, the PFU/EFU Review was considered to be free of coding error, given available data.

For A.C.E. Revision II, this high-quality coding was needed for subsamples of the A.C.E. P and E samples that were large enough to provide accurate subgroup estimates of net coverage. Twelve weeks coding time were allotted to clerically code approximately 25,000 cases. However, there were over 100,000 cases needing codes. To assign the highest quality codes, while meeting a demanding schedule, keyed data from both the PFU and EFU forms were used to augment clerical coding procedures. An automated coding algorithm, based on specific responses to the PFU and 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 that describes the reason why the particular code was assigned. There were more than 60 possible 'Why' code categories. A final code was assigned to each case using the following three-step process:

- **Validation.** Determine for each category of 'Why' code if the automated coding is of high quality using the PFU/EFU Review as a truth deck.
- **Targeting.** Target only those 'Why' code categories that have low levels of agreement between the automated coding and the PFU/EFU Review data.
- **Clerical Review.** Clerically recode only those cases in the targeted 'Why' code categories. The clerical recoding takes advantage of handwritten interviewer comments.

### Validation of Keyed Data

To validate the quality of coding produced by the keyed data algorithm, skip patterns for both questionnaires were programmed to determine an appropriate match code and 'Why' code for each case. Then, for both the PFU and EFU forms, the percentage agreement with the original coding (either production coding or the coding of the EFU form) for the respective form, the percentage agreement with the PFU/EFU Review, and the residual risk were examined. That is, the following calculations were performed twice - once for PFU and once for EFU.

The residual risk of disagreement (i.e. potential bias) represented the number of cases at risk for being coded wrong due to accepting the automated code for categories defined by questionnaire responses. Cases subject to risk were those where the automated code and original code agreed. If they disagreed, the automated code was rejected and the case was sent for clerical review. The risk for the cases agreeing is calculated as follows:

$$\text{risk} = \text{Agree}_{K} - \text{Agree}_{Rev}$$

where

$\text{Agree}_{K}$ = The weighted number of cases whose code from the keyed data agreed with the original production code.

$\text{Agree}_{Rev}$ = Of those cases where the code from the keyed data agreed with the original production code, the weighted number of cases whose code from the keyed data agreed with the PFU/EFU Review code.

The term risk, rather than an error, is used because some potential coding changes may not have had an effect on the DSE. For example, people who were in group quarters have a residual risk of 26,517 after computer coding. These represent cases that probably should have been coded as erroneous enumerations, but were not. However, some of the 26,517 cases could be unresolved, which have a probability less than one of being correct.

The automated coding results for a given 'Why' code category were rejected if the residual risk was too high or if there were not enough cases to make an informed decision. The exception to this rule was the category consisting of cases without any indication of living in a group quarters or other residence. This group was, by far, the largest category for both the PFU and EFU, so a higher residual risk[6] was expected.

### Targeting Cases for Clerical Review

After the decision was made to accept or reject the automated code for each 'Why' code category, cases were targeted for clerical review. Analysts, who were the highest level of clerical matchers, performed the clerical review. Due to their experience and additional training, they were assumed to make negligible errors in coding.

[6]Absolute risk, rather than relative risk, is used. Therefore, larger categories tended to have higher risks.

In general, cases did not go to clerical review if both the PFU and EFU automated codes agree, the mover statuses agree, and the 'Why' code category was deemed to be of high enough quality. In some instances, cases are 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. These cases included:

- **Census Usual Home Elsewhere.** If the person claimed a Usual Home Elsewhere on certain types of census forms, they were counted as a correct[7] enumeration within the cluster and did not need clerical review.
- **Geocoding Errors from Initial Housing Unit Matching.** If a case should not have been sent to PFU or EFU and was only sent due to clerical error in the initial production matching, then it did not need clerical review.

In contrast, some cases are automatically sent to clerical review. For example, this includes cases in the PFU/EFU Review that resulted in a conflicting status, noninterview cases, or cases where mover dates could not be determined from the EFU keyed data. Some of the cases that went to clerical review 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 review because the automated coding process was not reliable for that 'Why' code category.

For P-sample inmovers, there was no validation data. Cases where the original EFU mover status did not match the mover status from the keyed data, or the residence status from the keyed data did not match the original EFU residence status, were sent to clerical review. Noninterview cases or cases where mover dates could not be determined from the keyed data were also sent to clerical review.

Cases with the following attributes were sent to clerical review:

- the code from the keyed data for either form was not accepted for that case.
- the code from the keyed data was accepted for both forms, but at least one of the codes from the keyed data did not agree with its original code (i.e., the PFU code from the keyed data did not agree with production or the EFU code from the keyed data did not agree with the original EFU code).
- for P-sample people, the mover status from the keyed data did not agree with mover status assigned during the EFU coding.
- there was write-in information in open-ended questions on the form that could not be coded.
- the case was a possible match in before follow-up matching and the production and original EFU code disagreed.
- the case was a duplicate in either the original EFU coding or production after follow-up coding.
- the case was not yet flagged for clerical review and the PFU code from the keyed data did not agree with EFU code from the keyed data, and one of the cases was not unresolved for certain reasons.
- the case was in the PFU/EFU Review and was conflicting or had a mover status disagreement between the keyed data and the original EFU mover status.

### Clerical Review

The clerical review for A.C.E. Revision II was an analyst-only operation. The following data were collected:

- Match Code for each form
- 'Why' Code for each form
- Respondent for each form
- Whether the respondents are the same for the two interviews
- Best Code. A code indicating which form is the better of the two forms
- Smooshed Code. Information from both forms combined to make a code to represent the true situation
- Mover Status. Mover Status from the EFU form for P-sample people

The match codes were assigned using the census residence rules to construct coding rules for the flow of the questionnaire.

The best code could be one of four values:

- Both = The enumeration statuses were the same
- PFU = The PFU form provided better information
- EFU = The EFU form provided better information
- Conflicting = Similar caliber respondents (e.g., husband and wife; two neighbors) provided contradictory information for the case

To ensure reproducibility, computer edits were applied to the best code. If the analyst did not follow pre-specified rules, then the analyst had to review the case again or leave a note indicating the situation.

[7] A person can claim a usual home elsewhere if he or she is enumerated on certain types of census forms in group quarters (e.g. military, shipboard, and certain types of special places like shelters). If a person on one of these forms claims a usual home elsewhere, then that person is counted at the address they indicate is their usual home. These people are part of the E sample because they are part of the housing unit universe.

## CORRECTION OF MOVER STATUS ASSIGNMENT ERRORS

For each P-sample case, mover status was based on the EFU. This was used to determine whether or not the person needed clerical review.

## CORRECTION OF MATCHING ERRORS

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 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.

## DATA OUTPUTS

After the clerical operation was completed, two files were assembled - one for the P sample and another for the E sample. The files contain match codes and 'Why' codes (where appropriate) for original March 2001 A.C.E., EFU, PFU/EFU Review, Keyed Data, and A.C.E. Revision II Clerical Review. A final code is also assigned in the following hierarchy: A.C.E. Revision II Clerical Review, PFU/EFU Review, Keyed Data. This code reflects the final match, residence, and enumeration status for the A.C.E. Revision II process.

## LIMITATIONS

There were several limitations on the data for the A.C.E. Revision II:

- **Sample Size.** The sample used to estimate measurement error is 2,259 clusters, containing about 10 percent of the persons in the sample used in the production A.C.E. Due to the smaller sample size, some subgroup estimates are subject to higher variances compared to those for the original March 2001 A.C.E.
- **Conflicting Cases.** Conflicting cases occurred when the PFU and EFU interviews had respondents of the same caliber (either both nonproxy or proxy respondents who were in the position to have similar knowledge about the household, e.g. two neighbors) who gave contradictory information. Since an additional field follow-up was not possible, these cases were coded as conflicting, were reviewed separately, and imputed.
- **Data Collection Error.** Cases were coded as best as possible. However, there was no attempt to correct for any residual data collection error. Any remaining respondent and interviewer errors could not be rectified without an additional field follow-up.

# Chapter 4.
# A.C.E. Revision II Missing Data Methods

## BACKGROUND

Missing data arises because it is not possible to obtain interviews for all sample cases or to obtain answers to all interview questions. This was as true for the A.C.E. Revision II, as it was for the A.C.E. To put the A.C.E. Revision II missing data methods in perspective, a brief summary of the A.C.E. missing data adjustments is presented. For the A.C.E. P sample, a household noninterview adjustment compensated for noninterviewed households. Imputation methods were implemented to handle missing characteristics such as age or tenure. Further, match and residency probabilities were assigned when the respective match and residency statuses could not be definitively determined. There was no noninterview adjustment for the A.C.E. E sample, nor was there an imputation for missing characteristics as the census imputations were used. However, E-sample cases with unresolved enumeration status were assigned probabilities of correct enumeration. See Ikeda and McGrath (2001) for details on the A.C.E. missing data methodology.

As will be discussed in Chapter 6, the A.C.E. Revision II estimation utilizes both the original A.C.E. coding results on the Full E and P samples and the Revision coding results on the smaller Revision Samples. Note that the A.C.E. Revision II subsample of the A.C.E. is referred to as the Revision Sample and the new coding operation is called the Revision coding. The missing data adjustments for the A.C.E. E and P samples were unchanged from those used to produce A.C.E. estimates, with the exception of the imputation for missing age. It was necessary to impute age again for the Full A.C.E. P sample because the A.C.E. Revision II post-strata had different age groupings.

The Revision P sample used the same imputations for missing characteristics that the A.C.E. did, including the new age imputation. However, since A.C.E. Revision II measurement methodology had important differences from the A.C.E. measurement methods, it was necessary to develop new missing data methods. The A.C.E. Revision II missing data confronted three general types of new missing data problems:

1. New noninterviewed households: Revision P-sample households that were considered interviews in the A.C.E. were identified as noninterviews in the Revision coding when it was determined that none of the P-sample people there were valid Census Day residents.
2. Revision E- and P-sample cases with unresolved match, enumeration, or residency status, because of incomplete or ambiguous interview data from the Person Follow-up (PFU) or the Evaluation Follow-up (EFU).
3. Revision E- or P-sample cases with 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.

## AGE IMPUTATION

For the original A.C.E., P-sample people with missing age were assigned to age categories defined by the post-stratification plan. The A.C.E. Revision P-sample post-stratification divided the original A.C.E. post-stratification group of 0-17 year olds into two age groups: 0-9 and 10-17. 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.

## HOUSEHOLD NONINTERVIEW ADJUSTMENT

The A.C.E. household noninterview adjustment generally spread the weights of P-sample noninterviewed housing units over interviewed housing units in the same block cluster with the same housing unit structure type. Housing units were determined to be noninterviews in two ways: 1) an interview was not conducted during the A.C.E. person interview operation, and 2) 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 noninterview adjustments were implemented for Census Day and A.C.E. Interview Day.

The A.C.E. Revision II noninterview adjustment methodology for A.C.E. Interview Day was essentially unchanged from the A.C.E. There was, however, an important change from the A.C.E. methodology for the noninterview adjustment for Census Day residency. In A.C.E. Revision II, a new imputation cell was defined. It included new noninterviews due to whole households of A.C.E. nonmovers who were determined to be inmovers or nonresident outmovers by the Revision coding. The new noninterview cell

spread the weights of these noninterviewed units over housing units with at least one person who: 1) indicated he/she lived at another address, or 2) was identified as potentially fictitious in the A.C.E. These new noninterviews were assumed to have both a low match rate and a low residency rate similar to this group. Otherwise, the noninterview adjustment for Census Day used methodology similar to that of the A.C.E.

## 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 occurred in one of two ways. Firstly, the A.C.E. person interview may not have provided sufficient information for match and follow-up. Secondly, the A.C.E. PFU may not have collected adequate information for determining a person's Census Day residency status or their match status. Inadequate data collection can also result in unresolved enumeration statuses for A.C.E. E-sample people. In the A.C.E. Revision II, the EFU was also the source of unresolved cases. How a case was imputed depended on how it became unresolved.

### Imputation for People with Insufficient Information for Match and Follow-Up

The Revision P-sample people with insufficient information for match and follow-up tended to be the same people who had insufficient information for match and follow-up 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 not sent to EFU. There were about three million weighted people with insufficient information for match and follow-up in both the Full and Revision P samples.

In the A.C.E., P-sample people with insufficient information for match and follow-up were assigned a probability of Census Day residency equal to the residency rate of P-sample people who went to PFU. This methodology was improved in the A.C.E. Revision II by defining finer imputation cells that accounted for whether or not the housing unit was matched, nonmatched, or had a conflicting household. A conflicting household existed when the P- and E-sample households had no people in common.

The probability of match was assigned based on the overall match rate, divided into groups based on mover status and housing unit match status, as was done in the A.C.E., and additionally on conflicting household status.

### Imputation for People with Incomplete or Ambiguous Follow-Up

In contrast to P-sample people with insufficient information, the residency status for Revision P-sample people and the correct enumeration status for Revision E-sample people often changed from what it was in the A.C.E. These statuses changed because the Revision coding processed new information from the EFU, in addition to the original information from the PFU. Thus, while the EFU information resolved many cases that were unresolved in the A.C.E. because of the PFU, EFU cases with incomplete or ambiguous information were a new source of unresolved cases. There were about the same number of weighted E-sample unresolved cases in the Revision sample as in the A.C.E., more than six million, with about half of these representing new unresolved cases. In contrast, the Revision coding generated substantially more P-sample unresolved cases than the A.C.E., 4.6 million compared to 2.7 million. This increase was due to the fact that all Revision P-Sample cases (except those with insufficient information) went to EFU, including whole households of nonmatched people who had not gone to PFU. These people were assumed to be resolved in the A.C.E. and could have become unresolved because of the EFU.

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

The after follow-up group definitions were based on keyed responses to the PFU and EFU questionnaire checkboxes and the 'Why' codes. 'Why' codes were clerically-applied codes that reflected responses in the questionnaire checkboxes, as well as handwritten notes. See Adams and Krejsa (2002) for a detailed description. The keyed results and 'Why' codes helped identify the following:

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

PFU after follow-up groups were defined for those cases that were unresolved as a result of the PFU.

Similarly, EFU after follow-up groups were defined for those cases unresolved because of the EFU. It was necessary to define separate groups for the PFU and EFU, because their interviews and questionnaires were different. However, the same after follow-up groups were

employed for the P- and E-sample unresolved cases, as the PFU and EFU questions about Census Day residency were the same as the PFU and EFU questions about enumeration status.

It is useful to distinguish between uninformative and informative unresolved cases:

- uninformative unresolved: the follow-up was a noninterview or an incomplete interview, though there was no evidence of an erroneous enumeration or nonresident.
- informative unresolved: a follow-up interview was conducted, and there was evidence of an erroneous enumeration or nonresident.

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

It often happened that both the PFU and EFU interviews were unresolved. To assign this case to an imputation cell, the unresolved interview that was more informative was selected. When both interviews had the same level of information, the EFU was typically selected over the PFU, because questions on the EFU questionnaire were more sharply defined.

Consider the following example of an after follow-up group. One cell of unresolved E-sample people or recipients was defined as people with evidence from the EFU interview that they had moved in since Census Day, or moved out before Census Day, though the EFU interview did not provide the address they moved to or from. It was impossible to determine the enumeration status of these people, since it was unclear if their Census Day address was in the A.C.E. cluster. The corresponding donor pool consisted of those resolved people who indicated in the EFU that they had moved in after Census Day or moved out before Census Day. Generally, these people provided their mover address in the EFU. An analogous after follow-up group was formed for people unresolved because they indicated they were movers in the PFU interview. These groups are characterized as informative, because the follow-up provided evidence of an erroneous enumeration.

Table 4-1 shows the nine EFU after follow-up groups, while Table 4-2 shows the nine PFU after follow-up groups. People who moved in after Census Day or moved out before Census Day were the largest informative after follow-up group. Another important informative after follow-up group consisted of people who, according to the follow-up, had another residence such as a vacation home, though the follow-up did not indicate whether the other residence or the sample address was the Census Day residence. The noninterview groups and 'didn't answer other residence questions' group were the larger uninformative groups.

Table 4-1. **EFU After Follow-up Groups**

| Informative groups |
|---|
| 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 the sample address or other residence was the Census Day residence. |
| Followed up person moved in or moved out, but no move dates given. |
| **Uninformative groups** |
| The respondent indicated the followed up person 'Lived here' at the sample residence, but did not answer the 'other residence' question. |
| The respondent answered the current residence question, but did not answer the group quarters and 'other residence' questions. |
| 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. |

Some of the larger EFU groups 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 follow-up groups tended to have imputed probabilities of correct enumeration or residence close to one, typically in the range of 0.92 to 0.99. In contrast, the informative after follow-up groups had smaller probabilities, often less than 0.25. The probability of correct enumeration is calculated as the weighted proportion of correct enumerations in the donor pool. For example,

$$\text{Probability of correct enumeration} = \frac{\text{Weighted CE's in Donor Pool}}{\text{Weighted Resolved Enumerations in Donor Pool}}.$$

For the P sample, probabilities of residency and match status were calculated analogously.

Table 4-2. **PFU After Follow-up Groups**

| Informative groups |
|---|
| 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 was not given. |
| The respondent indicated the followed up person 'did not live here' at the sample address, but did not indicate the other address and did not answer the group quarters and 'other residence' questions. |
| The followed up person had an 'other residence,' but did not indicate where the usual residence was. |
| **Uninformative groups** |
| The respondent indicated the followed up person 'Lived here' at the sample residence, but did not answer the 'other residence' question. |
| The respondent answered the usual residence question, but did not answer the group quarters and 'other residence' questions. |
| The 'lived here' question is Don't Know/refused, and the group quarters and 'other residence' questions were not answered. |
| Blank questionnaire. |
| Potentially fictitious person, no respondents knew of the followed up person. |

### Imputation for Conflicting Coding Cases

When the A.C.E. EFU and PFU interviews had contradictory information, the Revision coding procedure assigned the case a conflicting code. Note that a conflicting code is different than a conflicting household. All conflicting cases in the Revision coding process were sent to analysts for clerical review. By examining the handwritten notes of interviewers, analysts could often determine which of the two interviews was better and assign the appropriate code. 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 proxies of equal caliber. For these conflicting cases, the interviews seemed equally likely to be correct based on the analyst's expertise. Therefore, the probability of correct enumeration for Revision E-sample conflicting cases and the probability of Census Day residency status for Revision P-sample conflicting cases were assigned to be 0.5. It should be noted that the Revision coding resulted in considerably fewer conflicting cases than the PFU/EFU Review Sample. According to Adams and Krejsa (2001), the PFU/EFU Review Sample had about 2.6 million weighted conflicting people in contrast to only about 100,000 weighted conflicting people in the Revision Samples.

# Chapter 5.
# Further Study of Person Duplication in Census 2000

## INTRODUCTION

Evaluations of the March 2001 coverage estimates indicated the A.C.E. failed to detect a large number of erroneous census enumerations. One type of these census erroneous enumerations was duplicate census enumerations; that is, 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 search area. However, there was an expectation that the A.C.E. would detect that these E-sample enumerations had another residence, and that, roughly half the time this other residence was the usual residence. Feldpausch (2001) showed this expectation was not met.

For purposes of constructing A.C.E. Revision II estimates, matching and modeling techniques were used to identify duplicate links between the Full E and P samples to census enumerations. 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 are identical, modeling techniques were used to assign a measure of confidence, the duplicate probability, that the two records refer to the same individual. These duplicate probabilities were used in the A.C.E. Revision II estimates.

This chapter documents the matching and modeling methods that were used to identify duplicate links and to produce duplicate probabilities. Note that this study was not intended to identify which enumeration was in the correct location. Chapter 6 describes how to compute the conditional probability that the sample case was in the correct location given that it had a link to a census enumeration outside the A.C.E. search area. This calculation impacts the correct enumeration status in the E sample and the residence status in the P sample. A full discussion of the estimation components is given in Chapter 6.

## BACKGROUND

Mule (2001) reported results for initial attempts at measuring the extent of person duplication in Census 2000. This work was conducted by an inter-divisional group as part of the further research to inform the ESCAP II decision on adjusting census data products. This study is referred to as the ESCAP II duplicate study in this chapter. The ESCAP II duplicate study used conservative computer matching rules to minimize the number of false matches that could be introduced when doing a nationwide search, since there was no clerical review of the results. As a consequence of the matching rules, comparisons to benchmarks indicated that the ESCAP II duplicate estimates were a lower bound. Specifically, comparing the ESCAP II results within the A.C.E. sample area to the A.C.E. clerical matching results showed that only 37.8 percent of the census duplicates were identified. Fay (2001, 2002) estimated the matching efficiency at 75.7 percent when accounting for the census records out-of-scope for the A.C.E. duplicate search. The out-of-scope records were those that were reinstated and deleted from the Housing Unit Duplication Operation, documented in Nash (2000).

The ESCAP II matching was a two-step process. First, the sample of census records were matched to the full census on first name, last name, month of birth, day of birth and computed age. Age was allowed to vary by one year. Middle initials and suffixes being scanned into the first name field were accounted for; however, the other characteristics had to be exact matches at this stage. This first-stage match established a link between households. In the second stage, all person records in the linked households from the first stage were statistically matched using first name, middle initial, last name, month of birth, day of birth, and computed age. The matching parameters used in the statistical matching were borrowed from other Census 2000 matching operations. Mule (2001) describes this matching algorithm in more detail.

To reduce the impact of false matches, particularly with respect to persons with common names and the same month and day of birth, model weights were applied to each set of linked records as a measure of confidence that the linked records were indeed duplicates. Due to schedule constraints, a national, Poisson model was used in lieu of a probability model.

The ESCAP II census duplicate methodology satisfied the intended project goals and provided a valuable evaluation of the census by showing that person duplication existed. However, limitations of the methodology made it difficult to estimate the magnitude of person duplication in the census.

## OVERVIEW OF THE DUPLICATE STUDY PLAN

Like the ESCAP II study, the A.C.E. Revision II duplicate plan involved matching the Full E and P samples to the census to establish potential duplicate links. Then, modeling techniques were used to identify the links most likely

to be duplicate enumerations and to assign a measure of confidence that the links are duplicates. Key differences with the ESCAP II study include extending the use of statistical matching and developing models to assign a duplicate probability to the links. An advantage of duplicate probabilities over the Poisson model weights used in the ESCAP II study is that all duplicate links outside the A.C.E. search area could be reflected in the A.C.E. Revision II estimates. Fay (2001, 2002) used a subset of the ESCAP II duplicate links to produce a lower bound on the level of erroneous enumerations that the A.C.E. did not measure.

Estimates of census duplication were based on matching and modeling E-sample cases to the census. For purposes of A.C.E. Revision II estimation, the P sample was also matched to the census. However, these results did not contribute to estimates of person duplication in the census. The A.C.E. Revision II estimation methodology adjusted the A.C.E. correct enumeration rate for E-sample cases with links outside the A.C.E. search area. Further, the A.C.E. Revision II estimation methodology adjusted the A.C.E. match rate for P-sample cases that linked to census cases outside the search area.

The matching algorithm consisted of two stages. The first stage was a national match of persons using statistical matching as described in Winkler (1995). Statistical matching attempted to link records based on similar characteristics or close agreement of characteristics. Exact matching required exact agreement of characteristics. Statistical matching allowed two records to link in the presence of missing data and typographical or scanning errors.

Six characteristics common to both files, called matching variables, were used to link records in the Full E and P sample with records in the census. Matching parameters associated with each matching variable were used to measure the degree to which the matching variables agreed between the two records, ranging from "full agreement" to "full disagreement." The measurement of the degree to which each matching variable agreed was called the variable match score. The overall match score for the linked records was the sum of the variable match scores.

Full agreement of at least four characteristics was required to be considered a duplicate link. Because this study was a computer process without the benefit of a clerical review, this limitation of the statistical matching was necessary in order to minimize linking records with similar characteristics that represented different people. This was a particular concern when looking for duplicate enumerations across the entire country. The need to use statistical matching at the first stage was apparent after the limited success of the ESCAP II exact matching procedure to identify A.C.E. duplicates in the A.C.E. sample areas. The statistical matching yielded better identification of the A.C.E. duplicates, but to identify all of the A.C.E. duplicates would have required fewer characteristics to be exact matches. This could potentially lead to a high number of false links.

The search for duplicate links between the Full E and P samples and the census was limited to those pairs that agreed on certain identifiers, or blocking criteria. Blocking criteria were sort keys that were used to increase the computer processing efficiency by searching for links where they were most likely to be found. For instance, to search only for duplicates when the first and last names agreed, both the sample and census files would have been sorted by the blocking criteria of first and last name. Then, all possible pairs within each first name/last name combination would have been searched for duplicate links. Although true matches can be missed by using blocking criteria, multiple sets of blocking criteria minimize the number of missed matches. The A.C.E. Revision II duplicate study utilized four sets of blocking criteria.

At the first stage of matching, it was possible for one sample case to link to multiple census records. All of these links were retained for the second stage of matching.

The second stage of matching was limited to matching persons within households. If an E- or P-sample case linked to a census record in a group quarter, the case did not go to the second stage. Using results from the first stage of matching, a link between two housing units was established. The second stage was a statistical match of all household members in the sample housing unit to all household members in the census housing unit. The second-stage matching variables were the same as the first stage; however, the matching parameters differed. Using a subset of the first-stage links, the second-stage matching parameters were derived using the Expectation-Maximization (EM) algorithm. See Winkler (1995) for a more detailed explanation. A key difference between the first- and second- stage parameters was the reduced emphasis on requiring last names to agree in the second stage. This intuitively makes sense, since second stage matching was within a given household.

The household was the only set of blocking criteria used at the second stage of matching. Sample records were allowed to link to only one census record within the household. As a consequence, this limited the ability to identify within-household duplicate links. Each link had an overall match score based on the second-stage matching.

The set of linked records from the second-stage matching and the links to group quarter enumerations from the first stage consisted of both duplicate enumerations and person records with common characteristics. Two modeling approaches were used to estimate the probability that the linked records were duplicates. 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 referred to as the statistical match modeling and the exact match modeling, respectively. These two approaches were combined so that each sample case with a link to a census enumeration had an estimated probability of being a duplicate.

The statistical match modeling was used when two or more duplicate links were found between housing units in the second stage. After the second-stage matching, each duplicate link between a sample household and census household had an overall match score. So, for each sample household, a set of match scores was observed. For any resulting set of match scores, a probability of not observing this set of match scores was estimated. See the attachment for details. The higher this probability, the more likely that the set of linked records in the household were duplicates.

The estimate of the probability of not observing this set of match scores assumed independence of the individual match scores within each household. This assumption was based on using the EM algorithm to determine the second-stage matching parameters. The probability of observing the individual match scores was estimated from the empirical distribution of individual match scores resulting from the second-stage matching. Further, this measure accounted for the number of times that a unique sample household was matched to different census households within a given level of geography. The probability of not observing this set of match scores was translated into a "statistical match" duplicate probability of 0 or 1 based on critical values that varied by level of geography.

The exact match modeling relied on an exact match of the census to itself. The methodology accounted for the overall distribution of births, frequency of names, and population size in a specific geographic area. Duplicate probabilities were computed separately by geographical distance of the links. Further, duplicate links were modeled separately by how common the last name was, as well as for Hispanic names.

The two approaches were combined to assign an estimated probability that the linked records were duplicates. 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 nonexact matches, this was always from the statistical match modeling. For exact matches, adjustments were made to account for the integration of these two methods.

Based on the results of this matching and modeling, an overall estimate of census duplicates was derived from the E-sample links. Further, for each Full E- and P-sample person who linked outside the A.C.E. search area, these results provided the probability that they were in fact the same person. These duplicate probabilities were used in the A.C.E. Revision II estimates.

## MATCHING ALGORITHM

Efforts to increase matching efficiency over the ESCAP II duplicate study included implementing statistical matching of persons at the first stage and the use of more discriminating matching parameters at the second stage.

### Inputs

Both the Full E and P samples were matched to the census records. The E-sample records reflected any updates made by the clerical staff during the A.C.E. matching operation when the census characteristics were incorrectly transcribed or scanned. The P sample included all nonmovers, outmovers, and inmovers. The same matching algorithm was used for the Full E and P samples.

The census files consisted of data-defined person records for both the household and group quarters populations. Both the reinstated and deleted records from the Housing Unit Duplication Operation described in Nash (2000) were included in the matching, so these links could be reflected in the A.C.E. Revision II estimates.

### First Stage: Person-Level Matching

The first stage was a statistical match of the Full E and P samples to the census. This was a national match where each Full sample case was compared with census records across the nation to assess how well the matching variables agreed.

The matching variables were first name, last name, middle initial, month of birth, day of birth, and computed age. The matching variables and parameters are given in Table 5-1. The agreement weight and the disagreement weight are the matching parameters of each variable. Standard matching parameters were used at the first stage. The relationship of the agreement and disagreement parameters translated into the match score for each variable. For example, the full agreement value for first name was 2.1972; whereas, the full disagreement match score was -2.1972. The sum of the variable match scores was the total match score. When the match score was 9.4006, this indicated full agreement of all variables. A match score of -9.4006, on the other hand, indicated full disagreement.

Table 5-1. **First-Stage Matching Parameters**

| Matching variables | Type of comparison | Matching parameters | | Match score | |
|---|---|---|---|---|---|
| | | Agreement weight (m) | Disagreement weight (u) | Agreement ln (m/u) | Disagreement ln ((1-m)/(1-u)) |
| First name | String (uo) | 0.9 | 0.1 | 2.1972 | -2.1972 |
| Last name | String (uo) | 0.9 | 0.1 | 2.1972 | -2.1972 |
| Middle initial | Exact | 0.7 | 0.3 | 0.8473 | -0.8473 |
| Month of birth | Exact | 0.8 | 0.2 | 1.3863 | -1.3863 |
| Day of birth | Exact | 0.8 | 0.2 | 1.3863 | -1.3863 |
| Computed age | Age (p) | 0.8 | 0.2 | 1.3863 | -1.3863 |
| **Total** | | | | **9.4006** | **-9.4006** |

The type of comparison indicated the statistical matching method for comparing the variables. For example, the string comparitor was used for first name and last name. This method addressed typographical errors in names. For example, "Tim" and "Tum" can yield a positive agreement score. An exact match algorithm would have treated these as a disagreement. For age, the age values could have been off by ± one year and still receive a full agreement score on computed age.

The Statistical Research Division matching software called BigMatch documented in Yancey (2002) was used in the first stage. This software allowed a sample record to link to more than one census record. This capability was important, since it was possible for there to be more than two enumerations of the same person in the census.

Four blocking criteria were used. Blocking restricted the comparisons of records to only those that exactly agreed on certain values. Most records that did not agree on the values below are probably not duplicates. The blocking criteria were:

- 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

All possible links within each blocking criteria were compared. For each comparison, the variable match score and the total match score were computed. The first-stage matching decision rules were as follows. First, a match must have had at least four of the match variables in full agreement. This meant that four of the variables had to have a match score equal to the agreement match score in Table 5-1. The one exception was the middle initial. When the middle initial was blank, it was considered to be in full agreement in this study since the middle initial was often missing on the sample and census records. In this case, the middle initial score was zero. Second, the total match score had to be 4.7 or greater. This minimum score was about half the total score for full agreement of all matching variables.

Table 5-2 shows the distribution of A.C.E. links within cluster that were identified by the resulting number of matching variables in full agreement. There were a total of 10,559 duplicate links identified by the A.C.E. clerical staff that agreed on the first letter of the first and last name. The table shows the number of identified A.C.E. duplicates as the number of matching variables in full agreement decreased. The table also displays the number of total links that were identified. The percent of A.C.E. links in each row of the table decreases as the number of matching variables in full agreement decreases.

By requiring at least four matching variables to be in full agreement, 68.4 percent of these A.C.E. duplicates were identified. On the other hand, when only four of the six variables fully agreed, only 30.4 percent of the total links identified by this criteria were A.C.E. Revision II duplicates. Note that it was tempting to require that only three variables be in full agreement, since this would increase the number of A.C.E. duplicates by 20 percent. However, this change would substantially increase the number of false matches.

Table 5-3 shows that introducing a minimum total score greatly increased the density of A.C.E. links identified. Note that some A.C.E. duplicate links were dropped by using this criteria. This was a consequence of applying rules that reduced the false link rate.

### Second Stage: Household-Level Matching

The second stage of matching was restricted to the household population. The person links from the first stage established a link between two housing units. The second stage was a statistical match of the household members from the two housing units. A sample household was included in the second stage multiple times, if the sample-household had persons with links to multiple census households in the first stage. This was the same approach used for the ESCAP II duplicate study.

Table 5-2. **Distribution of Links Within A.C.E. Clusters by Full Agreement**

[Percentages may not add due to rounding]

| Number of variables in full agreement | A.C.E. links | | | Total links | Percent of A.C.E. links in row |
|---|---|---|---|---|---|
| | Count | Percent | Cumulative percent | | |
| 6 | 2,348 | 22.2 | 22.2 | 2,451 | 95.8 |
| 5 | 2,895 | 27.4 | 49.7 | 3,983 | 72.7 |
| 4 | 1,983 | 18.8 | 68.4 | 6,520 | 30.4 |
| 3 | 2,211 | 20.9 | 89.4 | 40,891 | 5.4 |
| 2 | 954 | 9.0 | 98.4 | 180,324 | 0.5 |
| 1 | 164 | 1.6 | 99.9 | 601,370 | <0.1 |
| 0 | 4 | <0.1 | 100 | 350,987 | <0.1 |
| **Total** | **10,559** | **100** | **100** | **1,186,526** | **0.9** |

Table 5-3. **Distribution of A.C.E. and Total Links Within A.C.E. Clusters**

[Only include links with score ≥ 4.7]

| Number of variables in full agreement | A.C.E. links | Total links | Percent of A.C.E. links in row |
|---|---|---|---|
| 6 | 2,348 | 2,451 | 95.8 |
| 5 | 2,868 | 3,763 | 76.2 |
| 4 | 1,680 | 2,670 | 62.9 |
| 3 | 0 | 0 | n/a |
| 2 | 0 | 0 | n/a |
| 1 | 0 | 0 | n/a |
| 0 | 0 | 0 | n/a |
| **Total** | **6,896** | **8,884** | **77.6** |

Table 5-4. **Second-Stage Matching Parameters**

| Matching variables | Type of comparison | Matching parameters | | Match score | |
|---|---|---|---|---|---|
| | | Agreement weight (m) | Disagreement weight (u) | Agreement ln (m/u) | Disagreement ln ((1-m)/(1-u)) |
| First name | String (uo) | 0.9500 | 0.0125 | 4.3307 | -2.9832 |
| Last name | String (uo) | 0.9600 | 0.5700 | 0.5213 | -2.3749 |
| Middle initial | Exact | 0.0840 | 0.0220 | 1.3398 | -0.0655 |
| Month of birth | Exact | 0.6000 | 0.0600 | 2.3026 | -0.8544 |
| Day of birth | Exact | 0.3000 | 0.0200 | 2.7081 | -0.3365 |
| Computed age | Age (p) | 0.9750 | 0.1325 | 1.9959 | -3.5467 |
| **Total** | | | | **13.1984** | **-4.1948** |

The matching variables were the same as the first stage: first name, last name, middle initial, month of birth, day of birth, and age. Table 5-4 gives the matching parameters. The data in this table have similar meaning as that for the first stage parameters in Table 5-1. Using a subset of the first-stage links, the second-stage matching parameters were derived using the EM algorithm as described in Winkler (1995). These parameters were anticipated to be more discriminating than the set used for the ESCAP II study.

Since the first-stage matching established a link between two housing units, first name had more discriminating power than last name in the second stage. When first name fully agreed, it contributed 4.3307 toward the total score, while last name only contributed 0.5213 when it was in full agreement. Further, month of birth and day of birth were more powerful than computed age. This was expected since adults in a housing unit often have similar ages, but not the same month and day of birth.

The Statistical Research Division Record Linkage software described in Winkler (1999) was used for the second stage. Each sample record was linked to only one census record within the household, a one-to-one matching. There was no additional blocking criteria beyond household; all possible links within households were compared. Each link had a total match score ranging from -4.1948 to 13.1984. This second-stage match score was used for the modeling.

All links with a second-stage match score greater than 0.3419 were retained as input to the modeling.

### Reverse Name Matching

Occasionally, first and last name was captured in reverse order on the data files. The first name was in the last name field and the last name was in the first name field. When the data was in reverse-order on one file but not the other, it was difficult to identify these duplicate links since the variable match scores for first and last name disagreed for both the first and second stage. To attempt to identify these cases, the first and last name fields were reversed and then matched to the census files a second time. The duplicate links from both runs, name in the usual order and in reverse order, were input to the modeling. When both methods identified the same duplicate link, the higher of the two match scores was retained and used in the modeling.

## MODELING LINKS

Since the goal of this study was to provide duplicate information for calculating A.C.E. Revision II estimates, it was important to provide a measure of confidence that could be incorporated into the estimation methodology. Consequently, modeling efforts focused on methods for estimating the probability that two linked records were duplicate enumerations. An advantage of duplicate probabilities over the Poisson model weights used in ESCAP II was that all duplicate links outside the A.C.E. search area could be reflected in the A.C.E. Revision II estimates. The statistical and exact match modeling approaches were combined to yield an estimated duplicate probability for the linked records from the statistical matching of the E and P samples to the census.

### Statistical Match Probability

The statistical match modeling was used when the second stage matching resulted in two or more duplicate links. After the second-stage matching, each duplicate link between a sample household and census household had an overall match score. So, for each sample housing unit to census housing unit match, a set of match scores was observed. For any resulting set of match scores, a probability of not observing this set of match scores, Pr(NT), was estimated for each link within the sample household. The higher this probability, the more likely that the set of linked records in the household were duplicates.

Since a sample housing unit could have been matched to more than one census housing unit during the second stage, there were multiple sets of duplicate links and match scores for each sample housing unit. Each set of duplicate links for a sample housing unit was assigned a separate Pr(NT) since the match scores differ for each matching attempt. Further, the Pr(NT) for each set of duplicate links for a sample housing unit varied because of the geographic distance of the duplicate links. As shown in the attachment, Pr(NT) was estimated by

$$Pr\,(\mathrm{NT}) = \left[1 - \prod_{d=1}^{p} Pr\,(X_d \geq x_d)\right]^n$$

where

$Pr(X_d \geq x_d)$ was the probability of getting a total match score $X_d$ greater than or equal to $x_d$,

p was the number of duplicate links in the sample household, and

n is the number of census housing units the sample household was matched with in the second stage within a given geographic area.

The estimate of the probability of not observing this set of match scores assumed independence of the individual match scores within each household. This assumption was based on using the EM algorithm to determine the second-stage matching parameters. 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 unique sample household was matched to different census households within a given level of geography. The geographical levels were block, tract, same county (outside tract), same state (outside county), and different state.

For the E sample, this analysis was done at the E-sample household level. For the P sample, a household consisted of any combination of nonmovers, outmovers, and inmovers. To account for this, the duplicate links were analyzed separately by mover status when looking at patterns of match scores.

The probability of not observing this set of match scores was translated into a "statistical match" duplicate probability of 0 or 1 based on critical values that varied by level of geography. Table 5-5 shows the minimum value of Pr(NT) for assigning a statistical match duplicate probability of 1 for E and P samples.

Table 5-5. **Minimum Value for Assigning Statistical Match Probability**

| Geographic distance of linked records | Minimum Pr(NT) | |
|---|---|---|
| | E sample | P sample |
| Same block | 0.00 | 0.25 |
| Same tract (different block) | 0.70 | 0.35 |
| Same county (different tract) | 0.97 | 0.60 |
| Same state (different county) | 0.97 | 0.60 |
| Different state | 0.97 | 0.60 |

Duplicate links with a Pr(NT) greater than or equal to the minimum value in Table 5-5 were assigned a statistical match duplicate probability of 1. All other links were assigned a statistical match duplicate probability of 0.

### Exact Match Probability

Given exact matching of the census to itself, duplicate probabilities were assigned to linked records by taking 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. Further, duplicate links were modeled separately by how common the last name was, as well as for Hispanic names. Fay (2002b) gives the model and preliminary results. The following are excerpts from Fay (2002b) to give the reader a general idea of the approach.

Like the Poisson model, the new approach uses frequencies of occurrences of combinations of first and last name. The result is an estimated probability of duplication for most matches, except for matches of frequently occurring names, where the probability of duplication is low and difficult to estimate with high relative precision.

This work results in a series of probability models, with parameters that can be estimated statistically from observed census data. A core model characterizes probabilities of duplication, triple enumeration (apparent enumeration of the same person three times), and other forms of multiple enumeration within a given geographic area. The other models account for duplication across domain.

The first part of the core model expresses the probability of coincidentally sharing a birthday. A second set of expressions, a model for census duplication, is built on top of the model for coincidental sharing of date of birth. The core model combines the two models to account for observed patterns of exact computer matches of census enumerations. The core model provides a basis to estimate a probability that a given computer match links the same person instead of two persons coincidentally sharing a birthday. An approximate argument allows the core model to be extended to nested geographic categories, such as (1) counties, (2) other counties within state, and (3) other states.

The result of the exact match model is a duplicate probability greater than or equal to zero, but less than one for census records that agree exactly on first name, last name, month and day of birth and two-year age intervals.

### Combining the Two Models

The two approaches were combined to give one duplicate probability to each E- and P-sample duplicate link. Table 5-6 summarizes the results of combining the two models. 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, as indicated by the shaded cells in Table 5-6. For nonexact matches, the duplicate probability assignment was always based on the statistical match modeling.

For exact matches in sample households with two or more persons, adjustments were made to account for the integration of these two methods. The exact match probabilities were determined conditionally, requiring a downward adjustment of the exact probabilities for the links, which the statistical match modeling assigned a probability of zero. The amount of the downward adjustment was based on the upward adjustment made when using the statistical match probability of one instead of the exact match probability.

Table 5-6. **Combining the Two Modeling Results**

| Type of Link | Size of sample HU | Type of match | Statistical match probability | Exact match probability |
|---|---|---|---|---|
| Housing Unit | 1 | Exact | – | [0, 1) |
| | | Nonexact | – | – |
| | 2+ | Exact | 1 | [0, 1) |
| | | Exact | 0 | [0, 1) |
| | | Nonexact | 1 | – |
| | | Nonexact | 0 | – |
| Group Quarter | | Exact | – | [0, 1) |
| | | Nonexact | – | – |

– Modeling did not assign a value.

The results of this modeling provided, for each Full E- and P-sample person who linked to a census person outside the A.C.E. search area, the probability that they were in fact the same person. These probabilities, referred to as $p_t$ in Chapter 6, were used to obtain A.C.E. Revision II estimates.

### Reinstated and Deleted Census Records

For the exact match modeling, separate probabilities were computed based on population distributions with and without the reinstated and deleted records. One computed duplicate probability allowed sample records to link to reinstated and deleted census records, while a second duplicate probability did not allow links to reinstated and deleted records. Under this second scenario, any links to reinstated or deleted records were assigned a duplicate probability of zero. The duplicate probabilities used in the A.C.E. Revision II estimation were those that allowed links to reinstated and deleted census records.

## ASSESSMENT OF LINKS

Throughout the development of the Further Study of Person Duplication in Census 2000, the A.C.E. duplicate links found during production were the benchmark used to

gauge whether the matching algorithm did a good job of finding true duplicates and minimizing the number of false links found within the block cluster.

This study and the ESCAP II duplicate study documented in Fay (2001, 2002) utilized the same method for estimating efficiency for the E sample. Basically, the method estimated the effectiveness of identifying A.C.E. clerical duplicates within the A.C.E. sample area and accounted for duplicate links to reinstated and deleted records that were out-of-scope for A.C.E. Instead of producing one overall efficiency measure, several measures were computed for various levels of detail including size of sample household and number of links between the units.

## FORMING ESTIMATES OF DUPLICATES

Estimates of census duplicates were formed by summing the product of the sampling weight for the E-sample person, the duplicate probability, and the multiplicity factor. Since a sample of the census (E sample) was matched to the census, a naive approach would treat each duplicate link of A to B as one duplicate. However, had a different sample been drawn, it could have contained the B to A link. Applying a multiplicity factor of 1/2 in this simple case ensured that the estimate of this example was only one duplicate. See Mule (2002b) for more details on the computation of the multiplicity factor.

# Attachment.
# Probability of Not Observing a Set of Match Scores

Each E-sample household had a set of duplicate links to a particular census household. Each duplicate link had a corresponding overall match score from the second-stage matching resulting in a pattern of match scores for the sample household. The task was to assess whether this observed set of match scores occurred because the links were duplicates or because the records had characteristics in common but were different people.

Objective: To estimate the probability of not observing this set of match scores or better for each E-sample household.

The hypothesis is that the higher the probability of not observing this set of match scores or better, the more likely the links represent duplicate enumerations.

Suppose a particular E-sample household has $p \geq 2$ duplicate links with observed match scores $x_1, x_2, ..., x_p$.

Define Pr(NT) to be the probability of not observing the set of match scores or better, $(X_1 \geq x_1, X_2 \geq x_2, ..., X_p \geq x_p)$.

This probability can be expressed as

$$Pr(\mathrm{NT}) = [1 - Pr(X_1 \geq x_1, X_2 \geq x_{2,\ldots,} X_p \geq x_p)]^n \qquad (1)$$

where n was the number of different census housing units that the E-sample housing unit was linked to during the second-stage match. This calculation accounted for the fact that the more times the E-sample housing unit matched to different housing units, the greater the chance of obtaining this outcome.

Individual match scores $X_1, X_2, ..., X_p$ were assumed to be independent, since the second-stage matching parameters gave more emphasis to first name rather than last name. Further, the parameters gave more emphasis to month and day of birth rather than age. Under the independence assumption, (1) can be written as follows:

$$Pr(\mathrm{NT}) = \left[1 - \prod_{d=1}^{p} Pr\,(X_d \geq x_d)\right]^n \qquad (2)$$

The probability of observing a match score $X_d$ greater than or equal to $x_d$, $Pr(X_d \geq x_d)$, was obtained from the empirical distribution of second-stage match scores. The probability in (2) was used for the P-sample households as well.

# Chapter 6.
# A.C.E. Revision II Estimation

The A.C.E. Revision II Dual System Estimate (DSE) methodology was developed with the following objectives in mind:

- Integration of the corrections for measurement errors so that measurement errors identified by both the evaluations and the duplicate study are not over-corrected.
- Separate estimation for both E- and P- sample persons based on whether or not they linked to a census enumeration outside the search area.
- Flexibility in the post-stratification design, because the factors that affect correct enumeration (as measured by the E sample) were not necessarily the same as those that affect coverage (as measured by the P sample).
- Adjustment for correlation bias.

This chapter presents how this additional information was incorporated into the DSE for A.C.E. Revision II estimates. The reader is assumed to be familiar with the basic dual system model and how it was used to produce the March 2001 A.C.E. estimates. See Haines (2001) for a detailed description of this methodology and Davis (2001) for the original dual system estimation results. This chapter describes the approach to A.C.E. Revision II dual system estimation. The chapter discusses estimation of the term accounting for persons in the census who are not in the E sample. The correct enumeration rate from the E-sample data is described. Then, the estimation of the match rate from the P-sample data is addressed. The census, E-sample, and P-sample data are combined to form a single DSE formula. Next, the post-stratification variables used for the A.C.E. Revision II Full and Revision Samples are defined. The chapter then discusses adjustment for correlation bias using demographic analysis sex ratios and concludes with a discussion of synthetic estimation.

## DUAL SYSTEM ESTIMATION

The basic form of the dual system estimate (DSE) is:

$$\mathrm{DSE} = (\mathrm{C}en' - \mathrm{II}') \times \frac{\mathrm{CE}}{\mathrm{E}} \times \frac{\mathrm{P}}{\mathrm{M}} \qquad (1)$$

where

Cen´ = the census count, excludes late adds
II' = non-data-defined census records, excludes late adds
LA = "late additions" to the census, i.e. records included too late for A.C.E. processing; primarily reinstated cases from the housing unit duplication operation
CE = E-sample weighted estimate of census correct enumerations
E = E-sample weighted estimate of census total enumerations (includes insufficient information for matching and followup cases, excludes non–data–defined cases and late adds)
P = P-sample weighted estimate of total persons
M = P-sample weighted estimate of matches to census persons

DSEs were computed separately within post-strata. A post-stratum is a group of people defined by demographic and geographic characteristics who are assumed to have the same probabilities of inclusion in the census. Post-strata can also be defined to have equal probabilities of correct enumeration in the census.

The DSE in (1) can be written as a function of the final census count, *Cen*, which includes late adds and the following three rates:

$$\mathrm{DSE} = \mathrm{C}en \times r_{\mathrm{DD}} \times \frac{r_{\mathrm{CE}}}{r_{\mathrm{M}}} \qquad (2)$$

where

$r_{\mathrm{DD}} = (\mathrm{C}en' - \mathrm{II}') / \mathrm{C}en$ is the census data-defined rate. The numerator excludes late adds, but the denominator includes late adds.
$r_{\mathrm{CE}} = \mathrm{CE} / \mathrm{E}$ is the E-sample correct enumeration rate
$r_{\mathrm{M}} = \mathrm{M} / \mathrm{P}$ is the P-sample match rate.

The three rates can be interpreted as estimates of probabilities. Thus, within post-stratum,

- $r_{DD}$ estimates the probability that a census person record has sufficient (and timely) information for inclusion in A.C.E. processing,
- $r_{CE}$ estimates the probability that an E-sample universe person is a correct enumeration, and
- $r_{M}$ estimates the probability that a person in the P sample is included in the census.

The interpretation of $r_M$ may be less obvious than the other two; it is the sample-weighted proportion of P-sample persons who were also found in the census. The general independence assumption underlying DSE is that either the census or the A.C.E. inclusion probabilities are the same (both are not required). Assuming causal independence, the match rate $r_M$ estimates the probability of census inclusion for the post-stratum.

Equation (2) also gives an interpretation of how the DSE constructs population estimates within a post-stratum.

- Multiply the census count (*Cen*) by the data-defined rate, $r_{DD}$ , to estimate the number of census persons who are data-defined and, therefore, eligible for inclusion in the E sample.
- Reduce this product by multiplying it by the estimated probability of correct enumeration, $r_{CE}$
- Increase this result by dividing it by the estimated probability of census inclusion, $r_M$

The primary tasks in developing DSEs at the post-stratum level are the estimation of the three rates involved. The estimate $r_{DD}$ is straightforward because it is based on 100-percent census tabulations. More detail is provided for the estimates $r_{CE}$ and $r_M$, since they are more challenging.

The different estimation tasks can be tackled one term at a time. Basically, the goal is to estimate the numerators and denominators of the terms $r_{CE}$ and $r_M$. Since *E*, the estimated number of total census data-defined enumerations, is a simple, direct sample-weighted estimate, the challenges relate mostly to developing the estimates *CE*, *P*, and *M*. The estimation challenges for A.C.E. Revision II focus on accounting for: (*i*) information from the revised coding of the A.C.E. Revision Sample, (*ii*) information from the A.C.E. Revision II study of census duplicates, and (*iii*) different post-stratification schemes for the Full E and P samples. The most difficult issue is (*ii*).

Before proceeding to a detailed discussion of the A.C.E. Revision II DSE components, consider the general nature of the estimator. While the basic DSE shown in equation (1) was applied in the 1990 PES (Hogan, 1993), the March 2001 A.C.E. incorporated the modification called PES-C estimation. See Haines (2001) and Mule (2001b) for details. This DSE had the general form:

$$\mathrm{DSE}^{\mathrm{C}} = (\mathrm{C}en' - \mathrm{II}') \times \frac{\mathrm{CE}}{\mathrm{E}} \times \frac{\mathrm{P}_{nm} + \mathrm{P}_{im}}{\mathrm{M}_{nm} + \frac{\mathrm{M}_{om}}{\mathrm{P}_{om}}\mathrm{P}_{im}} \qquad (3)$$

where the following quantities are all P-sample weighted estimates for the given post-stratum:

$M_{nm}$ = estimate of matches to census persons for nonmovers
$M_{om}$ = estimate of matches to census persons for outmovers
$P_{nm}$ = estimate of total nonmovers
$P_{om}$ = estimate of total outmovers
$P_{im}$ = estimate of total inmovers

Nonmovers, outmovers, and inmovers were defined with reference to their status in the period of time between Census Day (April 1, 2000) and the A.C.E. interview. Nonmovers were those who did not move during this period, outmovers were those persons who moved out of a sample block during this period, and inmovers are those who moved into a sample block during this period. Equation (3) estimated P-sample matches (*M*) as the sum of estimated matches among nonmovers ($M_{nm}$) and estimated matches among movers. The number of mover matches was estimated as the product of an estimated number of movers ($P_{im}$) and an estimate of the mover match rate ($M_{om} / P_{om}$). Thus, P-sample outmovers were used to estimate the mover match rate while P-Sample inmovers were used to estimate the number of movers. This approach implies that $P_{nm} + P_{im}$ should be used for the estimated total of P-sample persons (*P*).

Equation (3) can be further expanded to include post-stratification subscripts. The Full E- and P-sample post-strata are denoted by subscripts *i* and *j*, respectively. The census term was calculated for the cross-classification of *i* and *j* post-strata, denoted *ij*. The DSE formula, using procedure C for movers, with different post-strata for the E and P samples is:

$$\mathrm{DSE}^{\mathrm{C}}_{ij} = \mathrm{C}en_{ij} \times r_{\mathrm{DD},ij} \times \frac{\left[\frac{\mathrm{CE}_i}{\mathrm{E}_i}\right]}{\left[\frac{\mathrm{M}_{nm,j} + \left[\frac{\mathrm{M}_{om,j}}{\mathrm{P}_{om,j}}\right]\mathrm{P}_{im,j}}{\mathrm{P}_{nm,j} + \mathrm{P}_{im,j}}\right]}$$

## ESTIMATION OF $r_{DD}$

Recall the general form of the DSE in equation (2). This section discusses the estimation of the data-defined rate, or "DD-rate."

The DD-rate estimate ($r_{\mathrm{DD}}$) is defined as $(\mathrm{C}en' - \mathrm{II}') / \mathrm{C}en$ for a given detailed *ij* post-stratum, where $\mathrm{C}en'$, $\mathrm{II}'$, and *Cen* are defined from 100-percent census tabulations. At the post-stratum level, $\mathrm{C}en \times r_{\mathrm{DD}}$ reduces to $\mathrm{C}en' - \mathrm{II}'$. This suggests that an alternative to computing $r_{\mathrm{DD}}$ at a post-stratum level is to compute $\mathrm{C}en' - \mathrm{II}'$ for all levels (e.g., demographic post-stratum groups within small geographic areas) for which estimates were to be computed, and then to adjust these quantities by the appropriate $r_{\mathrm{CE}} / r_{\mathrm{M}}$ factors. This approach may be problematic, especially when applied to very small areas.

The problem with direct computation of $\mathrm{C}en' - \mathrm{II}'$ for very small areas can be seen with the following hypothetical example. Suppose a particular small geographic area (e.g.,

a collection of blocks) has a high rate of imputation in the census, say 15.0 percent. Imputation rates will vary geographically, and high rates could result from a number of factors, such as difficulties getting access to housing units in secure communities or difficulties in hiring and retaining census enumerators in a particular area. In this hypothetical example, removing all imputations from the census count for the area by computing $Cen' - II'$ would reduce the census count by 15.0 percent. Subsequent multiplication by the $r_{CE} / r_M$ factors and summing the resulting DSEs over post-strata may increase the population estimate from this base, but perhaps by no more than two or three percent (depending on the post-stratum composition of the area). The net synthetic DSE would, thus, be 12.0 or 13.0 percent lower than the census count. While this estimate could make sense if almost all the housing units for which persons were imputed were actually vacant (and this fact were not discovered in the census enumeration), it would not make sense if most of the units were occupied and the high rate of imputation resulted from other factors such as those suggested above. Calculating $r_{DD}$ for post-strata and applying it synthetically avoids such problems in small area estimates, though perhaps incurring some error for larger areas for which the direct tabulation of $Cen' - II'$ would be sensible.

The data-defined rates, $r_{DD}$ , are computed at the detailed post-stratum obtained as the intersection of the E- and P-sample post-strata.

### ESTIMATION OF $r_{CE}$

This section discusses the estimation of the correct enumeration rate, $r_{CE} = CE/E$. The Full E-sample post-strata are denoted by the subscript *i*. The Revision E sample has post-strata denoted by $i'$, where $i'$ is based on collapsed post-strata *i*. This means that the Revision Sample post-strata were obtained by collapsing the Full Sample post-strata *i*. The correct enumeration rate is written:

$$r_{CE,\,i} = \frac{CE_i^{ND} f_{l,\,i'} + \tilde{CE}_i^{D}}{E_i} \qquad (4)$$

Note that the numerator term separates the E-sample enumerations with a duplicate link to a census enumeration outside the A.C.E. search area, as identified in the duplicate study, from those enumerations without a link. As discussed in Chapter 5, 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, therefore, were never included in the census.

The term $CE_i^{ND}$ estimates the number of correct enumerations in the Full E sample without duplicate links in post-stratum *i*. This term includes the probability of not being a duplicate, $1\text{-}p_t$.

The component $\tilde{CE}_i^{D}$ represents the estimated number of correct enumerations in the Full E sample with duplicate links in post-stratum *i*, which are retained after unduplication. This term includes the probability of being a duplicate, $p_t$, as well as the conditional 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.

The total weighted number of persons in post-stratum *i* in the E sample are denoted by $E_i$.

The double-sampling ratio factor $f_{l,\,i'}$ corrects for measurement error based on the Revision E sample. It is a ratio of an estimate that uses the revised coding (indicated by *) to an estimate that uses the original coding. These adjustments, which are calculated for measurement error post-strata $i'$, are represented by:

$$f_{l,\,i'} = \frac{\dfrac{CE_{i'}^{ND*}}{E_{i'}^{ND}}}{\dfrac{CE_{i'}^{ND}}{E_{i'}^{ND}}} = \frac{CE_{i'}^{ND*}}{CE_{i'}^{ND}} \ .$$

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 $p_t$ value 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.

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 no additional data were collected for this purpose. 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. Thus, it is necessary to estimate the following conditional probability:

$z_t$ 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.

### E-Sample Links

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: $\tilde{CE}_i^D$. 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. Attachment 1 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 Attachment 1. 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 off-campus apartment. This is represented by Linked Situations 2a. and 2b. in Attachment 1.

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 indicated in Attachment 1, by:

- 3 Race/Hispanic Origin Domains
- Tenure

These resulting control cells are given in Attachment 2. 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 is calculated as:

$$\hat{z}_t = \frac{0.5 \sum_t \mathrm{W}_t\, p_t}{\sum_t \mathrm{W}_t\, p_t\, \mathrm{Pr\,(CE)}}$$

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

The components of equation (4) are defined below.

$$\tilde{\mathrm{CE}}_i^{\mathrm{D}} = \sum_{t \in i} \mathrm{W}^{\mathrm{E}}_{\pi,\,t}\, p_t\, z_t\, \mathrm{PR}ce_{\pi,\,t}$$

is the estimated number of correct enumerations with duplicate links in post-stratum *i* who were retained after unduplication.

$$\mathrm{CE}_i^{\mathrm{ND}} = \sum_{t \in i} \mathrm{W}^{\mathrm{E}}_{\pi,\,t}\,(1 - p_t)\,\mathrm{PR}ce_{\pi,\,t}$$

is the number of correct enumerations without duplicate links in post-stratum *i*, where the summation is taken over all enumerations in the A.C.E. E sample in post-stratum *i*.

$\mathrm{W}^{\mathrm{E}}_{\pi,\,t}$ is the production A.C.E. sampling weight for E-sample person *t*.

$p_t$ is the probability that person *t* has a duplicate link outside the search area. This is usually 0 or 1, but could be between these two values for probability matches, where the accuracy of the link was uncertain.

$\mathrm{PR}ce_{\pi,\,t}$ is the probability that person *t* is a correct enumeration in the original production coding. This is either 0 or 1 unless it was not possible to code the E-sample case a correct or erroneous enumeration. In these cases, a probability of correct enumeration was imputed.

$$f_{l,\,i'} = \frac{\mathrm{CE}_{i'}^{\mathrm{ND}*}}{\mathrm{CE}_{i'}^{\mathrm{ND}}} = \frac{\sum_{t \in i'} \mathrm{W}^{\mathrm{E}}_{\mathrm{RR},\,t}\,(1 - p_t)\,\mathrm{PR}ce_{\mathrm{R},\,t}}{\sum_{t \in i'} \mathrm{W}^{\mathrm{E}}_{\mathrm{R}\pi,\,t}\,(1 - p_t)\,\mathrm{PR}ce_{\pi,\,t}}$$

where

$\mathrm{W}^{\mathrm{E}}_{\mathrm{RR},t}$ is the A.C.E. Revision Sample weight for person *t* to be used for Revision Sample coding.

$\mathrm{W}^{\mathrm{E}}_{\mathrm{R}\pi,\,t}$ is the A.C.E. Revision Sample weight for person *t* to be used with production coding. These two weights could differ slightly depending on TES status and noninterview adjustment.

$\mathrm{PR}ce_{\mathrm{R},\,t}$ is the probability that person *t* is a correct enumeration in the A.C.E. Revision Sample coding.

$$\mathrm{E}_i = \sum_{t \in i} \mathrm{W}^{\mathrm{E}}_{\pi,\,t}$$

is the total weighted number of persons in the E sample in post-stratum *i*.

**ESTIMATION OF $\mathbf{r_M}$**

This section discusses the estimated match rate in equation (2). E-sample post-strata are indexed by *i*, while the P-sample post-strata are indexed by *j*. The match rate for post-stratum *j* is represented as:

$$r_{\mathrm{M},\,j} = \frac{\mathrm{M}^{\mathrm{ND}}_{nm,\,j} f_{2,\,j'} + \tilde{\mathrm{M}}^{\mathrm{D}}_{nm,\,j} + \left[\dfrac{\mathrm{M}_{om,\,j} f_{3,\,j'}}{\mathrm{P}_{om,\,j} f_{4,\,j'}}\right](\mathrm{P}_{im,\,j} f_{5,\,j'} + g\,(\mathrm{P}^{\mathrm{D}}_{nm,\,j} - \tilde{\mathrm{P}}^{\mathrm{D}}_{nm,\,j}))}{\mathrm{P}^{\mathrm{ND}}_{nm,\,j} f_{6,\,j'} + \tilde{\mathrm{P}}^{\mathrm{D}}_{nm,\,j} + \mathrm{P}_{im,\,j} f_{5,\,j'} + g\,(\mathrm{P}^{\mathrm{D}}_{nm,\,j} - \tilde{\mathrm{P}}^{\mathrm{D}}_{nm,\,j})} \qquad (5)$$

The residence status of P-sample movers was adjusted for coding error. The computer matching results were not used. Outmovers in the P sample were collected by a proxy interview, which made it difficult to obtain date of birth and age information. Since date of birth and age were important characteristics used in the computer matching, the movers were only adjusted for coding error.

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 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 the following conditional probability:

$h_t$ 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.

### 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. Attachment 1 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 those 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 Attachment 2. The $h_t$'s are then included in the weighted tallies, along with the $p_t$, to calculate the duplicate contribution to the Full P-sample nonmovers and nonmover matches.

The terms in equation (5) are defined below. Summation $t \in j$ denotes summation over A.C.E. Full P-Sample post-stratum j, while summation $t \in j'$ denotes summation over Revision Sample post-stratum $j'$. The summation notation also indicates whether the sum is taken over nonmovers, outmovers, or inmovers, and if the Production ($\pi$) or Revision (R) Sample coding is used.

$$\mathrm{M}^{\mathrm{ND}}_{nm,\,j} = \sum_{\substack{t \in j \\ t\ nonmover \\ production}} \mathrm{W}^{\mathrm{P}}_{\pi,\,t}(1-p_t)\,\mathrm{PR}res_{\pi,t}\,\mathrm{PR}m_{\pi,\,t}$$

where

$\mathrm{W}^{\mathrm{P}}_{\pi,\,t}$ is the P-sample production weight of person *t*.

$p_t$ is the probability that person *t* has a duplicate link outside the search area.

$\mathrm{PR}m_{\pi,\,t}$ is the probability that person *t* is a match in the production coding.

$\mathrm{PR}res_{\pi,\,t}$ is the probability that person *t* is a resident in the production coding.

$$f_{2,j'} = \frac{\mathrm{M}^{\mathrm{ND*}}_{nm,\,j'}}{\mathrm{M}^{\mathrm{ND}}_{nm,\,j'}} = \frac{\sum_{\substack{t \in j' \\ t\ nonmover \\ revision}} \mathrm{W}^{\mathrm{P}}_{\mathrm{RR},\,t}(1-p_t)\,\mathrm{PR}res_{\mathrm{R},\,t}\,\mathrm{PR}m_{\mathrm{R},t}}{\sum_{\substack{t \in j' \\ t\ nonmover \\ production}} \mathrm{W}^{\mathrm{P}}_{\mathrm{R}\pi,\,t}(1-p_t)\,\mathrm{PR}res_{\pi,\,t}\,\mathrm{PR}m_{\pi,t}}$$

is the double-sampling adjustment for nonmover matches.

$\mathrm{PR}m_{\mathrm{R},\,t}$ is the probability that person *t* is a match in the Revision Sample coding.

$\mathrm{PR}res_{\mathrm{R},\,t}$ is the probability that person t is a resident in the Revision Sample coding.

$\mathrm{W}^{\mathrm{P}}_{\mathrm{RR},\,t}$ is the A.C.E. Revision Sample weight for person *t* to be used for Revision Sample coding.

$\mathrm{W}^{\mathrm{P}}_{\mathrm{R}\pi,\,t}$ is the A.C.E. Revision Sample weight for person *t* to be used with production coding. These two weights could differ slightly depending on TES status and the noninterview adjustment.

$$\mathrm{M}_{om,\,j} = \sum_{\substack{t \in j \\ t\ outmover \\ production}} \mathrm{W}^{\mathrm{P}}_{\pi,\,t}\,\mathrm{PR}res_{\pi,\,t}\,\mathrm{PR}m_{\pi,\,t}$$

is the number of matched outmovers in the Full Sample in post-stratum *j*.

$$f_{3,j'} = \frac{M^{*}_{om,j'}}{M_{om,j'}} = \frac{\sum_{\substack{t \in j' \\ t\ outmover \\ revision}} W^{P}_{RR,t}\, PRres_{R,t}\, PRm_{R,t}}{\sum_{\substack{t \in j' \\ t\ outmover \\ production}} W^{P}_{R\pi,t}\, PRres_{\pi,t}\, PRm_{\pi,t}}$$

is the double-sampling ratio for matched outmovers for post-stratum *j´*.

$$P_{om,j} = \sum_{\substack{t \in, j \\ t\ nonmover \\ production}} W^{P}_{\pi,t}\, PRres_{\pi,t}$$

is the number of outmovers in the Full Sample for post-stratum *j*.

$$f_{4,j'} = \frac{P^{*}_{om,j'}}{P_{om,j'}} = \frac{\sum_{\substack{t \in j' \\ t\ outmover \\ revision}} W^{P}_{RR,t}\, PRres_{R,t}}{\sum_{\substack{t \in j' \\ t\ outmover \\ production}} W^{P}_{R\pi,t}\, PRres_{\pi,t}}$$

is the double-sampling ratio for outmovers for post-stratum *j´*.

$$P_{im,j} = \sum_{\substack{t \in j \\ t\ nonmover \\ production}} W^{P}_{\pi,t}$$

is the number of inmovers in the Full Sample post-stratum *j*.

$$f_{5,j'} = \frac{P^{*}_{im,j'}}{P_{im,j'}} = \frac{\sum_{\substack{t \in j' \\ t\ inmover \\ revision}} W^{P}_{RR,t}\, PRinmover_{R,t}}{\sum_{\substack{t \in j' \\ t\ inmover \\ production}} W^{P}_{R\pi,t}}$$

is the double-sampling ratio for inmovers for post-stratum *j´*.

$PRinmover_{R,t}$ is the probability that person *t* in the Revision Sample is an inmover.

$$g\,(P^{D}_{nm,j} - \tilde{P}^{D}_{nm,j})$$

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 to 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 is 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. The term g is estimated using the Revision Sample and both the original A.C.E. and the revision coding as follows:

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

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

A couple of important assumptions are:

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

$$\tilde{M}^{D}_{nm,j} = \sum_{\substack{t \in j \\ t\ inmover \\ production}} W^{P}_{\pi,t}\; p_t\, h_t\, PRm_{\pi,t}\, PRres_{\pi,t}$$

is the number of duplicate persons determined to have been Census Day residents who matched to the census in post–stratum *j*.

$$P^{ND}_{nm,j} = \sum_{\substack{t \in j \\ t\ inmover \\ production}} W^{P}_{\pi,t}\,(1 - p_t)\, PRres_{\pi,t}$$

is the number of nonmovers without links outside the search area in post-stratum *j*.

$$f_{6,j'} = \frac{P^{ND*}_{nm,j'}}{P^{ND}_{nm,j'}} = \frac{\sum_{\substack{t \in j' \\ t\ nonmover \\ revision}} W^{P}_{RR,t}\,(1 - p_t)\, PRres_{R,t}}{\sum_{\substack{t \in j' \\ t\ nonmover \\ production}} W^{P}_{R\pi,t}\,(1 - p_t)\, PRres_{\pi,t}}$$

is the double-sampling adjustment for nonmovers in post-stratum j´.

$$\tilde{P}^{D}_{nm,j} \sum_{\substack{t \in j \\ t\ nonmover \\ production}} = W^{P}_{\pi,t}\, p_t\, h_t\, PRres_{\pi,t}$$

is the estimated number of nonmover persons with duplicate links who were residents after unduplication.

$$P^{D}_{nm,j} = \sum_{\substack{t \in j \\ t\ nonmover \\ production}} W^{P}_{\pi,t}\, P_t\, PRres_{\pi,t}$$

is the number of P-sample persons with duplicate links, regardless of whether they were determined to be residents by the unduplication process.

## THE A.C.E. REVISION II DSE FORMULA

The A.C.E. Revision II DSE formula, using procedure C for movers, separate E- and P-sample post-strata, measurement error corrections from the E- and P- Revision Samples, and duplicate study results is:

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

| Notation | | |
|---|---|---|
| Terms | CE | Correct enumerations |
| | E | E-sample total |
| | M | Matches |
| | P | P-sample total |
| | f | 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 procedure 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 |

In some small post-strata, the number of inmovers was substantially larger than the number of outmovers. If there were only a few outmovers, the outmover match rate was subject to high sampling error. In these post-strata, it was not considered appropriate to apply a suspect match rate to what could be a relatively large number of inmovers, so PES-A was used. PES-A uses only outmovers. PES-A was applied for post-strata with nine or fewer P-sample outmovers. For these post-strata, it was assumed that some of the duplicate links determined not to have been residents were really outmovers.

The DSE formula that uses procedure A for movers with different post-strata for the E- and P-samples is:

$$DSE^{A}_{ij} = Cen_{ij} \times r_{DD,ij} \times \frac{\frac{CE_i}{E_i}}{\left[\frac{M_{nm,j} + M_{om,j}}{P_{nm,j} + P_{om,j}}\right]}$$

The A.C.E. Revision II DSE formula, using procedure A for movers, separate E- and P-sample post-strata, measurement error corrections from the E- and P- Revision Samples, and duplicate study results is written:

$$DSE^{A}_{ij} = Cen_{ij} \times r_{DD,ij} \times \frac{\left[\frac{CE^{ND}_{i} f_{1,i'} + \tilde{CE}^{D}_{i}}{E_i}\right]}{\left[\frac{M^{ND}_{nm,j} f_{2,j'} + \tilde{M}^{D}_{nm,j} + M_{om,j} f_{3,j'} + g\,(M^{D}_{nm,j} - \tilde{M}^{D}_{nm,j})}{P^{ND}_{nm,j} f_{6,j'} + \tilde{P}^{D}_{nm,j} + P_{om,j} f_{4,j'} + g\,(P^{D}_{nm,j} - \tilde{P}^{D}_{nm,j})}\right]}$$

This version of the formula is used only when the sample size for outmovers in the Full P sample is strictly less than 10. This formula was used 93 times in the A.C.E. Revision II production process. The new term introduced in this formula is defined as follows:

$$M^{D}_{nm,j} = \sum_{\substack{t \in j \\ t\ nonmover \\ production}} W^{p}_{\pi,t}\ p_t\, PRres_{\pi,t}\, PRm_{\pi,t}$$

is the number of matched P-sample persons with duplicate links, regardless of whether they were determined to be residents by the unduplication process.

## A.C.E. REVISION II POST-STRATIFICATION DESIGN

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

### Full P Sample

For the Full P sample, the new post-strata were nearly identical to those used for the March 2001 A.C.E. estimates. The only difference was that the 0-17 age group

was split into two groups, 0-9 and 10-17, which resulted in some collapsing differences. The Full P sample, consisting of 480 post-strata, was 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
- Return Rate Indicator (Low vs. High)
- Region
- Age
- Sex

For the Full P sample, the post-stratum groups either retained all eight Age/Sex categories or were collapsed to four Age/Sex categories as shown below:

Figure 6-1. **P-Sample Age/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-sample post-strata (j), only the Revision P-sample post-strata (j′).

Table 6-1 shows the 64 Full P-sample post-stratum groups. The number in each cell represents the number of Age/Sex categories in each post-stratum group.

Table 6-1. **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 |
| | | Nonowner | 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<br>Medium MSA MO/MB | 8 | | | | 8 | | | |
| | | | Small MSA & Non-MSA MO/MB<br>All other TEAs | 8 | | | | 8 | | | |
| | | Nonowner | Large MSA MO/MB<br>Medium MSA MO/MB | 8 | | | | 8 | | | |
| | | | Small MSA & Non-MSA MO/MB<br>All other TEAs | 8 | | | | 4 | | | |
| Domain 3 (Hispanic) | | Owner | Large MSA MO/MB<br>Medium MSA MO/MB | 8 | | | | 8 | | | |
| | | | Small MSA & Non-MSA MO/MB<br>All other TEAs | 8 | | | | 8 | | | |
| | | Nonowner | Large MSA MO/MB<br>Medium MSA MO/MB | 8 | | | | 8 | | | |
| | | | Small MSA & Non-MSA MO/MB<br>All other TEAs | 8 | | | | 4 | | | |
| Domain 5 (Native Hawaiian or Pacific Islander) | | Owner | | 4 | | | | | | | |
| | | Nonowner | | 4 | | | | | | | |
| Domain 6 (Non-Hispanic Asian) | | Owner | | 8 | | | | | | | |
| | | Nonowner | | 8 | | | | | | | |
| American Indian or Alaska Native | Domain 1 (On Reservation) | Owner | | 8 | | | | | | | |
| | | Nonowner | | 8 | | | | | | | |
| | Domain 2 (Off Reservation) | Owner | | 8 | | | | | | | |
| | | Nonowner | | 8 | | | | | | | |

### Full E Sample

For the A.C.E. Revision II Full E sample, the post-strata definitions have undergone major revisions. Some of the original post-stratification variables were omitted and additional variables were added. Logistic regression models identified several variables, not included in the Full P-sample post-stratification, that were good indicators of correct enumeration. The Full E sample, consisting of 525 post-strata, was defined using 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

The new variables proxy status, household relationship and size, and type (mailback/nonmailback) and date (early/late) of census return are described generally below.

- **Proxy Status.** Nonproxy includes those housing unit persons for whom census data were provided by a household member. Proxy includes those housing unit persons for whom census data were provided by a non-household member, such as a neighbor or rental agent.
- **Household Relationship.** The Householder/Nuclear (HHer/Nuclear) relationship category includes persons in housing units consisting only of the householder with spouse or own children (I7 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.
- **Household Size.** Household size, or number of persons residing in the housing unit.
- **Early/Late Mailback.** Persons in mailback housing units with an earliest form processing date. On or before March 24 is early and after March 24 is late.
- **Early/Late Nonmailback.** Persons in nonmailback housing units with an earliest form processing date. On or before June 1 is early and after June 1 is late.

For the Full E sample, the post-stratum groups either retained all eight Age/Sex categories or were collapsed to four, two, or one Age/Sex groups, based on sample sizes, as shown below:

Figure 6-2. **E- Sample Age/Sex Groupings**

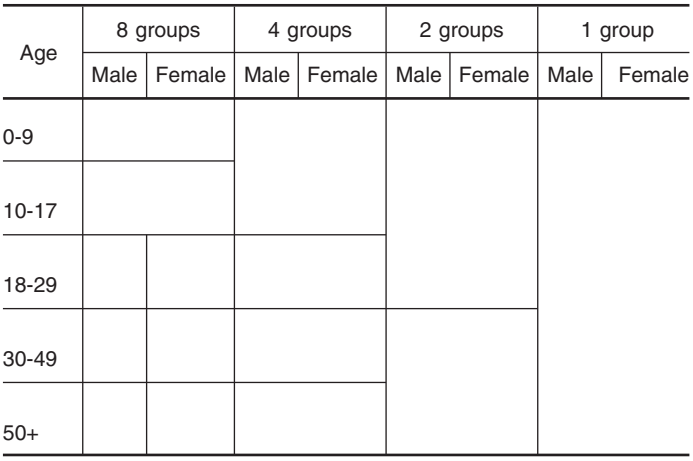



Table 6-2 shows the 93 Full E-sample post-stratum groups. The number in each cell represents the number of Age/Sex categories in each post-stratum group.

Table 6-2. **Full E-Sample Post-Stratum Groups and Number of Age and Sex Groupings**

| Proxy status & domain | | Tenure | Relationship | HH Size | Early mailback | Late mailback | 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 (America Indian or Alaska Native On Reservation) | | | | | 4 | | | |
| Proxy: Domain 2 (American Indian or Alaska Native Off Reservation) | | | | | 1 | | | |
| Nonproxy: 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 |
| | | Nonowner | HHer/Nuclear | | 8 | 8 | 8 | 8 |
| | | | Other | | 8 | 8 | 8 | 8 |
| Nonproxy: Domain 4 (Non-Hispanic Black) | | Owner | HHer/Nuclear | | 4 | 4 | 2 | 4 |
| | | | Other | | 8 | 8 | 4 | 8 |
| | | Nonowner | HHer/Nuclear | | 8 | 8 | 8 | 8 |
| | | | Other | | 8 | 8 | 8 | 8 |
| Nonproxy: Domain 3 (Hispanic) | | Owner | HHer/Nuclear | | 8 | 8 | 4 | 8 |
| | | | Other | | 8 | 8 | 4 | 8 |
| | | Nonowner | HHer/Nuclear | | 8 | 8 | 8 | 8 |
| | | | Other | | 8 | 8 | 8 | 8 |
| Nonproxy: Domain 5 (Native Hawaiian or Pacific Islander) | | Owner & Nonowner | HHer/Nuclear | | 2 | 2 | 2 | 2 |
| | | | Other | | 2 | 2 | 1 | 2 |
| Nonproxy: Domain 6 (Non-Hispanic Asian) | | Owner & Nonowner | HHer/Nuclear | | 8 | 8 | 4 | 4 |
| | | | Other | | 4 | 4 | 2 | 4 |
| Nonproxy: (American Indian or Alaska Native) | Domain 1 On Reservation | Owner & Nonowner | HHer/Nuclear | | 8 | | | |
| | | | Other | | 8 | | | |
| | Domain 2 Off Reservation | Owner & Nonowner | HHer/Nuclear | | 2 | 2 | 2 | 2 |
| | | | Other | | 2 | 2 | 1 | 2 |

### Revision P Sample

The Revision P sample is a subsample of the Full P sample and is comprised of over 60,000 sample persons. The Revision P sample has 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 P sample has undergone extensive recoding using all available interview data and matching results. Missing data adjustments have also been applied to the Revision P sample. This recoded data are used to correct for measurement error in the Full P sample.

The measurement error correction post-stratum definitions ($j'$) depend on a person's mover status. Both inmovers and outmovers are subdivided into Owner and Nonowner groups. For nonmovers, the measurement error correction post-strata are: American Indians on Reservations (AIR) and, for the Non-AIR cases, a cross of Tenure (Owner versus Nonowner) with eight Age and Sex categories. The Age/Sex collapsing pattern from the Full P sample is retained when defining the measurement error correction post-strata. The Revision P-sample post-strata ($j'$) are defined as follows:

Figure 6-3. **Revision P-Sample Post-Strata ($j'$)**

| Mover Status & Domain | Tenure | Age | 8 groups | | 1 group |
|---|---|---|---|---|---|
| | | | Male | Female | |
| Movers: Domains 1 thru 7 | Owner | | | | |
| | Nonowner | | | | |
| Nonmovers: Domains 2 thru 7 | Owner | 0-9 | | | N/A |
| | | 10-17 | | | |
| | | 18-29 | | | |
| | | 30-49 | | | |
| | | 50+ | | | |
| | Nonowner | 0-9 | | | N/A |
| | | 10-17 | | | |
| | | 18-29 | | | |
| | | 30-49 | | | |
| | | 50+ | | | |
| Nonmovers: Domain 1 (American Indian or Alaska Native On Reservation) | | | | | |

N/A means not applicable.

### Revision E Sample

The Revision E sample is a subsample of the Full E sample and is comprised of over 75,000 sample persons. The Revision E sample has 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 E sample has undergone extensive recoding using all available interview data and matching results. Missing data adjustments have also been applied to the Revision E sample. These recoded data are used to correct for measurement error in the Full E Sample.

For the Revision E sample, the measurement error correction post-strata are: Proxies, American Indians on Reservations (AIR) and, for the Nonproxy/Non-AIR cases, a cross of a two-level Relationship variable with eight Age/Sex categories. Note that Household Size is collapsed out of the Household Relationship/Size variable. The Age/Sex collapsing pattern from the Full E sample is retained when defining the measurement error correction post-strata. The Revision E sample post-strata ($i'$) are defined as follows:

Figure 6-4. **Revision E-Sample Post-Strata ($i'$)**

| Proxy Status & Domain | Relationship | Age | 8 groups | | 1 group |
|---|---|---|---|---|---|
| | | | Male | Female | |
| Proxy:<br>Domain 7 (Non-Hispanic White or ''Some Other Race'')<br>Domain 4 (Non-Hispanic Black)<br>Domain 3 (Hispanic)<br>Domain 5 (Native Hawaiian or Pacific Islander)<br>Domain 6 (Non-Hispanic Asian)<br>Domain 1 (American Indian or Alaska Native On Reservation)<br>Domain 2 (American Indian or Alaska Native Off Reservation) | | | | | |
| Nonproxy:<br>Domains 2 thru 7 | HHer/Nuclear | 0-9 | | | N/A |
| | | 10-17 | | | |
| | | 18-29 | | | |
| | | 30-49 | | | |
| | | 50+ | | | |
| | Other | 0-9 | | | N/A |
| | | 10-17 | | | |
| | | 18-29 | | | |
| | | 30-49 | | | |
| | | 50+ | | | |
| Nonproxy:<br>Domain 1 (American Indian or Alaska Native On Reservation) | | | | | |

N/A means not applicable.

## ADJUSTMENT FOR CORRELATION BIAS USING DEMOGRAPHIC ANALYSIS

The dual system 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; no correlation bias adjustment for non-Black males aged 18-29; and that Black males have a relative correlation bias that is different than the relative correlation bias for non-Black 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 non–Black.

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 group quarters people from the DA totals using Census 2000 data. The reason for making this adjustment is that the group quarters 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 non-Black domain are shown in Attachment 3.

In general the correlation bias adjustment factor, $c_k$, is defined for $k = 3$ age groups such that:

$\mathrm{E}\,[c_k\,\mathrm{DSE}_k^m]$ = True male population for age group $k$,

where

$\mathrm{DSE}_k^m$ 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 $c_k$ was not allowed to be less than one.

### Correlation Bias Adjustment for Black and Non-Black Males 18 Years and Older

The correlation bias adjustment for Black and non-Black 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 non-Blacks. This correlation bias adjustment is calculated as:

$$c_{\mathrm{R},\,k} = \left( \frac{\sum_{ij\,\epsilon\,k} \mathrm{DSE}_{ij}^{\mathrm{R}f}}{\sum_{ij\,\epsilon\,k} \mathrm{DSE}_{ij}^{\mathrm{R}m}} \right) r_{\mathrm{DAR},\,k}$$

where

$\mathrm{DSE}_{ij}^{\mathrm{R}f}$ = DSE for race, R=Black or non-Black, female post-strata *ij*.

$\mathrm{DSE}_{ij}^{\mathrm{R}m}$ = DSE for race, R=Black or non-Black, male post-strata *ij*.

$r_{\mathrm{DAR},k}$ = DA sex ratio for race, R=Black or non-Black, for age group *k* as given in Attachment 3.

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 age-race group *k* is calculated as:

$$\tilde{\mathrm{DSE}}_{ij}^{m} = c_k\,\mathrm{DSE}_{ij}^{m}$$

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:

$$\tilde{\mathrm{DSE}}_{ij}^{f} = \mathrm{DSE}_{ij}^{f}$$

The $\tilde{\mathrm{DSE}}_{ij}$'*s* are then used to form synthetic estimates.

## SYNTHETIC ESTIMATION

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

$$\tilde{\mathrm{CCF}}_{ij} = \frac{\tilde{\mathrm{DSE}}_{ij}}{\mathrm{Cen}_{ij}}$$

where the $\tilde{\mathrm{DSE}}_{ij}$ are the correlation bias-adjusted DSEs for post-stratum *ij*.

$\mathrm{Cen}_{ij}$'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, except those in group quarters or Remote Alaska. Effectively, these persons have a coverage correction factor of 1.0. 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:

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

Note that the coverage correction factor can be expressed as:

$$\tilde{\mathrm{CCF}}_{ij} = \left(\frac{\mathrm{DD}_{ij}}{\mathrm{C}en_{ij}}\right)\left(\frac{r_{\mathrm{CE},i}}{r_{\mathrm{M},j}}\right) c_k$$

where

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

$r_{\mathrm{M},j}$ 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 non-Black groups and *k* age cells.

$\mathrm{DD}_{ij} / \mathrm{C}en_{ij}$ is the data-defined rate, varying over the *ij* post-strata.

# Attachment 1.
# Rules for Assigning $z_t$ & $h_t$ for Full P- and E-Sample Duplicate Links

The Linked Situations and assignment of $z_t$'s and $h_t$'s occur in the order 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.

# Attachment 2.
# Control Cells for Linked E Sample

| Race/Hispanic Origin Domain | Tenure | Linked situation | Control cell |
|---|---|---|---|
| Domain 4 (Non-Hispanic Black) | Owner | 3. | |
| | | 4. | |
| | | 5. | |
| | Nonowner | 3. | |
| | | 4. | |
| | | 5. | |
| Domain 3 (Hispanic) | Owner | 3. | |
| | | 4. | |
| | | 5. | |
| | Nonowner | 3. | |
| | | 4. | |
| | | 5. | |
| Domain 7 (Non-Hispanic White or ''Some Other Race'')<br>Domain 5 (Native Hawaiian or Pacific Islander)<br>Domain 6 (Non-Hispanic Asian)<br>Domain 1 (American Indian or Alaska Native On Reservation)<br>Domain 2 (American Indian or Alaska Native Off Reservation) | Owner | 3. | |
| | | 4. | |
| | | 5. | |
| | Nonowner | 3. | |
| | | 4. | |
| | | 5. | |

# Attachment 3.
# Correlation Bias Adjustment Groupings and Factors

| Race/Hispanic Origin Domain | Age | DA sex ratios | Adjustment factor |
|---|---|---|---|
| Black:<br>Domain 4 (Non-Hispanic Black) | 18-29 | 0.90 | 1.08 |
| | 30-49 | 0.89 | 1.10 |
| | 50+ | 0.76 | 1.05 |
| Non-Black:<br>Domain 3 (Hispanic)<br>Domain 7 (Non-Hispanic White or ''Some Other Race'')<br>Domain 5 (Native Hawaiian or Pacific Islander)<br>Domain 6 (Non-Hispanic Asian)<br>Domain 1 (American Indian or Alaska Native On Reservation)<br>Domain 2 (American Indian or Alaska Native Off Reservation) | 18-29 | 1.04 | 1.00* |
| | 30-49 | 1.01 | 1.02 |
| | 50+ | 0.86 | 1.01 |

*This number set to 1.00 due to the inconsistency between DA and A.C.E. Revision II results.

# Chapter 7.
# Assessing the Estimates

## INTRODUCTION

The evaluations of the A.C.E. Revision II estimates may be divided into two categories. One category contains the evaluations that focus on individual error components. The other group consists of comparisons of the relative error between the census and the A.C.E. Revision II estimator.

This chapter provides a brief description of the evaluation studies. The component errors examined by separate studies are sampling error, error from imputation model selection, error due to using inmovers to estimate outmovers in PES-C, synthetic error, error in the identification of the census duplicates as determined by administrative records, error in the identification of computer duplicates as determined by a clerical review, error from inconsistent post-stratification variables, and potential error arising from the automated coding of some cases, called the at-risk coding, in the Revision Sample. The comparisons of relative error between the census and the A.C.E. Revision II estimator include a comparison with Demographic Analysis, the construction of confidence intervals that account for bias as well as random error, and loss function analyses. Also in this category is an examination of the consistency of the estimates of coverage error measured by the A.C.E. Revision II estimator and the Housing Unit Coverage Study (HUCS). Although an adjustment for correlation bias is included in the A.C.E. Revision II estimates, no evaluations address the error in the level of correlation bias or the model used to distribute it across post-strata. The reason is that examining alternative models only accounts for differences in models. Those differences would reflect the variations in how the several models correct the original DSEs for correlation biases, but would not reflect the presence or absence of correlation bias in the corrected DSEs.

## SAMPLING ERROR

Sampling error gives rise to random error, which is quantified by sampling variance. The sampling variance is present in any estimate based on a sample instead of the whole population. The variance estimation methodology is a simplified jackknife with the block clusters being the primary sampling unit. The effect of within-cluster subsampling is implicitly captured in the weighting.

The March 2001 A.C.E. data showed that the simplified jackknife method produces satisfactory variance estimates. Since a correlation bias adjustment was included in the A.C.E. Revision II estimates, the adjustment for correlation bias was recalculated for each replicate. An alternative variance estimation procedure assumed that the form of the correlation bias adjustment was a scalar times the double-sampling estimator. The replication method also accounts for the A.C.E. block cluster sampling.

## SYNTHETIC ERROR EVALUATION

The A.C.E. Revision II has several potential sources of synthetic error. One source involves correcting the individual post-stratum estimates for error estimates at more aggregate levels, such as corrections for correlation bias and measurement coding errors. However, the evaluation of synthetic error focuses on error in small area estimation. Synthetic estimation bias arises when areas in a post-stratum have different coverage error rates, but have the same census coverage correction factor. To assess synthetic estimation bias for a given area, an estimate based on data from the area alone, called a direct estimate, must be developed. Such an estimate is possible for only large areas. In lieu of direct estimates, synthetic estimation bias in undercount estimates is estimated from analysis of "artificial populations" or "surrogate" variables whose geographic distributions are known. These surrogate variables are constructed as best as possible to have patterns similar to coverage error. Sensitivity analyses assess the impact of synthetic estimation bias for these variables.

The evaluation of synthetic error within post-strata uses an artificial population analysis similar to those conducted for ESCAP I and ESCAP II. These studies are documented in Griffin and Malec (2001, 2001b). This time, however, the evaluation compares the A.C.E. Revision II estimates and Census 2000. The study uses loss functions for assessing the effect of synthetic error. The major products are:

- Estimates of the bias in the difference between census loss and A.C.E. Revision II estimator loss.
- Indicator of whether the decision to use the A.C.E. Revision II estimator would have changed due to synthetic error.

## ERROR DUE TO USING INMOVERS TO ESTIMATE OUTMOVERS IN PES-C

The error due to using inmovers to estimate outmovers is unique to the PES-C model for dual system estimation used in the original A.C.E. and the A.C.E. Revision II. For the PES-C model, the members of the P sample are the

residents of the housing units on Census Day. There is some difficulty in identifying all the residents of all the housing units on Census Day because some move prior to the A.C.E. interview. The A.C.E. interview relies on the respondents to identify those who have moved out, the outmovers. Since the outmovers are identified by proxies, many of the outmovers are not recorded. Therefore, the estimate of outmovers is too low. To avoid a bias caused by an underestimate of the number of movers, PES-C uses the number of inmovers to estimate the number of outmovers. The inmovers are those who did not live in the sample blocks on Census Day, but moved in prior to the A.C.E. interview. Theoretically, the number of inmovers in the whole country should equal the number of outmovers. However, the number of inmovers may not equal the number of outmovers in a post-stratum because of circumstances such as economic conditions causing more people to move out of an area than to move into an area.

The first step of the methodology consists of raking the number of outmovers to total inmovers. The distribution of the raked outmovers may better describe the outmovers than the distribution of the inmovers. The A.C.E. Revision II estimates formed by using the number of inmovers are compared with the A.C.E. Revision II estimates calculated using the raked number.

## ERROR FROM IMPUTATION MODEL SELECTION

This project estimates the uncertainty due to choice of imputation model by drawing on the analysis of reasonable alternatives to the imputation model conducted in 2001. See Keathley et al. (2001) for details. The ideal approach would be to repeat the very time-consuming analysis of reasonable alternatives for the A.C.E. Revision II estimator. However, this analysis was not conducted due to limited resources. Instead, an estimate of the additional variance due to the choice of imputation model is developed using the previous A.C.E. work.

Estimates of the variance component for census coverage correction factors that account for the missing data error component due to the imputation of enumeration status, residency status, match status, and the P-sample noninterview adjustment are formed. The replicates used to estimate the missing data variance are used in the loss function analysis to represent the random error due to the choice of the model's imputation for missing data.

## EXAMINING THE QUALITY OF THE COMPUTER DUPLICATES WITH ADMINISTRATIVE RECORDS

Administrative records provide an opportunity to examine the quality of the estimates of duplicate enumerations used in the A.C.E. Revision II estimates. This study uses the Statistical Administrative Records System (StARS) 2000 (Leggieri et al., 2002; Judson, 2000) to assess the effectiveness of the automated methodology used in the Further Study of Person Duplication (FSPD) to identify duplicate enumerations. Secondary goals are to provide data that can be analyzed to determine the nature of the census duplication, so that the information may be used in reducing census duplication in 2010 and to aid in the evaluation of the methodology for the construction of StARS 2000. The study produces a comparison of the estimated amount of census duplication based on administrative records with the estimate from FSPD.

StARS is new methodology that compiles seven administrative records files, including files from IRS, Medicare, HUD, and Selective Service[1]. The evaluation uses a previous match between the census and StARS 2000 to assign an Identification (ID) Number to as many census records as possible. The process of assigning ID Numbers was based on name and address. One pass through the census files used both the address and the name to assign ID Numbers. A second pass used only the name and birth date. A census record was assigned an ID Number only if it was linked with exactly one ID Number.

Census enumerations with the same ID Number are considered duplicates. The method accounts for coincidental agreement of names by requiring assignment of ID Numbers only when exactly one ID Number was linked to the enumeration. In most cases, two people with very similar names and characteristics would have linked to each others' ID Number and would not have been assigned a unique ID Number.

## CLERICAL REVIEW OF COMPUTER DUPLICATES

The study examines accuracy of the FSPD computer identification of duplication in the census by having clerks review the enumerations that the computer designates as duplicates. The clerks determine whether the sets of two enumerations appear to be the same persons. In addition, census enumerations identified as duplicates by administrative records, but not by the computer, also have a clerical review. The potential census duplicates identified by administrative records are a by-product of the evaluation of the computer duplicates using administrative records.

The review is restricted to duplicates between enumerations in the E sample in the A.C.E. blocks and census enumerations outside the search area. Links between P-sample nonmatches and enumerations outside the search area also are reviewed.

The clerical review produces the following:

- Number of E-sample enumerations with false duplicate links identified by the computer.

[1]The Census Bureau obtains administrative data for its StARS database as authorized by Title 13 U.S.C., section 6 and supported by provisions of the Privacy Act of 1974. Under Title 13, the Census Bureau is required to protect the confidentiality of all the information it receives directly from respondents or indirectly from administrative agencies and is permitted only to use that information for statistical purposes.

- Number of E-sample enumerations with missed duplicates identified by administrative records that are correct.
- Number of P-sample nonmatches with false duplicate links identified by the computer.
- Number of P-sample nonmatches with missed duplicates identified by administrative records that are correct.

With these results, the accuracy rate for the computer identification of duplicates in the census and between the P-sample nonmatches and the census can be computed.

## AT-RISK CODING

The study assesses the amount of error at risk due to not having each and every case in the Evaluation Follow-up (EFU) sample reviewed clerically (Adams and Krejsa, 2002). The data collected in the Evaluation Follow-up of the A.C.E. found errors in the coding of E-sample census enumeration status and P-Sample residence and match status that needed to be corrected for the A.C.E. Revision II estimator. Ideally, this would mean recoding the entire A.C.E. sample, but that was not possible because the Evaluation Follow-up collected data in only 2,259 out of the 11,303 A.C.E. sample clusters. Even clerically recoding the 70,000 cases in the Evaluation Follow-up sample was not feasible because of time constraints. A new strategy was devised to provide the most high quality data in the time allowed by restricting the clerical review to the more difficult cases. This strategy reduced the clerical workload to about 25,000, which could be done, and ensured the largest sample possible for the A.C.E. Revision II estimates.

Since the Person Follow-up (PFU) and the Evaluation Follow-up (EFU) questionnaires had been keyed and were available in electronic form, data were combined using an algorithm based on the keyed data and a clerical coding of the categories of cases where the computer did not appear to do a good job.

The method compares the code assigned based on the PFU questionnaire to the code assigned based on the EFU questionnaire, and then, determines the best code. The effectiveness of the computer algorithm is assessed by the agreement between the two new codes, and a comparison with recodes assigned in the fall of 2001 to a subsample of the EFU E sample called the Person Follow-up/ Evaluation Follow-up (PFU/EFU) Review. The PFU/EFU Review is believed to have been the best A.C.E. coding operation.

For the P sample in the Evaluation Follow-up, a coding algorithm for the keyed data from the PFU and EFU questionnaires also was developed. Assessing the quality was not as easy for the nonmatches and unresolved cases as for the matches. Although recodes from the PFU/EFU Review were available for the matches in the P sample, none of the nonmatches or unresolved cases were included.

The categories of cases not sent for clerical review had a high agreement rate between the PFU and EFU codes assigned by the computer algorithm. For the cases in these categories where the PFU and EFU disagreed, the selected code came from the form with more detailed information. Therefore, there are three types of cases in the estimation:

1. The PFU and EFU codes assigned by computer agree.
2. The PFU and EFU codes assigned by computer disagree, but are in a category where there is high consistency between the PFU and EFU codes, and either the PFU form or the EFU form does not have answers to all the questions. The code for the form with complete data is selected.
3. Clerically assigned codes.

The first group is called the "at-risk" cases. These cases may have a higher risk of error than the others because the lack of clerical review, even though the codes assigned by the computer algorithm agree. However, cases in the second group may also have error, although they are in a category with high consistency between the PFU and EFU. For these cases, there is no way to assess the risk of error due to the lack of information on one of the forms.

To assess the potential for error, the at-risk cases are assumed to have the same error rate as cases in their category in the PFU/EFU Review. The potential impact is assessed by comparing the A.C.E. Revision II double-sampling adjustment factors with the double-sampling ratios under the assumption that incorporates the error rates. The double-sampling adjustment factors are described in Chapter 6.

## INCONSISTENCY OF POST-STRATIFICATION VARIABLES

Inconsistency in the E- and P-sample reporting of the characteristics used in defining the post-strata may create a bias in the dual system estimate (DSE). This bias affects the estimation of the P-sample match rate.

The analysis of the post-stratification variables for the A.C.E. Revision II estimator was similar to the investigation done for the original A.C.E. The basic approach was to estimate the inconsistency in the post-stratification variables using the matches, then assume that the rates also held for the nonmatches. The models used for the inconsistency analysis of the original A.C.E. post-strata, described in Haberman and Spencer (2001), were fitted in two steps: (1) models for inconsistency of basic variables,

and (2) derivation of inconsistency probabilities for post-stratification given the inconsistency probabilities of the basic variables. The inconsistency probabilities led to an estimate of the bias in the P-sample match rate that was used to estimate the bias in the DSE. The approach taken for the A.C.E. Revision II estimator is to re-calculate the models in (1) and (2) to reflect revisions in the P-sample post-stratification and repeat the analysis.

To assess the bias due to inconsistency in the post-stratification variables, the A.C.E. Revision II estimates are calculated with a correction to the match rate for the inconsistency. Estimates with and without the correction are then compared.

## CONSISTENCY BETWEEN THE A.C.E. REVISION II ESTIMATOR AND HUCS

The study examines the validity of the A.C.E. Revision II estimates by assessing the consistency in the results from the A.C.E. Revision II estimates and the Housing Unit Coverage Study (HUCS) described in Barrett et al. (2001). Since the A.C.E. Revision II estimates could have been used in the post-censal estimates program that utilizes the average household size in many calculations, it is important to consider the consistency between the A.C.E. Revision II estimates and the HUCS data.

A.C.E. Revision II estimates census coverage for people and HUCS estimates census coverage for housing units. Patterns in the differential coverage for demographic and geographic groups were examined. Similar patterns in the measures of change in census coverage between 1990 and 2000 for demographic and geographic groups are expected. If there is a substantial difference in the census coverage error caused by missing whole households and by missing people within households, the patterns of differential coverage of people and of housing units may not have similar patterns.

If there are demographic or geographic groups where the differential coverage from the A.C.E. Revision II estimator and HUCS is substantially different, the study attempts to describe whether the disagreement is a symptom of problems with the A.C.E. Revision II estimator or HUCS, or the result of legitimate differences in coverage.

## RELATIVE ACCURACY OF THE CENSUS AND A.C.E. REVISION II ESTIMATOR USING DEMOGRAPHIC ANALYSIS

Demographic Analysis (DA) uses vital records, immigration statistics, and Medicare data to obtain an estimate of the population size. Since the methods are somewhat independent of the census, DA provides a method for assessing the relative quality of the census and the A.C.E. Revision II. The consistency of estimates of differential census coverage from the A.C.E. Revision II estimator and DA are assessed for demographic groups.

Estimates of differential census coverage are compared by demographic characteristics, including race, sex, and age. The estimates of population size based on DA are not viewed with as much confidence as the estimates of differential coverage. DA does a better job of measuring differences in coverage between groups than population size.

In addition, sex ratios from the A.C.E. Revision II estimates and DA are compared. The sex ratio is the ratio of males to females and provides a measure of differential coverage of males and females, especially when calculated for race groups.

These comparisons are repeated with 1990 Post-Enumeration Survey and DA estimates to provide a context for viewing the comparisons with the 2000 data. An assessment is conducted to determine whether both methods measure the same change in differential net undercounts from 1990 to 2000.

## RELATIVE ACCURACY OF THE CENSUS AND THE A.C.E. REVISION II ESTIMATOR USING CONFIDENCE INTERVALS AND LOSS FUNCTION ANALYSIS

Two additional methods of assessing the relative accuracy of the census and the A.C.E. Revision II estimates are using confidence intervals for the net undercount rate and a loss function analysis. Confidence intervals for net undercount rates are formed using estimates of net bias and variance. Since most of the data available on the quality of the original A.C.E. is being incorporated in the A.C.E. Revision II estimates, the estimation of the net bias uses the data that were not included. In the loss function analysis, the mean squared error weighted by the reciprocal of the census count is used to estimate loss for levels and shares for counties and places across the nation and within state.

Confidence intervals that incorporate the net bias as well as the variance for the net undercount rate $\hat{U}$ provide a method for comparing the relative accuracy of the census and the A.C.E. Revision II estimates. The net bias in the census coverage correction factor is estimated for each post-stratum. With the estimated bias and variance for each census coverage correction factor, the bias $\hat{B}(\hat{U})$ and variance $\hat{V}$ in the net undercount rate $\hat{U}$ are estimated. Also, 95 percent confidence intervals for the net undercount rate are constructed by

$$(\hat{U} - \hat{B}(\hat{U}) - 2\sqrt{\hat{V}},\ \hat{U} - \hat{B}(\hat{U}) + 2\sqrt{\hat{V}}).$$

Since $\hat{U}=0$ corresponds to no adjustment of the census, one comparison of the relative accuracy of the census and the A.C.E. Revision II estimates is based on an assessment of whether the confidence intervals for the evaluation post-strata cover 0 and $\hat{U}$.

A loss function analysis for levels and shares compares the census and the A.C.E. Revision II estimator for counties and places across the nation and within state. The measure of accuracy used by the loss functions is the weighted mean squared error with the weights set to the reciprocal of the census count for levels and the reciprocal of census share for shares. The motivation for the selected groupings for the loss functions is their potential use in the post-censal estimates. These groupings are:

- Levels
  - All counties with population of 100,000 or less
  - All counties with population greater than 100,000
  - All places with population at least 25,000 but less than 50,000
  - All places with population at least 50,000 but less than 100,000
  - All places with population greater than 100,000
- Shares within state
  - All counties
  - All places
- Shares within U.S.
  - All places with population at least 25,000 but less than 50,000
  - All places with population at least 50,000 but less than 100,000
  - All places with population greater than 100,000
  - All states

# Section II. References


Adams, T. and Krejsa, E. (2001). "ESCAP II: Results of the Person Followup and Evaluation Followup Forms Review," Executive Steering Committee for A.C.E. Policy II, Report 24.

Adams, T. and Krejsa, E. (2002). "A.C.E. Revision II Measurement Subgroup Documentation," DSSD A.C.E. Revision II Memorandum Series #PP-6.

Adams, T. and Liu, X. (2001). "ESCAP II: Evaluation of Lack of Balance and Geographic Errors Affecting Person Estimates," Executive Steering Committee for A.C.E. Policy II, Report 2.

Barrett, D., Beaghen, M., Smith, D., and Burcham, J. (2001). "ESCAP II: Census 2000 Housing Unit Coverage Study," Executive Steering Committee for A.C.E. Policy II, Report 17.

Bean, S. (2001). "ESCAP II: Accuracy and Coverage Evaluation Matching Error," Executive Steering Committee for A.C.E. Policy II, Report 7.

Cantwell, P. and 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, #Q-44.

Childers, D. (2001). "Accuracy and Coverage Evaluation: The Design Document," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised.

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

ESCAP I (2001). "Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy," March 1, 2001. (See www.census.gov/dmd/www/pdf/Escap2.pdf)

ESCAP II (2001). "Report of the Executive Steering Committee for Accuracy and Coverage Evaluation Policy on Adjustment for Non-Redistricting Uses," October 17, 2001. (See www.census.gov/dmd/www/pdf/Recommend2.pdf)

Fay, R. (2001). "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee for A.C.E. Policy II, Report 9, Preliminary Version, October 26, 2001.

Fay, R. (2002). "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee for A.C.E. Policy II, Report 9, Revised Version, March 27, 2002.

Fay, R. (2002b). "Probabilistic Models for Detecting Census Person Duplication," Proceedings of the Survey Research Methods Section, American Statistical Association.

Feldpausch, R. (2001). "ESCAP II: Census Person Duplication and the Corresponding A.C.E. Enumeration Status," Executive Steering Committee for A.C.E. Policy II, Report 6.

Griffin, R. and Malec, D. (2001). "Accuracy and Coverage Evaluation: Assessment of Synthetic Assumption," DSSD Census 2000 Procedures and Operations Memorandum Series, B-14*.

Griffin, R and Malec, D. (2001b). "ESCAP II: Sensitivity Analysis for the Assessment of the Synthetic Assumption," Executive Steering Committee for A.C.E. Policy II, Report 23.

Haberman, S. and Spencer, B. (2001). "Estimation of Inconsistent Post-stratification in the 2000 A.C.E." Paper prepared by Abt Associates Inc. and Spencer Statistics, Inc. under Task Number 46-YABC-7-00001, Contract Number 50-YABC-7-66020.

Haines, D. (2001). "Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Dual System Estimation (U.S.) - Re-issue of Q-37," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-48.

Hogan, H. (1993). "The 1990 Post-Enumeration Survey: Operations and Results," Journal of the American Statistical Association, 88, 1047-1060.

Hogan, H. (2002). "Five Challenges in Preparing Improved Post Censal Population Estimates," DSSD A.C.E. Revision II Memorandum Series #PP-1.

Hogan, H., Kostanich, D., Whitford, D., and Singh, R. (2002). "Research Findings of the Accuracy and Coverage Evaluation and Census 2000 Accuracy," Proceedings of the Survey Research Methods Section, American Statistical Association.

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 #Q-77.

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

Judson, D. (2000). "The Statistical Administrative Records System: System Design and Challenges," Paper presented at the NISS/Telcordia Data Quality Conference, November, 2000.

Keathley, D., Kearney, A., and Bell, W. (2001). "ESCAP II: Analysis of Missing Data Alternatives for the Accuracy and Coverage Evaluation," Executive Steering Committee for A.C.E. Policy II, Report 12.

Kostanich, D. (2003). "A.C.E. Revision II: Summary of Methodology," DSSD A.C.E. Revision II Memorandum Series #PP-35.

Krejsa, E. and Adams, T. (2002). "Results of the A.C.E. Revision II Measurement Coding," DSSD A.C.E. Revision II Memorandum Series #PP-55.

Krejsa, E. and Raglin, D. (2001). "ESCAP II: Evaluation Results for Changes in A.C.E. Enumeration Status," Executive Steering Committee for A.C.E. Policy II, Report 3.

Leggieri, C., Pistiner, A., and Farber, J. (2002). "Methods for Conducting an Administrative Records Experiment in Census 2000," Proceedings of the Survey Research Methods Section, American Statistical Association.

Mule, T. (2001). "ESCAP II: Person Duplication in Census 2000," Executive Steering Committee for A.C.E. Policy II, Report 20.

Mule, T. (2001b). "Accuracy and Coverage Evaluation: Decomposition of Dual System Estimate Components," DSSD Census 2000 Procedures and Operations Memorandum Series #B-8*.

Mule, T. (2002). "Revised Preliminary Estimates of Net Undercounts for Seven Race/Ethnicity Groupings," DSSD A.C.E. Revision II Memorandum Series #PP-2.

Mule, T. (2002b). "Further Study of Person Duplication Statistical Matching and Modeling Methodology," DSSD A.C.E. Revision II Memorandum Series #PP-51.

Mulry, M. and Petroni, R. (2002). "Error Profile for PES-C as Implemented in the 2000 A.C.E.," Proceedings of the Survey Research Methods Section, American Statistical Association.

Nash, F. (2000). "Overview of the Duplicate Housing Unit Operations," Census 2000 Information Memorandum Number 78.

Raglin, D. and Krejsa, E. (2001). "ESCAP II: Evaluation Results for Changes in Mover and Residence Status in the A.C.E.," Executive Steering Committee for A.C.E. Policy II, Report 16.

Robinson, J. G. (2001). "ESCAP II: Demographic Analysis Results," Executive Steering Committee for A.C.E. Policy II, Report 1.

Thompson, J., Waite, P., Fay, R. (2001). "Basis of 'Revised Early Approximation' of Undercounts Released October 17, 2001," Executive Steering Committee for A.C.E. Policy II, Report 9a.

U.S. Census Bureau (2003). "Technical Assessment of A.C.E. Revision II," March 12, 2003. (See www.census.gov/dmd/www/pdf/ACETechAssess.pdf)

Winkler, W. (1995). "Matching and Record Linkage," Business Survey Methods, ed. B. G. Cox et al. (New York: J. Wiley and Sons), 355-384.

Winkler, W. (1999). "Documentation for Record Linkage Software," U.S. Census Bureau, Statistical Research Division.

Yancey, W. (2002). "BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage," U.S. Census Bureau, Statistical Research Division.