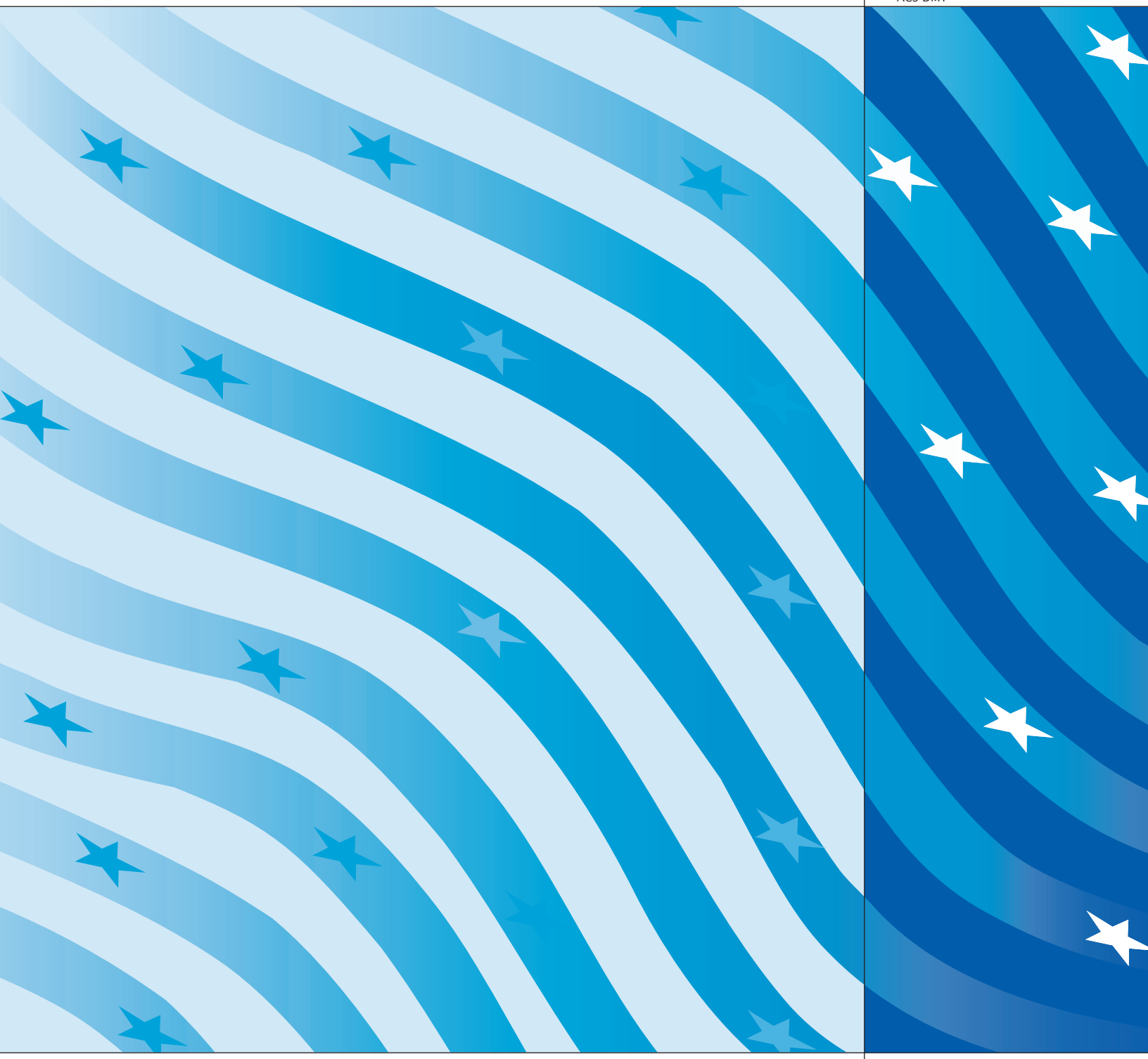
Design and Methodology

American Community Survey

Issued April 2009

ACS-DM1



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

United States Census 2010

ACKNOWLEDGMENTS

The updating of the May 2006 unedited version of this technical report was conducted under the direction of **Susan Schechter**, Chief, American Community Survey Office. **Deborah H. Griffin**, Special Assistant to the Chief, American Community Survey Office, provided overall management and coordination. The American Community Survey program is under the direction of **Arnold A. Jackson**, Associate Director for Decennial Census, and **Daniel H. Weinberg**, Assistant Director for American Community Survey and Decennial Census.

Major contributing authors for this updated 2008 report include **Herman A. Alvarado**, **Mark E. Asiala**, **Lawrence M. Bates**, **Judy G. Belton**, **Grace L. Clemons**, **Kenneth B. Dawson**, **Deborah H. Griffin**, **James E. Hartman**, **Steven P. Hefter**, **Douglas W. Hillmer**, **Jennifer L. Holland**, **Cynthia Davis Hollingsworth**, **Todd R. Hughes**, **Karen E. King**, **Debra L. U. Klein**, **Pamela M. Klein**, **Alfredo Navarro**, **Susan Schechter**, **Nicholas M. Spanos**, **John G. Stiller**, **Anthony G. Tersine, Jr.**, **Nancy K. Torrieri**, **Kai T. Wu**, and **Matthew A. Zimolzak**.

The U. S. Census Bureau is also grateful to staff from Mathematica Policy Research, Inc., who provided valuable comments and revisions to an earlier draft of this report.

Assisting in the production of this report were **Cheryl V. Chambers**, **Destiny D. Cusick**, **Susan L. Hostetter**, **Clive Richmond**, and **Sue Wood**.

The May 2006 unedited version was produced through the efforts of a number of individuals, primarily **Mark E. Asiala**, **Lisa Blumerman**, **Sharon K. Boyer**, **Maryann M. Chapin**, **Thomas M. Coughlin**, **Barbara N. Diskin**, **Donald P. Fischer**, **Brian Gregory**, **Deborah H. Griffin**, **Wendy Davis Hicks**, **Douglas W. Hillmer**, **David L. Hubble**, **Agnes Kee**, **Susan P. Love**, **Lawrence McGinn**, **Marc Meyer**, **Alfredo Navarro**, **Joan B. Peacock**, **David Raglin**, **Nicholas M. Spanos**, and **Lynn Weidman.**

Catherine M. Raymond, **Christine E. Geter**, **Crystal Wade**, and **Linda Chen**, of the Administrative and Customer Services Division (ACSD), **Francis Grailand Hall**, Chief, provided publications and printing management, graphics design and composition, and editorial review for the print and electronic media. **Claudette E. Bennett**, Assistant Division Chief, and **Wanda Cevis**, Chief, Publications Services Branch, provided general direction and production management.

Design and Methodology

American Community Survey

Issued April 2009

ACS-DM1

U.S. Department of Commerce
Gary Locke,
Secretary

Vacant,
Deputy Secretary

Economics and Statistics Administration
Vacant,
Under Secretary for Economic Affairs

U.S. CENSUS BUREAU
Thomas L. Mesenbourg,
Acting Director

SUGGESTED CITATION

U.S. CENSUS BUREAU
Design and Methodology
American Community Survey

U.S. Government Printing Office,
Washington, DC,
2009.

Economics and Statistics Administration

Vacant,
Under Secretary
for Economic Affairs

U.S. CENSUS BUREAU

Thomas L. Mesenbourg,
Acting Director

Thomas L. Mesenbourg,
Deputy Director and
Chief Operating Officer

Arnold A. Jackson,
Associate Director
for Decennial Census

Daniel H. Weinberg,
Assistant Director
for ACS and Decennial Census

Foreword

The American Community Survey—A Revolution in Data Collection

The American Community Survey (ACS) is the cornerstone of the U.S. Census Bureau's effort to keep pace with the nation's ever-increasing demands for timely and relevant data about population and housing characteristics. The new survey provides current demographic, social, economic, and housing information about America's communities every year—information that until now was only available once a decade. Implementation of the ACS is viewed by many as the single most important change in the way detailed decennial census information is collected since 1940, when the Census Bureau introduced statistical sampling as a way to collect "long-form" data from a sample of households.

The ACS and the reengineering of the decennial census will affect data users and the public for decades to come. Beginning with the survey's full implementation in 2005, the ACS has replaced the census long-form questionnaire that was sent to about one-in-six addresses in Census 2000. As with the long form, information from the ACS will be used to administer federal and state programs and distribute more than $300 billion a year in federal funds. Obtaining more current data throughout the decade from the ACS will have long-lasting value for policy and decision-making across federal, state, local, and tribal governments, the private sector, and virtually every local community in the nation.

The Beginning. In 1994, the Census Bureau started developing what became the ACS with the idea of continuously measuring the characteristics of population and housing, instead of collecting the data only once a decade with each decennial census. Testing started in four counties across the country and with encouraging results, the testing expanded to 31 test sites by 1999. Realizing that a continuous program would also be collecting information during a decennial census, the sample was increased to about 800,000 addresses in 2000 and continued its demonstration period through 2004. This was a national sample that yielded results for the country, states, and most geographic areas with 250,000 or more population.

Comparing the 2000 ACS data with the results from the Census 2000 long form proved that the idea of a monthly survey was feasible and would generate quality data. With some changes to the sample design and other methodologies, the ACS was fully implemented in 2005 with a sample of three million addresses each year. A sample also was implemented in Puerto Rico, where the survey is known as the Puerto Rico Community Survey (PRCS). In 2006, a sample of group quarters facilities was included so that estimates from the ACS and the PRCS would reflect complete characteristics of all community residents.

Annual results will be available for all areas by 2010. Currently, the ACS publishes single-year data for all areas with populations of 65,000 or more. Among the roughly 7,000 areas that meet this threshold are all states, all congressional districts, more than 700 counties, and more than 500 places. Areas with populations less than 65,000 will require the use of multiyear estimates to reach an appropriate sample size for data publication. In 2008, the Census Bureau will begin releasing 3-year estimates for areas with populations greater than 20,000. And, we plan to release the first 5-year estimates for all census tracts and block groups starting in 2010. These multiyear estimates will be updated annually, with data published for the largest areas in both 1-, 3-, and 5-year formats, and for those meeting the 3-year threshold in both 3- and 5-year formats. Of course, even the smallest communities will be able to obtain ACS data based on 5-year estimates annually.

The 2008 release of the ACS Design and Methodology Report. This *ACS Design and Methodology Report* is an update of the first unedited version that was released in 2006. We released that draft version because of the need to provide data users with information about the first full sample year of the survey. The version released in 2006 provided design and methodology information for the 2005 ACS only.

This version of the *ACS Design and Methodology Report* includes updated information reflecting survey changes, modifications, and improvements through the end of 2007. Many portions of each chapter have been revised. We hope that data users find this report helpful and that it will aid in improving the public's understanding of the ACS statistical design and the methods it uses.

Success of the Program. The ACS program has been successful in large part because of the innovation and dedication of many people who have worked so hard to bring it to this point in time. With this publication of the *ACS Design and Methodology Report*, many individuals—both past and current—deserve special congratulations. From those early beginnings with a handful of designers, survey methodologists, and technical experts, through full implementation, countless individuals have contributed to the survey's successful implementation.

All of the primary survey activities are designed and managed by the staff at Census Bureau headquarters in Suitland, MD, who continually strive to improve the accuracy of the ACS estimates, streamline its operations, analyze its data, conduct important research and evaluation to achieve greater efficiencies and effectiveness, and serve as educational resources and experts for the countless data users who come to the Census Bureau in need of technical assistance and help. In addition, the Census Bureau's field partners provide many of the critical day-to-day activities that are the hub of the ACS existence. The ACS, which is the largest household survey conducted by the federal government, could not be accomplished without the dedication and effort of staff at the Census Bureau's National Processing Center (NPC) in Jeffersonville, IN; the Census Bureau telephone call centers in Jeffersonville, IN; Hagerstown, MD; and Tucson, AZ; and the thousands of field representatives across the country who collect ACS data. In addition, the ACS field operations are run by Census Bureau survey managers in the NPC, telephone call centers and the twelve Regional Offices, all of whom add immeasurably to the smooth and efficient running of a very complex and demanding survey operation.

Finally, the ACS would not have achieved its success without the continued cooperation of millions of Americans who willingly provide the data that are collected each year. The data they provide are invaluable and contribute daily to the survey's exceptional accomplishments. Sincere thanks are extended to each and every respondent who took the time and effort to participate in this worthwhile endeavor.

We invite you to suggest ways in which we can enhance this report in the future. Also, please remember to look for updated versions of this report as the ACS continues in the coming years.

CONTENTS

CONTENTS

CONTENTS

CONTENTS

Tables—Con.

Chapter 1.
Introduction

The American Community Survey (ACS) is a relatively new survey conducted by the U.S. Census Bureau. It uses a series of monthly samples to produce annually updated data for the same small areas (census tracts and block groups) formerly surveyed via the decennial census long-form sample. Initially, 5 years of samples will be required to produce these small-area data. Once the Census Bureau has collected 5 years of data, new small-area data will be produced annually. The Census Bureau also will produce 3-year and 1-year data products for larger geographic areas. The ACS includes people living in both housing units (HUs) and group quarters (GQs). The ACS is conducted throughout the United States and in Puerto Rico, where it is called the Puerto Rico Community Survey (PRCS). For ease of discussion, the term ACS is used here to represent both surveys.

This document describes the basic ACS design and methodology as of the 2007 data collection year. The purpose of this document is to provide data users and other interested individuals with documentation of the methods used in the ACS. Future updates of this report are planned to reflect additional design and methodology changes. This document is organized into 15 chapters. Each chapter includes an overview, followed by detailed documentation, and a list of references.

Chapter 2 provides a short summary of the history and evolution of the ACS, including its origins, the development of a survey prototype, results from national testing, and its implementation procedures for the 2007 data collection year.

Chapters 3 and 4 focus on the ACS sample. Chapter 3 describes the survey frame, including methods for updating it. Chapter 4 documents the ACS sample design, including how samples are selected.

Chapters 5 and 6 describe the content covered by the ACS and define several of its critical basic concepts. Chapter 5 provides information on the survey's content development process and addresses the process for considering changes to existing content. Chapter 6 explains the interview and residence rules used in ACS data collection and includes definitions of key concepts covered in the survey.

Chapters 7, 8, and 9 cover data collection and data capture methods and procedures. Chapter 7 focuses on the methods used to collect data from respondents who live in HUs, while Chapter 8 focuses on methods used to interview those living in GQs. Chapter 9 discusses the ACS language assistance program, which serves as a critical support for data collection.

Chapters 10, 11, and 12 focus on ACS data processing, weighting and estimation, and variance estimation methods. Chapter 10 discusses data preparation activities, including the coding required to produce files for certain data processing activities. Chapter 11 is a technical discussion of the process used to produce survey weights, while Chapter 12 describes the methods used to produce variance estimates.

Chapters 13 and 14 cover the definition, production, and dissemination of ACS data products. Chapter 13 explains the process used to produce, review, and release ACS data. Chapter 14 explains how to access ACS data products and provides examples of each type of data product.

Chapter 15 documents the methods used in the ACS to control for nonsampling error, and includes examples of measures of quality produced annually to accompany each data release.

A glossary of terms and acronyms used in this report appear at the end. Also, note that the first release of this report, issued May 2006, contained an extensive list of appendixes that included copies of forms and letters used in the data collection operations for the ACS. The size of these documents and the changing nature of some of them precludes their inclusion here. Readers are encouraged to review the ACS Web site <www.census.gov> if data collection materials are needed or are of interest.

Chapter 2.
Program History

2.1 OVERVIEW

Continuous measurement has long been viewed as a possible alternative method for collecting detailed information on the characteristics of population and housing; however, it was not considered a practical alternative to the decennial census long form until the early 1990s. At that time, demands for current, nationally consistent data from a wide variety of users led federal government policymakers to consider the feasibility of collecting social, economic, and housing data continuously throughout the decade. The benefits of providing current data, along with the anticipated decennial census benefits in cost savings, planning, improved census coverage, and more efficient operations, led the Census Bureau to plan the implementation of continuous measurement, later called the American Community Survey (ACS). After years of testing, outreach to stakeholders, and an ongoing process of interaction with key data users—especially those in the statistical and demographic communities—the Census Bureau expanded the ACS to full sample size for housing units (HUs) in 2005 and for group quarters (GQs) in 2006.

The history of the ACS can be divided into four distinct stages. The concept of continuous measurement was first proposed in the 1990s. Design proposals were considered throughout the period 1990 to 1993, the design and early proposals stage. In the development stage (1994 through 1999), the Census Bureau tested early prototypes of continuous measurement for a small number of sites. During the demonstration stage (2000 to 2004), the Census Bureau carried out large-scale, nationwide surveys and produced reports for the nation, the states, and large geographic areas. The full implementation stage began in January 2005, with an annual HU sample of approximately 3 million addresses throughout the United States and 36,000 addresses in Puerto Rico. And in 2006, approximately 20,000 group quarters were added to the ACS so that the data fully describe the characteristics of the population residing in geographic areas.

Design Origins and Early Proposals

In 1981, Leslie Kish introduced the concept of a rolling sample design in the context of the decennial census (Kish 1981). During the time that Kish was conducting his research, the Census Bureau also recognized the need for more frequently updated data. In 1985, Congress authorized a mid-decade census, but funds were not appropriated. In the early 1990s, Congress expressed renewed interest in an alternative to the once-a-decade census. Based on Kish's research, the Census Bureau began developing continuous measurement methods in the mid-1990s.

The Census Bureau developed a research proposal for continuous measurement as an alternative to the collection of detailed decennial census sample data (Alexander 1993g), and Charles Alexander, Jr. developed three prototypes for continuous measurement (Alexander 1993i). Based on staff assessments of operational and technical feasibility, policy issues, cost, and benefits (Alexander 1994e), the Census Bureau selected one prototype for further development. Designers made several decisions during prototype development. They knew that if the survey was to be cost-efficient, the Census Bureau would need to mail it. They also determined that like the decennial census, response to the survey would be mandatory and therefore, a nonresponse follow-up would be conducted. It was decided that the survey would use both telephone and personal visit nonresponse follow-up methods. In addition, the designers made critical decisions regarding the prototype's key definitions and concepts (such as the residence rule), geographic makeup, sampling rates, and use of population controls.

With the objective of producing 5-year cumulations for small areas at the same level of sampling reliability as the long-form census sample, a monthly sample size of 500,000 HUs was initially suggested (Alexander 1993i), but this sample size drove costs into an unacceptable range. When potential improvements in nonsampling error were considered, it was determined that a monthly sample size of 250,000 would generate an acceptable level of reliability.

Development

Development began with the establishment of a permanent Continuous Measurement Staff in 1994. This staff continued the development of the survey prototype and identified several design elements that proved to be the foundation of the ACS:

- Data would be collected continuously by using independent monthly samples.
- Three modes of data collection would be used: mailout, telephone nonresponse follow-up, and personal visit nonresponse follow-up.
- The survey reference date for establishing HU occupancy status, and for many characteristics, would be the day the data were collected. Certain data items would refer to a longer reference period (for example, "last week," or "past 12 months").
- The survey's estimates would be controlled to intercensal population and housing estimates.
- All estimates would be produced by aggregating data collected in the monthly surveys over a period of time so that they would be reported annually based on the calendar year.

The documentation of early development took several forms. Beginning in 1993, a group of 20 reports, known as the Continuous Measurement Series (Alexander 1992; 1993a–1993i; 1994a–1994f; and 1995a–1995b; Alexander and Wetrogan 1994; Cresce 1993), documented the research that led to the final prototype design. Plans for continuous measurement were introduced formally at the American Statistical Association's (ASA) Joint Statistical Meetings in 1995. Love et al. (1995) outlined the assumptions for a successful survey, while Dawson et al. (1995) reported on early feasibility studies of collecting survey information by telephone. Possible modifications of continuous measurement data also were discussed (Weidman et al. 1995).

Operational testing of the ACS began in November 1995 at four test sites: Rockland County, NY; Brevard County, FL; Multnomah County, OR; and Fulton County, PA. Testing was expanded in November 1996 to encompass areas with a variety of geographic and demographic characteristics, including Harris County, TX; Fort Bend County, TX; Douglas County, NE; Franklin County, OH; and Otero County, NM. This testing was undertaken to validate methods and procedures and to develop cost models for future implementation; it resulted in revisions to the prototype design and identified additional areas for research. Further research took place in numerous areas, including small-area estimation (Chand and Alexander 1996), estimation methods (Alexander et al. 1997), nonresponse follow-up (Salvo and Lobo 1997), weighting in ACS tests (Dahl 1998), item nonresponse (Tersine 1998), response rates (Love and Diffendal 1998), and the quality of rural data (Kalton et al. 1998).

Operational testing continued, and in 1998 three counties were added: Kershaw County, SC; Richland County, SC; and Broward County, FL. The two counties in South Carolina were included to produce data to compare with the 1998 Census Dress Rehearsal results, and Broward County was substituted for Brevard County. In 1999, testing expanded to 36 counties in 26 states (U.S. Census Bureau 2004e). The sites were selected to represent different combinations of county population size, difficulty of enumeration, and 1990–1995 population growth. The selection incorporated geographic diversity as well as areas representing different characteristics, such as racial and ethnic diversity, migrant or seasonal populations, American Indian reservations, changing economic conditions, and predominant occupation or industry types. Additionally, the Census Bureau selected sites with active data users who could participate in evaluating and improving the ACS program. Based on the results of the operational tests, revisions were made to the prototype and additional areas for research were identified.

Tests of methods for the enumeration of people living in GQs also were held in 1999 and 2001. These tests focused on the methodology for visiting GQs, selecting resident samples, and conducting interviews. The tests selected GQ facilities in all 36 test counties and used the procedures developed in the prototyping stage. Results of the tests led to modification of sampling techniques and revisions to data collection methods.

While the main objective of the development phase testing was to determine the viability of the methodologies utilized, it also generated usable data. Data tables and profiles were produced and released in 1999, providing data on demographic, social, economic, and housing topics. Additionally, public use microdata sample (PUMS) files were generated for a limited number of locations during the period of 1996 through 1999. PUMS files show data for a sample of all HUs, with information on the housing and population characteristics of each selected unit. All identifying information is removed and other disclosure avoidance techniques are used to ensure confidentiality.

Demonstration

In 2000, a large-scale demonstration was undertaken to assure Congress and other data users that the ACS was capable of producing the demographic, social, economic, and housing data previously obtained from the decennial census long-form sample.

The demonstration stage of the ACS was initially called the Census 2000 Supplementary Survey (C2SS). Its primary goal was to provide critical assessments of feasibility, quality, and comparability with Census 2000 so as to demonstrate the Census Bureau's ability to implement the ACS fully. Although ACS methods had been successful at the test sites, it was vital to demonstrate national implementation. Additional goals included refining procedures, improving the understanding of the cost structure, improving cost projections, exploring data quality issues, and assuring users of the reliability and usefulness of ACS data.

The C2SS was conducted in 1,239 counties, of which 36 were ACS test counties and 1,203 were new to the survey. It is important to note that only the 36 ACS test counties used the proposed ACS sample design. The others used a primary sampling unit stratified design similar to the Current Population Survey (CPS). The annual sample size increased from 165,000 HUs in 1999 to 866,000 HUs in 2000. The test sites remained in the sample throughout the C2SS, and through 2004 were sampled at higher rates than the C2SS counties. This made 3-year estimates from the ACS in these counties comparable to the planned 5-year period estimates of a fully implemented ACS, as well as to data from Census 2000.

Eleven reports issued during the demonstration stage analyzed various aspects of the program. There were two types of reports: methodology and data quality/comparability. The methodology reports reviewed the operational feasibility of the ACS. The data quality/comparability reports compared C2SS data with the data from Census 2000, including comparisons of 3 years of ACS test site data with Census 2000 data for the same areas.

Report 1 (U.S. Census Bureau 2001) found that the C2SS was operationally successful, its planned tasks were completed on time and within budget, and the data collected met basic Census Bureau quality standards. However, the report also noted that certain areas needed improvement. Specifically, due to their coinciding with the decennial census, telephone questionnaire assistance (TQA) and failed-edit follow-up (FEFU) operations were not staffed sufficiently to handle the large workload increase. The evaluation noted that the ACS would improve planning for the 2010 decennial census and simplify its design, and that implementing the ACS, supported by an accurate Master Address File (MAF) and Topologically Integrated Geographic Encoding and Referencing (TIGER®) database, promised to improve decennial census coverage. Report 6 (U.S. Census Bureau 2004c) was a follow-up evaluation on the feasibility of utilizing data from 2001 and 2002. The evaluation concluded that the ACS was well-managed, was achieving the desired response rates, and had functional quality control procedures.

Report 2 (U.S. Census Bureau 2002) concluded that the ACS would provide a reasonable alternative to the decennial census long-form sample, and added that the timeliness of the data gave it advantages over the long form. This evaluation concluded that, while ACS methodology was sound, its improvement needed to be an ongoing activity.

A series of reports compared national, state, and limited substate 1-year period estimates from the C2SS and Census 2000. Reports 4 and 10 (U.S. Census Bureau 2004a; 2004g) noted differences; however, the overall conclusion was that the research supported the proposal to move forward with plans for the ACS.

Report 5 (U.S. Census Bureau 2004b) analyzed economic characteristics and concluded that estimates from the ACS and the Census 2000 long form were essentially the same. Report 9 (U.S. Census Bureau 2004f) compared social characteristics and noted that estimates from both methods were consistent, with the exceptions of disability and ancestry. The report suggested the completion of further research on these and other issues.

A set of multiyear period estimates (1999–2001) from the ACS test sites was created to help demonstrate the usability and reliability of ACS estimates at the county and census tract geographic levels. Results can be found in Reports 7 and 8 (U.S. Census Bureau 2004d; 2004e). These comparisons with Census 2000 sample data further confirmed the comparability of the ACS and the Census 2000 long-form estimates and identified potential areas of research, such as variance reduction in subcounty estimates.

At the request of Congress, a voluntary methods test also was conducted during the demonstration phase. The test, conducted between March and June of 2003, was designed to examine the impact that a methods change from mandatory to voluntary response would have on mail response, survey quality, and costs. Reports 3 and 11 (U.S. Census Bureau 2003b; 2004h) examined the results. These reports identified the major impacts of instituting voluntary methods, including reductions in response rates across all three modes of data collection (with the largest drop occurring in traditionally low response areas), reductions in the reliability of estimates, and cost increases of more than $59 million annually.

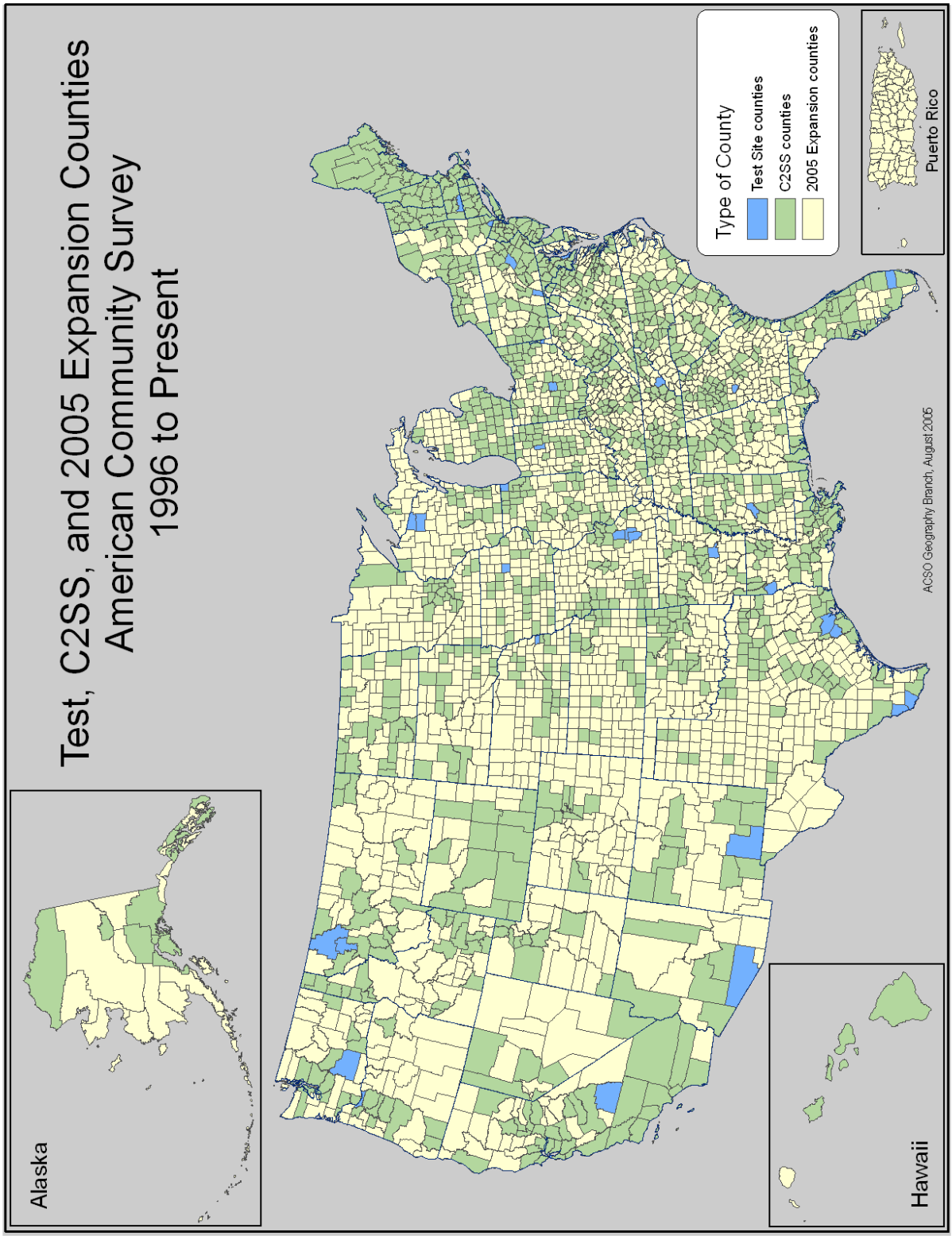
Full Implementation

In 2003, with full implementation of the ACS approaching, the American Community Survey Office (ACSO) came under the direction of the Associate Director for the Decennial Census. While the Census Bureau's original plan was to implement the ACS fully in 2003, budget restrictions pushed back full HU implementation of the ACS and PRCS to January 2005. The GQ component of the ACS was implemented fully in January 2006.

With full implementation, the ACS expanded from 1,240 counties in the C2SS and ACS test sites to all 3,141 counties in the 50 states and the District of Columbia, and to all 78 municipios in Puerto Rico (Figure 2.1). The annual ACS sample increased from 800,000 addresses in the demonstration phase to 3 million addresses in full implementation. Workloads for all ACS operations increased by more than 300 percent. Monthly mailouts from the National Processing Center (NPC) went from approximately 67,000 to 250,000 addresses per month. Telephone nonresponse follow-up workloads, conducted from three telephone call centers, expanded from 25,000 calls per month to approximately 85,000. More than 3,500 field representatives (FRs) across the country conducted follow-up visits at 40,000 addresses a month, up from 1,200 FRs conducting follow-ups at 11,000 addresses each month in 2004. And, approximately 36,000 addresses in Puerto Rico were sampled every year, using the same three modes of data collection as the ACS. Beginning in 2006, the ACS sampled 2.5 percent of the population living in GQs. This included approximately 20,000 GQ facilities and 195,000 people in GQs in the United States and Puerto Rico.

With full implementation beginning in 2005, population and housing profiles for 2005 first became available in the summer of 2006 and have been available every year thereafter for specific geographic areas with populations of 65,000 or more. Three-year period estimates, reflecting combined data from the 2005–2007 ACS, will be available for the first time late in 2008 for specific areas with populations of 20,000 or more, and 5-year period estimates, reflecting combined data from the 2005–2009 ACS, will be available late in 2010 for areas down to the smallest block groups, census tracts, and small local governments. Beginning in 2010, and every year thereafter, the nation will have a 5-year period estimate available as an alternative to the decennial census long-form sample; this will serve as a community information resource that shows change over time, even for neighborhoods and rural areas.

Figure 2.1 **Test, C2SS, and 2005 Expansion Counties, American Community Survey, 1996 to Present**



2.2 STAKEHOLDERS AND CONTRIBUTORS

Consultations with stakeholders began early in the ACS development process, with the goals of gaining feedback on the overall approach and identifying potential pitfalls and obstacles. Stakeholders included data users, federal agencies, and others with an interest in the survey. A wide range of contacts encompassed federal, state, tribal, and local governments, advisory committees, professional organizations, and other data users at many levels. These groups provided their insights and expertise to the staff charged with developing the ACS.

The Census Bureau established special-purpose advisory panels in partnership with the Committee on National Statistics of the National Academies of Science (NAS) to identify issues of relevance in survey design. The ACS staff undertook meetings, presentations, and other activities to support the ACS in American Indian and Alaska Native areas. These activities included meetings with tribal officials and liaisons, attendance at the National Conference of American Indians, and continued interactions with the Advisory Committee for the American Indian and Alaska Native Populations. A Rural Data Users Conference was held in May 1998 to discuss issues of concern to small areas and populations. Numerous presentations were made at annual meetings of the ASA and other professional associations.

Data users also were given opportunities to learn more about the ACS through community workshops held during the development phase. From March 1996 to November 1999, 31 town hall-style meetings were held throughout the country, with more than 600 community members attending the meetings. A series of three regional outreach meetings, in Dallas, TX; Grand Rapids, MI; and Seattle, WA, was held in mid-2004, with an overall attendance of more than 200 individuals representing data users, academicians, the media, and local governments.

Meetings with the Decennial Census Advisory Committee, the Census Advisory Committee of Professional Associations, and the Race and Ethnic Advisory Committees provided opportunities for ACS staff to discuss methods and receive specific advice on methods and procedures to improve the quality of the survey and the value of the ACS data. The Census Bureau's Field Division Partnership and Data Services Staff and regional directors all played prominent roles in communicating the message of the ACS. These groups provided valuable input to the decision-making process. Further, the ACS staff regularly briefed several oversight groups, including the Office of Management and Budget (OMB), the Government Accountability Office (GAO), and the Inspector General of the U.S. Department of Commerce (DOC). The Census Bureau also briefed Congress regularly on multiple aspects of the ACS; these briefings began during the early states of the ACS and continued on a regular basis.

Changes based on stakeholder input were important in shaping the design and development of the ACS and continue to influence its future form, including questionnaire content and design. For example, a "Symposium on the ACS: Data Collectors and Disseminators" took place in September 2000. It focused on the data uses and needs of the private sector. A periodic newsletter, the *ACS Alert*, was established to share program information and solicit feedback. The Interagency Committee for the ACS was formed in 2000 to discuss the content and methods of the ACS and how the survey meets the needs of federal agencies. In 2003, the ACS Federal Agency Information Program was developed to ensure that federal agencies having a current or potential use for data from the ACS would have the assistance they need in using the data. In 2007, the Committee on National Statistics issued an important report, "Using The American Community Survey: Benefits and Challenges," which reflected the input of many stakeholders and addressed the interpretation of ACS data by a wide variety of users. Finally, the Census Bureau senior leadership, as well as the ACS staff, routinely participated in conferences, meetings, workshops, and panels to build support and understanding of the survey and to ensure that users' needs and interests were being met.

Efforts were also made toward the international sharing of the Census Bureau's experiences with the development and implementation of the ACS. Presentations were given to many international visitors who came to the Census Bureau to learn about surveys and censuses. Papers were shared and presentations have been made at many international conferences' working sessions and meetings. Outreach to stakeholders was a key component of launching and gaining support for the ACS program, and its importance and prominence continue.

2.3 REFERENCES

Alexander, C. H. (1992). "An Initial Review of Possible Continuous Measurement Designs." Internal Census Bureau Reports CM-2. Washington, DC: U.S. Census Bureau, 1992.

Alexander, C. H. (1993a). "A Continuous Measurement Alternative for the U.S. Census." Internal Census Bureau Reports CM-10. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993b). "Determination of Sample Size for the Intercensal Long Form Survey Prototype." Internal Census Bureau Reports CM-8. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993c). "Including Current Household Surveys in a 'Cumulated Rolling Sample' Design." Internal Census Bureau Reports CM-5. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993d). "Overview of Continuous Measurement for the Technical Committee." Internal Census Bureau Reports CM-4. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993e). "Overview of Research on the 'Continuous Measurement' Alternative for the U.S. Census." Internal Census Bureau Reports CM-11. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993f). "Preliminary Conclusions About Content Needs for Continuous Measurement." Internal Census Bureau Reports CM-6. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993g). "Proposed Technical Research to Select a Continuous Measurement Prototype." Internal Census Bureau Reports CM-3. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993h). "A Prototype Design for Continuous Measurement." Internal Census Bureau Reports CM-7. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1993i). "Three General Prototypes for a Continuous Measurement System." Internal Census Bureau Reports CM-1. Washington, DC: U.S. Census Bureau, 1993.

Alexander, C. H. (1994a). "An Idea for Using the Continuous Measurement (CM) Sample as the CPS Frame." Internal Census Bureau Reports CM-18, Washington, DC: U.S. Census Bureau, 1994.

Alexander, C. H. (1994b). "Further Exploration of Issues Raised at the CNSTAT Requirements Panel Meeting." Internal Census Bureau Reports CM-13. Washington, DC: U.S. Census Bureau, 1994.

Alexander, C. H. (1994c). "Plans for Work on the Continuous Measurement Approach to Collecting Census Content." Internal Census Bureau Reports CM-16. Washington, DC: U.S. Census Bureau, 1994.

Alexander, C. H. (1994d). "Progress on the Continuous Measurement Prototype." Internal Census Bureau Reports CM-12. Washington, DC: U.S. Census Bureau, 1994.

Alexander, C. H. (1994e). "A Prototype Continuous Measurement System for the U.S. Census of Population and Housing." Internal Census Bureau Reports CM-17. Washington, DC: U.S. Census Bureau, 1994.

Alexander, C. H. (1994f). "Research Tasks for the Continuous Measurement Development Staff." Internal Census Bureau Reports CM-15. Washington, DC: U.S. Census Bureau, 1994.

Alexander, C. H. (1995a). "Continuous Measurement and the Statistical System." Internal Census Bureau Reports CM-20. Washington, DC: U.S. Census Bureau, 1995.

Alexander, C. H. (1995b). "Some Ideas for Integrating the Continuous Measurement System into the Nation's System of Household Surveys." Internal Census Bureau Reports CM-19. Washington, DC: U.S. Census Bureau, 1995.

Alexander, C. H., S. Dahl, and L. Weidmann (1997). "Making Estimates from the American Community Survey." Paper presented to the Annual Meeting of the American Statistical Association (ASA), Anaheim, CA, August 1997.

Alexander, C. H. and S. I.Wetrogran (1994). "Small Area Estimation with Continuous Measurement: What We Have and What We Want." Internal Census Bureau Reports CM-14. Washington, DC: U.S. Census Bureau, 1994.

Chand, N. and C. H. Alexander (1996). "Small Area Estimation with Administrative Records and Continuous Measurement." Presented at the Annual Meeting of the American Statistical Association, 1996.

Cresce, Art (1993). "'Final' Version of JAD Report and Data Tables from Content and Data Quality Work Team." Internal Census Bureau Reports CM-9. Washington, DC: U.S. Census Bureau, 1993.

Dahl, S. (1998a). "Weighting the 1996 and 1997 American Community Surveys." Presented at American Community Survey Symposium, 1998.

Dahl, S. (1998b). "Weighting the 1996 and 1997 American Community Surveys." *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 1998, pp.172–177.

Dawson, Kenneth, Susan Love, Janice Sebold, and Lynn Weidman (1995). "Collecting Census Long Form Data Over the Telephone: Operational Results of the 1995 CM CATI Test." Presented at 1996 Annual Meeting of the American Statistical Association, 1995.

Kalton, G., J. Helmick, D. Levine, and J. Waksberg (1998). "The American Community Survey: The Quality of Rural Data, Report on a Conference." Prepared by Westat, June 29, 1998.

Kish, Leslie (1981). "Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau: An Analysis, Review, and Response." Washington, DC: U.S. Government Printing Office, 1981.

Love, S., C. Alexander, and D. Dalzell (1995). "Constructing a Major Survey: Operational Plans and Issues for Continuous Measurement." Proceedings of the Survey Research Methods Section. Alexandria, VA: American Statistical Association, pp.584–589.

Love, S. and G. Diffendal (1998). "The 1996 American Community Survey Monthly Response Rates, by Mode." Presented to the American Community Survey Symposium, 1998.

Salvo, J. and J. Lobo (1997). "The American Community Survey: Non-Response Follow-Up in the Rockland County Test Site." Presented to the Annual Meeting of the American Statistical Association, 1997.

Tersine, A. (1998). "Item Nonresponse: 1996 American Community Survey." Paper presented to the American Community Survey Symposium, March 1998.

U.S. Census Bureau (2001). "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: July 2001, Report 1: Demonstrating Operational Feasibility." Washington, DC, July 2001.

U.S. Census Bureau (2002b). "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: May 2002, Report 2: Demonstrating Survey Quality." Washington, DC, May 2002.

U.S. Census Bureau (2003b). "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 3: Testing the Use of Voluntary Methods." Washington, DC, December 2003.

U.S. Census Bureau (2004a). "Census 2000 Topic Report No. 8: Address List Development in Census 2000." Washington, DC, 2004.

U.S. Census Bureau (2004a). "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 4: Comparing General Demographic and Housing Characteristics With Census 2000." Washington, DC, May 2004.

U.S. Census Bureau (2004a). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey, Report 6: The 2001–2002 Operational Feasibility Report of the American Community Survey. Washington, DC, 2004.

U.S. Census Bureau (2004b). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 5: Comparing Economic Characteristics With Census 2000. Washington, DC, May 2004.

U.S. Census Bureau (2004b). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 7: Comparing Quality Measures: The American Community Survey's Three-Year Averages and Census 2000's Long Form Sample Estimates. Washington, DC, June 2004.

U.S. Census Bureau 2004c. Housing Recodes 2004. Internal U.S. Census Bureau data processing specification, Washington, DC.

U.S. Census Bureau (2004e). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 8: Comparison of the ACS 3-year Average and the Census 2000 Sample for a Sample of Counties and Tracts. Washington, DC, June 2004.

U.S. Census Bureau (2004f). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 9: Comparing Social Characteristics with Census 2000. Washington, DC, June 2004.

U.S. Census Bureau (2004g). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 10: Comparing Selected Physical and Financial Housing Characteristics with Census 2000. Washington, DC, July 2004.

U.S. Census Bureau (2004h). Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 11: Testing Voluntary Methods—Additional Results. Washington, DC, December 2004.

Weidman, L., C. Alexander, G. Diffendahl, and S. Love. (1995). Estimation Issues for the Continuous Measurement Survey. *Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association, pp. 596–601, <www.census.gov/acs/www/AdvMeth/Papers/ACS/Paper5.htm>.

Chapter 3.
Frame Development

3.1 OVERVIEW

The sampling frame used for the American Community Survey (ACS) is an extract from the national Master Address File (MAF), which is maintained by the U.S. Census Bureau and is the source of addresses for the ACS, other Census Bureau demographic surveys, and the decennial census. The MAF is the Census Bureau's official inventory of known living quarters (housing units [HUs] and group quarters [GQs] facilities) and selected nonresidential units (public, private, and commercial) in the United States and Puerto Rico. It contains mailing and location address information, geocodes, and other attribute information about each living quarter. (A geocoded address is one for which state, county, census tract, and block have been identified.)

The MAF is linked to the Topologically Integrated Geographic Encoding and Referencing (TIGER®) system. TIGER® is a database containing a digital representation of all census-required map features and related attributes. It is a resource for the production of maps, data tabulation, and the automated assignment of addresses to geographic locations in geocoding.

The initial MAF was created for Census 2000 using multiple sources, including the 1990 Address Control File, the U.S. Postal Service's (USPS's) Delivery Sequence File (DSF), field listing operations, and addresses supplied by local governments through partnership operations. The MAF was used as the initial frame for the ACS, in its state of existence at the conclusion of Census 2000. The Census Bureau continues to update the MAF using the DSF and various automated, clerical, and field operations, such as the Demographic Area Address Listing (DAAL).

The remainder of this chapter provides detailed information on the development of the ACS sampling frame. Section B provides basic information about the MAF and its contents. Sections C and D describe the MAF development and update activities for HUs in the United States and Puerto Rico. Section E describes the MAF development and ACS GQ data collection activities. Finally, Section F describes the ACS extracts from the MAF.

3.2 MASTER ADDRESS FILE CONTENT

The MAF is the Census Bureau's official inventory of known HUs and GQs in the United States and Puerto Rico. Each HU and GQ is represented by a separate MAF record that contains some or all of the following information: geographic codes, a mailing and/or location address, the physical state of the unit or any relationship to other units, residential or commercial status, latitude and longitude coordinates, and source and history information indicating the operation(s) (see Section C) that add/update the record. This information is gathered from the MAF and provided to ACS in files called MAF extracts (see Section F).

The geographic codes in the MAF, some of which come from the TIGER® database, identify a variety of areas, including states, counties, county subdivisions, places,[1] American Indian areas, Alaska Native areas, Hawaiian Homelands, census tracts, block groups, and blocks. Two of the MAF's important geographic code sets are the Census 2000 tabulation geography set, based on the January 1, 2000, legal boundaries, and the current geography set, based on the January 1 legal boundaries of the most recent year (for example, MAF extracts received in July 2007 reflect legal boundaries as of January 1, 2007). The geographic codes associated with each MAF record

[1] "Place" is defined by the Census Bureau as "A concentration of population either legally bounded as an incorporated place, or delineated for statistical purposes as a census designated place (in Puerto Rico, a comunidad or zona urbana). See census designated place, consolidated city, incorporated place, independent city, and independent place." From <http://www.census.gov/geo/www/tiger/glossary.html#glossary>.

are assigned by the TIGER® database. Because each record contains a variety of geographic codes, it is possible to sort MAF records according to different geographic hierarchies. ACS operations generally require sorting by state, county, census tract, and block.

The MAF contains both city-style and non-city-style mailing addresses. A city-style address is one that uses a structure number and street name format; for example, 201 Main Street, Anytown, ST 99988. Additionally, city-style addresses usually appear in a numeric sequence along a street and often follow parity conventions, such as all odd numbers occurring on one side of the street and even numbers on the other side. They often contain information used to uniquely identify individual units in multiple-unit structures, such as apartment buildings or rooming houses. These are known as unit designators, and are part of the mailing address.

A non-city-style mailing address is one that uses a rural route and box number format, a post office (PO) box format, or a general delivery format. Examples of these types of addresses are RR 2, Box 9999, Anytown, ST 99988; P.O. Box 123, Anytown, ST 99988; and T. Smith, General Delivery, Anytown, ST 99988.

In the United States, city-style addresses are most prevalent in urban and suburban areas, and accounted for 94.4 percent of all residential addresses in the MAF at the conclusion of Census 2000. Most city-style addresses represent both the mailing and location addresses of the unit. City-style addresses are not always mailing addresses, however. Some residents at city-style addresses receive their mail at those addresses, while others use non-city-style addresses (Census 2000b). For example, a resident could have a location address of 77 West St. and a mailing address of P.O. Box 123. In other cases, city-style addresses ("E-911 addresses") have been established so that state emergency service providers can find a house even though mail is delivered to a rural route and box number.

Non-city-style mailing addresses are prevalent in rural areas and represented approximately 2.5 percent of all residential addresses in the MAF at the conclusion of Census 2000. Because these addresses do not provide specific information about the location of a unit, finding a rural route and box number address in the field can be difficult. To help locate non-city-style addresses in the field, the MAF often contains a location description of the unit and its latitude and longitude coordinates.[2] The presence of this information in the MAF makes field follow-up operations possible.

Both city-style and non-city-style addresses can be either residential or nonresidential. A residential address represents a housing unit in which a person or persons live or could live. A nonresidential address represents a structure, or a unit within a structure, that is used for a purpose other than residence. While the MAF includes many nonresidential addresses, it is not a comprehensive source of such addresses (Census 2000b).

The MAF also contains some address records that are classified as incomplete because they lack a complete city-style or non-city-style address. Records in this category often are just a description of the unit's location, and usually its latitude and longitude. This incomplete category accounted for the remaining 3.1 percent of the United States residential addresses in the MAF at the conclusion of Census 2000.

For details on the MAF, including its content and structure, see Census (2000b).

3.3 MASTER ADDRESS FILE DEVELOPMENT AND UPDATING FOR THE UNITED STATES HOUSING UNIT INVENTORY

MAF Development in the United States

For the 1990 decennial and earlier censuses, address lists were compiled from several sources (commercial vendors, field listings, and others). Before 1990, these lists were not maintained or updated after a census was completed. Following the 1990 census, the Census Bureau decided to develop and maintain a master address list to support the decennial census and other Census Bureau survey programs in order to avoid the need to rebuild the address list prior to each census.

[2] For example, "E side of St. Hwy, white house with green trim, garage on left side."

The MAF was created by merging city-style addresses from the 1990 Address Control File;[3] field listing operations;[4] the USPS's DSF; and addresses supplied by local governments through partnership operations, such as the Local Update of Census Addresses (LUCA)[5] and other Census 2000 activities, including the Be Counted Campaign.[6] At the conclusion of Census 2000, the MAF contained a complete inventory of known HUs nationwide.

MAF Improvement Activities and Operations

MAF maintenance is an ongoing and complex task. New HUs are built continually, older units are demolished, and the institution of addressing schemes to allow emergency response personnel to find HUs with noncity mailing addresses render many older addresses obsolete. Maintenance of the MAF occurs through a coordinated combination of automated, clerical, and field operations designed to improve existing MAF records and keep up with the nation's changing housing stock and associated addresses. With the completion of Census 2000, the Census Bureau implemented several short-term, one-time operations to improve the quality of the MAF. These operations included count question resolution (CQR), MAF/TIGER® reconciliation, and address corrections from rural directories. For the most part, these operations were implemented to improve the addresses recognized in Census 2000 and their associated characteristics.

Some ongoing improvement operations are designed to deal with errors remaining from Census 2000, while others aim to keep pace with post-Census 2000 address development. In the remainder of this section, several ongoing operations are discussed, including DSF updates, Master Address File Geocoding Office Resolution (MAFGOR), ACS nonresponse follow-up updates, and Demographic Area Address Listing (DAAL) updates. We also discuss the Community Address Updating System (CAUS), which has been employed in rural areas. Table 3.1 summarizes the development and improvement activities.

Table 3.1 **Master Address File Development and Improvement**

Initial Input	Improvements (POST-2000)
1990 Decennial Census address control file	DSF updates
USPS Delivery Sequence File (DSF)	Master Address File Geocoding Office Resolutions (MAFGOR)
Local government updates	ACS nonresponse follow-up
Other Census 2000 activities	Community Address Updating System (CAUS)
	Other Demographic Area Address Listing (DAAL) Operations

Delivery Sequence File. The DSF is the USPS's master list of all delivery-point addresses served by postal carriers. The file contains specific data coded for each record, a standardized address and ZIP code, and codes that indicate how the address is served by mail delivery (for example, carrier route and the sequential order in which the address is serviced on that route). The DSF record for a particular address also includes a code for delivery type that indicates whether the address is business or residential. After Census 2000, the DSF became the primary source of new city-style addresses used to update the MAF. DSF addresses are not used for updating non-city-style addresses in the MAF because those addresses might provide different (and unmatchable) address representations for HUs whose addresses already exist in the MAF. New versions of the DSF are shared with the Census Bureau twice a year, and updates or refreshes to the MAF are made at those times.

[3] The Address Control File is the residential address list used in the 1990 Census to label questionnaires, control the mail response check-in operation, and determine the response follow-up workload (Census 2000, pp. XVII–1).

[4] In areas where addresses were predominantly non-city-style, the Census Bureau created address lists through a door-to-door canvassing operation (Census 2000, pp. VI–2).

[5] The 1999 phase of the LUCA program occurred from early March through mid-May 1999 and involved thousands of local and tribal governments that reviewed more than 10 million addresses. The program was intended to cover more than 85 percent of the living quarter addresses in the United States in advance of Census 2000. The Census Bureau validated the results of the local or tribal changes by rechecking the Census 2000 address list for all blocks in which the participating governments questioned the number of living quarter addresses.

[6] The Be Counted program provided a means to include in Census 2000 those people who may not have received a census questionnaire or believed they were not included on one. The program also provided an opportunity for people who had no usual address on Census Day to be counted. The Be Counted forms were available in English, Spanish, Chinese, Korean, Tagalog, and Vietnamese. For more information, see Carter (2001).

When DSF updates do not match an existing MAF record, a new record is created in the MAF. These new records, which could be new HUs, are then compared to the USPS Locatable Address Conversion Service (LACS), which indicates whether the new record is merely an address change or is new housing. In this way, the process can identify duplicate records for the same address.

For additional details on the MAF update process via the DSF, see Hilts (2005).

MAFGOR. MAFGOR is an ongoing clerical operation in all Census Bureau regional offices, in which geographic clerks examine groups of addresses, or "address clusters" representing addresses that do not geocode to the TIGER® database. Reference materials available commercially, from local governments and on the Internet, are used to add or correct street features, street feature names, or the address ranges associated with streets in the TIGER® database. This process increases the Census Bureau's ability to assign block geocodes to DSF addresses. At present, MAFGOR operations are suspended until the 2010 Census Address Canvassing and field follow-up activities are completed.

Address Updates From ACS Nonresponse Follow-Up. Field representatives (FRs) can obtain address corrections for each HU visited during the personal visit nonresponse follow-up phase of the ACS. This follow-up is completed for a sample of addresses. The MAF is updated to reflect these corrections.

For additional details on the MAF update process for ACS updates collected at time of interview, see Hanks, et al. (2008).

DAAL. DAAL is a combination of operations, systems, and procedures associated with coverage improvement, address list development, and automated listing for the CAUS and the demographic household surveys. The objective of DAAL is to update the inventory of HUs, GQs, and street features in preparation for sample selection for the ACS and surveys such as the Current Population Survey (CPS), the National Health Interview Survey (NHIS), and the Survey of Income and Program Participation (SIPP).

In a listing operation such as DAAL, a defined land area—usually a census tabulation block—is traveled in a systematic manner, while an FR records the location and address of every structure where a person lives or could live. Listings for DAAL are conducted on laptop computers using the Automated Listing and Mapping Instrument (ALMI) software. The ALMI uses extracts from the current MAF and TIGER® databases as inputs. Functionality in the ALMI allows users to edit, add, delete, and verify addresses, streets, and other map features; view a list of addresses associated with the selected geography; and view and denote the location of HUs on the electronic map. Compared to information once collected by paper and pencil, ALMI allows for the standardization of data collected through edits and defined data entry fields, standardization of field procedures, efficiencies in data transfer, and timely reflection of the address and feature updates in MAF and TIGER®. For details on DAAL, see Perrone (2005).

CAUS. The CAUS program is designed specifically to address ACS coverage concerns. The Census Bureau recognized that the DSF, being the primary source of ACS frame updates, does not adequately account for changes in predominantly rural areas of the nation where city-style addresses generally are not used for mail delivery. CAUS, an automated field data collection operation, was designed to provide a rural counterpart to the update of city-style addresses received from the DSF. CAUS improved coverage of the ACS by (1) adding addresses that exist but do not appear in the DSF, (2) adding non-city-style addresses in the DSF that do not appear on the MAF, (3) adding addresses in the DSF that also appear in the MAF but are erroneously excluded from the ACS frame, and (4) deleting addresses that appear in the MAF but are erroneously included in the ACS frame.

Implemented in September 2003, CAUS focused its efforts on census blocks with high concentrations of non-city-style addresses and suspected growth in the HU inventory. Of the approximately 8.2 million blocks nationwide, the CAUS universe comprised the 750,000 blocks where DSF updates are not used to provide adequate coverage. CAUS blocks were selected by a model-based method that used information gained from previous field data collection efforts and administrative records to predict where CAUS work was needed. At present, the CAUS program is suspended until the 2010 Census Address Canvassing and field follow-up activities are completed. For details on the CAUS program and its block selection methodology, see Dean (2005).

All of these MAF improvement activities and operations contribute to the overall update of the MAF. Its continual evaluation and updating are planned and will be described in future releases of this report.

It is expected that the 2010 Census address canvassing and enumeration operations will improve the coverage and quality of the MAF. Field operations to support the 2010 Census will enable HU and GQ updates, additions, and deletions to be identified, collected, and used to update the MAF. The Census Bureau began its Census 2010 operations in 2007. The operations will include several nationwide field canvassing and enumeration operations and will obtain address data through cooperative efforts with tribal, county, and local governments to enhance the MAF. The MAF extracts used by the ACS for sample selection will be improved by these operations. ACS and Census 2010 planners are working together closely to assess the impact of the decennial operations on the ACS.

3.4 MASTER ADDRESS FILE DEVELOPMENT AND UPDATING FOR PUERTO RICO

The Census Bureau created an initial MAF for Puerto Rico through field listing operations. This MAF did not include mailing addresses because, in Puerto Rico, Census 2000 used an Update/Leave methodology through which a census questionnaire was delivered by an enumerator to each living quarter. The MAF update activities that took place from 2002 to 2004 were focused on developing mailing addresses, updating address information, and improving coverage through yearly updates.

MAF Development in Puerto Rico

MAF development in Puerto Rico also used the Census 2000 operations as its foundation. These operations in Puerto Rico included address listing, Update/Leave, the LUCA, and the Be Counted Campaign.

For details on the Census 2000 for Puerto Rico, see Census Bureau (2004b).

The Census 2000 procedures and processing systems were designed to capture, process, transfer, and store information for the conventional three-line mailing address. Mailing addresses in Puerto Rico generally incorporate the urbanization name (neighborhood equivalent), which creates a four-line address. Use of the urbanization name eliminates the confusion created when street names are repeated in adjacent communities. In some instances, the urbanization name is used in lieu of the street name.

The differences between the standard three-line address and the four-line format used in Puerto Rico created problems during the early MAF building stages. The resulting file structure for the Puerto Rico MAF was the same as that used for states in the United States, so it did not contain the additional fields required to handle the more complex Puerto Rico mailing address. These processing problems did not adversely impact Census 2000 operations in the United States because the record structure was designed to accommodate the standard U.S. three-line address. However, in Puerto Rico, where questionnaire mailout was originally planned as the primary means of collecting data, the three-line address format turned out to be problematic. As a result, it is not possible to calculate the percentage of city-style, non-city-style, and incomplete addresses in Puerto Rico from Census 2000 processes.

MAF Improvement Activities and Operations in Puerto Rico

Because of these address formatting issues, the MAF for Puerto Rico as it existed at the conclusion of Census 2000 required significant work before it could be used by the ACS. The Census Bureau had to revise the address information in the Puerto Rico MAF. This effort involved splitting the address information into the various fields required to construct a mailing address using Puerto Rico addressing conventions.

The Census Bureau contracted for updating the list of addresses in the Puerto Rico MAF. Approximately 64,000 new Puerto Rico HUs have been added to the MAF since Census 2000, with each address geocoded to a municipio, tract, and block. The Census Bureau also worked with the USPS

DSF for Puerto Rico to extract information on new HU addresses. Matching the USPS file to the existing MAF was only partially successful because of inconsistent naming conventions, missing information in the MAF, and the existence of different house numbering schemes (USPS versus local schemes).

Data collection activities in Puerto Rico began in November 2004. The Census Bureau is pursuing options for the ongoing collection of address updates in Puerto Rico. This may include operations comparable to those that exist in the United States, such as DSF updates, MAFGOR, and CAUS. Future versions of this document will include discussions of these operations and MAF development and updating in Puerto Rico.

3.5. MASTER ADDRESS FILE DEVELOPMENT AND UPDATING FOR SPECIAL PLACES AND GROUP QUARTERS IN THE UNITED STATES AND PUERTO RICO

MAF Development for Special Places and GQs

In preparation for Census 2000, the Census Bureau developed an inventory of special places (SPs) and GQs. SPs are places such as prisons, hotels, migrant farm camps, and universities. GQs are contained within SPs, and include college and university dormitories and hospital/prison wards. The SP/GQ inventory was developed using data from internal Census Bureau lists, administrative lists obtained from various federal agencies, and numerous Census 2000 operations such as address listing, block canvassing, and the SP/GQ Facility Questionnaire operation. Responses to the SP/GQ Facility Questionnaire identified GQs and any HUs associated with the SP. Similar to the HU MAF development process, local and tribal governments had an opportunity to review the SP address list. In August 2000, after the enumeration of GQ facilities, the address and identification information for each GQ was incorporated into the MAF.

MAF Improvement Activities and Operations for Special Places and GQs

As with the HU side of the MAF, maintenance of the GQ universe is an ongoing and complex task. The earlier section on MAF Improvement Activities and Operations for HUs mentions short-term/one-time operations (such as CQR and MAF/TIGER® reconciliation) that also updated GQ information. Additionally, the Census Bureau completed a GQ geocoding correction operation to fix errors (mostly census block geocodes) associated with college dormitories in the MAF and TIGER®.

Information on the new GQ facilities and updated address information for existing GQ facilities are collected on an ongoing basis by listing operations such as DAAL, which also includes the CAUS in rural areas. This information is used to update the MAF. Additionally, it is likely that DSF updates of city-style address areas are providing the Census Bureau with new GQ addresses; however, the DSF does not identify such an address as a GQ facility.

A process to supplement these activities was developed to create an updated GQ universe from which to select the ACS sample. The ACS GQ universe for 2007 was constructed by merging the updated SP/GQ inventory file, extracts from the MAF, and a file of those seasonal GQs that were closed on April 1, 2000 (but might have been open if visited at another time of year). To supplement the ACS GQ universe, the Census Bureau obtained a file of federal prisons and detention centers from the Bureau of Prisons and a file from the U.S. Department of Defense containing military bases and vessels. The Census Bureau also conducted Internet research to identify new migrant worker locations, new state prisons, and state prisons that had closed.

ACS FRs use the Group Quarters Facility Questionnaire (GQFQ) to collect updated address and geographic location information. The ACS will use the updates collected via the GQFQ to provide more accurate information for subsequent visits to a facility, as well as to update the ACS GQ universe. For more information about the GQFQ, see the section titled Group Quarters Facility Questionnaire—Initial GQ Contact in Section B.2 of Chapter 8.

In addition to the major decennial operations that will collect and provide updates for GQs, ACS and Census 2010 planners are evaluating the feasibility of a repeatable operation to extract information on new GQ facilities from administrative sources, including data provided by members of

the Federal and State Cooperative Program for Population Estimates (FSCPE). If this approach is successful, it likely will provide a cost-effective mechanism for updating the GQ universe for the ACS during the intercensal years. For more information on SP and GQ issues, see Bates (2006a).

3.6 AMERICAN COMMUNITY SURVEY EXTRACTS FROM THE MASTER ADDRESS FILE

The MAF data are provided to ACS in files called MAF extracts. These MAF extracts contain a subset of the data items in the MAF. The major classifications of variables included in the MAF extracts are: address variables, geocode variables, and source and status variables (see Section B).

The MAF, as an inventory of living quarters (HUs and GQs) and some nonresidential units, is a dynamic entity. It contains millions of addresses that reflect ongoing additions, deletions, and changes; these include current addresses, as well as those determined to no longer exist. MAF users, such as the ACS, define the set of valid addresses for their programs.

Since the ACS frame must be as complete as possible, filtering rules are applied during the creation of the ACS extracts to minimize both overcoverage and undercoverage and obtain an inclusive listing of addresses. For example, the ACS includes units that represent new construction units, some of which may not exist yet. The ACS also includes other housing units that are not geocoded, which means that the address is one that cannot be linked to a county, census tract, and block. In addition, the ACS includes units that are "excluded from delivery statistics" (EDS); these units often are those under construction, i.e., the housing unit is being constructed and has an address, but the USPS is not yet delivering to the address. In this regard, the ACS filtering rules differ from those for the Census 2000 and the 2004 Census Test, both of which excluded EDS and ungeocoded addresses. The 2006 Census Test filter included EDS, but excluded ungeocoded records.

The filter is reviewed each year and may be enhanced as the ACS learns about its sample addresses and more about the coverage and content of the MAF. For a record to be eligible for the ACS survey, it must meet the conditions set forth in the filter. In general, the ACS sampling frame contains several classes of units, including HUs that existed during Census 2000, post-census additions from the DSF, additions from the DAAL, CQR additions and reinstatements, additions from special censuses and census tests, and Census 2000 deletions that persist in the DSF.

Filtering rules change, and with them, the ACS frame. One change was implemented in 2003 when ungeocoded addresses in counties not part of mail-out/mail-back areas (areas where mail is the major mode of data collection) were excluded from the ACS sample.

As discussed above, the ACS attempts to create a sampling frame that is as accurate as possible by minimizing both overcoverage and undercoverage. In the process, the ACS filter rules can lead to net overcoverage, reflecting some duplicate and ineligible units. This overcoverage has been estimated to be approximately 2.0 to 3.7 percent for the years 2002–2006, see Hakanson (2007).

For details on the ACS requirements for MAF extracts, see Bates (2006b). For more information on the ACS sample selection, see Chapter 4. For a description of data collection procedures for these different kinds of addresses, see Chapter 7. For details on the MAF, its coverage, and the implications of extract rules on the ACS frame, see Shapiro and Waksberg (1999) and Hakanson (2007).

3.7 REFERENCES

Bates, Lawrence M. (2006a). "Creating the Group Quarters Universe for the American Community Survey for Sample Year 2007." Internal U.S. Census Bureau Memorandum From D. Whitford to L. Blumerman, Draft, Washington, DC, October 30, 2006.

Bates, Lawrence M. (2006b). "Geographic Products Requirements for the American Community Survey. REVISED for July 2006 Delivery." Internal U.S. Census Bureau Memorandum From D. Kostanich to R. LaMacchia, Draft, Washington, DC, June 19, 2006.

Carter, Nathan E. (2001). "Be Counted Campaign for Census 2000." *Proceedings of the Annual Meeting of the American Statistical Association*, August 5–9, 2001. Washington, DC: U.S. Census Bureau, DSSD.

Dean, Jared (2005). “Updating the Master Address File: Analysis of Adding Addresses via the Community Address Updating System.” Washington, DC.

Hakanson, Amanda (2007). “National Estimate of Coverage of the MAF for 2006,” Internal U.S. Census Bureau Memorandum From D. Whitford to R. LaMacchia, Washington, DC, September 28, 2007.

Hanks, Shawn C., Jeremy Hilts, Daniel Keefe, Paul L. Riley, Daniel Sweeney, and Alicia Wentela (2008). “Software Requirements Specification for Address Updates From the Demographic Area Address Listing (DAAL) Operations.” Version 1.0, Washington, DC, March 26, 2008.

Hilts, Jeremy (2005). “Software Requirement Specification for Updating the Master Address File From the U.S. Postal Service’s Delivery Sequence File.” Version 7.0, Washington, DC, April 18, 2005.

Perrone, Susan (2005). “Final Report for the Assessment of the Demographic Area Address Listing (DAAL) Program.” Internal U.S. Census Bureau Memorandum From R. Killion to R. LaMacchia, Washington, DC, November 9, 2005.

Shapiro, Gary and Joseph Waksberg (1999). “Coverage Analysis for the American Community Survey Memo.” Final Report Submitted by Westat to the U.S. Census Bureau, Washington, DC, November 1999.

U.S. Census Bureau (2000). “Census 2000 Operational Plan.” Washington, DC, December 2000.

U.S. Census Bureau (2000b). “MAF Basics.” Washington, DC, 2000.

U.S. Census Bureau (2004b). “Census 2000 Topic Report No. 14: Puerto Rico.” Washington, DC, 2004.

Chapter 4.
Sample Design and Selection

4.1 OVERVIEW

The American Community Survey (ACS) and Puerto Rico Community Survey (PRCS) each consist of two separate samples: housing unit (HU) addresses and persons in group quarters (GQ) facilities. As described in Chapter 3, the sampling frames from which these samples are drawn are derived from the Census Bureau's Master Address File (MAF). The MAF is the Census Bureau's official inventory of known living quarters and selected nonresidential units in the United States and Puerto Rico. Independent HU address samples are selected for each of the 3,141 counties and county equivalents in the United States, including the District of Columbia, for the ACS. Similarly, for the PRCS, address samples are selected for each of the 78 municipalities in Puerto Rico. The first full-implementation county-level samples of HU addresses were selected in 2004 and fielded in 2005.[1] Each year, approximately 3 million HU addresses in the United States and 36,000 HU addresses in Puerto Rico are selected. The first full-implementation samples of GQ facilities and persons were selected independently within each state, as well as the District of Columbia and Puerto Rico, for use in 2006. Each year, approximately 2.5 percent of the expected number of residents in GQ facilities are included in the ACS and the PRCS, respectively. Details of the data collection methods are provided in Chapters 7 and 8.

This chapter presents details on the selection of the HU address and GQ samples. In some hard-to-reach areas in Alaska, referred to as Remote Alaska, several sampling and data collection processes have been modified. The section on Remote Alaska sampling at the end of this chapter describes the differences in sampling and data collection methodology for Remote Alaska.

4.2 HOUSING UNIT SAMPLE SELECTION

There are two phases of HU address sampling for each county.[2] First-phase sampling includes two stages and involves a series of processes that result in the annual ACS sample of addresses. First-phase sampling is performed twice a year and these two annual processes are referred to as main and supplemental sampling, respectively. During first-phase sampling, blocks are assigned to the sampling strata, the sampling rates are calculated, and the sample is selected.[3] During the second phase of sampling, a sample of addresses for which neither a mail questionnaire nor a telephone interview has been completed is selected for computer-assisted personal interviewing (CAPI). This is referred to as the CAPI sample. Figure 4.1 provides a visual overview of the HU address sampling process.

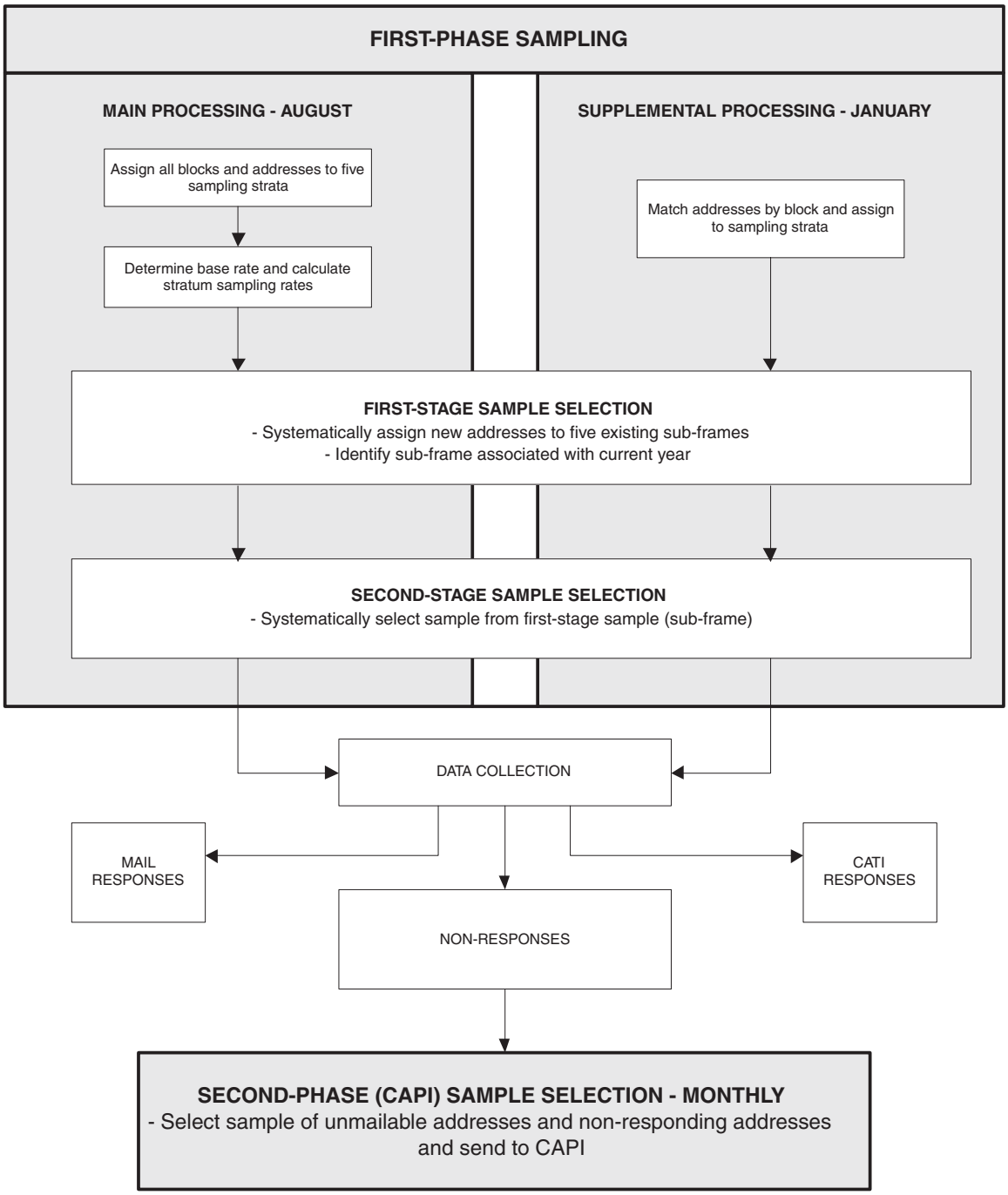
First-Phase Sample

The first step of sampling is to assign each address on the sampling frame to one of the five sampling strata by block. This process is discussed in detail in section B.1.b. Also included in this process are two separate stages of sampling. The first-stage of sampling maintains five distinct partitions of the addresses on the sampling frame for each county. This is accomplished by systematically sorting and assigning addresses that are new to the frame to one of the five partitions or subframes.[4] Each subframe is a representative county sample. These subframes have been assigned to specific years and are rotated each year. The subframes maintain their annual designation over time. Finally the sampling rates are determined for each stratum for the current

[1] In the remainder of this chapter, the term "county" refers to counties, county equivalents, and municipalities.
[2] Throughout this chapter, "addresses" refers to valid ACS addresses that have met the filter criteria (Bates, 2006).
[3] Note that the second-stage sampling rates are calculated once annually during main sampling and these rates are used in supplemental sampling also.
[4] All existing addresses retain their previous assignment to one of the 5-year subframes. The five subframes were created to meet the requirement that no addresses can be in sample more than once in a 5-year period.

sample year. This is discussed in Section B.1.c. During the second stage of sampling, a sample of the addresses in the current year's subframe is selected and allocated to different months for data collection. This process is described in Section B.1.d. and B.1.e.

FIGURE 4.1
SELECTING THE SAMPLES OF HOUSING UNIT ADDRESSES



Main and Supplemental Sampling

Two separate sampling operations are carried out at different times of the year: (1) main sampling occurs in August and September preceding the sample year, and (2) supplemental sampling occurs in January and February of the sample year. This allows an opportunity for new addresses to have a chance of selection during supplemental sampling. The ACS sampling frames for both main and supplemental sampling are derived from the most recently updated MAF, so the sampling frames for the main and supplemental sample selections differ for a given year. The MAF available at the time of main sampling, obtained in the July preceding the sample year, reflects address updates from October of the preceding year through March of that year. The MAF available at the time of the supplemental sample selection, obtained in January of the sample year, reflects address updates from April through September of the preceding year.

For the main sample, addresses are selected from the subframe assigned to the sample year. These sample addresses are allocated systematically, in a predetermined sort order, to all 12 months of the sample year. During supplemental sampling, addresses new to the frame are systematically assigned to the five subframes. The new addresses in the current year's subframe are sampled and are systematically assigned to the months of April through December of the sample year for data collection.

Assigning Addresses to the Second-Stage Sampling Strata. Before the first stage of address sampling can proceed for each year's main sampling, each block must be assigned to one of the five sampling strata. The ACS produces estimates for geographic areas having a wide range of population sizes. To ensure that the estimates for these areas have the desired level of reliability, areas with smaller populations must be sampled at higher rates relative to those areas with larger populations. To accomplish this, each block and its constituent addresses are assigned to one of five sampling strata, each with a unique sampling rate. The stratum assignment for a block is based on information about the set of geographic entities—referred to as sampling entities—which contain the block, or on information about the size of the census tract that the block is located in, as discussed below. Sampling entities are defined as:

- Counties.
- Places with active and functioning governments.[5]
- School districts.
- American Indian Areas/Alaska Native Areas/Hawaiian Home Lands (AIANHH).
- American Indian Tribal Subdivisions with active and functioning governments.
- Minor civil divisions (MCDs) with active and functioning governments in 12 states.[6]
- Census designated places (in Hawaii only).

The sampling stratum for most blocks is based on the measure of size (MOS) for the smallest sampling entity to which any part of the block belongs. To calculate the MOS for a sampling entity, block-level counts of addresses are derived from the main MAF. This count is converted to an estimated number of occupied HUs by multiplying it by the proportion of HUs in the block that were occupied in Census 2000. For American Indian and Alaska Native Statistical Areas (AIANSA[7]) and Tribal Subdivisions, the estimated number of occupied HUs is also multiplied by the proportion of its population that responded as American Indian or Alaska Native (either alone or in combination) in Census 2000. For each sampling entity, the estimate is summed across all blocks in the entity and is referred to as the MOS for the entity. In AIANSAs if the sum of these estimates across all

[5] Functioning governments have elected officials who can provide services and raise revenue.
[6] The 12 states are considered "strong" MCD states and are: Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Wisconsin.
[7] AINSA is a general term used to describe American Indian and Alaska Native Village statistical areas. For detailed technical information on the Census Bureau's American Indian and Alaska Native Area's Geographic Program for Census 2000, see Federal Register Notice Vol. 65, No. 121, June 22, 2000.

blocks is nonzero, then this sum becomes the MOS for the AIANSA. If it is zero (due to a zero census count of American Indians or Alaska Natives), the occupied HU estimate for the AIANSA is the MOS for the AIANSA (see Hefter, 2006a, for additional details). Each block is then assigned the smallest MOS of all the sampling entities in which the block is contained and is referred to as Smallest Entity Measure of Size, or SEMOS.

If the SEMOS is greater than or equal to 1,200, the stratum assignment for the block is based on the MOS for the census tract that contains it. The MOS for each tract (TMOS) is obtained by summing the estimated number of occupied HUs across all of its blocks. Using SEMOS and TMOS, blocks are assigned to the five strata as defined in Table 4.1 below. These strata are consistent with the sampling categories used in Census 2000 except for the category for sampling entities with MOS less than 800, which has been split into two categories for ACS.

Table 4.1 **Sampling Strata Thresholds for the ACS/PRCS**

Stratum	Smallest Entity Measure of Size (SEMOS) and Tract Measure of Size (TMOS)
Blocks in large sampling entities (SEMOS >1,200) and large tracts	TMOS >2,000
Blocks in large sampling entities (SEMOS >1,200) and small tracts	TMOS ≤2,000
Blocks in small sampling entities	800 ≤SEMOS ≤1,200
Blocks in smaller sampling entities	200 ≤SEMOS <800
Blocks in smallest sampling entities	SEMOS < 200

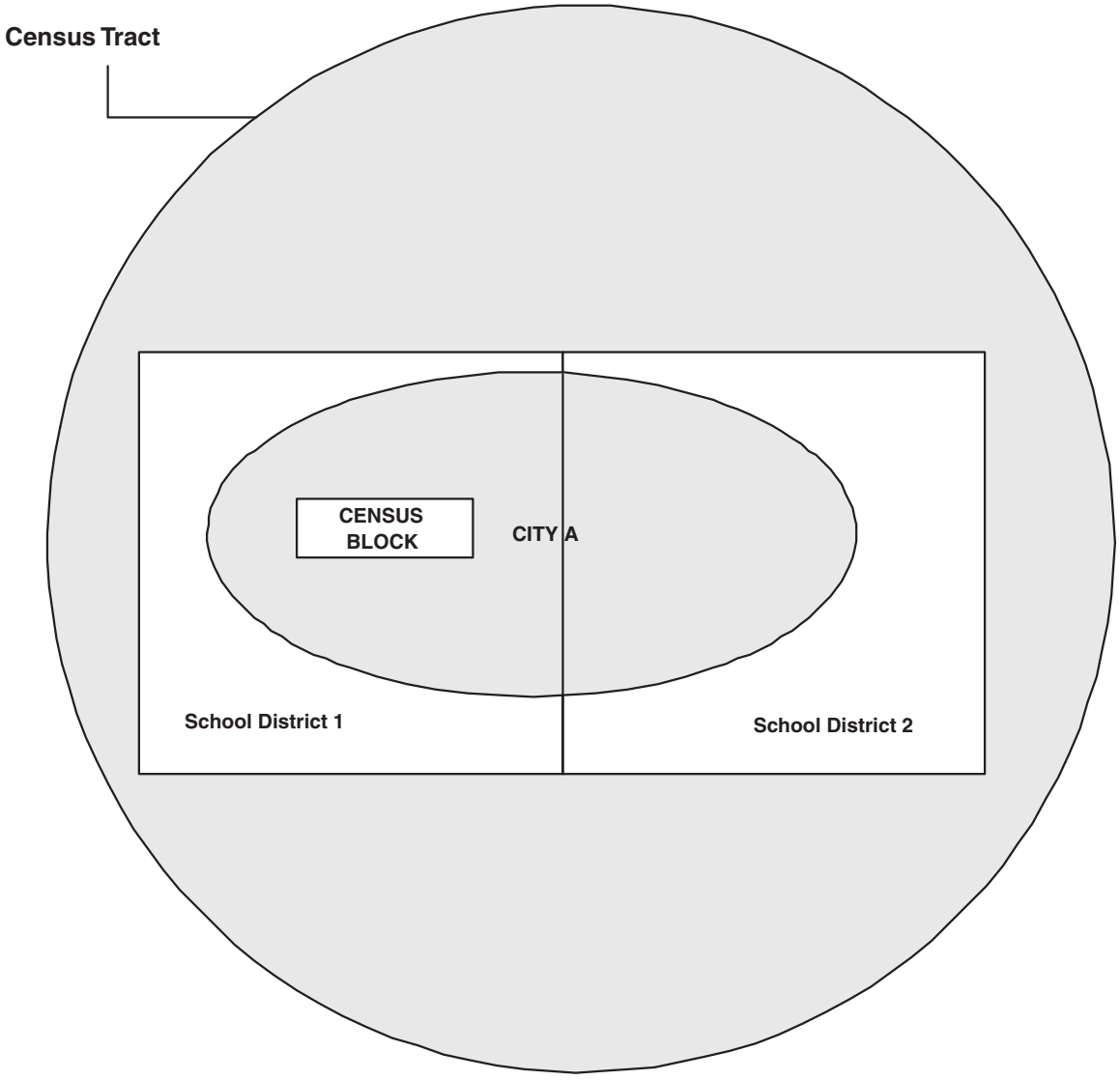
The figure shows a census block that is in City A and is also contained in School District 1. Therefore, it is contained wholly in three sampling entities:

- County (not shown).
- Place with active and functioning government—City A.
- School district.

FIGURE 4.2

ASSIGNMENT OF BLOCKS (AND THEIR ADDRESSES) TO SECOND-STAGE SAMPLING STRATA

(Note that the land area of a sampling entity does not necessarily correlate to its MOS)



Example 1: Suppose the MOS for City A is 600 and the MOS for School District 1 is 1,100. Then the SEMOS for the census block is 600 and it is placed in the 200 ≤SEMOS ≤800 stratum.

Example 2: Suppose the MOS for City A is 1,300 and the MOS for School District 1 is 1,400, then the SEMOS for the block is 1,300. Since the SEMOS for the block is greater than 1,200, the block will be assigned to one of the two strata with SEMOS >1,200 depending on the size of the census tract (TMOS—not shown in the diagram). In this example, suppose the TMOS is 1,800, then the census block will be placed in the 1,200 <SEMOS and TMOS ≤2000 stratum.

Determining the Sampling Rates

Each year, the specific set of sampling rates is determined for each of the five sampling strata defined in Table 4.1. Before this can be done, the following three steps are performed. The first step is to calculate a base rate (BR) for the current year. Four of the five sampling rates are a function of a base sampling rate, and the fifth is fixed at 10 percent. Table 4.2 shows the relationship between the base rate and the five sampling rates.

Table 4.2 **Relationship Between the Base Rate and the Sampling Rates**

Stratum	Sampling rates	
	United States	Puerto Rico
Blocks in large tracts (SEMOS >1,200, TMOS >2,000)	0.735 x BR	0.75 x BR
Blocks in small tracts (SEMOS >1,200, TMOS ≤2,000)	BR	BR
Blocks in small sampling entities (800 ≤SEMOS ≤1,200)	1.5 x BR	1.5 x BR
Blocks in smaller sampling entities (200 ≤SEMOS <800)	3 x BR	3 x BR
Blocks in smallest sampling entities (SEMOS <200)	10 percent	10 percent

The distribution of addresses by sampling stratum, coupled with the target sample size of three million, allows a simple algebraic equation to be set up and solved for BR. The BR for 2007 was 2.23 percent for the United States and 2.7 percent for Puerto Rico.

The second step is the calculation of the sampling rates using the value of BR and the equations in Table 4.2. The third step reduces these sampling rates for certain blocks and is discussed in the following subsection.

First-Phase Sampling Rates. The sampling rates for the 2007 ACS are given in columns 2 and 4 of Table 4.3, for the United States and Puerto Rico respectively (Hefter, 2006b). Since the design of the ACS calls for a target annual address sample of approximately three million in the United States and 36,000 in Puerto Rico, the sampling rates for all but the smallest sampling entities' stratum (SEMOS <200) are reduced each year as the number of addresses in the United States and Puerto Rico increases. However, as shown in Table 4.2, among the strata where the rates are decreasing, the relationship of the sampling rates will remain proportionally constant. The sampling rate for the smallest sampling entities will remain at 10 percent.

The sampling rates that are used to select the sample are obtained after the sampling rates are reduced for blocks in specific strata that are in certain census tracts in the United States. These tracts are predicted to have the highest rates of completed questionnnaires by mail and via a telephone follow-up operation, computer-assisted telephone interviewing (CATI). This adjustment is to compensate for the increase in costs due to increasing the CAPI sampling rates in tracts predicted to have the lowest rate of completed interviews by mail and CATI.

Specifically, the sampling rates are multiplied by 0.92 for some blocks in the United States in the two strata in which the SEMOS was greater than 1,200. This adjustment is made for blocks in tracts that were predicted to have a level of completed mail and CATI interviews of at least 60 percent, and at least 75 percent of the block's addresses were defined as mailable.

Projections of the combined mail and CATI rates were used because ACS rates of completed questionnaires by mail and CATI were not available for all census tracts in the country prior to 2005. For census tracts included in the 2000–2003 ACS, these projections were based on ACS operational data from those years. In the remaining tracts, the rates were projections based on a model that also used information from Census 2000 long-form operational data. Each census tract was assigned to a CAPI sampling stratum, and this designation has been used since 2005.

As a result of this adjustment, there are a total of seven sampling rates used in the United States, and five in Puerto Rico, as shown in columns 3 and 4 of Table 4.3. A brief description of the relationship between this reduction and the CAPI sampling rates is given in Section B.2. (For full details, see Asiala, 2005.) This reduction does not occur in Puerto Rico, so there are five rates used in Puerto Rico.

Table 4.3 **2007 ACS/PRCS Sampling Rates Before and After Reduction**

Stratum (1)	Sampling rates		
	United States		Puerto Rico
	Before reduction[1] (2)	After reduction[1] (3)	No reduction[1] (4)
Blocks in large tracts (SEMOS >1,200, TMOS >2,000)	1.6	(NA)	2.0
Mailable addresses ≥75 percent and predicted levels of completed interviews prior to CAPI sampling >60 percent	(NA)	1.5	(NA)
Mailable addresses <75 percent or predicted levels of completed interviews prior to CAPI sampling ≤60 percent	(NA)	1.6	(NA)
Blocks in small tracts (SEMOS >1,200, TMOS ≤2,000)	2.2	(NA)	2.7
Mailable addresses ≤75 percent and predicted levels of completed interviews prior to CAPI sampling >60 percent	(NA)	2.1	(NA)
Mailable addresses <75 percent or predicted levels of completed interviews prior to CAPI sampling ≤60 percent	(NA)	2.2	(NA)
Blocks in small sampling entities 800 ≤SEMOS ≤1,200)	3.3	3.3	4.0
Blocks in smaller sampling entities (200 ≤SEMOS <800)	6.7	6.7	8.1
Blocks in smallest sampling entities (SEMOS <200)	10.0	10.0	10.0

NA Not applicable.

[1] In percent.

Note: The rates in the table have been rounded to one decimal place.

First-Stage Sample: Random Assignment of Addresses to a Specific Year

One of the ACS design requirements is that no HU address can be in a sample more than once in any 5-year period. To accommodate this restriction, the addresses in the frame are assigned systematically to five subframes, each containing roughly 20 percent of the frame, and each being a representative sample. Addresses from only one of these subframes are eligible to be in the ACS sample in each year and each subframe is used every fifth year. For example, 2011 will have the same addresses in its subframe as did 2006, with the addition of all new addresses that have been assigned to that subframe during the 2007–2011 time period. As a result, both the main and supplemental sample selection is performed in two stages. The first stage partitions the sampling frame into the five subframes and determines the subframe for the current year, and the second selects addresses to be included in the ACS from the subframe eligible for the sample year.

Prior to the ACS 2005 selection, there was a one-time allocation of all addresses then present on the ACS frame to the five subframes. In subsequent years, only addresses new to the frame have been systematically allocated to these five subframes. This is accomplished by sorting the addresses in each county by stratum and geographical order including tract, block, street name, and house number. Addresses are then sequentially assigned to each of the five existing subframes. This procedure is similar to the use of a systematic sample with a sampling interval of five, in which the first address in the interval is assigned to year one, the second address in the interval to year two, and so on. Specifically, during main sampling, only the addresses new to the MAF since the previous year's supplemental MAF are eligible for first-stage sampling and go through the process of being assigned to a subframe. Similarly, during supplemental sampling, only addresses new to the MAF since main sampling go through first-stage sampling. The addresses to be included in the ACS will be selected from the subframe allocated to the sample year during the second stage of sampling. (For additional details about HU address sampling, see Asiala, 2004 and Hefter, 2006b.)

Second-Stage Sampling: Selection of Addresses

This sampling process selects a subset of the addresses from the subframe that is assigned to the sample year. This is the final annual ACS sample. These addresses are selected from the subframe in each of the 3,141 counties. The addresses in each county are sorted by stratum and the first-stage order of selection. After sorting, systematic samples of addresses are selected using a sampling rate approximately equal to its final sampling rate divided by 20 percent.[8]

Sample Month Assignment for Address Samples

Each sample address for a particular year is assigned to a data collection month. The set of all addresses assigned to a specific month is referred to as the month's sample or panel. Addresses selected during main sampling are sorted by their order of selection and assigned systematically to the 12 months of the year. However, addresses that have also been selected for one of several Census Bureau household surveys in specified months (which vary by survey) are assigned to an ACS data collection month based on the interview month(s) for these other household surveys.[9] The goal of the assignments is to reduce the respondent burden of completing interviews for both the ACS and another survey during the same month.

The supplemental sample is sorted by order of selection and assigned systematically to the months of April through December. Since this sample is only approximately 1 percent of the total ACS sample, very few addresses are also in one of the other household surveys in the specified months. Therefore the procedure described above to move the ACS data collection month for cases in common with the current surveys is not implemented during supplemental first-phase sampling.

4.3 SECOND-PHASE SAMPLING FOR CAPI FOLLOW-UP

As discussed earlier, the ACS uses three modes of data collection—mail, telephone, and personal visit in consecutive months. (See Chapter 7 for more information on data collection.) An interview for an HU and its residents can be completed during the month it was mailed out or during the two subsequent months. All addresses mailed a questionnaire can return a completed questionnaire during this 3-month time period.

All mailable addresses with available telephone numbers for which no response is received during the assigned month are sent to CATI for follow-up. The CATI follow-up for these cases is conducted during the following month. Cases where neither a completed mail questionnaire has been received nor a CATI interview completed are eligible for CAPI in the third month, as are the unmailable addresses. An address is considered unmailable if the address is incomplete or directs mail to only a post office box. Table 4.4 summarizes the eligibility of addresses.

Table 4.4 **Addresses Eligible for CAPI Sampling**

Mailable address	Responds to mailing	Responds to CATI	Eligible for CAPI
No	(NA)	(NA)	Yes
Yes	No	No	Yes
Yes	No	Yes	No (completed)
Yes	Yes	(NA)	No (completed)

NA Not applicable.

During the CAPI sample selection, a systematic sample of these addresses is selected for CAPI data collection each month, using the rates shown in Table 4.5. The selection is made after sorting within county by CAPI sampling rate, mailable versus unmailable, and geographical order within the address frame. See Hefter (2005) for details of CAPI sampling.

[8]Since the first-stage sampling rate is approximately 20 percent, and the first-stage rate times the second-stage rate equals the sampling rate, the second-stage rate is approximately equal to the sampling rate divided by 20 percent. An adjustment is made to account for uneven distributions of addresses in the subframe.
[9]These surveys include the Survey of Income and Program Participation, the National Crime Victimization Survey, the Consumer Expenditures Quarterly and Diary Surveys, the Current Population Survey, and the State Child Health Insurance Program Surveys.

The variance of estimates for HUs and people living in them in a given area is a function of the number of interviews completed within that area. However, due to sampling for nonresponse follow-up, CAPI cases have larger weights than cases completed by mail or CATI. The variance of the estimates for an area will tend to increase as the proportion of mail and CATI responses decreases. Large differences in these proportions across areas of similar size may result in substantial differences in the reliability of their estimates. To minimize this possibility, tracts in the United States that are predicted to have low levels of interviews completed by mail and CATI have their CAPI sampling rates adjusted upward from the default 1-in-3 rate for mailable addresses. This tends to reduce variances for the affected areas both by potentially increasing their total numbers of completed interviews and by decreasing the differences in weights between their CAPI cases and mail/CATI interviews.

No information was available to reliably predict the levels of completed interviews prior to second-phase sampling for CAPI follow-up in Puerto Rico prior to 2005, so the sampling rates of 1-in-3 for mailable and 2-in-3 for unmailable addresses were used initially. On the basis of early response results observed during the first months of the ACS in Puerto Rico, the CAPI sampling rate for mailable addresses in all Puerto Rico tracts was changed to 1-in-2 beginning in June 2005.

Table 4.5 **2007 CAPI Sampling Rates**

Address and tract characteristic	CAPI sampling rate (percent)
United States	
Unmailable addresses and addresses in Remote Alaska	66.7
Mailable addresses in tracts with predicted levels of completed interviews prior to CAPI subsampling between 0 percent and 35 percent	50.0
Mailable addresses in tracts with predicted levels of completed interviews prior to CAPI subsampling greater than 35 percent and less than 51 percent	40.0
Mailable addresses in other tracts	33.3
Puerto Rico	
Unmailable addresses	66.7
Mailable addresses	50.0

4.4 GROUP QUARTERS SAMPLE SELECTION

GQ facilities include such places as college residence halls, residential treatment centers, skilled nursing facilities, group homes, military barracks, correctional facilities, workers' dormitories, and facilities for people experiencing homelessness. Each GQ facility is classified according to its GQ type. (For more information on GQ facilities, see Chapter 8.) As noted previously, GQ facilities were not included in the 2005 ACS, but have been included since 2006. The GQ sample for a given year is selected during a single operation carried out in August and September of the previous year. The sampling frame of GQ facilities and their locations is derived from the most recently available updated MAF and lists from other sources and operations. The ultimate sampling units for the GQ sample are the GQ residents, not the facilities. The GQ samples are independent state-level samples. Certain GQ types are excluded from the ACS sampling and data collection operations. These are domestic violence shelters, soup kitchens, regularly scheduled mobile food vans, targeted nonsheltered outdoor locations, crews of commercial maritime vessels, natural disaster shelters, and dangerous encampments. There are several reasons for their exclusion and they vary by GQ type. Concerns about privacy and the operational feasibility of repeated interviewing for a continuing survey, rather than once a decade for a census led to the decision to exclude these GQ types. However, ACS estimates of the total population are controlled to be consistent with the Population Estimates Program estimate of the GQ resident population from all GQs, even those excluded from the ACS.

All GQ facilities are classified into one of three groups: (1) small GQ facilities (having 15 or fewer people according to Census 2000 or updated information); (2) large GQ facilities (with an expected population of more than 15 people); and (3) GQ facilities closed on Census Day (April 1, 2000) or new to the sampling frame since Census Day (with no information regarding the expected population size). There are approximately 105,000 small GQ facilities, 77,000 large GQ

facilities, and 3,000 facilities with an unknown population count on the GQ sampling frame. Two sampling strata are created to sample the GQ facilities. The first stratum includes both small GQ facilities and those with no population count. The second includes large facilities. In the remainder of this chapter, these strata will be referred to as the small GQ stratum and the large GQ stratum, respectively. A GQ measure of size (GQMOS) is computed for use in sampling the large GQ facilities. The GQMOS for each GQ is the expected population count divided by 10.

Different sampling procedures are used for these two strata. GQ in the small GQ stratum are sampled like the HU address sample, and data are collected for all people in the selected GQ facilities. Like HU addresses, small GQ facilities are eligible to be in the sample only once in a 5-year period. Groups of ten people are selected for interview from GQ facilities in the large GQ stratum, and the number of these groups selected for a large GQ facility is a function of its GQMOS. Unlike HU addresses, large GQ facilities are eligible for sampling each year. (For details on GQ sampling, see Hefter, 2006c.)

Small Group Quarters Stratum Sample

For the small GQ stratum, a two-phase, two-stage sampling procedure is used. In the first phase, a GQ facility sample is selected using a method similar to that used for the first-phase HU address sample. Just as we saw in the HU address sampling, the first phase has two stages. Stage one systematically assigns small GQ facilities to a subframe associated with a specific year. During the second stage, a systematic sample of the small GQ facilities is selected. In the second phase of sampling, all people in the facility are interviewed as long as there are 15 or fewer at the time of interview. Otherwise, a subsample of ten people is selected and interviewed.

First Phase of Small GQ Sampling—Stage One: Random Assignment of GQ Facilities to Subframes

The sampling procedure for 2006 assigned all of the GQ facilities in the small stratum to one of five 20 percent subframes. The GQ facilities within each state are sorted by small versus closed on Census Day, new versus previously existing, GQ type (such as skilled nursing facility, military barracks, or dormitory), and geographical order (county, tract, block, street name, and GQ identifier) in the small GQ frame. In each year subsequent to 2006, new GQ facilities are assigned systematically to the five subframes. So the subframe for 2007 GQ sample selection contains the facilities previously designated to the subframe for calendar year 2007 and the 20 percent of new small GQ facilities added since the 2006 sampling. The small GQ facilities in the 2007 subframe will not be eligible for sampling again until 2012, since the 1-in-5-year period restriction also applies to small GQ facilities.

First Phase of Small GQ Sampling Stage Two: Selection of Facilities

The second-stage sample is a 1-in-8 systematic sample of the GQ facilities from the assigned subframe within each state. The GQs are sorted by new versus previously existing addresses and order of selection. Regardless of their actual size, all of these small GQ facilities have the same probability of selection. This 1-in-8 second-stage sampling rate combined with the 1-in-5 first-stage sampling rate yields an overall first-phase-sampling rate of 1-in-40, or 2.5 percent.

Second Stage of Small GQ Sampling: Selection of Persons Within Selected Facilities

Every person in the GQ facilities selected in this sample is eligible to be interviewed. If the number of people in the GQ facility exceeds 15, a field subsampling operation is performed to reduce the total number of sampled people to ten, similar to the groups of ten selected in the large GQ stratum.

4.5 LARGE GROUP QUARTERS STRATUM SAMPLE

Unlike the HU address and small GQ samples, the large GQ facilities are not divided into five subframes. The ultimate sampling unit for large GQ facilities is people, with interviews collected in groups of ten, not the facility itself. A two-phase sampling procedure is used to select these groups: The first indirectly selects the GQ facilities by selecting groups of ten within the facilities

and the second selects the people for each facility's group(s) of ten. The number of groups of ten eligible to be sampled from a large GQ facility is equal to its GQMOS. For example, if a facility had 550 people in Census 2000, its GQMOS is 55 and there are 55 groups of ten eligible for selection in the sample.

First Phase of Large GQ Sampling: Selection of Groups of Ten (and Associated Facilities)

All the large GQ facilities in a state are sorted by GQ type and geographical order in the large GQ frame, and a systematic sample of 1-in-40 groups of ten is selected. For this reason, a GQ facility with fewer than 40 groups (or roughly 400 individuals) may or may not have one of its groups selected for the sample. GQ facilities with between 40 and 80 groups will have at least one group selected. GQ facilities with between 80 and 120 groups will have at least two groups selected, and so forth.

Second Phase of Large GQ Sampling: Selection of Persons Within Facilities

The second phase of sampling takes place within each GQ facility that has at least one group selected in the first stage. When a field representative visits a GQ facility to conduct interviews, an automated listing instrument is used to randomly select the ten people to be included in each group of ten being interviewed. The instrument is preloaded with the number of expected person interviews (ten times the number of groups selected), and a random starting number. The field representative then enters the actual number of people in the facility, as well as a roster of their names. To achieve a group size of ten, the instrument computes the appropriate sampling interval based on the observed population at the time of interviewing and then selects the actual people for interviewing using a preloaded random start and a systematic algorithm. If the large GQ has an observed population of 15 or fewer people, the instrument selects a group size of ten or the observed population if less than ten.

For most GQ types, if multiple groups are selected within a GQ facility, their groups of ten are assigned to different sample months for interviewing. Very large GQ facilities with more than 12 groups selected have multiple groups assigned to some sample months. In these cases, an attempt is made to avoid selecting the same person more than once in a sample month. However, there is no attempt made to avoid selection of someone more than once across sample months within a year. Thus someone in a very large GQ facility could be interviewed in consecutive months. All GQ facilities in this stratum are eligible for selection every year, regardless of their sample status in previous years.

4.6 SAMPLE MONTH ASSIGNMENT FOR SMALL AND LARGE GROUP QUARTER SAMPLES

The selected small GQ facilities and groups of ten for large GQ facilities are assigned to months using a procedure similar to the one used for sampled HU addresses. All GQ samples from a state are combined and sorted by small versus large stratum and first-phase order of selection. Consecutive samples are assigned to the 12 months in a predetermined order, starting with a randomly determined month.

Due to operational and budgeting constraints, the same month is assigned to all sample groups of ten within certain types of correctional GQs or military barracks. All samples in federal prisons are assigned to September, and data collection may take up to 4.5 months, an exception to the 6 weeks allowed for all other GQ types. For the samples in nonfederal correctional facilities, state prisons, local jails, halfway houses, military disciplinary barracks, and other correctional institutions or military barracks, individual GQ facilities are randomly assigned to months throughout the year.

4.7 REMOTE ALASKA SAMPLE

Remote Alaska is a set of rural areas in Alaska that are difficult to access and for which all HU addresses are treated as unmailable. Due to the difficulties in field operations during specific months of the year, and the extremely seasonal population in these areas, data collection operations in Remote Alaska differ from the rest of the country. In both the main and supplemental HU address samples, the month assigned for each Remote Alaska HU address is based on the place,

AIANSA, block group, or county (in that order) in which it is contained. All designated addresses located in each of these geographical entities are assigned to either January or September. These month assignments are done in such a way as to balance workloads between the months, and to keep groups of cases together geographically. The addresses for each month are sorted by county and geographical order in the address frame, and a sample of 2-in-3 is sent directly to CAPI (no mail or CATI) in the appropriate month. The GQ sample in Remote Alaska is assigned to January or September using the same procedure. Up to 4 months is allowed to complete the HU and GQ data collection for each of the two data collection periods.

4.8 REFERENCES

Asiala, M. (2004). "Specifications for Selecting the ACS 2005 Main HU Sample." 2005 American Community Survey Sampling Memorandum Series #ACS-S-40, Census Bureau Memorandum to L. McGinn from R.P. Singh, Washington, DC, August 8, 2005.

Asiala, M. (2005). "American Community Survey Research Report: Differential Sub-Sampling in the Computer Assisted Personal Interview Sample Selection in Areas of Low Cooperation Rates." 2005 American Community Survey Documentation Memorandum Series #ACS05-DOC-2, Census Bureau Memorandum to R.P. Singh from D. Hubble, Washington, DC, February 15, 2005.

Bates, L. M. (2006). "Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey." 2007 American Community Survey Universe Creation Memorandum Series #ACS07-UC-1, Census Bureau Memorandum to L. Blumerman from D. Kostanich, Washington, DC, September 20, 2006.

Federal Register Notice (2000). "American Indian and Alaska Native Areas Geographic Program for Census 2000; Notice." Department of Commerce, Bureau of the Census, Volume 65, Number 121, Washington, DC, June 22, 2000.

Hefter, S. P. (2005). "American Community Survey: Specifications for Selecting the Computer Assisted Personal Interview Samples." 2005 American Community Survey Sampling Memorandum Series #ACS-S-45, Census Bureau Memorandum to L. McGinn from R.P. Singh, Washington, DC, May 23, 2005.

Hefter, S. P. (2006a). "Creating the Governmental Unit Measure of Size (GUMOS) Datasets for the American Community Survey and the Puerto Rico Community Survey." 2007 American Community Survey Sampling Memorandum Series #ACS07-S-1, Census Bureau Memorandum to S. Schechter from D. Whitford, Washington, DC, August 8, 2006.

Hefter, S. P. (2006b). "Specifications for Selecting the Main and Supplemental Housing Unit Address Samples for the American Community Survey." 2007 American Community Survey Sampling Memorandum Series #ACS07-S-3, Census Bureau Memorandum to S. Schechter from D. Whitford, Washington, DC, August 23, 2006.

Hefter, S. P. (2006c). "Specifications for Selecting the American Community Survey Group Quarters Sample." 2007 American Community Survey Sampling Memorandum Series #ACS07-S-6, Census Bureau Memorandum to S. Schechter from D. Whitford, Washington, DC, October 27, 2006.

Chapter 5.
Content Development Process

5.1 OVERVIEW

American Community Survey (ACS) content is designed to meet the needs of federal government agencies and is a rich source of local area information useful to state and local governments, universities, and private businesses. The U.S. Census Bureau coordinates the content development and determination process for the ACS with the Office of Management and Budget (OMB) through an interagency committee comprised of more than 30 federal agencies. All requests for content changes are managed by the ACS Content Council, which provides the Census Bureau with guidelines for pretesting, field testing, and implementing new content and changes to existing ACS content. This chapter provides greater detail on the history of content development for the ACS, current survey content, and the content determination process and policy.

5.2 HISTORY OF CONTENT DEVELOPMENT

The ACS is part of the 2010 Decennial Census Program and is an alternative method for collecting the long-form sample data collected in the last five censuses. The long-form sample historically collected detailed population and housing characteristics once a decade through questions asked of a sample of the population.[1] Beginning in 2005, the ACS collects this detailed information on an ongoing basis, thereby providing more accurate and timely data than was possible previously. Starting in 2010, the decennial census will include only a short form that collects basic information for a total count of the nation's population.[2]

Historically, the content of the long form was constrained by including only the questions for which:

- There was a current federal law calling for the use of decennial census data for a particular federal program (mandatory).
- A federal law (or implementing regulation) clearly required the use of specific data, and the decennial census was the historical or only source; or the data are needed for case law requirements imposed by the U.S. federal court system (required).
- The data were necessary for Census Bureau operational needs and there was no explicit requirement for the use of the data as explained for mandatory or required purposes (programmatic).

Constraining the content of the ACS was, and still is, critical due to the mandatory reporting requirement and respondent burden. To do this, the Census Bureau works closely with the OMB and the Interagency Committee for the ACS, co-chaired by the OMB and the Census Bureau. This committee was established in July 2000, and includes representatives from more than 30 federal departments and agencies that use decennial census data. Working from the Census 2000 long-form justification, the initial focus of the committee was to verify and confirm legislative justifications for every 2003 ACS question. The agencies were asked to examine each question and provide the Census Bureau with justification(s) by subject matter, the legal authority for the use, the lowest geographic level required, the variables essential for cross-tabulation, and the frequency

[1] Sampling began in the 1940 census when a few additional questions were asked of a small sample of people. A separate long-form questionnaire was not implemented until 1960.

[2] In addition to counting each person in every household, the basic information planned for the Census 2010 short form will include a very select set of key demographic characteristics needed for voting rights and other legislative requirements. Currently, the plan is to ask for data on tenure at residence, sex, age, relationship, Hispanic origin, and race.

with which the data are needed. They were asked to cite the text of statutes and other legislative documentation, and to classify their uses of the ACS questions as "mandatory," "required," or "programmatic," consistent with the constraints of the traditional long form.

In the summer of 2002, the U.S. Department of Commerce General Counsel's Office asked each federal agency's General Counsel to examine the justifications submitted for its agency and, if necessary, to revise the information so that the agency would be requesting only the most current material necessary to accomplish the statutory departmental missions in relation to census data. This step ensured that the highest-ranking legal officer in each agency validated its stated program requirements and data needs.

Only questions on those subjects classified as either "mandatory" or "required" were asked on the 2003 ACS questionnaire, along with questions on two programmatic subjects (fertility and seasonal residence). The end result of this review was a 2003 ACS questionnaire with content almost identical to the Census 2000 long form. In 2002, the ACS questionnaire was approved for 3 years by the OMB in its role of implementing the 1995 Paperwork Reduction Act.

5.3 2003–2007 CONTENT

ACS Content

In 2003–2007, the ACS consisted of 25 housing and 42 population questions (6 basic and 36 detailed population questions). (See Table 5.1 for a complete list of ACS topics.) The ACS GQ questionnaire contains all population questions in the population column of Table 5.1, except the question on relationship to householder. One housing question, food stamp benefit, is on the ACS GQ questionnaire.

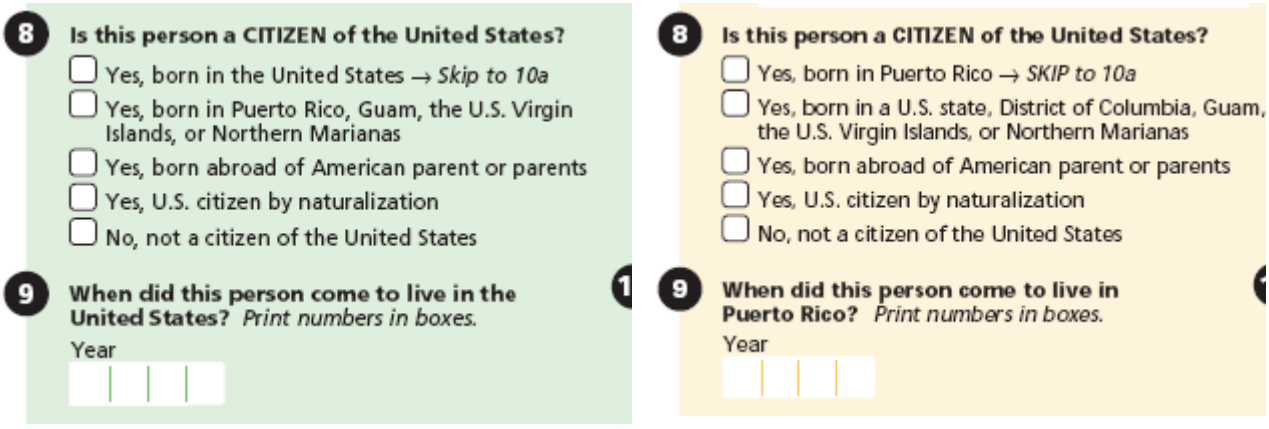
Table 5.1 **2003–2007 ACS Topics Listed by Type of Characteristic and Question Number**

Housing	Population
Household size	**Name**
H1 Units in Structure	P1 Sex
H2 Year Structure Built	P2 Age and Date of Birth
H3 Year Householder Moved Into Unit	P3 Relationship to Householder
H4 Acreage	P4 Marital Status
H5 Agricultural Sales	P5 Hispanic Origin
H6 Business on Property	P6 Race
H7 Rooms	P7 Place of Birth
H8 Bedrooms	P8 Citizenship
H9 Plumbing Facilities	P9 Year of Entry
H10 Kitchen Facilities	P10 Type of School and School Enrollment
H11 Telephone Service Available	P11 Educational Attainment
H12 Vehicles Available	P12 Ancestry
H13 House Heating Fuel	P13 Language Spoken at Home, Ability to Speak English
H14 Cost of Utilities	P14 Residence 1 Year Ago (Migration)
H15 Food Stamp Benefit	P15 Disability: Sensory, Physical
H16 Condominium Status and Fee	P16 Disability: Mental, Self-care
H17 Tenure	P17 Disability: Going out Alone, Ability to Work
H18 Monthly Rent	P18 Fertility
H19 Value of Property	P19 Grandparents as Caregivers
H20 Real Estate Taxes	P20 Veteran Status
H21 Insurance for Fire, Hazard, and Flood	P21 Period of Military Service
H22 Mortgage Status, Payment, Real Estate Taxes	P22 Years of Military Service
	P23 Worked Last Week
H23 Second or Junior Mortgage Payment or Home Equity Loan	P24 Place of Work
	P25 Means of Transportation
H24 Mobile Home Costs	P26 Private Vehicle Occupancy
H25 Seasonal Residence	P27 Time Leaving Home to Go to Work
	P28 Travel Time to Work
	P29 Layoff, Temporarily Absent, Informed of Recall or Return Date
	P30 Looking for Work
	P31 Available to Work
	P32 When Last Worked
	P33 Weeks Worked
	P34 Usual Hours Worked Per Week
	P35 Class of Worker
	P36 Employer
	P37 Type or Kind of Business
	P38 Industry
	P39 Occupation
	P40 Primary Job Activity
	P41 Income in the Past 12 Months (by type of income)
	P42 Total Income

Puerto Rico Community Survey (PRCS) Content

The content for the PRCS is identical to that used in the United States. The PRCS includes six questions that are worded differently from those on the ACS to accommodate cultural and geographic differences between the two areas. (See Figure 5.1 for an example of ACS questions that were modified for the PRCS.)

Figure 5.1 **Example of Two ACS Questions Modified for the PRCS**



ACS (2005) **PRCS (2005)**

5.4 CONTENT POLICY AND CONTENT CHANGE PROCESS

The ACS is designed to produce detailed demographic, housing, social, and economic data every year. Because it accumulates data over time to obtain sufficient levels of reliability for small geographic areas, the Census Bureau must minimize content changes. Consistency must be maintained throughout all ACS data collection operations, including HUs and GQ facilities. Introducing changes could affect data quality and result in only partial releases of data for a given year if a question changes significantly, or has not been asked for long enough to accumulate 3 or 5 years' worth of data.

In 2006, the OMB, in consultation with Congress and the Census Bureau, adopted a more flexible approach to content determinations for the ACS. In making content determinations, the OMB, in consultation with the Census Bureau, will consider issues such as frequency of data collection, the level of geography needed to meet the required need, and other sources of data that could meet a requestor's need in lieu of ACS data. In some cases, legislation still may be needed for a measure to be justified for inclusion in the ACS. In other cases, OMB may approve a new measure based on an agency's justification and program needs.

The Census Bureau recognizes and appreciates the interests of federal partners and stakeholders in the collection of data for the ACS. Because participation in the ACS is mandatory, only necessary questions will be approved by OMB and asked by the Census Bureau. The OMB's responsibility under the Paperwork Reduction Act requires that the practical utility of the data be demonstrated and that the respondent burden be minimized (especially for mandatory collections).

The Census Bureau's ACS Content Policy is used as a basic guideline for all new question proposals from federal agencies, the Congress, and the Census Bureau. The Content Change Process is part of a risk management strategy to ensure that each new or modified question has been tested fully and will collect quality data without reducing overall response rates.

The policy provides guidance for ongoing ACS content development. To implement this policy, the Census Bureau coordinates input from internal and external groups, while the Interagency Committee for the ACS obtains broad input from all federal agencies. The Census Bureau also coordinates the creation of subject area subcommittee groups that include representatives from the Interagency Committee and the Census Bureau; these groups provide expertise in designing sets of questions and response categories so that the questions will meet the needs of all agencies. Census Bureau staff review the subcommittee proposals and provide comments and internal approval of content changes.

The ACS Content Change Process provides guidance for Census Bureau pretesting, including a field test, for all new or modified questions prior to incorporating them into ACS instruments; this

guidance is based on the standards outlined in the *Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses* (DeMaio, Bates, Ingold, and Willimack 2006). New pretested questions will be added to the ACS only after OMB approval has been given to the Census Bureau.

Content Change Factors

The OMB and the Census Bureau consider several factors when new content is proposed. Federal agencies must provide both agencies with specific information about the new data collection need(s).

The uses of the data must be identified to determine the appropriateness of collecting it through a national mandatory survey. Other Census Bureau surveys or other sources of data are reviewed and considered. Because ACS data are collected and tabulated at the tract or block-group level, the response burden for the majority of respondents must be considered.

Federal agencies interested in content changes must be able to demonstrate that they require detailed data with the frequency of ACS data collection, and that failure to obtain the information with this frequency will result in a failure to meet agency needs. Requests for new ACS content will be assessed relative to the impact on the requesting agency if the data are not collected through the ACS. Federal agencies requesting new content must demonstrate that they have considered legitimate alternative data sources, and why those alternatives do not meet their needs.

Content Change Requirements

Federal agency or Census Bureau proposals for new content and/or changes to existing ACS questions due to identified quality issues are subject to the following requirements:

- ACS content can be added to or revised only once a year, due to the annual nature of the survey and the number of operations that also must be revised. New content will be incorporated into the ACS only after pretesting, including a field test, has been completed, and the OMB has provided final approval.
- The requesting federal agency will assist with the development of a draft question(s), work with the Census Bureau and other agencies to develop or revise the question, and submit the proposal to the OMB and Census Bureau for further review. In addition, a plan to pretest new or modified content, including a field test, must be developed in accordance with the *Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses*.
- Pretesting must be conducted to detect respondent error and to determine whether or not a change would increase or decrease a respondent's understanding of what is being asked. Alternative versions of questions are pretested to identify the version most likely to be answered accurately by respondents, and then are field tested.

5.5 2006 CONTENT TEST

In 2004, planning began for the 2006 ACS Content Test, so that the content changes in the ACS could be field tested before the 2008 ACS instrument was finalized. The OMB and the Census Bureau first asked members of the ACS Interagency Committee to review the legislative authority for current or proposed ACS questionnaire content and to identify any questions that needed to be reworded or reformatted.

The 2006 ACS Content Test was the first opportunity to test revisions to the long-form sample questions used in Census 2000. The content of the 2006 ACS Content Test included new questions on the subjects of marital history, health insurance and coverage, and veterans' service-connected disability ratings.

The test methodology for the 2006 ACS Content Test was designed to be similar to ACS data collection in the production phase, and incorporated the prenotice letter, initial mailing package, reminder postcard, and potential second mailing package (due to nonresponse). A computer-assisted personal interview follow-up was conducted. To measure response error, a computer-assisted telephone interview content reinterview also was conducted. Simple response variance and gross difference rates, along with other data quality measures, such as item nonresponse rates and measures of distributional changes, served as indicators of the quality of the test questions relative to current ACS questions.

5.6 REFERENCES

DeMaio, Theresa J., Nancy Bates, Jane Ingold, and Diane Willimack (2006). "Pretesting Questionnaires and Related Materials for Surveys and Censuses." Washington, DC: U.S. Census Bureau, 2006.

Chapter 6.
Survey Rules, Concepts, and Definitions

6.1 OVERVIEW

Interview and residence rules define the universe, or target population, for a survey, and so identify the units and people eligible for inclusion. The 2006–2007 ACS interviewed the resident population living in both housing units (HUs) and group quarters (GQ) facilities. The ACS uses residence rules based on the concept of current residence.

Sections B and C in this chapter detail the interview and residence rules. Section D describes the full set of topics included in the ACS, and is organized into four sections to parallel the organization of the ACS questionnaire: address, HU status, and household information; basic demographic information; detailed housing information; and detailed population information.

6.2 INTERVIEW RULES

The Census Bureau classifies all living quarters as either HUs or GQ facilities. An HU is a house, an apartment, a group of rooms, or a single room either occupied or intended for occupancy as separate living quarters. GQ facilities are living quarters owned and managed by an entity or organization that provides housing and/or services for the residents. GQ facilities include correctional facilities and such residences as group homes, health care and treatment facilities, and college dormitories.

Interview rules define the scope of data collection by defining the types of places included in the sample frame, as well as the people eligible for inclusion. Beginning in 2006, the ACS included HUs and GQ facilities (only HUs and those living in HUs were included in the 2005 ACS). Like the decennial census, the ACS interviews the resident population without regard to legal status or citizenship, and excludes people residing in HUs only if the residence rules (see below) define their current residence as somewhere other than the sample address.

6.3 RESIDENCE RULES

Residence rules are the series of rules that define who (if anyone) should be interviewed at a sample address, and who is considered, for purposes of the survey or census, to be a resident. Residence rules decide the occupancy status of each HU and the people whose characteristics are to be collected.

ACS data are collected nearly every day of the year. The survey's residence rules are applied and its reference periods are defined as of the date of the interview. For mail returns, this is when the respondent completes the questionnaire; for telephone and personal visit interviews, it is when the interview is conducted.

Housing Units

The ACS defined the concept of current residence to determine who should be considered residents of sample HUs. This concept is a modified version of a de facto rule in which a time interval is used to determine residency.[1] The basic idea behind the ACS current residence concept is that everyone who is currently living or staying at a sample address is considered a current resident of that address, except for those staying there for only a short period of time. For the purposes of the ACS, the Census Bureau defines this short period of time as less than 2 consecutive months (often described as the 2-month rule). Under this rule, anyone who has been or will be living for

[1] A de facto rule would include all people who are staying at an address when an interview is conducted, regardless of the time spent at this address. It would exclude individuals away from a regular residence even in they are away only for that one day.

2 months or less in the sample unit when the unit is interviewed (either by mail, telephone, or personal visit) is not considered a current resident. This means that their expected length of stay is 2 months or less, not that they have been staying in the sample unit for 2 months or less. In general, people who are away from the sample unit for 2 months or less are considered to be current residents, even though they are not staying there when the interview is conducted, while people who have been or will be away for more than 2 months are considered not to be current residents. The Census Bureau classifies as vacant an HU in which no one is determined to be a current resident.

As noted earlier, residency is determined as of the date of the interview. A person who is living or staying in a sample HU on interview day and whose actual or intended length of stay is more than 2 months is considered a current resident of the unit. That person will be included as a current resident unless he or she, at the time of interview, has been or intends to be away from the unit for a period of more than 2 months. There are three exceptions:

- Children (below college age) who are away at boarding school or summer camp for more than 2 months are always considered current residents of their parents' home.
- Children who live under joint custody agreements and move between residences are always considered current residents of the sample unit where they are staying at the time of the interview.
- People who stay at a residence close to work and return regularly to another residence to be with their families are always considered current residents of the family residence.

A person who is staying at a sample HU when the interview is conducted, but has no place where he or she stays for periods of more than 2 months, is considered to be a current resident. A person whose length of stay at the sample HU is for 2 months or less and has another place where he or she stays for periods of more than 2 months is not considered a current resident.

Group Quarters

Residency in GQ facilities is determined by a purely de facto rule. All people staying in the GQ facility when the roster of residents is made and sampled are eligible for selection to be interviewed in the ACS. The GQ sample universe will include all people residing in the selected GQ facility at the time of interview. Data are collected for all people sampled, regardless of their length of stay. Children (below college age) staying at a GQ facility functioning as a summer camp are not considered GQ residents.

Reference Period

As noted earlier, the survey's reference periods are defined relative to the date of the interview. Specifically, the survey questions define the reference periods and always include the date of the interview. When the question does not specify a time frame, respondents are told to refer to the situation on the interview day. When the question mentions a time frame, it refers to an interval that includes the interview day and covers a period before the interview. For example, a question that asks for information about the "past 12 months" would be referring to the previous 12 months relative to the date of the interview.

6.4 STRUCTURE OF THE HOUSING UNIT QUESTIONNAIRE

The ACS questionnaires and survey instruments used to collect data from the HU population are organized into four sections, with each section collecting a specific type of information. The first section verifies basic address information, determines the occupancy status of the HU, and identifies who should be interviewed as part of the ACS household. The second section of the questionnaire collects basic demographic data. The third section collects housing information, and the final section collects population data.

There are data collection instruments for all three data collection modes (mail, telephone, and in-person interviews). A paper questionnaire is used in the mail mode. For telephone, there is a computer-assisted telephone interview (CATI) instrument; for personal interviews, there is a computer-assisted personal interview (CAPI) instrument. This section describes the basic data collection process from a personal visit perspective, but the same basic process is followed in the mail and telephone modes.

Address, Housing Unit Status, and Household Information

During personal visit follow-up, the field representative (FR) first must verify that he or she has reached the sample address, and then determine if the sample address identifies an HU. If an HU is not identified, the address is not eligible and is considered out of scope. Out-of-scope addresses include those determined to be nonexistent because the HU has been demolished, or because they identify a business and not a residential unit. Interviewers use the residence rules to determine whether the sample HU is occupied (at least one person staying in the unit is a current resident) or vacant (no one qualifies as a current resident). Interviewers also apply the residence rules to create a household roster of current occupants to interview. The name of the household respondent and the telephone number are collected in case followup contact is needed. The terms below are key for data collection.

Housing Unit (HU). An HU may be a house, an apartment, a mobile home or trailer, a group of rooms, or a single room that is occupied (or, if vacant, intended for occupancy) as separate living quarters.

Housing Unit Status. All sample addresses are assigned a status as either an occupied, vacant, or temporarily occupied HU, or are assigned a status of delete, indicating that the address does not identify an HU. A temporarily occupied unit is an HU where at least one person is staying, but where no people are current residents; this is considered a type of vacant unit. Deleted units are addresses representing commercial units or HUs that either have been demolished or are nonexistent.

Household. A household is defined as all related or unrelated individuals whose current residence at the time of the ACS interview is the sample address.

Household Roster. This roster is a list of all current residents of the sample address; all of these people will be interviewed.

Household Respondent. One person may provide data for all members of the household. The Census Bureau refers to this person as the household respondent. ACS interviewers try to restrict their household respondents to members who are at least 18 years old but, if necessary, household members who are 15 and older can be interviewed. If no household member can be found to provide the survey information, the interviewer must code the case as a noninterview.

Basic Demographic Information

The basic demographic data of sex, age, relationship, marital status, Hispanic origin, and race are collected at the outset and are considered the most critical data items. They are used in many of the survey's tabulations. Age defines the critical paths and skip patterns used in the instrument/questionnaire. Name also is collected for all household members. One individual in the household must be identified as a reference person to define relationships within the household. The section below provides details of the concept (Person 1) and definitions associated with the basic demographic data.

Reference Person or Householder. One person in each household is designated as the householder. Usually this is the person, or one of the people, in whose name the home is owned, being bought, or rented, and who is listed as "Person 1" on the survey questionnaire. If there is no such person in the household, any adult household member 15 and older can be designated.

Sex. Each household member's sex is marked as "male" or "female."

Age and Date of Birth. The age classification is based on the age of the person in complete years at the time of interview. Both age and date of birth are used to calculate each person's age on the interview day.

Relationship. The instrument/questionnaire asks for each household member's relationship to the reference person/householder. Categories include both relatives and nonrelatives.

Marital Status. The marital-status question is asked of everyone responding via mail, but only of people 15 and older responding through CATI or CAPI interviews. The response categories are "now married," "widowed," "divorced," "separated," or "never married." Couples who live together (unmarried people, people in common-law marriages) report the marital status they consider the most appropriate.

Hispanic Origin. A person is of Spanish/Hispanic/Latino origin if the person's origin (ancestry) is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, Argentinean, Colombian, Costa Rican, Dominican, Ecuadoran, Guatemalan, Honduran, Nicaraguan, Peruvian, Salvadoran, from other Spanish-speaking countries of the Caribbean or Central or South America, or from Spain. People who identify their origin as Spanish, Hispanic, or Latino may be of any race. Like the concept of race, Hispanic origin is based on self-identification.

Race. According to the Office of Management and Budget (OMB), and as used by the Census Bureau, the concept of race reflects self-identification by people according to the race or races with which they most closely identify. These categories are socio-political constructs and should not be interpreted as scientific or anthropological in nature. The minimum race categories are determined by OMB and required for use in all federal information collections.

Detailed Housing Information

The ACS housing section collects data on physical and financial characteristics of housing. The 2003–2007 ACS questionnaire includes 25 detailed housing questions. For temporarily occupied HUs, selected housing data are collected from the occupants. For vacant units, selected housing data are collected from information given by neighbors, or determined by observation or from another source. This section of the chapter details the concepts associated with some of the housing items.

Units in Structure. All HUs are categorized by the type of structure in which they are located. A structure is a separate building that either has open spaces on all sides, or is separated from other structures by dividing walls that extend from ground to roof. In determining the number of units in a structure, all HUs—both occupied and vacant—are counted. Stores and office space are excluded.

Year Structure Built. This question determines when the building in which the sample address is located was first constructed, not when it was remodeled, added to, or converted. The information is collected for both occupied and vacant HUs. Units that are under construction are not considered housing units until they meet the HU definition—that is, when all exterior windows, doors, and final usable floors are in place. This determines the year of construction. For mobile homes, houseboats, and recreational vehicles, the manufacturer's model year is taken as the year the unit was built.

Year Householder Moved Into Unit. This question is collected only for occupied HUs, and refers to the year of the latest move by the householder. If the householder moved back into an HU he or she previously occupied, the year of the last move is reported. If the householder moved from one apartment to another within the same building, the year the householder moved into the present apartment is reported. The intent is to establish the year the current occupancy of the unit by the householder began. The year that the householder moved in is not necessarily the same year other members of the household moved in.

Acreage. This question determines a range of the acres on which the house or mobile home is located. A major purpose of this item is to identify farm units.

Agricultural Sales. This item refers to the total amount (before taxes and expenses) received from the sale of crops, vegetables, fruits, nuts, livestock and livestock products, and nursery and forest products produced on the property in the 12 months prior to the interview. This item is used to classify HUs as farm or nonfarm residences.

Business on Property. A business must be easily recognizable from the outside. It usually will have a separate outside entrance and the appearance of a business, such as a grocery store, restaurant, or barbershop. It may be attached either to the house or mobile home, or located elsewhere on the property.

Rooms. The intent of this question is to determine the number of whole rooms in each HU that are used for living purposes. Living rooms, dining rooms, kitchens, bedrooms, finished recreation rooms, enclosed porches suitable for year-round use, and lodger's rooms are included. Excluded are strip or Pullman kitchens, bathrooms, open porches, balconies, halls or foyers, half rooms, utility rooms, unfinished attics or basements, or other unfinished spaces used for storage. A partially divided room is considered a separate room only if there is a partition from floor to ceiling, but not if the partition consists solely of shelves or cabinets.

Bedrooms. Bedrooms include only rooms designed to be used as bedrooms; that is, the number of rooms that the respondent would list as bedrooms if the house, apartment, or mobile home were on the market for sale or rent. Included are all rooms intended for use as bedrooms, even if currently they are being used for another purpose. An HU consisting of only one room is classified as having no bedroom.

Plumbing Facilities. Answers to this question are used to estimate the number of HUs that do not have complete plumbing facilities. Complete plumbing facilities include: hot and cold piped water, a flush toilet, and a bathtub or shower. All three facilities must be located inside the house, apartment, or mobile home, but not necessarily in the same room. HUs are classified as lacking complete plumbing facilities when any of the three facilities is not present.

Kitchen Facilities. Answers to this question are used to estimate the number of HUs that do not have complete kitchen facilities. A unit has complete kitchen facilities when it has all three of the following: a sink with piped water, a range or cook top and oven, and a refrigerator. All kitchen facilities must be located in the house, apartment, or mobile home, but not necessarily in the same room. An HU having only a microwave or portable heating equipment, such as a hot plate or camping stove, is not considered to have complete kitchen facilities.

Telephone Service Available. For an occupied unit to be considered as having telephone service available, there must be a telephone in working order and service available in the house, apartment, or mobile home that allows the respondent both to make and receive calls. Households whose service has been discontinued for nonpayment or other reasons are not considered to have telephone service available. Beginning in 2003, the instructions that accompanied the ACS mail questionnaire advised respondents to answer that the house or apartment has telephone service available if cellular telephones are used by household members.

Vehicles Available. These data show the number of passenger cars, vans, and pickup or panel trucks of one-ton capacity or less kept at home and available for the use of household members. Vehicles rented or leased for 1 month or more, company vehicles, and police and government vehicles are included if kept at home and used for nonbusiness purposes. Dismantled or immobile vehicles are excluded, as are vehicles kept at home but used only for business purposes.

House Heating Fuel. House heating fuel information is collected only for occupied HUs. The data show the type of fuel used most to heat the house, apartment, or mobile home.

Selected Monthly Owner Costs. Selected monthly owner costs are the sum of payments for mortgages, deeds of trust, contracts to purchase, or similar debts on the property; real estate taxes; fire, hazard, and flood insurance; utilities (electric, gas, water, and sewer); and fuels (such as oil, coal, kerosene, or wood). These costs also encompass monthly condominium fees or mobile home costs.

Food Stamp Benefit. The Food and Nutrition Service of the U.S. Department of Agriculture (USDA) administers the Food Stamp Program through state and local welfare offices. The Food Stamp Program is the major national income-support program for which all low-income and low-resource households, regardless of household characteristics, are eligible. This question estimates the number of households that received food stamp benefits at any time during the 12-month period before the ACS interview.

Tenure. All occupied HUs are divided into two categories—owner-occupied and renter-occupied. An HU is owner-occupied if the owner or co-owner lives in the unit, even if it is mortgaged or not fully paid for. All occupied HUs that are not owner-occupied, whether they are rented for cash rent or occupied without payment of rent, are classified as renter-occupied.

Contract Rent. Contract rent is the monthly rent agreed to or contracted for, regardless of any furnishings, utilities, fees, meals, or services that may be included.

Gross Rent. Gross rent is the contract rent plus the estimated average monthly cost of utilities and fuels, if these are paid by the renter.

Value of Property. The survey estimates of value of property are based on the respondent's estimate of how much the property (house and lot, mobile home and lot, or condominium unit) would sell for. The information is collected for HUs that are owned or being bought, and for vacant HUs that are for sale. If the house or mobile home is owned or being bought, but the land on which it sits is not, the respondent is asked to estimate the combined value of the house or mobile home and the land. For vacant HUs, value is defined as the price asked for the property. This information is obtained from real estate agents, property managers, or neighbors.

Mortgage Status. Mortgage refers to all forms of debt where the property is pledged as security for repayment of the debt.

Mortgage Payment. This item provides the regular monthly amount required to be paid to the lender for the first mortgage on the property.

Detailed Population Information

Detailed population data are collected for all current household members. Some questions are limited to a subset, based on age or other responses. The 2003–2007 ACS included 36 detailed population questions. In Puerto Rico, the place of birth, residence 1 year ago (migration), and citizenship questions differ from those used in the United States. The definitions below refer specifically to the United States. This section describes concepts and definitions for the detailed population items.

Place of Birth. Each person is asked whether he or she was born in or outside of the United States. Those born in the United States are then asked to report the name of the state; people born elsewhere are asked to report the name of the country, or Puerto Rico and U.S. Island Areas.

Citizenship. The responses to this question are used to determine the U.S. citizen and non-U.S. citizen populations and native and foreign-born populations. The foreign-born population includes anyone who was not a U.S. citizen at birth. This includes people who indicate that they are not U.S. citizens, or are citizens by naturalization.

Year of Entry. All respondents born outside of the country are asked for the year in which they came to live in the United States, including people born in Puerto Rico and U.S. Island Areas, those born abroad of an American (U.S. citizen) parent(s), and foreign-born people.

Type of School and School Enrollment. People are classified as enrolled in school if they have attended a regular public or private school or college at any time during the 3 months prior to the time of interview. This question includes instructions to "include only nursery or preschool, kindergarten, elementary school, and schooling which leads to a high school diploma, or a college degree" as a regular school or college. Data are tabulated for people 3 years and older.

Educational Attainment. Educational attainment data are tabulated for people 18 years and older. Respondents are classified according to the highest degree or the highest level of school completed. The question includes instructions for people currently enrolled in school to report the level of the previous grade attended or the highest degree received.

Ancestry. Ancestry refers to a person's ethnic origin or descent, roots or heritage, place of birth, or place of parents' ancestors before their arrival in the United States. Some ethnic identities, such as "Egyptian" or "Polish" can be traced to geographic areas outside the United States, while other ethnicities such as "Pennsylvania German" or "Cajun" evolved within the United States.

Language Spoken at Home. Respondents are instructed to mark "Yes" if they sometimes or always speak a language other than English at home, but "No" if the language is spoken only at school or is limited to a few expressions or slang. Respondents are asked the name of the non-English language spoken at home. If the person speaks more than one language other than English at home, the person should report the language spoken most often or, if he or she cannot determine the one spoken most often, the language learned first.

Ability to Speak English. Ability to speak English is based on the person's self-response.

Residence 1 Year Ago (Migration). Residence 1 year ago is used in conjunction with location of current residence to determine the extent of residential mobility and the resulting redistribution of the population across geographic areas of the country.

Disability. Disability is defined as a long-lasting sensory, physical, mental, or emotional condition that makes it difficult for a person to perform activities such as walking, climbing stairs, dressing, bathing, learning, or remembering. It may impede a person from being able to go outside of the home alone or work at a job or business; the definition includes people with severe vision or hearing impairments.

Fertility. This question asks if the person has given birth in the previous 12 months.

Grandparents as Caregivers. Data are collected on whether a grandchild lives with a grandparent in the household, whether the grandparent has responsibility for the basic needs of the grandchild, and the duration of that responsibility.

Veteran Status. A "civilian veteran" is a person aged 18 years and older who has served (even for a short time), but is not now serving, on active duty in the U.S. Army, Navy, Air Force, Marine Corps, or Coast Guard, or who served in the U.S. Merchant Marine during World War II. People who have served in the National Guard or military reserves are classified as veterans only if they were called or ordered to active duty at some point, not counting the 4 to 6 months of initial training or yearly summer camps. All other civilians aged 18 and older are classified as nonveterans.

Work Status. People aged 16 and older who have worked 1 or more weeks are classified as having "worked in the past 12 months." All other people aged 16 and older are classified as "did not work in the past 12 months."

Place of Work. Data on place of work refer to the location (street address, city/county, state) at which workers carried out their occupational activities during the reference week.

Means of Transportation to Work. Means of transportation to work refers to the principal mode of travel or type of conveyance that the worker usually used to get from home to work during the reference week.

Time Leaving Home to Go to Work. This item covers the time of day that the respondent usually left home to go to work during the reference week.

Travel Time to Work. This question asks the total number of minutes that it usually took the worker to get from home to work during the reference week.

Labor Force Status. These questions on labor force status are designed to identify: (1) people who worked at any time during the reference week; (2) people on temporary layoff who were available for work; (3) people who did not work during the reference week but who had jobs or businesses from which they were temporarily absent (excluding layoffs); (4) people who did not work but were available during the reference week, and who were looking for work during the last 4 weeks; and (5) people not in the labor force.

Industry, Occupation, Class of Worker. Information on industry relates to the kind of business conducted by a person's employing organization; occupation describes the kind of work the person does. For employed people, the data refer to the person's job during the previous week. For those who work two or more jobs, the data refer to the job where the person worked the greatest number of hours. For unemployed people, the data refer to their last job. The information on class of worker refers to the same job as a respondent's industry and occupation, and categorizes people according to the type of ownership of the employing organization.

Income. "Total income" is the sum of the amounts reported separately for wage or salary income; net self-employment income; interest, dividends, or net rental or royalty income, or income from estates and trusts; social security or railroad retirement income; Supplemental Security Income; public assistance or welfare payments; retirement, survivor, or disability pensions; and all other income. The estimates are inflation-adjusted using the Consumer Price Index.

6.5 STRUCTURE OF THE GROUP QUARTERS QUESTIONNAIRES

The 2006–2007 GQ questionnaire includes all of the population items included on the HU questionnaire, except for relationship. One housing question, food stamp benefit, is asked. Address information is for the GQ facility itself and is collected as part of the automated GQ Facility Questionnaire. The survey information collected from each person selected to be interviewed is entered on a separate questionnaire. The number of questionnaires completed for each GQ facility is the same as the number of people selected, unless a sample person refuses to participate.

Chapter 7.
Data Collection and Capture for Housing Units

7.1 OVERVIEW

A key measure of the success of a data collection effort is the final response rate. The American Community Survey (ACS) achieves a high total response rate each year, due in part to the data collection design, which in turn reflects the experience and research in data collection strategies drawn from the U.S. Census Bureau's decennial census and demographic survey programs. Success, however, would not be possible without the high quality of the actual data collection, which is due to the efforts of the interviewing staff in the telephone centers and regional offices. This success also is related to the mandatory nature of the survey. Title 13 of the United States Code [U.S.C.] authorizes the Census Bureau to conduct the ACS, requires households to participate, and requires the Census Bureau to keep confidential all information collected.

The data collection operation for housing units (HUs) consists of three modes: mail, telephone, and personal visit. For most HUs, the first phase includes a questionnaire mailed to the sample address, with a request to the household to complete the questionnaire and return it by mail. If no response is received, the Census Bureau follows up with computer-assisted telephone interviewing (CATI) when a telephone number is available. If the Census Bureau is unable to reach an occupant using CATI, or if the household refuses to participate, the address may be selected for computer-assisted personal interviewing (CAPI).

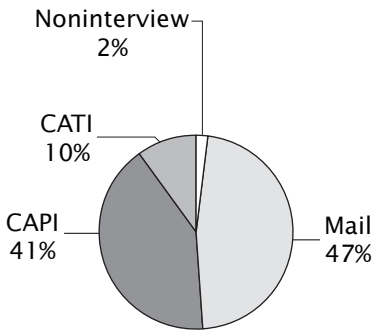
Figure 7.1 **ACS Data Collection Consists of Three Overlapping Phases**

ACS sample panel	Month of data collection							
	2005		2006					
	November	December	January	February	March	April	May	June
November 2005	Mail	Phone	Personal visit					
December 2005		Mail	Phone	Personal visit				
January 2006			Mail	Phone	Personal visit			
February 2006				Mail	Phone	Personal visit		
March 2006					Mail	Phone	Personal visit	
April 2006						Mail	Phone	Personal visit
May 2006							Mail	Phone
June 2006								Mail

The ACS includes 12 monthly independent samples. Data collection for each sample lasts for 3 months, with mail returns accepted during this entire period, as shown in Figure 7.1. This three-phase process operates in continuously overlapping cycles so that, during any given month, three samples are in the mail phase, one is in the CATI phase, and one is in the CAPI phase.

Figure 7.2 summarizes the distribution of interviews and noninterviews for the 2007 ACS. Among the ACS sample addresses eligible for interviewing in the United States, approximately 47 percent were interviewed by mail, 10 percent by CATI, and 41 percent were represented by CAPI interviews. Two percent were noninterviews.

Figure 7.2 **Distribution of ACS Interviews and Noninterviews**



Source: 2007 ACS Sample.

7.2 MAIL PHASE

Mail is the least expensive method of data collection, and the success of the program depends on high levels of mail response. Sample addresses are reviewed to determine whether the available information is sufficient for mailing. The requirement for a "mailable" address in the United States is met if there is either a complete city-style or rural route address. A complete city-style address includes a house number, street name, and ZIP Code. (The town or city and state fields are not required because they can be derived from the ZIP Code.) A complete rural-route address includes a rural-route number, box number, and ZIP Code. About 95 percent of the 2007 sample addresses in the United States met these criteria and were designated as mailable.

The requirement for a mailable address differs slightly in Puerto Rico. In addition to the criteria for the United States, sample city-style addresses in Puerto Rico also must have an "urbanización" name, building name, or condominium name to be considered mailable. About 72 percent of the addresses in Puerto Rico were considered mailable in 2007.

Examples of unmailable addresses include those with only physical descriptions of an HU and its location, or with post office (P.O.) box addresses, as well as addresses missing place names and ZIP Codes. P.O. box addresses are considered unmailable because of the unknown location of the HU using the P.O. box. Addresses missing ZIP Codes are considered unmailable when the place name is also missing. HU addresses not meeting one of the completeness criteria are still included in the sample frame, but they bypass the mail and telephone phases.

Mailout

Because a high level of mail response is critical, the mail phase used in the ACS consists of three to four mailings to each sample address, depending on when a return is received. ACS materials for U.S. addresses are printed in English, and Puerto Rico Community Survey (PRCS) materials sent to Puerto Rico are printed in Spanish. U.S. respondents can request Spanish mailing packages, and Puerto Rico respondents can request English mailing packages, via telephone questionnaire assistance (TQA). The address label file that includes all mailable sample addresses defines the universe for the first three mailings: a prenotice letter, an initial mail package, and a reminder postcard. A replacement mail package is sent to sample addresses when there is no response 3 weeks after mailing the initial mail package. (Details of each are provided below, and samples are available at <http://www.census.gov/acs/www/SBasics/>.)

Prenotice Letter. The first mailing consists of a prenotice letter, signed by the Census Bureau's director, alerting residents that they will receive the ACS questionnaire in a few days and encouraging them to return the questionnaire promptly. The prenotice letter is mailed on the Thursday before the last Monday of the month, unless that last Monday is one of the last 2

days of the month, in which case the mailout schedule begins 1 week earlier. The prenotice letter is one of two ACS items printed in-house using print-on-demand technology, which merges the letter text and the sample address from the address label file.

Initial Mail Package. The next mailing is the initial mail package. On the front of the envelope is a boxed message informing recipients that the ACS form is enclosed, and stating in bold, uppercase type that a response is required by law. This initial mail package is mailed on the last Monday of the month or on the previous Monday if the last day of the month is a Monday or a Tuesday. The first mail package includes a cover letter, the questionnaire, an instructional guide, a brochure, and a return envelope.

Cover Letter. The cover letter is signed by the Census Bureau's director. It reminds householders that they received the prenotice letter a few days earlier and encourages them to return the completed questionnaire as soon as possible. The letter then explains the purpose of the ACS and how the data are used. Finally, a toll-free telephone number is included for respondents if they have questions or need help completing the questionnaire.

ACS Questionnaire. The 2006 and 2007 ACS questionnaires are 24-page, two-color booklet-style forms. They are printed on white paper with colored ink—green for the U.S. form, yellow for the Puerto Rico form. The cover of the questionnaire includes information in English and Spanish on how to obtain assistance. The questionnaire includes questions about the HU and the people living in it. Space is provided for detailed information for up to five people. Follow-up by telephone is used for households that return their questionnaires by mail and report that six or more people reside in the household.

Guide to the ACS. The guide instructs respondents how to complete the survey.

Frequently Asked Questions (FAQs) Brochure. This color brochure, available in both English and Spanish, provides answers to frequently asked questions about the ACS. Examples include "What is the American Community Survey?," "Do I have to answer the questions on the American Community Survey?," and "Will the Census Bureau keep my information confidential?" A similar brochure about the PRCS is used in packages mailed to Puerto Rico.

Return Envelope. The postage-paid envelope is for returning the questionnaire to the Census Bureau.

Reminder Postcard. The third mailing is a postcard, also signed by the director of the Census Bureau. The postcard is mailed on Thursdays, 3 days after the initial mail package, and reminds respondents to return their questionnaires. The reminder postcard also is printed in-house, using print-on-demand technology to merge text and addresses.

Replacement Mail Package. The last mailing is sent only to those sample addresses from which the initial questionnaire has not been returned. It is mailed about 3½ weeks after the initial mail package. The contents are the same except that it contains a different cover letter. Signed by the director of the Census Bureau, it reminds the household of the importance of the ACS, and asks them to respond soon.

The Census Bureau's National Processing Center (NPC) assembles and mails the packages for the selected addresses. All of the components of the mail packages except the prenotice letter and reminder postcard are printed under contract by outside vendors. As the vendors print the materials, NPC quality control staff monitor the work and reject materials that do not meet contractual quality standards.

The NPC is responsible for labeling the outgoing mail packages. Several months before each sample's mailings, Census Bureau headquarters staff provides an address file to the NPC for use in creating address labels for the first three mailings. An updated address file is provided to the NPC about 3 days before the mailing of the replacement mail package. This file excludes addresses from which a questionnaire was returned during the first 3 weeks; these usually amount to about 25 to 30 percent of the sample addresses for the United States, and about 10 percent of the sample addresses for Puerto Rico.

Most mail responses are received within 5 weeks after the initial mail package is sent, but the NPC will continue to accept questionnaires for 3 months from the start of each monthly sample. After a specified cutoff date, late mail returns will not be included in the data set.

Check-In

The United States Postal Service (USPS) returns all completed ACS questionnaires to the NPC. The check-in unit receives mail deliveries two or three times each business day. Each questionnaire contains a unique bar code in the address label area. The mail returns are sent through a laser sorter, where the bar code is scanned; this allows sorting by and within monthly sample and by location. During this step, the return envelopes are opened mechanically.

After clerks remove the forms from the return envelopes, the forms are taken to a unit where another set of clerks looks at each page of every returned questionnaire. They also look for enclosed correspondence, which they forward to headquarters, if necessary. The clerks then scan the bar code on each questionnaire to officially check in the form, and organize the forms into batches of 50. Staff have 3 days to check in a form, although usually they check in all the forms they receive within 1 day. Each day, NPC staff transmit a file of the checked-in cases, and headquarters staff update the status of each case in the control file.

Some of the forms are returned to the NPC as "undeliverable as addressed" (UAA) by the USPS. UAAs occur for many reasons, including bad or unknown addresses, vacant HUs, or residents' refusals to accept mail delivery. Sample addresses that are UAAs initially remain eligible for the replacement mail package because the delivery process for an address often is successful on the second attempt without any change to the address. UAAs are eligible for the CATI and CAPI operations.

Telephone Questionnaire Assistance (TQA)

TQA is a toll-free, interactive voice recognition (IVR) telephone system that respondents can call if they have questions about completing the questionnaire, or to request one in another language. The TQA telephone number is listed on the questionnaire, as well as on all of the letters, brochures, and postcards. Alternate TQA numbers are listed on the questionnaire for Spanish speakers and for a telephone device for the deaf (TDD).

When respondents call TQA, they enter the IVR system, which provides some basic information on the ACS and directions on using the IVR. Respondents may obtain recorded answers to FAQs, or they can speak directly to an agent during business hours. Respondents can furnish their ACS identification number from any of the mailing pieces, which allows them to hear a customized message about the current status of their questionnaire. The IVR can indicate whether the NPC has received a questionnaire for the sample address and, if not, can state that an ACS interviewer may call or visit. If a respondent chooses to speak directly to an agent, the agent answers the caller's questions and encourages the respondent to complete the questionnaire over the telephone. Agents use an automated survey instrument to capture the respondent's answers.

Household members from approximately 6 percent of the mailable addresses called the toll-free number for assistance in 2006 and 2007. For less than 1 percent of the mailable addresses in 2006 and 2007, household members agreed to complete the survey over the telephone. All calls are logged, and the system can record up to five reasons for each call. Even though TQA interviews are conducted by telephone, they are considered mail responses because the call was initiated by the sample household upon receiving the questionnaire in the mail.

Data Capture

After the questionnaires have been checked in and batched into groups of 50, they move to the data entry (keying) unit in the NPC. The keying unit has the goal of keying the responses from the questionnaires within 3 weeks of receipt. Data keyers enter the information from the forms into a data capture file. Each day, NPC staff transmit a file with the keyed data, and headquarters staff update the status of each case in the control file. The NPC's data keying operation uses stringent quality assurance procedures to minimize nonsampling errors.

Data keyers move through three levels of quality assurance verification. When new keyers begin data entry for ACS questionnaires, they are in a training stage, during which 100 percent of their work is checked for correctness. An experienced keyer independently rekeys the same batch of 50 questionnaires, and the work of the two keyers is compared to check for keying errors, defined as incorrectly keyed data items. If the new keyer's error rate (the percentage of all keyed data items that are in error) in one of the first two batches of questionnaires is equal to or less than 1.5 percent, the keyer is moved to the prequalified stage. If the keyer's error rate is greater than 1.5 percent, the keyer is retrained immediately, reassessed, and then advances to the prequalified stage. (These keyers are still subject to 100-percent verification.)

Once prequalified keyers key a batch at an error rate equal to or less than 1.5 percent, they are moved to the qualified stage. If these keyers exceed the error rate of 1.5 percent, they receive immediate feedback. A supervisor eventually decides whether to move them to the qualified stage by verifying a sample of their work, with an acceptable error rate of 1.5 percent or less. Keyers at all levels are subject to removal from the project and administrative action if they fail to maintain an error rate of less than 0.80 percent, but most have a much lower rate.

In mid-2007, the Census Bureau moved to a key-from-image (KFI) data capture system for the HU questionnaires, which involves imaging the questionnaire, interpreting the check box entries with optical mark recognition (OMR), and keying write-in responses from the images using a computerized system. The advantages of KFI include the potential for reduced costs and increased data-capture accuracy.

Failed-Edit Follow-Up

After the data are keyed, the data files are processed in batches through a computerized edit to check coverage consistency and content completeness. This edit identifies cases requiring additional information. Cases that fail are eligible for the telephone failed-edit follow-up (FEFU) operation, and become part of the FEFU workload if a telephone number for the sample address is available. This operation is designed to improve the final quality of mail-returned questionnaires.

Cases failing the edit fall into two broad categories: coverage failures and content failures. Coverage failures can take two forms. First, since the ACS questionnaire is designed to accommodate detailed answers for households with five or fewer people, a case will fail when a respondent indicates that there are more than five people living in the household, or if the reported number of people differs from the number of people for whom responses are provided. Content failures occur if the edit determines that two or more critical items, or a specific number of other required items, have not been answered.

Approximately 33 percent of the keyed mail-return questionnaires in 2006 and 2007 failed either the coverage or content edits and required FEFU. A new set of FEFU cases is generated each business day, and telephone center staff call respondents to obtain the missing data. The interview period for each FEFU case is 3 weeks.

7.3 TELEPHONE PHASE

The second data collection phase is the telephone phase, or CATI. The automated data collection instrument (the set of questions, the list of response categories, and the logic that presents the next appropriate question based on the response to a given question) is written in BLAISE, an open-source scripting software language. The CATI instrument is available in English and Spanish in both the United States and Puerto Rico.

To be eligible for CATI, an HU that did not respond by mail must have a mailable address and a telephone number. The Census Bureau contracts with vendors who attempt to match the ACS sample addresses to their databases of addresses and then provide telephone numbers. There are two vendors for United States addresses and one for Puerto Rico addresses and, since the vendors use different methodologies and sources, one may be able to provide a telephone number while another may not. This matching operation occurs each month before a sample is mailed. About a month later, just prior to the monthly CATI work, headquarters staff transmit a file of the CATI-eligible sample addresses and telephone numbers to a common queue for all three telephone call centers.

The Census Bureau conducts CATI from its three telephone call centers located in Jeffersonville, Indiana; Hagerstown, Maryland; and Tucson, Arizona. The CATI operation begins about 5 weeks after the first mail package is sent out. A control system, WebCATI, is used to assign the cases to individual telephone interviewers. As CATI interviewers begin contacting the households, the WebCATI system evaluates the skills needed for each case (for example, language or refusal conversion skills) and delivers the case to those interviewers who possess the requisite skill(s).

Once a CATI interviewer reaches a person, the first task is to verify that the interviewer has contacted the correct address. If so, the interviewer attempts to complete the interview. If the householder refuses to participate in the CATI interview, a different CATI interviewer trained in dealing with refusals will call the household after a few days. If the household again refuses, CATI contact attempts are stopped, and the case is coded as a noninterview. If a household's questionnaire is received at any time during the CATI operation, that case is removed from the CATI sample and is considered a mail response. Each day, NPC staff transmit a file with the status of each case, and headquarters staff update the status on the control file.

The CATI operation has a strong quality assurance program, including CATI software-related quality assurance and monitoring of telephone interviewers. The CATI instrument has a sophisticated, integrated set of checks to prevent common errors. For example, a telephone interviewer cannot input out-of-range responses, skip questions that should have been asked, or ask questions that should have been skipped. Both new and experienced telephone interviewers are subject to random monitoring by supervisors to ensure that they follow procedures for asking questions and effectively probe for answers, and to verify that the answers they enter match the answers provided by the respondent.

Approximately 650 interviewers conduct CATI interviews from the Census Bureau's three telephone call centers. Interviewers participate in a 3-day classroom training session to learn and practice the appropriate interviewing procedures. They have 25 to 26 calendar days to complete the monthly CATI caseload, which averaged in 2006 and 2007 about 95,000 cases each month. At the end of the CATI interview cycle, all cases receive a CATI outcome code in one of three general categories: interview, noninterview, or ineligible for CATI. This last category includes cases with incorrect telephone numbers. Cases in the last two categories are eligible for the personal visit phase.

7.4 PERSONAL VISIT PHASE

The last phase of ACS data collection is the personal visit phase, or CAPI. This phase usually begins on the first day of the third month of data collection for each sample, and typically lasts for the entire month.

After mail and CATI operations have been completed, a CAPI subsample is selected from two categories of cases. Mailable addresses with neither a response to the mailout nor a telephone interview are sampled at a rate of 1 in 2, 2 in 5, or 1 in 3 based on the expected rate of completed interviews at the tract level. Unmailable addresses are sampled at a rate of 2 in 3 (U.S Census Bureau 2007).

The CAPI operation is conducted by Census Bureau field representatives (FRs) operating from the Census Bureau's 12 regional offices (ROs). The sampled cases are distributed among the 12 ROs based on their geographic boundaries. The Boston RO is responsible for CAPI data collection in Puerto Rico.

After the databases containing the sample addresses are distributed to the appropriate RO, the addresses are assigned to FRs. FRs can conduct interviews by telephone or personal visit, using laptop PCs loaded with a survey instrument similar to the one used in the CATI operation. The CAPI instrument is available in English and Spanish in the United States and Puerto Rico.

If a telephone number is available, the FR will first attempt to call the sample address. There are two exceptions: (1) unmailable addresses, because an FR would not be able to verify the location of the address over the telephone; and (2) refusals from the CATI phase, because these residents already have refused a telephone interview. The FR will call and confirm that he or she has

reached the sample address. If so, the FR uses the automated instrument and attempts to conduct the interview. If an FR cannot reach a resident after calling three to five times at different times of the day during the first few days of the interview period, he or she must make a personal visit.

Approximately 80 percent of CAPI cases require an FR visit. In addition to trying to obtain an interview, a visit is needed to determine whether the HU exists and to determine the occupancy status. If an HU does not exist at the sample address, that status is documented. If an FR verifies that an HU is vacant, he or she will interview a knowledgeable respondent, such as the owner, building manager, real estate agent, or a neighbor, and conduct a "vacant interview" to obtain some basic information about the HU. If the HU is currently occupied, the FR will conduct an "occupied" or "temporarily occupied" interview. An FR conducts a temporarily occupied interview when there are residents living in the HU at the time of the FR's visit, but no resident has been living there or plans to live there for more than 2 months.

The FRs are trained to remain polite but persistent when attempting to obtain responses. They also are trained in how to handle almost any situation, from responding to a household that claims to have returned its questionnaire by mail to conducting an interview with a non-English speaking respondent.

When FRs cannot obtain interviews, they must indicate the reason. Such noninterviews are taken seriously, because they have an impact on both sampling and nonsampling error. Noninterviews occur when an eligible respondent cannot be located, is unavailable, or is unwilling to provide the survey information. Additional noninterviews occur when FRs are unable to confirm the status of a sample HU due to restricted access to an area because of a natural disaster or nonadmission to a gated community during the interview period. Some sample cases will be determined to be ineligible for the survey. These include sample addresses of structures under construction, demolished structures, and nonexistent addresses.

One of the tasks for an FR is to check the geographic codes (state, county, tract, and block) for each address he or she visits. The FR either confirms that the codes are correct, corrects them, or records the codes if they are missing.

Approximately 3,500 FRs conduct CAPI interviews across the United States and Puerto Rico. Interviewers have almost the entire month to complete the monthly CAPI caseload, which averages more than 40,000 cases each month. Each day, FRs transmit a file with the status of all personal visit cases, and headquarters staff update the statuses on the control file.

FRs participate in a 4-day classroom training session to learn and practice the appropriate interviewing procedures. Supervisors travel with FRs during their first few work assignments to observe and reinforce the procedures learned in training. In addition, a sample of FRs is selected each month and supervisors reinterview a sample of their cases. The primary purpose of the reinterview program is to verify that FRs are conducting interviews, and doing so correctly.

DATA COLLECTION IN REMOTE ALASKA

Remote areas of Alaska provide special difficulties when interviewing, such as climate, travel, and seasonality of the population. To address some of these challenges, the Census Bureau has designated some of these areas to use different procedures for ACS interviewing.

For areas of Alaska that the Census Bureau defines as remote, ACS operations are different from those operations in the rest of the country. The Census Bureau does not mail questionnaires to Remote Alaska sample units and Remote Alaska respondents do not complete any interviews on a paper questionnaire. We do not attempt to conduct interviews with households in Remote Alaska via Census Bureau telephone center interviewers. All interviews for Remote Alaska are conducted using personal visit procedures only.

In order to allow FRs in Alaska adequate time to resolve some of the transportation and logistical challenges associated with conducting interviews in Remote Alaska areas, the normal period for interviewing is extended from 1 month to 4 months. There are two 4-month interview periods every year in Remote Alaska. The first starts in January and stops at the end of April. The second

starts in September and stops at the end of December. These months were identified as most effective in allowing FRs to gain access to remote areas, and in finding residents of Native Villages at home who might be away during the remaining months participating in subsistence activities.

For some boroughs designated as partially remote by the Census Bureau, hub cities in these boroughs are not included in these Remote Alaska procedures. These cities would have cases selected for sample each month of the year, and would be eligible to receive a mail questionnaire, or to be contacted by a telephone center or personal visit interviewer. Table 7.1 provides a list of Remote Alaska areas and their associated interview periods.

Table 7.1 **Remote Alaska Areas and Their Interview Periods**

Borough name	All or part of borough designated remote	Interview period for the remote portion of the borough	
		January–April	September–December
Aleutians East	All	(X)	
Aleutian Islands	All		(X)
Bethel	Part	½	½
Bristol Bay	All	(X)	
Denali	All		(X)
Dillingham	Part	(X)	
Lake and Peninsula	All		(X)
Nome	Part	½	½
North Slope	Part	(X)	
Northwest Arctic	All	½	½
Southeast	All	½	½
Valdez-Cordova	Part	½	½
Wade Hampton	All	½	½
Yukon-Koyukuk	All	½	½

Note: An X indicates that all workload falls in the interview period.

7.5 REFERENCES

U.S. Census Bureau. 2007. "Accuracy of the Data (2006)." Washington, DC, 2005, <www.census.gov/acs/www/Downloads/ACS/accuracy2006.pdf>.

Chapter 8.
Data Collection and Capture for Group Quarters

8.1 OVERVIEW

All living quarters are classified as either housing units (HUs) or group quarters (GQ). An HU is a house, an apartment, a mobile home, a group of rooms, or a single room occupied or intended for occupancy as separate living quarters. Separate living quarters are those in which the occupants live separately from any other people in the building and that are directly accessible from outside the building or through a common hall.

GQs are places where people live or stay, in a group living arrangement that is owned or managed by an entity or organization providing housing and/or services for the residents. These services may include custodial or medical care, as well as other types of assistance, and residency is commonly restricted to those receiving these services. People living in GQs usually are not related to each other. GQs include such places as college residence halls, residential treatment centers, skilled nursing facilities, group homes, military barracks, correctional facilities, workers' dormitories, and facilities for people experiencing homelessness. GQs are defined according to the housing and/or services provided to residents, and are identified by census GQ type codes.

In January 2006, the American Community Survey (ACS) was expanded to include the population living in GQ facilities. The ACS GQ sample encompasses 12 independent samples; like the HU sample, a new GQ sample is introduced each month. The GQ data collection lasts only 6 weeks and does not include a formal nonresponse follow-up operation. The GQ data collection operation is conducted in two phases. First, U.S. Census Bureau Field Representatives (FRs) conduct interviews with the GQ facility contact person or administrator of the selected GQ (GQ level), and second, the FR conducts interviews with a sample of individuals from the facility (person level).

The GQ-level data collection instrument is an automated Group Quarters Facility Questionnaire (GQFQ). Information collected by the FR using the GQFQ during the GQ-level interview is used to determine or verify the type of facility, population size, and the sample of individuals to be interviewed. FRs conduct GQ-level data collection at approximately 20,000 individual GQ facilities each year.

During the person-level phase, an FR collects the GQ survey information from sampled residents using a bilingual (English/Spanish) GQ paper questionnaire to record detailed information for one person. FRs collect data from approximately 195,000 GQ sample residents each year. All of the methods described in this chapter apply to the ACS GQ operation in both the United States and Puerto Rico, where the survey is called the Puerto Rico Community Survey (PRCS). Samples of all forms and materials used in GQ data collection can be found at <www.census.gov/acs/www/SBasics/GQ/index.htm>.

8.2 GROUP QUARTERS (FACILITY)-LEVEL PHASE

The GQ data collection operation is primarily completed through FR interviews. The FRs may obtain the facility information by conducting either a personal visit or a telephone interview with the GQ contact. Each FR is assigned approximately two sample GQ facilities each month, and interviews are conducted for a period of 6 weeks.

The GQ-level interviews determine whether the FR samples all, some, or none of the residents at a sampled facility for person-level interviews. The FR verifies the sample GQ information and records up to two additional GQ types, if they exist at the same structure. The GQFQ is programmed to determine the appropriate GQ population to sample when more than one GQ type is identified, assigning the correct type code(s) based on GQ contact responses to the questions. The information obtained from GQ-level interviews is transmitted nightly to Census Bureau headquarters through a secure file transfer.

Previsit Mailings. This section provides details about the materials mailed to each GQ facility before the FR makes any contact.

GQ Introductory Letter. Approximately 2 weeks before the FRs begin each monthly GQ assignment, the Census Bureau's National Processing Center (NPC) mails an introductory letter to the sampled GQ facility. The letter explains that the FR will visit the facility to conduct GQ- and person-level data collection. It describes the information that will be asked for by the FR during the visit, the uses of the data, the Internet address where they can find more information about the ACS, and Regional Office (RO) contact information. This letter is printed at NPC using print-on-demand technology, which merges the letter text and the sample GQ name and address. There are 12 RO-specific letters generated for each sample month.

GQ Frequently Asked Questions (FAQ) Brochure. The color, trifold brochure contains FAQs about the ACS and GQ facilities, and is mailed with the GQ introductory letter. Examples of the FAQs are "What is the American Community Survey?," "Do I have to answer the questions on the American Community Survey?," and "Will the Census Bureau keep my information confidential?" Similar brochures are sent to sample GQ facilities in Puerto Rico and Remote Alaska.

GQ State and Local Correctional Facilities Letter. FRs may mail another letter to selected correctional facilities after the GQ introductory letter is sent, but before calling to schedule an appointment to visit. This letter was developed to assist FRs in gaining access to state and local correctional facilities, although the GQ operation does not require FRs to send the letter. The letter asks for the name and title of a person with the authority to schedule FR visits and to coordinate the GQ data collection. It also provides information about the ACS and the dual nature of the FR visit to the facility, and includes a form to return to the RO with the contact name, title, and phone number of a designated GQ contact. A separate letter is also mailed to sampled federal prisons, but it is mailed directly from the Bureau of Prisons (BoP). Special procedures are established for the BoP data collection through a Memorandum of Understanding (MOU) between the Census Bureau and the BoP.

Initial Contact With GQ Facility

In order to conduct the GQ-level interviews for the assigned facility, the FR is instructed to try first to make the initial contact by telephone. If successful in reaching the GQ contact (usually the facility administrator), the FR uses the automated GQFQ—which is available in both English and Spanish to collect information about the facility (such as verifying the name and address of the facility) and to schedule an appointment to visit and complete the GQ-level data collection phase.

If the GQ contact refuses to schedule an appointment for a visit, the FR notifies the RO and the RO staff again try to gain the GQ contacts cooperation. If this attempt at scheduling an appointment is unsuccessful, the FR then visits the GQ facility to try to get the information needed to generate the sample of residents and to conduct the person-level interviews. If still unsuccessful, the RO or FR explains the mandatory nature of the survey, what the FR is attempting to do at the facility, and why.

Visiting the GQ Facility

Upon arrival at the facility, the FR updates or verifies the Special Place[1] (SP) and GQ name, mailing and physical address, facility telephone number, contact name(s), and telephone number(s). Using a flashcard, the FR asks the GQ administrator to indicate which GQ-type code best describes the GQ facility. The GQ contact can identify up to three different GQ-type codes at one address.

The FR generates a person-level sample from all, some, or none of the residents at the facility, depending on the size of the facility and the GQ-type code or codes assigned during the visit. When multiple type codes are assigned to the facility, only those people in the sampled GQ-type code are included in the universe for person-level sampling. The FR records any other GQ-type

[1] A Special Place is the entity or organization providing housing and/or services for the residents of the group quarters. For example, it is the university with multiple dormitories or the correctional facility with units housing inmates. Sometimes the Special Place and the group quarters are one in the same, such as nursing homes or group homes.

codes identified at the sample GQ address, and the address information is updated for future ACS GQ sample selection. If none of the codes are the same as the sampled GQ-type code, the type code that identifies the largest population is used for determining the population for person-level sampling. If the GQ type code assigned during the visit is out of scope for data collection, no residents will be sampled.

After determining that the GQ facility is in scope for GQ data collection, the FR asks for a register of names and/or bed locations for everyone that is living or staying at the sample GQ facility on the day of the visit. This register is used to generate the sample of residents to be interviewed. If a register is not available, the FR creates one using a GQ listing sheet. The listing sheet contains preprinted GQ contact and facility address information.

The FR uses the sampling component of the GQFQ instrument to verify the register provided by the GQ contact person. The instrument proceeds automatically to the beginning of the sampling component after the FR has entered all required facility information and the GQ contact person verifies that there are people living or staying there at the time of the visit. If there are no residents living or staying at the GQ facility at that time, the FR completes the GQ-level interview to update the GQ information and determines the GQ type, but does not conduct person-level interviews.

The sample of GQ residents is generated from the GQFQ instrument through a systematic sample selection. (See Section C for information about data collection from individuals.) The FR matches the line numbers generated for the person sample to the register of current residents. A grid up to 15 lines long appears on the GQFQ laptop screen, along with a place for name, the sample person location description, the line number corresponding to the register, a telephone number, a telephone extension, and a GQ control number (assigned by the GQFQ sampling program). To complete the sampling process, the FR enters information into the GQFQ that specifically identifies the location of each sample person.

The FR must select an interim or final outcome to record the status of the GQ-level interview, and reasons for GQ refusals or noninterviews are specified. The FR can enter an interim GQ-level interview status reason to allow closure of a case and subsequent reentry. From a list in the GQFQ, the FR selects the appropriate reason for exiting an interview and the GQFQ assigns an outcome code that reflects the current interview status.

There are several reasons why GQ-level data collection may not be completed, such as the FR being unable to locate a facility, finding that there are no residents living or staying at the sample GQ facility during the data collection period, determining that there are now only housing units at the sample GQ facility, or finding that the facility no longer exists.

The FRs ask the GQ contact one reinterview question from the GQ-level GQFQ interview. The purpose of the reinterview question is to detect and deter falsification at the GQ-level.

All information collected during the GQ-level phase is transmitted nightly from each FR to the Census Bureau through secure electronic file transfer.

8.3 PERSON-LEVEL PHASE

This section describes person-level interviews at sample GQ facilities. During this phase, the FR collects data for 10 to 15 sample residents at each assigned GQ facility. The FR prepares person-level survey packages from the GQ-level survey packages assembled at NPC, interviews or distributes survey packages to sampled residents, reviews and edits completed questionnaires, and assigns a final outcome code to all questionnaires and GQ assignments.

Preparation

The NPC is responsible for assembling GQ survey packages and delivering them to the ROs 2 weeks before the start of each survey month. Most of the GQ materials are printed under contract by outside vendors; however, due to the smaller scale of the GQ data collection, forms that are needed only at the GQ level are printed in-house. Trained quality control staff from NPC monitor the work as the contractors print the materials. The NPC rejects batches of work if they do not meet contractual quality standards.

The NPC also is responsible for printing and/or addressing the GQ introductory letters, Survey Package Control List for Special Sworn Status (SSS) Individuals, Instruction Manual for SSS Individuals, listing sheets, and FR folder labels. Contractors print all the questionnaires, the questionnaire information guide booklet, brochures, information card booklet, and Privacy Act notices. The NPC labels ACS GQ sample questionnaires with addresses and control numbers.

On a monthly basis, the Census Bureau headquarters provides label/address files for DocuPrinted materials to the NPC. The NPC receives the files approximately 8 weeks prior to the sample months. On each FR assignment folder, NPC preprints a label containing the GQ name; GQ address, state, city, county, and tract-block; RO name; and GQ type code. Each of the 10 to15 personal interview survey packages included in the assignment folder contains a GQ questionnaire (preprinted with the previously described folder label information), questionnaire instruction guide, an unlabeled GQ introductory letter, a return envelope, and a supply of FAQ brochures and Privacy Act notices. Other materials the FR may need, such as the SSS form and the instruction manual for SSS individuals, are provided to the FRs by the ROs.

The FR prepares the number of survey packages needed; 10 sample residents are selected at large sample GQ facilities, while all residents are interviewed at GQ facilities identified as small GQs. The FRs use the register information from the GQFQ to prepare the survey packages needed for person-level interviews. The GQFQ also generates a questionnaire control number to track the questionnaires from the beginning of the person-level phase through keying. The GQ questionnaire contains blank lines below the preprinted GQ address where the FR manually records specific information to locate the sample residents (name and floor, wing, room, or bed locations). This information helps the FR to organize the order of personal interviews efficiently, and enables another FR to locate the sampled residents at the GQ facility if a case is reassigned.

Person-Level Survey Materials

This section provides details about the materials needed to conduct ACS GQ person-level interviews.

Introductory Letter for the Sample Resident. The FR gives each sampled person an introductory letter at the time of the person-level interview. It provides information about the ACS, describes why it is important that they complete the GQ questionnaire, describes uses of ACS data, stresses the confidentiality of their individual responses, and includes the Internet address for the ACS Web site.

ACS GQ Questionnaire. The FR uses a paper GQ questionnaire for person-level data collection. This questionnaire is a bilingual, 14-page, two-color, flip-style booklet. Seven blue pages make up the English language GQ questionnaire and, when flipped over, seven green pages make up the Spanish language version. The GQ questionnaire is designed to record detailed population information for one person. It does not include housing questions except for the food stamp benefit question. When a questionnaire is damaged or missing, the FR uses Case Management assignment software to obtain the control number, SP/GQ name, and address information and transcribes this information into the label area of a blank questionnaire, using this new copy for the data collection. A PRCS GQ bilingual questionnaire is used for person-level data collection in Puerto Rico.

GQ Questionnaire Instruction Guide. The FR provides a copy of the questionnaire Instruction Guide to sample residents when a personal interview cannot be conducted, and the resident is completing the questionnaire him/herself. This guide provides respondents with detailed information about how to complete the GQ questionnaire. It explains each question, with expanded instructions and examples, and instructs the respondent on how to mark the check boxes and record write-in responses.

GQ Question and Answer Brochure. When beginning person-level data collection, the FR has a supply of question and answer brochures to give sample residents. This brochure provides answers to questions about the ACS GQ program.

GQ Return Envelopes. The GQ envelopes are used to return completed questionnaires to the FR or GQ contact. These envelopes are not designed for delivery through the U.S. Postal Service.

Completing the GQ Questionnaire

There are several ways for an FR to obtain a completed GQ questionnaire. The preferred method is for the FR to fill out the questionnaire in a face-to-face interview with the sample resident. However, other data collection methods may be necessary. The FR may fill out the questionnaire during a telephone interview with the resident; conduct a face-to-face proxy interview with a relative, guardian, or GQ contact; leave the questionnaire with the resident to complete; or leave the questionnaires with the GQ contact to distribute to sampled residents and collect them when completed. If the questionnaires are left with sample residents to complete, the FR arranges with the resident or GQ contact to return and pick up the completed questionnaire(s) within 2 days. The FR must be certain that sample residents are physically and mentally able to understand and complete the questionnaires on their own.

Before a GQ contact or a GQ employee obtains access to the names of the sample residents and the sample residents' answers to the GQ questionnaire, they must take an oath to maintain the confidential information about GQ residents. By taking this oath, one attains SSS. Generally, an SSS individual is needed when the sample person is not physically or mentally able to answer the questions. An FR must swear in social workers, administrators, or GQ employees under Title 13, United States Code (U.S.C.) if these individuals need to see a sampled resident's responses. In taking the Oath of Nondisclosure, SSS individuals agree to abide by the same rules that apply to other Census Bureau employees regarding safeguarding of Title 13 respondent information and other protected materials, and acknowledge that they are subject to the same penalties for unauthorized disclosure. Legal guardians do not need to be sworn as SSS individuals. If the sample person gives a GQ employee permission to answer questions or help to answer on their behalf, the GQ employee does not need to be sworn in.

Questionnaire Review

After data collection has been completed for each sample resident, the FR conducts separate edit reviews of the person-level questionnaires and of all questionnaires within a GQ-level assignment. The first review is a manual edit check of the responses recorded on each questionnaire. The FR verifies that all responses are legible and that the write-in entries and check boxes contain appropriate responses according to the skip patterns on the questionnaire.

The FR determines whether a person-level interview is complete, a sufficient partial, or incomplete. An interview is considered complete when all or most of the questions have been answered, and a sufficient partial when enough questions have been answered to define the basic characteristics of the sample person. A case is classified as a noninterview when the answers do not meet the criteria of a complete or sufficient partial interview. The FR verifies that the correct outcome code has been assigned to each questionnaire, recording the status of the questionnaire review with an interim or final outcome code.

The FR conducts a GQ-level assignment review after completing the questionnaire review. This review is necessary to ensure that all questionnaires within each GQ assignment are accurately coded and accounted for. The FR determines if all questionnaires for the GQ facility have been completed, or if a return visit will be necessary. The FR marks any unused questionnaires with an "X" and ships both unused and completed questionnaires to the RO on a flow basis throughout each 6-week data collection period. The ROs conduct a final review of the questionnaires prior to sending completed questionnaires to NPC for keying.

8.4 CHECK-IN AND DATA CAPTURE

The RO checks in all questionnaires returned by the FRs. Based on the final outcome code recorded for each questionnaire, the RO separates blank questionnaires from those with data. Only questionnaires that contain data, identified by the outcome code assignment, are shipped each week to NPC for check-in and keying. The forms are sorted according to the sample month and location (United States or Puerto Rico).

Check-In

The NPC check-in staff are given 3 days to check in a form, although they usually check in all the forms they receive within 1 day. The check-in process results in batches of 50 questionnaires for data capture.

NPC accepts completed questionnaires shipped from the RO on a weekly basis, for a period of 6 weeks from the start of the sample month. Each RO closes out the sample month GQ assignments, accounts for all questionnaires, and sends the remaining completed questionnaires to NPC on the last day of the 6-week data collection period. NPC completes the sample month check-in within 7 days of receipt of the final shipment from each RO. Each questionnaire contains a unique bar code that is scanned; this permits forms to be sorted according to monthly sample panel and within each panel, by location. The forms for the United States and Puerto Rico contain slightly different formatting and are keyed in separate batches.

Clerks review each page of every returned ACS GQ questionnaire. They look for correspondence, which they forward to headquarters if necessary. They then scan each bar code to officially check in the form, retain the English or Spanish pages of the questionnaire, and organize the forms into batches of 50 questionnaires.

Data Capture

After the questionnaires have been checked in and batched, they move to the keying unit where the questionnaires are keyed using Key-From-Paper (KFP) technology. NPC clerical staff key the data from the questionnaires and transmit data files to Census Bureau headquarters each night. Final keying of each GQ sample month is scheduled for the last day of the month following the sample month. This schedule allows approximately 2½ weeks to complete all GQ keying after the final delivery of questionnaires for a sample month.

8.5 SPECIAL PROCEDURES

Some exceptions to the data collection procedures are necessary to collect data efficiently from all GQ facilities, such as those in remote geographic locations or those with GQ security requirements.

Biannual Data Collection in Remote Alaska

FRs conduct data collection at sample GQ facilities in Remote Alaska during two separate periods each survey year; they visit a sample of GQ facilities from January through mid-April, and from September through mid-January. This exception is needed because of difficulties in accessing these areas at certain times of the year. The two time periods designated for GQ interviewing are the same as those used for ACS data collection from sample housing units in Remote Alaska. Chapter 7, Section E, provides additional information about data collection in Remote Alaska.

Annual Data Collection Restrictions in Correctional and Military Facilities

Once each survey year, the FRs conduct all data collection at state prisons, local jails, halfway houses, military disciplinary barracks, and correctional institutions. These GQ types, when selected for the sample multiple times throughout the survey year, have each instance of selection clustered into 1 random month for data collection. (The Census Bureau agreed to a Department of Justice request to conduct data collection at each sampled state prison and local jail only once a year.)

When these GQ types are selected for the sample more than once in a year, the FR (or group of FRs) makes one visit and conducts all interviews at the GQ facilities during one randomly assigned month. The GQFQ automatically takes the FR to the person-level sample selection screen for each multiple sample occurrence of the GQ facility.

Survey Period and Security Restrictions in Federal Correctional Facilities

Person-level data collection for the Bureau of Prisons (BoP) operation is during a 4-month period (September through December) for selected federal prisons and detention centers. The BoP provides the Census Bureau with a file containing all federal prisons and detention centers and a full roster list of inmates for each federal facility. The Census Bureau updates the GQ-level information and generates the person-level samples for these GQ facilities.

Prior to the beginning of the BoP operation, the BoP conducts the security clearances of a list of FR names provided to them by the ROs. This process takes 8 to 10 weeks. FRs cannot contact any federal prison or detention center until informed by their RO that all clearances and BoP contact notifications have taken place. The BoP provides the GQ contact names and phone numbers to the ROs prior to the start of data collection. RO staff schedules an appointment with the GQ contact so the FR can make a personal visit to the GQ. Appointments may be scheduled in advance for any time during the federal prison/detention center data collection period, but FRs are not authorized to enter a prison or detention center without an appointment. Each facility has different periods of time when there is limited or no access. The RO contacts the FR after clearance, provides them with the contact information for their BoP assignments, and gives the FR permission to visit the GQ to drop off the questionnaires for the sampled persons. FRs prepare their survey packages before entering the federal prison.

The FR visits the GQ based on the agreed upon appointment and swears in the GQ contact person at the federal facility. The sworn GQ contact person then delivers and collects the completed GQ questionnaires. The contact person mails the completed forms to the RO in a trackable overnight envelope provided by the FR.

Chapter 9.
Language Assistance Program

9.1 OVERVIEW

The language assistance program for the American Community Survey (ACS) includes a set of methods and procedures designed to assist sample households with limited English proficiency in completing the ACS interview. Language assistance can be provided in many forms, including the development of translated instruments and other survey materials, the recruiting and training of bilingual interviewers, and the provision of telephone or Internet assistance in multiple languages. Providing language assistance is one of many ways that the ACS can improve survey quality by reducing levels of survey nonresponse, the potential for nonresponse bias, and the introduction of response errors; it ensures that individuals with limited English skills will more fully understand the survey questions.

The ACS language assistance program includes the use of several key tools to support each mode of data collection—mail, telephone, and personal visit. The development of these tools was based on research that assessed the current performance of the ACS for non-English speakers. McGovern (2004) found that, despite the limited availability of mail questionnaires in languages other than English, non-English speakers were successfully interviewed by telephone and personal visit follow-up. She also found that the level of item nonresponse for households speaking languages other than English was consistent with the low levels of item nonresponse in English-speaking households. These results led to a focus on improving the quality of data collected in the telephone and personal visit data collection modes. The program includes assistance in a wide variety of languages during the telephone and personal visit nonresponse follow-up stages.[1] Efforts to expand language assistance in the mail mode were postponed; the current focus in the mail mode is limited to supporting Spanish-language speakers.

This chapter provides greater detail on the current language assistance program. It begins with an overview of the language support, translation, and pretesting guidelines. It then discusses methods for all three modes. The chapter closes with a discussion of research and evaluation activities.

9.2 BACKGROUND

The 2010 Decennial Census Program has placed a priority on developing and testing tools to improve the quality of data collected from people with limited English proficiency; in fact, staff involved in the ACS and the 2010 Census have been working jointly to study language barriers and effective methods for data collection. People with limited English skills represent a growing share of the total population. The 2004 ACS found that 8.4 percent of the total population who speak a language other than English at home speak English less than "very well." This is an increase from 7.6 percent in 2000 (U.S. Census Bureau 2004b).

9.3 GUIDELINES

The U.S. Census Bureau does not require the translation of all survey instruments or materials. Each census and survey determines the appropriate set of translated materials and language assistance options needed to ensure high quality survey results. The Census Bureau does require that guidelines be followed whenever a decision is made to translate a data collection instrument or a respondent letter.

In 2004, the Census Bureau released guidelines for language support translation and pretesting. These state that data collection instruments translated from a source language into a target language should be reliable, complete, accurate, and culturally appropriate. Reliable translations convey the intended meaning of the original text. Complete translations should neither add new

[1] In 2005, interviewer language capabilities included English, Spanish, Portuguese, Chinese, Russian, French, Polish, Korean, Vietnamese, German, Japanese, Arabic, Haitian Creole, Italian, Navajo, Tagalog, Greek, and Urdu.

information nor omit information already provided in the source document. An accurate translation is free of both grammatical and spelling errors. Cultural appropriateness considers the culture of the target population when developing the text for translation. In addition to meeting these criteria, translated Census Bureau data collection instruments and related materials should have semantic, conceptual, and normative equivalence. The Census Bureau guidelines recommend the use of a translation team approach to ensure equivalence. The language support guidelines include recommended practices for preparing, translating, and revising materials, and for ensuring sound documentation (U.S. Census Bureau 2004a). The ACS utilizes Census Bureau guidelines in the preparation of data collection instruments, advance letters, and other respondent communications.

9.4 MAIL DATA COLLECTION

The Census Bureau currently mails out ACS questionnaires to each address in a single language. In the United States, English language forms are mailed, while in Puerto Rico, Spanish is used. The cover of the questionnaire of both the English and Spanish mailouts contains a message written in the other language requesting that people who prefer to complete the survey in that language call a toll-free assistance number to obtain assistance or to request the appropriate form. In 2005, the Census Bureau received requests for Spanish questionnaires from less than 0.01 percent of the mailout sample (Griffin 2006b).

Telephone questionnaire assistance is provided in both English and Spanish. A call to the toll-free Spanish help number reaches a Spanish speaker directly. The interviewer will either provide general assistance or conduct the interview. Interviewers are encouraged to convince callers to complete the interview over the phone.

9.5 TELEPHONE AND PERSONAL VISIT FOLLOW-UP

The call centers and regional offices that conduct the computer-assisted telephone interviewing (CATI) and computer-assisted personal interviewing (CAPI) nonresponse follow-up operations make every effort to hire bilingual staff. In addition, CAPI interviewers are instructed to search for interpreters within the sample household, or from the neighborhood, to assist in data collection. The regional offices maintain a list of interpreters who are skilled in many languages and are available to assist the CAPI interviewer in the respondent's preferred language. Interviewers use a flashcard to identify the specific language spoken when they cannot communicate with a particular household. CAPI interviewers can also provide respondents that speak Spanish, Chinese, Russian, Korean, or Vietnamese translated versions of some informational materials. These materials include an introductory letter and two brochures that explain the survey, as well as a letter that thanks the respondent for his or her participation. Future plans include expanding the number of languages that these CAPI informational materials are available in, and increasing the number of materials that are translated.

The ACS CATI and CAPI survey instruments currently are available in both English and Spanish. Interviewers can conduct interviews in additional languages if they have that capability. Because a translated instrument is not available in languages other than English and Spanish, interviewers translate the English version during the interview and record the results on the English instrument. The Census Bureau is exploring the possibility of creating translated instruments or guides for interviewer use in languages other than English and Spanish. Also, there are special procedures and an interviewer training module that deal with the collection of data from respondents who do not speak English. All ACS interviewers are given this training as part of their classroom interviewer training. The training is designed to improve the consistency of these procedures and to remind interviewers of the importance of collecting complete data for all households.

The CATI and CAPI instruments collect important data on language-related issues, including the frequency of the use of interpreters and of the Spanish instrument, which allows the Census Bureau to monitor how data are being collected. The instruments also record how often interviewers conduct translations of their own into different languages. For example, Griffin (2006b) found that in 2005, more than 86 percent of all CAPI interviews with Spanish-speaking households were conducted by a bilingual (Spanish/English) interviewer. She also found that about 8 percent of the interviews conducted with Chinese-speaking households required the assistance of an interpreter who was not a member of the household.

Additional data collected allow the call centers and the regional offices to identify CATI and CAPI cases that were not completed due to language barriers. A profile of this information by language highlights those languages needing greater support. Griffin (2006b) found that, out of 31,489 cases in the 2005 CATI workload that were identified as requiring a language other than English, 9.3 percent could not be interviewed due to a language barrier. The greatest language needs were for Spanish, Vietnamese, Korean, and Chinese. Call center managers used this information to identify specific language recruiting needs and hire additional staff with these skills. Similar information was used to improve CAPI.

Griffin and McGovern (2004) compared the language abilities of CAPI interviewers in each regional office with the needs of the population for that area. This assessment was based on 2003 ACS language data and regional office staffing information. The regional offices used these data to assist in recruiting support in anticipation of the full sample expansion in 2005. A planned update of this assessment for both CATI and CAPI will look at current staffing.

9.6 GROUP QUARTERS

Chapter 8 describes the data collection methodology for people living in group quarters (GQ) facilities. Two instruments are used in GQ data collection—a paper survey questionnaire for interviewing GQ residents, and an automated instrument for collecting administrative information from each facility. The Census Bureau designed and field-tested a bilingual (English/Spanish) GQ questionnaire in 2005. Interviewers used these questionnaires to conduct interviews with a small sample of GQ residents. An interviewer debriefing found that the interviewers had no problems with these questionnaires and, as a result, this form currently is used for GQ data collection. The Census Bureau will hire bilingual interviewers to conduct interviews with non-English speakers in Puerto Rican GQ facilities. The Group Quarters Facility Questionnaire is available in both English and Spanish.

9.7 RESEARCH AND EVALUATION

Due to limited resources, priorities were set for research and development activities related to the language assistance program. Of critical importance was a benchmarking of the effectiveness of current methods. The potential for nonresponse bias due to language barriers was assessed by McGovern (2004) and Griffin and Broadwater (2005). In addition, ACS staff created a Web site on quality measures, including annual information about the effect of language barriers on survey nonresponse. These evaluations and the Web site both show that current methods result in very low levels of noninterviews caused by the interviewer's inability to speak the respondent's language. These nonresponse levels remain low because of special efforts in the field to use interpreters and other means to conduct these interviews. Item level nonresponse also was assessed by McGovern. She found that the mail returns received from non-English speakers are nearly as complete as those from English speakers and that the interviews conducted by telephone and personal visit with non-English speakers are as complete as those from English speakers. The Census Bureau continues to monitor unit nonresponse due to language barriers.

Language barriers can result in measurement errors when respondents do not understand the questions, or when interviewers incorrectly translate a survey question. Staff are exploring options for developing either translated instruments or language guides for use by telephone and personal visit interviewers who conduct interviews in Chinese, Korean, Vietnamese, and Russian to reduce the potential for translation errors. Cognitive testing of the ACS Spanish instrument identified translation concerns (Carrasco 2003). The Census Bureau is planning a more complete assessment of the Spanish instrument to improve the quality of data collected from Spanish-speaking households.

Future research is planned to develop and test additional language assistance materials for the mail mode. Increasing levels of participation by mail can reduce survey costs and improve the quality of final ACS data.

9.8 REFERENCES

Carrasco, Lorena. (2003). "The American Community Survey en Espanol: Using Cognitive Interviews to Test the Functional Equivalency of Questionnaire Translations." Statistical Research Division Study Series Report. Washington, DC: U.S. Census Bureau, 2003.

Griffin, Deborah. (2006b). “Requests for Alternative Language Questionnaires.” American Community Survey Discussion Paper. Washington, DC: U.S. Census Bureau, 2006.

Griffin, Deborah, and Joan Broadwater. (2005). “American Community Survey Noninterview Rates Due to Language Barriers.” Paper presented at the Meetings of the Census Advisory Committee on the African-American Population, the American Indian and Alaska Native Populations, the Asian Population, the Hispanic Population, and the Native Hawaiian and Other Pacific Islander Populations on April 25–27, 2005.

Griffin, Deborah, and Pamela McGovern. (2003). “Language Action Plan for the American Community Survey.” Washington, DC: U.S. Census Bureau, 2003.

McGovern, Pamela, Deborah Griffin, and Larry McGinn. (2003). “Language Action Plan for the American Community Survey.” Meetings of the Census Advisory Committee on the African-American Population, the American Indian and Alaska Native Populations, the Asian Population, the Hispanic Population, and the Native Hawaiian and Other Pacific Islander Populations, May 5–7, 2003.

McGovern, Pamela D. (2004). “A Quality Assessment of Data Collected in the American Community Survey for Households With Low English Proficiency.” Washington, DC: U.S. Census Bureau, 2004.

U.S. Census Bureau. (2004a). “Census Bureau Guideline: Language Translation of Data Collection Instruments and Supporting Materials.” Internal U.S. Census Bureau document, Washington, DC, 2004.

U.S. Census Bureau. (2004b). “Housing and Population Edit Specifications.” Internal U.S. Census Bureau documentation, Washington, DC.

Chapter 10.
Data Preparation and Processing for Housing Units and Group Quarters

10.1 OVERVIEW

Data preparation and processing are critical steps in the survey process, particularly in terms of improving data quality. It is typical for developers of a large ongoing survey, such as the American Community Survey (ACS) to develop stringent procedures and rules to guide these processes and ensure that they are done in a consistent and accurate manner. This chapter discusses the actions taken during ACS data preparation and processing, provides the reader with an understanding of the various stages involved in readying the data for dissemination, and describes the steps taken to produce high-quality data.

The main purpose of data preparation and processing is to take the response data gathered from each survey collection mode to the point where they can be used to produce survey estimates. Data returning from the field typically arrive in various stages of completion, from a completed interview with no problems to one with most or all of the data items left blank. There can be inconsistencies within the interviews, such that one response contradicts another, or duplicate interviews may be returned from the same household but contain different answers to the same question.

Upon arrival at the U.S. Census Bureau, all data undergo data preparation, where responses from different modes are captured in electronic form creating Data Capture Files. The write-in entries from the Data Capture Files are then subject to monthly coding operations. When the monthly Data Capture Files are accumulated at year-end, a series of steps are taken to produce Edit Input Files. These are created by merging operational status information (such as whether the unit is vacant, occupied, or nonexistent) for each housing unit (HU) and group quarters (GQ) facility with the files that include the response data. These combined data then undergo a number of processing steps before they are ready to be tabulated for use in data products.

Figure 10.1 **American Community Survey (ACS) Data Preparation and Processing**

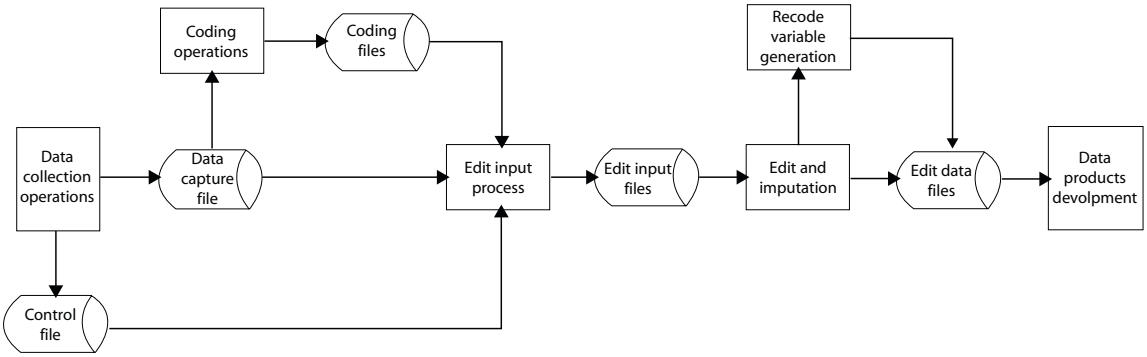


Figure 10.1 depicts the overall flow of data as they pass from data collection operations through data preparation and processing and into data products development. While there are no set definitions of data preparation versus data processing, all activities leading to the creation of the Edit Input Files are considered data preparation activities, while those that follow are considered data processing activities.

10.2 DATA PREPARATION

The ACS control file is integral to data preparation and processing because it provides a single database for all units in the sample. The control file includes detailed information documenting operational outcomes for every ACS sample case. For the mail operations, it documents the receipt and check-in date of questionnaires returned by mail. The status of data capture for these questionnaires and the results of the Failed-Edit Follow-up (FEFU) operation also are recorded in this file. Chapter 7 provides a detailed discussion of mail data collection, as well as computer-assisted telephone interview (CATI) and computer-assisted personal interview (CAPI) operations.

For CAPI operations, the ACS control file stores information on whether or not a unit was determined to be occupied or vacant. Data preparation, which joins together each case's control file information with the raw, unedited response data, involves three operations: creation and processing of data capture files, coding, and creation of edit input files.

Creation and Preparation of Data Capture Files

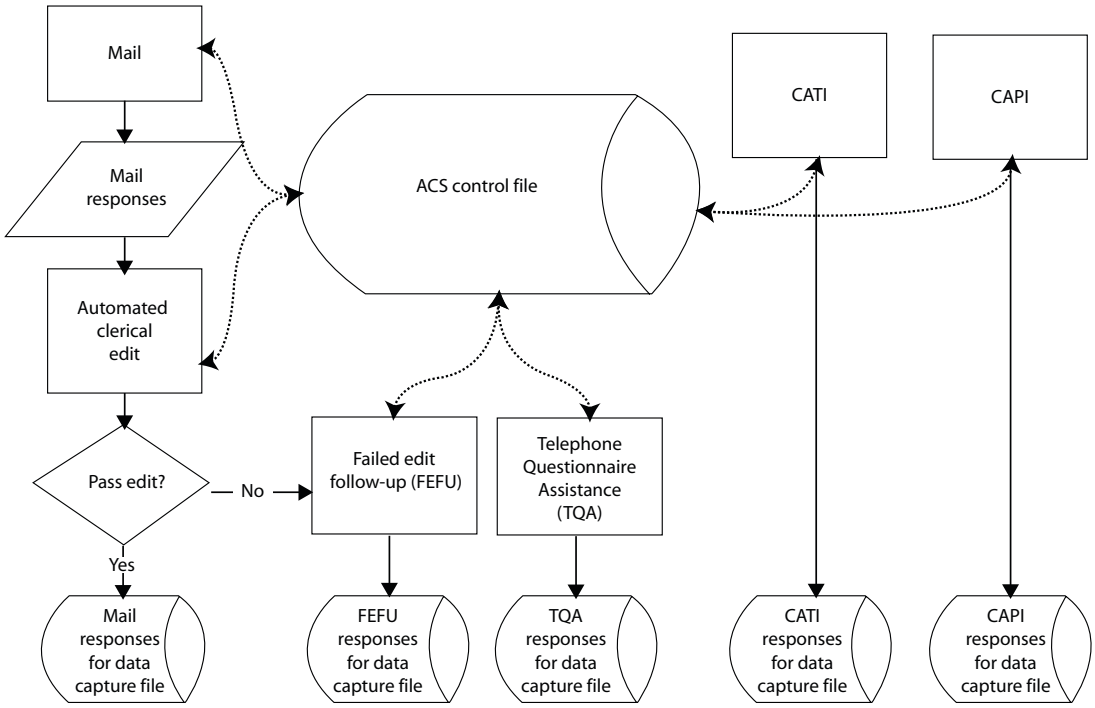
Many processing procedures are necessary to prepare the ACS data for tabulation. In this section, we examine each data preparation procedure separately. These procedures occur daily or monthly, depending on the file type (control or data capture) and the data collection mode (mail, CATI, or CAPI). The processing that produces the final input files for data products is conducted on a yearly basis.

Daily Data Processing

The HU data are collected on a continual basis throughout the year by mail, CATI, and CAPI. Sampled households first are mailed the ACS questionnaire; those households for which a phone number is available that do not respond by mail receive telephone follow-up. As discussed in Chapter 7, a sample of the noncompleted CATI cases is sent to the field for in-person CAPI interviews, together with a sample of cases that could not be mailed. Each day, the status of each sample case is updated in the ACS control file based on data from data collection and capture operations. While the control file does not record response data, it does indicate when cases are completed so as to avoid additional attempts being made for completion in another mode.

The creation and processing of the data depends on the mode of data collection. Figure 10.2 shows the monthly processing of HU response data. Data from questionnaires received by mail are processed daily and are added to a Data Capture File (DCF) on a monthly basis. Data received by mail are run through a computerized process that checks for sufficient responses and for large households that require follow-up. Cases failing the process are sent to the FEFU operation. As discussed in more detail in Chapter 7, the mail version of the ACS asks for detailed information on up to five household members. If there are more than five members in the household, the FEFU process also will ask questions about those additional household members. Telephone interviewers call the cases with missing or inconsistent data for corrections or additional information. The FEFU data are also included in the data capture file as mail responses. The Telephone Questionnaire Assistance (TQA) operation uses the CATI instrument to collect data. These data are also treated as mail responses, as shown in Figure 10.2.

Figure 10.2 **Daily Processing of Housing Unit Data**



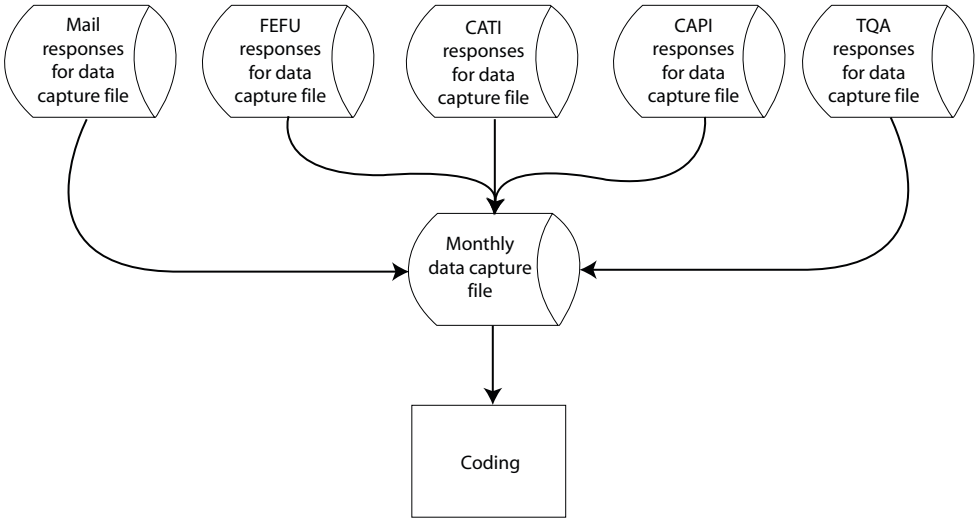
CATI follow-up is conducted at three telephone call centers. Data collected through telephone interviews are entered into a BLAISE instrument. Operational data are transmitted to the Census Bureau headquarters daily to update the control file with the current status of each case. For data collected via the CAPI mode, Census Bureau field representatives (FRs) enter the ACS data directly into a laptop during a personal visit to the sample address. The FR transmits completed cases from the laptop to headquarters using an encrypted Internet connection. The control file also is updated with the current status of the case. Each day, status information for GQs is transmitted to headquarters for use in updating the control file. The GQ data are collected on paper forms that are sent to the National Processing Center on a flow basis for data capture.

Monthly Data Processing

At the end of each month, a centralized DCF is augmented with the mail, CATI, and CAPI data collected during the past month. These represent all data collected during the previous month, regardless of the sample month for which the HU or GQ was chosen. Included in these files of mail responses are FEFU files, both cases successfully completed and those for which the required number of attempts have been made without successful resolution. As shown in Figure 10.3, monthly files from CATI and CAPI, along with the mail data, are used as input files in doing the monthly data capture file processing.

At headquarters, the centralized DCF is used to store all ACS response data. During the creation of the DCF, responses are reviewed and illegal values responses are identified. Responses of "Don't Know" and "Refused" are identified as "D" and "R." Illegal values are identified by an "I," and data capture rules cause some variables to be changed from illegal values to legal values (Diskin, 2007c). An example of an illegal value would occur when a respondent leaves the date of birth blank but gives "Age" as 125. This value is above the maximum allowable value of 115. This variable would be recoded as age of 115 (Diskin, 2007a). Another example would be putting a "19" in front of a four-digit year field where the respondent filled in only the last two digits as "76" (Jiles, 2007). A variety of these data capture rules are applied as the data are keyed in from mail questionnaires, and these same illegal values would be corrected by telephone and field interviewers as they complete the interview. Once the data capture files have gone through this initial data cleaning, the next step is processing the HU questions that require coding.

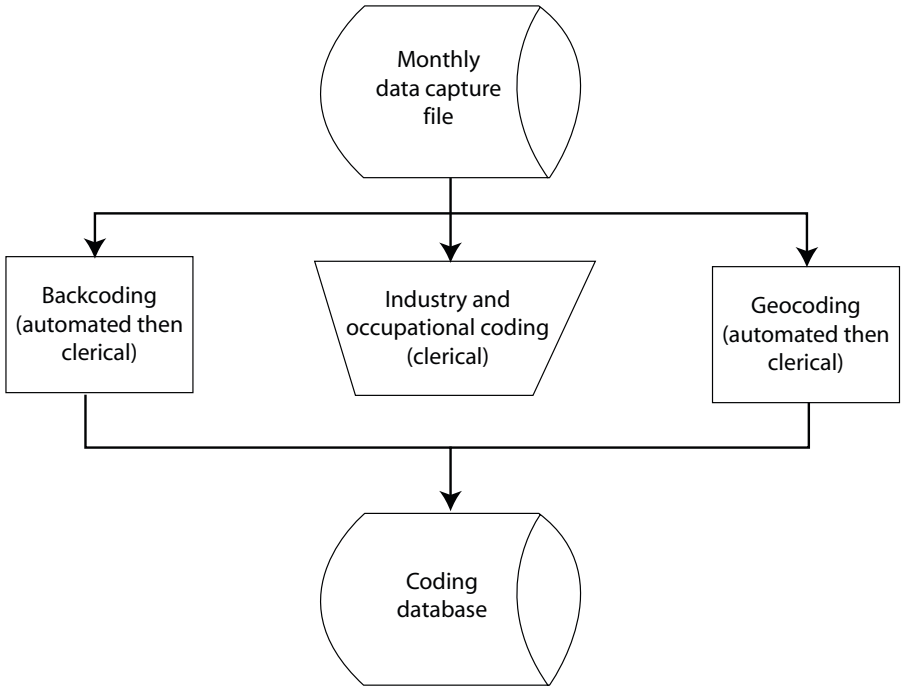
Figure 10.3 **Monthly Data Capture File Creation**



Coding

The ACS questionnaire includes a set of questions that offer the possibility of write-in responses, each of which requires coding to make it machine-readable. Part of the preparation of newly received data for entry into the DCF involves identifying these write-in responses and placing them in a series of files that serve as input to the coding operations. The DCF monthly files include HU and GQ data files, as well as a separate file for each write-in entry. The HU and GQ write-ins are stored together. Figure 10.4 diagrams the general ACS coding process.

Figure 10.4 **American Community Survey Coding**



During the coding phase for write-in responses, fields with write-in values are translated into a prescribed list of valid codes. The write-ins are organized into three types of coding: backcoding, industry and occupation coding, and geocoding. All three types of ACS coding are automated (i.e., use a series of computer programs to assign codes), clerically coded (coded by hand), or some combination of the two. The items that are sent to coding, along with the type and method of coding, are illustrated below in Table 10.1.

Table 10.1 **ACS Coding Items, Types, and Methods**

Item	Type of coding	Method of coding
Race	Backcoding	Automated with clerical follow-up
Hispanic origin	Backcoding	Automated with clerical follow-up
Ancestry	Backcoding	Automated with clerical follow-up
Language	Backcoding	Automated with clerical follow-up
Industry	Industry	Clerical
Occupation	Occupation	Clerical
Place of birth	Geocoding	Automated with clerical follow-up
Migration	Geocoding	Automated with clerical follow-up
Place of work	Geocoding	Automated with clerical follow-up

Backcoding

The first type of coding is the one involving the most items—backcoding. Backcoded items are those that allow for respondents to write in some response other than the categories listed. Although respondents are instructed to mark one or more of the 12 given race categories on the ACS form, they also are given the option to check "Some Other Race," and to provide write-in responses. For example, respondents are instructed that if they answer "American Indian or Alaska Native," they should print the name of their enrolled or principal tribe; this allows for a more specific race response. Figure 10.5 illustrates backcoding.

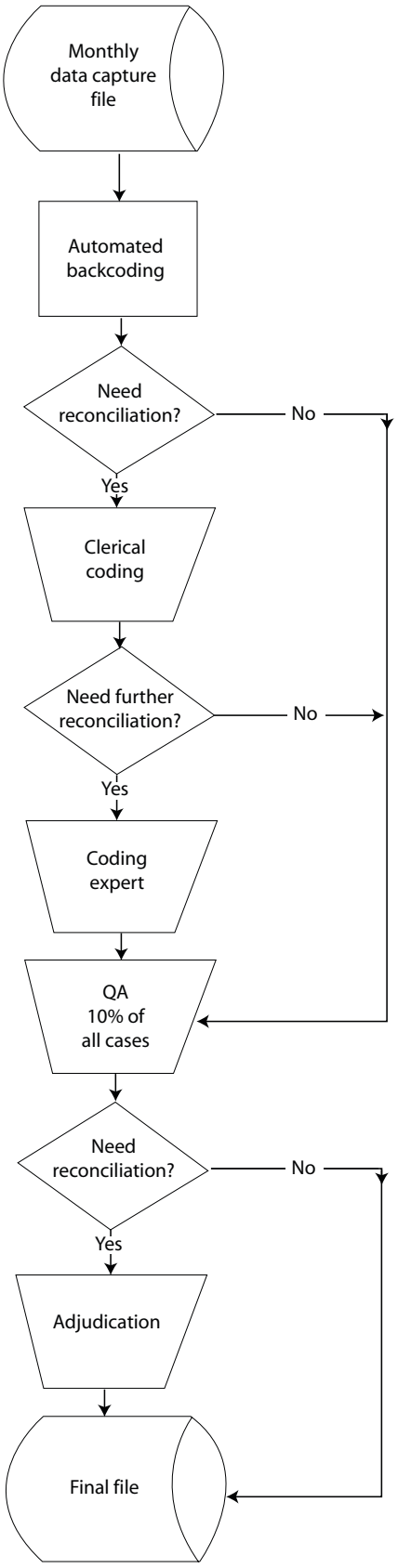
All backcoded items go through an automated process for the first pass of coding. The written-in responses are keyed into digital data and then matched to a data dictionary. The data dictionary contains a list of the most common responses, with a code attached to each. The coding program attempts to match the keyed response to an entry in the dictionary to assign a code. For example, the question of language spoken in the home is automatically coded to one of 380 language categories. These categories were developed from a master code list of 55,000 language names and variations. If the respondent lists more than one non-English language, only the first language is coded.

However, not all cases can be assigned a code using the automated coding program. Responses with misspellings, alternate spellings, or entries that do not match the data dictionary must be sent to clerical coding. Trained human coders will look at each case and assign a code.

One example of a combination of autocoding and follow-up clerical coding is the ancestry item. The write-in string for ancestry is matched against a census file containing all of the responses ever given that have been associated with codes. If there is no match, an item is coded manually. The clerical coder looks at the partial code assigned by the automatic coding program and attempts to assign a full code.

To ensure that coding is accurate, 10 percent of the backcoded items are sent through the quality assurance (QA) process. Batches of 1,000 randomly selected cases are sent to two QA coders who independently assign codes. If the codes they assign do not match one another, or the codes assigned by the automated coding program or clerical coder do not match, the case is sent to adjudication. Adjudicator coders are coding supervisors with additional training and resources. The adjudicating coder decides the proper code, and the case is considered complete.

Figure 10.5 **Backcoding**



Industry and Occupation Coding

The second type of coding is industry and occupation coding. The ACS collects information concerning many aspects of the respondents' work, including commute time and mode of transportation to work, salary, and type of organization employing the household members. To give a clear picture of the kind of work in which Americans are engaged, the ACS also asks about industry and occupation. Industry information relates to the person's employing organization and the kind of business it conducts. Occupation is the work the person does for that organization. To aid in coding the industry and occupation questions, two additional supporting questions are asked—one before the industry question and one after the occupation question. The wording for the industry and occupation questions are shown in Figures 10.6, 10.7, and 10.8.

Figure 10.6 **ACS Industry Questions**

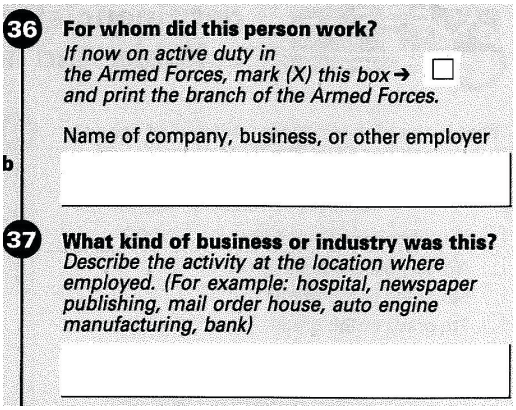


Figure 10.7 **ACS Industry Type Question**

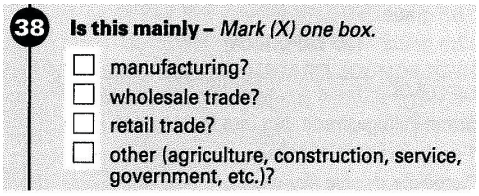
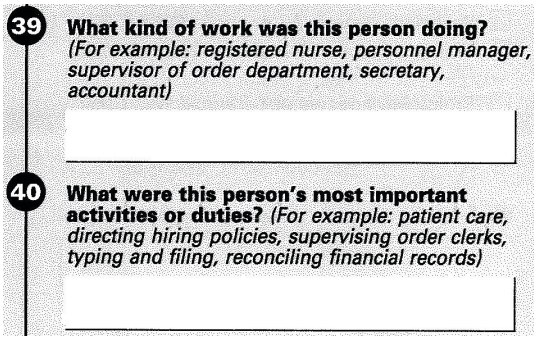
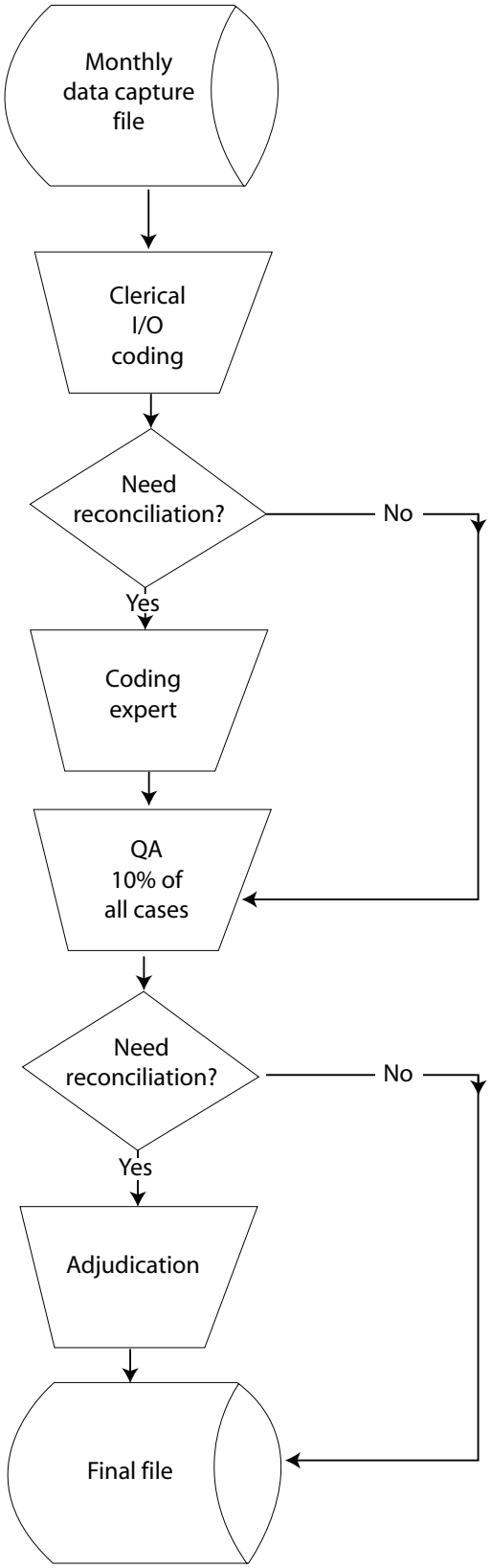


Figure 10.8 **ACS Occupation Questions**



From these questions, the specialized industry and occupation coders assign a code. Unlike back-coded items, industry and occupation items do not go through an automated assignment process. Automated coding programs were used for these items for the 2000 Decennial Census, but it was determined that using trained clerical coders would prove more efficient (Kirk, 2006). Figure 10.9 illustrates industry and occupation coding.

Figure 10.9 **Clerical Industry and Occupation (I/O) Coding**



Industry and occupation clerical coders are trained to use the Census Classification System to code responses. This system is based on the North American Industry Classification System (NAICS) and the Standard Occupational Classification (SOC) Manual. Both industry and occupation are coded to a specificity level of four digits. The Census Classification System can be bridged directly to the NAICS and SOC for comparisons (Kirk, 2006). The NAICS groups businesses into industries based upon their primary activity (U.S. Census Bureau, 2006a, pp. 52–53). The occupation system consists of 23 major occupational groups and 509 specific occupational categories.

To aid in the assigning of industry and occupation codes, coders are given access to additional responses from the respondent. The computer program displays responses to key items that can be used to assist coders in assigning the numeric industry or occupation codes. For example, along with the responses to both the industry and occupation questions, the program also displays the respondent's reported education level, age, and geographic location, all of which may be useful to coders in selecting the most accurate industry or occupation code. The software also includes an alphabetical index on the screen that coders can use for help in assigning codes. Codes are assigned directly into a computer database program. In addition, if respondents provide the name of the company or business for which they work, coders can compare that response with the Employer Name List (ENL), formerly known as the Company Name List, to see if the company name is listed. The Census Bureau developed the ENL from a publication that contains businesses and their NAICS codes. The ENL converts a company's NAICS designation to a Census Classification Code. Using this computerized system, as opposed to coding on the paper instrument itself, has greatly reduced the amount of resources needed to accomplish coding.

When industry and occupation clerical coders are unable to assign a code, the case is sent to an expert, or coding referralist, for a decision. Industry and occupation coding referralists receive an additional 16 hours of training, and are given access to more resources, including hardbound copies of the SOC and NAICS manuals, access to state registries, and use of the Internet for finding more information about the response. Approximately 18 percent of all industry and occupation responses are sent to coding referralists (Earle, 2007). Once these cases are assigned codes, they are placed in the general pool of completed responses.

From this general pool, a fixed percentage of cases are sent through an internal quality assurance verification process, also called the "weighted QA." Coders independently assign a code to a previously coded case; the codes then are reconciled to determine which is correct. Coders are required to maintain a monthly agreement rate of 95 percent or above and a 70 percent or above production rate to remain qualified to code (Earle, 2007). A coding supervisor oversees this process.

Geocoding

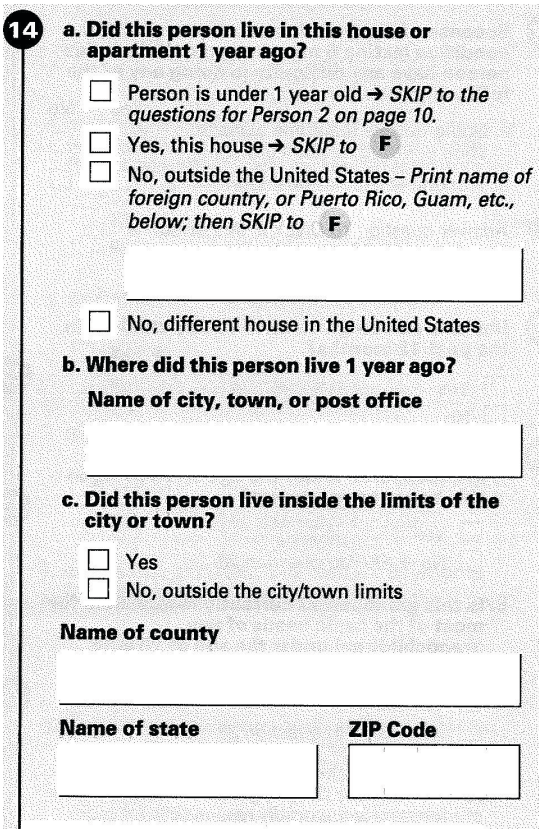
The third type of coding that ACS uses is geocoding. This is the process of assigning a standardized code to geographic data. Place-of-birth, migration, and place-of-work responses require coding of a geographic location. These variables can be as localized as a street address or as general as a country of origin (Boertlein, 2007b).[1]

The first category is place-of-birth coding, a means of coding responses to a U.S. state, the District of Columbia, Puerto Rico, a specific U.S. Island Area, or a foreign country where the respondents were born (Boertlein, 2007b). These data are gathered through a two-part question on the ACS asking where the person was born and in what state (if in the United States) or country (if outside the United States).

The second category of geocoding, migration coding, again requires matching the write-in responses of state, foreign country, county, city, inside/outside city limits, and ZIP code given by the respondent to geocoding reference files and attaching geographic codes to those responses. A series of three questions collects these data and are shown in Figure 10.10. First, respondents are asked if they lived at this address a year ago; if the respondent answers no, there are several follow-up questions, such as the name of the city, country, state, and ZIP code of the previous home.

[1] Please note: The following sections dealing with geocoding rely heavily on Boertlein (2007b).

Figure 10.10 **ACS Migration Question**



The goal of migration coding is to code responses to a U.S. state, the District of Columbia, Puerto Rico, U.S. Island Area or foreign country, a county (municipio in Puerto Rico), a Minor Civil Division (MCD) in 12 states, and place (city, town, or post office). The inside/outside city limits indicator and the ZIP code responses are used in the coding operations but are not a part of the final outgoing geographic codes.

The final category of geocoding is place-of-work (POW) coding. The POW coding questions and the question for employer's name are shown Figure 10.11. The ACS questionnaire first establishes whether the respondent worked in the previous week. If this question is answered "Yes," follow-up questions regarding the physical location of this work are asked.

The POW coding requires matching the write-in responses of structure number and street name address, place, inside/outside city limits, county, state/foreign country, and ZIP code to reference files and attaching geographic codes to those responses. If the street address location information provided by the respondent is inadequate for geocoding, the employer's name often provides the necessary additional information. Again, the inside/outside city limits indicator and ZIP code responses are used in the coding operations but are not a part of the final outgoing geographic codes.

Each of the three geocoding items is coded to different levels of geographic specificity. While place-of-birth geocoding concentrates on larger geographic centers (i.e., states and countries), the POW and migration geocoding tend to focus on more specific data. Table 10.2 is an outline of the specificity of geocoding by type.

Figure 10.11 **ACS Place-of-Work Questions**

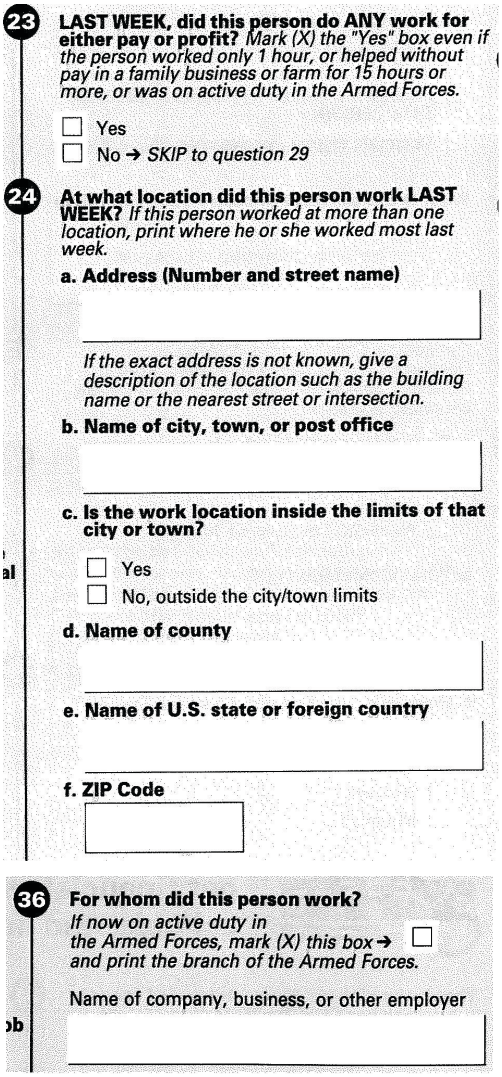


Table 10.2 **Geographic Level of Specificity for Geocoding**

Desired precision—geocoded items	Foreign countries (including: provinces, continents, and regions)	States and statistically equivalent entities	Counties and statistically equivalent entities	ZIP codes	Census designated places	Block levels
Place of birth	X	X				
Migration	X	X	X	X		
Place of work	X	X	X	X	X	X

The main reference file used for geocoding is the State and Foreign Country File (SFCF). The SFCF contains two key pieces of information for geocoding. They are:

- The names and abbreviations of each state, the District of Columbia, Puerto Rico, and the U.S. Island Areas.
- The official names, alternate names, and abbreviations of foreign countries and selected foreign city, state, county, and regional names.

Other reference files (such as a military installation list and City Reference File) are available and used in instances where "the respondent's information is either inconsistent with the instructions or is incomplete" (Boertlein, 2007b).

Responses do not have to match a reference file entry exactly to meet requirements for a correct geocode. The coding algorithm for this automated geocoding allows for equivocations, such as using Soundex values of letters (for example, m=n, f=ph) and reversing consecutive letter combinations (ie=ei). Each equivocation is assigned a numeric value, or confidence level, with exact matches receiving the best score or highest confidence (Boertlein, 2007b). A preference is given for matches that are consistent with any check boxes marked and/or response boxes filled. The responses have to match a reference file entry with a relatively high level of confidence for the automated match to be accepted. Soundex values are used for most types of geocoding and generally are effective at producing matches for given responses. Table 10.3 summarizes the properties of the geocoding workloads by category of codes that were assigned a code automatically.

Table 10.3 **Percentage of Geocoding Cases With Automated Matched Coding**

Characteristic	Percentage of cases assigned a code through automated geocoding
Place of birth	99
Migration	97
Place of work	53

The remaining responses that have not been assigned a code through the automated system are processed in computer-assisted clerical coding (CACC) operations. The CACC coding is separated, with one operation coding to place-level and one coding to block-level responses. Both the place- and block-level CACC operations involve long-term, specially trained clerks who use additional reference materials to code responses that cannot be resolved using the standard reference files and procedures. Clerks use interactive computer systems to search for and select reference file entries that best match the responses, and the computer program then assigns the codes associated with that geographic entity. The CACC operations also generally are effective at assigning codes.

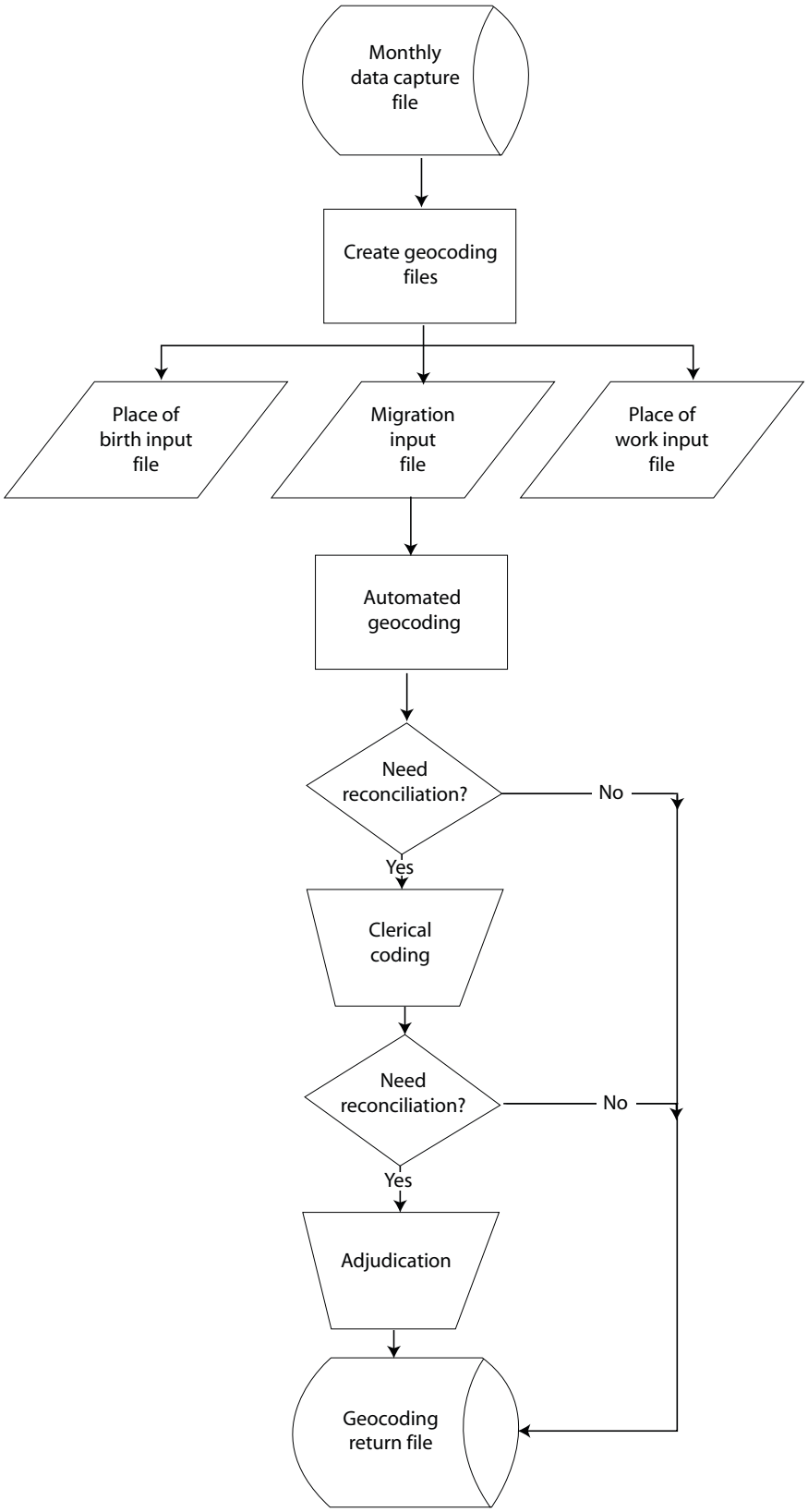
All three geocoding items—place of birth, migration, and place of work—require QA to ensure that the most accurate code has been assigned. The first step of assigning a geocode, the automated coding system, currently does not have a QA step. In both the 1990 and 2000 Decennial Censuses, the automated coding system had an error rate of less than 2.4 percent of all cases (Boertlein, 2007a); since then, the automated coder software has undergone revisions and has been shown to have an even lower error rate.

Among the place-of-birth, migration, and place-of-work cases that were not assigned geocodes by the automated coding system and that subsequently are sent to CACC, 5 percent will be sent to three independent clerical coders. If 2 out of 3 coders agree on a match, the third coder is assigned an error for the case. Coders must keep below a 5 percent error rate per month (Boertlein, 2007a).

For POW block-level coding, the QA protocol is slightly different. Block-level coders must maintain an error rate at or below 10 percent to continue coding. These coders also are expected to have 35 percent or less uncodeable rates. If block-level coders do not maintain these levels, 100 percent of their work is reviewed for accuracy, and additional training may be provided (Boertlein, 2007a).

The QA system for ACS geocoding also includes feedback to the coders. Those with high error rates or high uncodeable rates, as well as those who have low production rates or make consistent errors, may be offered additional training or general feedback on how to improve. Figure 10.12 illustrates automated geocoding.

Figure 10.12 **Geocoding**



10.3 PREPARATION FOR CREATING SELECT FILES AND EDIT INPUT FILES

The final data preparation operation involves creating Select Files and Edit Input Files for data processing. To create these files, a number of preparatory steps must be followed. By the end of the year, the response data stored in the DCF will have been updated 12 times and will become a principal source for the edit-input process. Coding input files are created from the DCF files of write-in entries. Edit Input Files combine data from the DCF files and the returned coding files, and operational information for each case is merged with the ACS control file. The resulting file includes housing and person data. Vacant units are included, as they may have some housing data.

Creation of the Select and Edit Input Files involves carefully examining several components of the data, each described in more detail below. First, the response type and number of people in the household unit are assessed to determine inconsistencies. Second, the return is examined to establish if there are enough data to count the return as complete, and third, any duplicate returns undergo a process of selection to assess which return will be used.

Response Type and Number of People in the HU

Each HU is assigned a response type that describes its status as occupied, temporarily occupied, vacant, a delete, or noninterview. Deleted HUs are units that are determined to be nonexistent, demolished, or commercial units, i.e., out of scope for ACS.

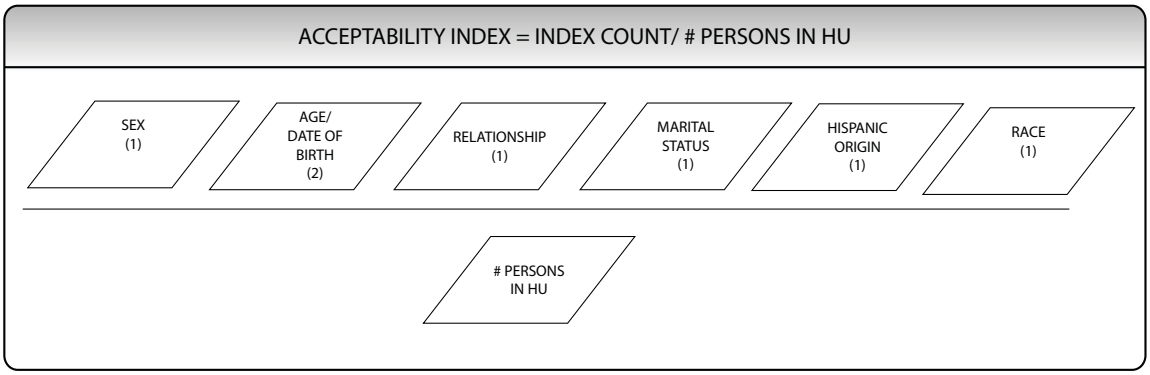
While this type of classification already exists in the DCF, it can be changed from "occupied" to "vacant" or even to "noninterview" under certain circumstances, depending on the final number of persons in the HU, in combination with other variables. In general, if the return indicates that the HU is not occupied and that there are no people listed with data, the record and number of people (which equals 0) is left as is. If the HU is listed as occupied, but the number of persons for whom data are reported is 0, it is considered vacant.

The data also are examined to determine the total number of people living in the HU, which is not always a straightforward process. For example, on a mail return, the count of people on the cover of the form sometimes may not match the number of people reported inside. Another inconsistency would be when more than five members are listed for the HU, and the FEFU fails to get information for any additional members beyond the fifth. In this case, there will be a difference between the number of person records and the number of people listed in the HU. To reconcile the numbers, several steps are taken, but in general, the largest number listed is used. (For more details on the process, see Powers [2006].)

Determining if a Return Is Acceptable

The acceptability index is a data quality measure used to determine if the data collected from an occupied HU or a GQ are complete enough to include a person record. Figure 10.13 illustrates the acceptability index. Six basic demographic questions plus marital status are examined for answers. One point is given for each question answered for a total of seven possible points that could be assigned to each person in the household. A person with a response to either age or date of birth scores two points because given one, the other can be derived or assigned. The total number of points is then divided by the total number of household members. For the interview to be accepted, there must be an average of 2.5 responses per person in the household. Household records that do not meet this acceptability index are classified as noninterviews and will not be included in further data processing. These cases will be accounted for in the weighting process, as outlined in Chapter 11.

Figure 10.13 **Acceptability Index**



If the Acceptability Index is $\geq$ than 2.5, the person record is accepted as a complete return.

If the Acceptability Index is < than 2.5, the person record is not accepted as a complete return.

Unduplicating Multiple Returns

Once the universe of acceptable interviews is determined, the HU data are reviewed to unduplicate multiple returns for a single HU. There are several reasons why more than one response can exist for an HU. A household might return two mail forms, one in response to the initial mailing and a second in response to the replacement mailing. A household might return a mailed form, but also be interviewed in CATI or CAPI before the mail form is logged in as returned. If more than one return exists for an HU, a quality index is used to select one as the final return. This index is calculated as the percentage of items with responses out of the total number of items that should have been completed. The index considers responses to both population and housing items.

The mode of each return also is considered in the decision regarding which of two returns to accept, with preference generally given to mail returns. If two mail returns are received, preference generally is given to the earliest return. For the more complete set of rules, see Powers (2006).

After the resolution of multiple returns, each sample case is assigned a value for three critical variables—data collection mode, month of interview, and case status. The month in which data were collected from each sample case is determined and then used to define the universe of cases to be used in the production of survey estimates. For example, data collected in January 2007 were included in the 2007 ACS data products, even if the returns were sampled in 2006, while ACS surveys sent out in November 2007 were included in the 2007 ACS data products if they were received by mail or otherwise completed by December 31, 2007. Surveys sent out in November 2007 that were received by mail or otherwise completed after December 31, 2007, will be included in the 2008 ACS data products.

10.4 CREATING THE SELECT FILES AND EDIT INPUT FILES

Select Files

Select Files are the series of files that pertain to those cases that will be included in the Edit Input File. As noted above, these files include the case status, the interview month, and the data collection mode for all cases. The largest select file, also called the Omnibus Select File, contains every available case from 14 months of sample—the current (selected) year and November and December of the previous year. This file includes acceptable and unacceptable returns. Unacceptable returns include initial sample cases that were subsampled out at the CAPI stage,[2] returns that were too incomplete to meet the acceptability requirements. In addition, while the "current year"

[2]See Chapter 7 for a full discussion of subsampling and the ACS.

includes all cases sampled in that year, not all returns from the sampled year were completed in that year. This file is then reduced to include only occupied housing units and vacant units that are to be tabulated in the current year. That is, returns that were tabulated in the prior year, or will be tabulated in the next year, are excluded. The final screening removes returns from vacant boats because they are not included in the ACS estimation universe.

Edit Input Files

The next step is the creation of the Housing Edit Input File and the Person Edit Input File. The Housing Edit Input file is created by first merging the Final Accepted Select File with the DCF housing data. Date variables then are modified into the proper format. Next, variables are given the prefix "U," followed by the variable name to indicate they are unedited variables. Finally, answers that are "Don't Know" and "Refuse" are set as missing blank values for the edit process.

The Person Edit Input File is created by first merging the DCF person data with the codes for Hispanic origin, race, ancestry, language, place of work, and current or most recent job activity. This file then is merged with the Final Accepted Select File to create a file with all person information for all accepted HUs. As was done for the housing items, the person items are set with a "U" in front of the variable name to indicate that they are unedited variables. Next, various name flags are set to identify people with Spanish surnames and those with "non-name" first names, such as "female" or "boy." When the adjudicated number of people in an HU is greater than the number of person records, blank person records are created for them. The data for these records will be filled in during the imputation process. Finally, as with the housing variables, "Don't Know" and "Refuse" answers are set as missing blank values for the edit process. When complete, the Edit Input Files encompass the information from the DCF housing and person files but only for the unduplicated response records with data collected during the calendar year.

10.5 DATA PROCESSING

Once the Edit Input Files have been generated and verified, the edit and imputation process begins. The main steps in this process are:

- Editing and imputation.
- Generating recoded variables.
- Reviewing edit results.
- Creating input files for data products.

10.6 EDITING AND IMPUTATION

Editing

As editing and imputation begins, the data file still contains blanks and inconsistencies. When data are missing, it is standard practice to use a statistical procedure called imputation to fill in missing responses. Filling in missing data provides a complete dataset, making analysis of the data both feasible and less complex for users. Imputation can be defined as the placement of one or more estimated answers into a field of a data record that previously had no data or had incorrect or implausible data (Groves et al., 2004). Imputed items are flagged so that analysts understand the source of these data.

As mentioned, the blanks come from blanked-out invalid responses and missing data on mail questionnaires that were not corrected during FEFU, as well as from CATI and CAPI cases with answers of "Refusal" or "Don't Know." The files also include the backcoded variables for the seven questions that allow for open-ended responses. As a preliminary step, data are separated by state because the HU editing and imputation operations are completed on a state-by-state basis.

Edit and imputation rules are designed to ensure that the final edited data are as consistent and complete as possible and are ready for tabulation. The first step is to address those internally inconsistent responses not resolved during data preparation. The editing process looks at internally contradictory responses and attempts to resolve them. Examples of contradictory responses are:

- A person is reported as having been born in Puerto Rico but is not a citizen of the United States.
- A young child answers the questions on wage or salary income.
- A person under the age of 15 reports being married.
- A male responds to the fertility question (Diskin, 2007a).

Subject matter experts at the Census Bureau develop rules to handle these types of responses. The application of such edit rules help to maintain data quality when contradictory responses exist. Some edits are more complex than others. For example, joint economic edits look at the combination of multiple variables related to a person's employment, such as most recent job activity, industry, type of work, and income. This approach maximizes information that can be used to impute any economic-related missing variables. As noted by Alexander et al. (1997),

> Editing the ACS data to identify for obviously erroneous values and imputing reasonable values when data were missing involved a complex set of procedures. Demographers and economists familiar with each specific topic developed the specific procedures for different sets of data, such as marital status, education, or income. The documentation of the procedures is over 1,000 pages long, so only a very general discussion will be given here.

As Alexander et al. (1997) note, edit checks encompass range and consistency. They also provide justification for the edit rules:

> The consistency edit for fertility ('how many babies has this person ever had') deletes response from anyone identified as Male or under age 15. In setting a cutoff like this, a decision must be made based on the data about which categories have more 'false positives' than 'true positives.' The consistency edit for housing value involves a joint examination of value, property taxes, and other variables. When the combination of variables is improbable for a particular area, several variables may be modified to give a plausible combination with values as close as possible to the original.

Another edit step relates to the income components reported by respondents for the previous 12 months. Because of general price-level increases, answers from a survey taken in January 2007 are not directly comparable to those of December 2007 because the value of the dollar declined during this period. Consumer Price Index (CPI) indexes are used to adjust these income components for inflation. For example, a household interviewed in March 2007 reports their income for the preceding 12 months—March 2006 through February 2007. This reported income is adjusted to the reference year by multiplying it by the 2007 (January–December 2007) CPI and dividing by the average CPI for March 2006–2007.

Imputation

There are two principal imputation methods to deal with missing or inconsistent data—assignment and allocation. Assignment involves looking at other data, as reported by the respondent, to fill in missing responses. For example, when determining sex, if a person reports giving birth to children in the past 12 months, this would indicate that the person is female. This approach also uses data as reported by other people in the household to fill in a blank or inconsistent field. For example, if the reference person and the spouse are both citizens, a child with a blank response to citizenship is assumed also to be a citizen. Assigned values are expected to have a high probability of correctness. Assignments are tallied as part of the edit output.

Certain values, such as whether a person has served in the military, are more accurate when provided from another HU or from a person with similar characteristics. This commonly used approach of imputation is known as hot-deck allocation, which uses a statistical method to supply responses for missing or inconsistent data from responding HUs or people in the sample who are similar.

Hot-deck allocation is conducted using a hot-deck matrix that contains the data for prospective donors and is called upon when a recipient needs data because a response is inconsistent or blank. For each question or item, subject matter analysts develop detailed specification outlines

for how the hot-deck matrices for that item are to be structured in the editing system. Classification variables for an item are used to determine categories of "donors" (referred to as cells) in the hot-deck. These donors are records of other HUs or people in the ACS sample with complete and consistent data. One or more cells constitute the matrix used for allocating one or more items. For example, for the industry, occupation, and place-of-work questions, some blanks still remain after backcoding is conducted. Codes are allocated from a similar person based on other variables such as age, sex, education, and number of weeks worked. If all items are blank, they are filled in using data allocated from another case, or donor, whose responses are used to fill in the missing items for the current case, the "recipient." The allocation process is described in more detail in U.S. Census Bureau (2006a).

Some hot-deck matrices are simple and contain only one cell, while others may have thousands. For example, in editing the housing item known as tenure (which identifies whether the housing unit is owned or rented), a simple hot-deck of three cells is used, where the cells represent responses from single-family buildings, multiunit buildings, and cases where a value for the question on type of building is not reported. Alternatively, dozens of different matrices are defined with thousands of cells specified in the joint economic edit, where many factors are used to categorize donors for these cells, including sex, age, industry, occupation, hours and weeks worked, wages, and self-employment income.

Sorting variables are used to order the input data prior to processing so as to determine the best matches for hot-deck allocation. In the ACS, the variables used for this purpose are mainly geographic, such as state, county, census tract, census block, and basic street address. This sequence is used because it has been shown that housing and population characteristics are often more similar within a given geographic area. The sorting variables for place of work edit, for example, are used to combine similar people together by industry groupings, means of transportation to work, minutes to work, state of residence, county of residence, and the state in which the person works.

For each cell in the hot-deck, up to four donors (e.g., other ACS records with housing or population data) are stored at any one time. The hot-deck cells are given starting values determined in advance to be the most likely for particular categories. Known as cold-deck values, they are used as donor values only in rare instances where there are no donors. Procedures are employed to replace these starting values with actual donors from cases with similar characteristics in the current data file. This step is referred to as hot-deck warming.

The edit and imputation programs look at the housing and person variables according to a predetermined hierarchy. For this reason, each item in a response record is edited and imputed in an order delineated by this hierarchy, which includes the basic person characteristics of sex, age, and relationship, followed by most of the detailed person characteristics, and then all of the housing items. Finally, the remainder of the detailed person items, such as migration and place of work, are imputed. For HUs, the edit and imputation process is performed for each state separately, with the exception of the place of work item, which is done at the national level. For GQ facilities, the data are processed nationally by GQ type, with facilities of the same type (e.g., nursing homes, prisons) edited and imputed together.

As they do with the assignment rules, subject matter analysts determine the number of cells and the variables used for the hot-deck imputation process. This allows the edit process to apply both assignment rules to missing or inconsistent data and allocation rules as part of the edit process.

In the edit and imputation system, a flag is associated with each variable to indicate whether or not it was changed and, if so, the nature of the change. These flags support the subject matter analysts in their review of the data and provide the basis for the calculation of allocation rates. Allocation rates measure the proportion of values that required hot-deck allocation and are an important measure of data quality. The rates for all variables are provided in the quality measures section on the ACS Web site. Chapter 15 also provides more information about these quality measures.

Generating Recoded Variables

New variables are created during data processing. These recoded variables, or recodes, are calculated based on the response data. Recoding usually is done to make commonly used, complex variables user-friendly and to reduce errors that could occur when users incorrectly recode their own data. There are many recodes for both housing and person data, enabling users to understand characteristics of an area's people, employment, income, transportation, and other important categories.

Data users' ease and convenience is a primary reason to create recoded variables. For example, one recode variable is "Presence of Persons 60 and Over." While the ACS also provides more precise age ranges for all people in a given county or state, having a recoded variable that will give the number and percentages of households in a region with one or more people aged 60 or over in a household provides a useful statistic for policymakers planning for current and future social needs or interpreting social and economic characteristics to plan and analyze programs and policies (U.S. Census Bureau, 2006a).

Reviewing Edit Results

The review process involves both review of the editing process and a reasonableness review. After editing and imputation are complete, Census Bureau subject matter analysts review the resulting data files. The files contain both unedited and edited data, together with the accompanying imputation flag variables that indicate which missing, inconsistent, or incomplete items have been filled by imputation methods. Subject matter analysts first compare the unedited and edited data to see that the edit process worked as intended. The subject analysts also undertake their own analyses, looking for problems or inconsistencies in the data from their perspectives. When conducting the initial edit review, they determine whether the results make sense through a process known as a reasonableness review. If year-to-year changes do not appear to be reasonable, they institute a more comprehensive review to reexamine and resolve the issues. Allocation rates from the current year are compared with previous years to check for notable differences. A reasonableness review is done by topic, and results on unweighted data are compared across years to see if there are substantial differences. The initial reasonableness review takes place with national data, and another final review compares data from smaller geographic areas, such as counties and states (Jiles, 2007).

These processes also are carried out after weighting and swapping data (discussed in Chapter 12). Analysts also examine unusual individual cases that were changed during editing to ensure accuracy and reasonableness.

The analysts also use a number of special reports for comparisons based on the edit outputs and multiple years of survey data. These reports and data are used to help isolate problems in specifications or processing. They include detailed information on imputation rates for all data items, as well as tallies representing counts of the number of times certain programmed logic checks were executed during editing. If editing problems are discovered in the data during this review process, it is often necessary to rerun the programs and repeat the review.

Creating Input Files for Data Products

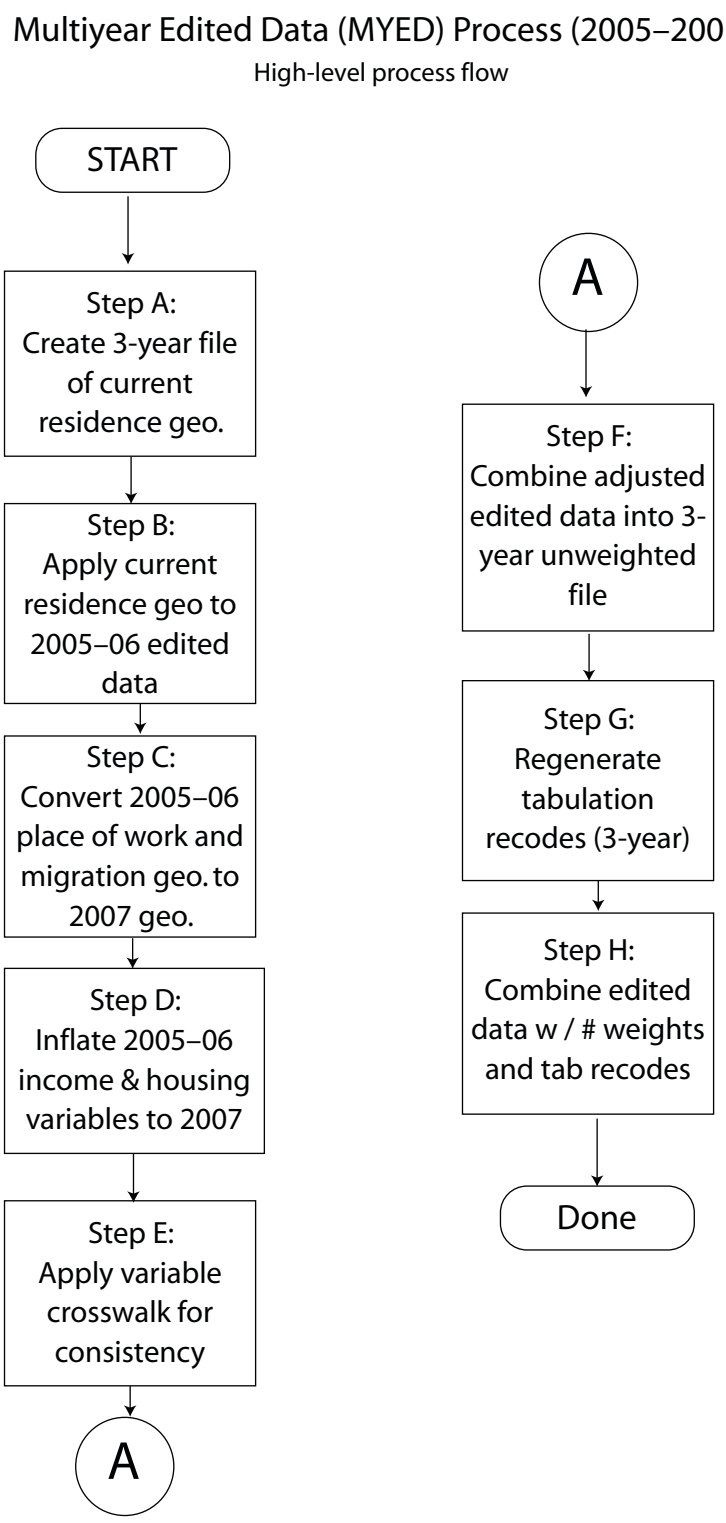
Once the subject matter analysts have approved data within the edited files, and their associated recodes, the files are ready to serve as inputs to the data products processing operation. If errors attributable to editing problems are detected during the creation of data products, it may be necessary to repeat the editing and review processes.

10.7 MULTIYEAR DATA PROCESSING

ACS multiyear estimates will be published for the first time in 2008 based on the 3-year combined file from the 2005 ACS, 2006 ACS, and 2007 ACS. To do this, multiyear edited data (or microdata) are needed as the basis for producing the 3-year ACS tabulated estimates for the multiyear period. This discussion will focus on this first 3-year tabulation period, the data collection years 2005–2007. A number of steps must be applied to the previous year's final edited data to make it consistent for multiyear processing. The first step is to update the current residence geography for

2005 and 2006 data to 2007 geography. The most involved step in the process pertains to how the vintage of geography in the "Place of Work" and "Migration" variables and recodes are updated to bring them up to the current year (2007). This step is necessary due to the fact that for the 2005 edited data for these variables and recodes would be in 2005 vintage geography, and in 2006 vintage geography for the 2006 edited data. The geocodes in these variables and recodes from prior years need to be converted in some way to current geography. This transformation was accomplished using a matching process to multiyear geographic bridge files (Boertlein, 2008) to update these variables to 2007 geography. Inflation adjustments also must be applied to monetary income and housing variables and recodes to inflate them up to a constant reference year of 2007 for the 2005–2007 edited file. Yet another step is needed to deal with variable changes across years, so that a consistent 3-year file may be created. A crosswalk table for the multiyear process attempts to map values of variables that changed across years into a consistent format. For the creation of the 2005–2007 file, only two recode variables were identified whose definition had changed over the period: Veteran's Period of Service (VPS) and Unmarried partner household (PARTNER). To make them consistent for the 3-year file, both recodes were recreated for the 2005 and 2006 data using the 2007 algorithm. When all of these modifications have been applied to the prior year's data, these data are combined with the 2007 data into an unweighted multiyear edited dataset. Tabulation recodes are then recreated from this file, and the outputs of that process joined with the 3-year weights and edited data to create the multiyear weighted and edited file. At this point the 3-year ACS edited and weighted data file will be suitable for input to the data products system. See Figure 10.14 for a flowchart showing high-level process flow.

Figure 10.14 **Multiyear Edited Data Process**



10.8 REFERENCES

Alexander, C. H., S. Dahl, and L. Weidmann. (1997). "Making Estimates From the American Community Survey." Paper presented to the Annual Meeting of the American Statistical Association (ASA), Anaheim, CA, August 1997.

Bennett, Aileen D. (2006). "Questions on Tech Paper Chapter 10." Received via e-mail, December 28, 2006.

Bennett, Claudette E. (2006). "Summary of Editing and Imputation Procedures for Hispanic Origin and Race for Census 2000." Washington, DC, December 2006.

Biemer, P., and L. Lyberg. (2003). *Introduction to Survey Quality*. Hoboken, NJ: John Wiley & Sons, Inc.

Boertlein, Celia G. (2007a). "American Community Survey Quality Assurance System for Clerical Geocoding." Received via personal e-mail, January 23, 2007.

Boertlein, Celia G. (2007b). "Geocoding of Place of Birth, Migration, and Place of Work—an Overview of ACS Operations." Received via personal e-mail, January 23, 2007.

Diskin, Barbara N. (2007a). Hand-edited review of Chapter 10. Received January 15, 2007.

Diskin, Barbara N. (2007b). Telephone interview. January 17, 2007.

Diskin, Barbara N. (2007c). "Additional data preparation questions—ACS Tech. Document," Received via e-mail January 30, 2007.

Earle, Katie. (2007). Edited review of Chapter 10. Received January 30, 2007.

Griffin, Deborah. (2006). "Question about allocation rates." Received via e-mail July 3, 2006.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, (2004). *Survey Methodology*. Hoboken, NJ: John Wiley & Sons, Inc.

Jiles, Michelle. (2007). Telephone interview. January 29, 2007.

Kirk, Mary. (2006). Telephone interview. July 12, 2006.

Powers, J. (2006). U.S. Census Bureau Memorandum, "Specification for Creating the Edit Input and Select Files, 2005." Washington, DC. Draft of 2006-10-02.

Raglin, David. (2004). "Edit Input Specification 2004." Internal U.S. Census Bureau technical specification, Washington, DC.

Tersine, A. (1998). "Item Nonresponse: 1996 American Community Survey." Paper presented to the American Community Survey Symposium, March 1998.

U.S. Census Bureau. (1997). "Documentation of the 1996 Record Selection Algorithm." Internal U.S. Census Bureau memorandum, Washington, DC.

U.S. Census Bureau. (2000). "Census 2000 Operational Plan." Washington, DC, December 2000.

U.S. Census Bureau. (2001a). "Meeting 21st Century Demographic Data Needs: Implementing the American Community Survey." Washington, DC, July 2001.

U.S. Census Bureau, Population Division, Decennial Programs Coordination Branch. (2001b). "The U.S. Census Bureau's Plans for the Census 2000 Public Use Microdata Sample Files: 2000." Washington, DC, December 2001.

U.S. Census Bureau. (2002). "Meeting 21st Century Demographic Data Needs: Implementing the American Community Survey: May 2002." Report 2, Demonstrating Survey Quality. Washington, DC.

U.S. Census Bureau. (2003a). "American Community Survey Operations Plan Release 1: March 2003." Washington, DC.

U.S. Census Bureau. 2003b. “Data Capture File 2003.” Internal U.S. Census Bureau technical specification, Washington, DC.

U.S. Census Bureau. 2003c. “Technical Documentation: Census 2000 Summary File 4.” Washington, DC.

U.S. Census Bureau. 2004a. “American Community Survey Control System Document.” Internal U.S. Census Bureau documentation, Washington, DC.

U.S. Census Bureau. 2004b. “Housing and Population Edit Specifications.” Internal U.S. Census Bureau documentation, Washington, DC.

U.S. Census Bureau. 2004c. “Housing Recodes 2004.” Internal U.S. Census Bureau data processing specification, Washington, DC.

U.S. Census Bureau. 2004d. “Hispanic and Race Edits for the 2004 American Community Survey.” Internal U.S. Census Bureau data processing specifications. Washington, DC.

U.S. Census Bureau. 2006a. “American Community Survey 2004 Subject Definitions.” Washington, DC, <www.census.gov/acs/www/Downloads/2004/usedata/Subject_Definitions.pdf>.

U.S. Census Bureau. 2006b. “Automated Geocoding Processing for the American Community Survey.” Internal U.S. Census Bureau Documentation.

U.S. Census Bureau, May 21, 2008. “Issues and Activities Related to the Migration and Place-of-Work Items in the Multi-Year Data Products.” Celia Boertlein, Kin Koerber, Journey to Work and Migration Staff, Housing and Household Economics Statistics Division.

Chapter 11.
Weighting and Estimation

11.1 OVERVIEW

Beginning in 2010, the U.S. Census Bureau plans to release three sets of American Community Survey (ACS) estimates annually for specified geographic areas, using data collected over three different periods. In general, the Census Bureau will produce and publish estimates for the same set of statistical, legal, and administrative entities as the previously published Census long form: the nation, states, American Indian and Alaska Native (AIAN) areas, counties (*municipios* in Puerto Rico), minor civil divisions (MCDs), incorporated places, and census tracts, among others (see Chapter 8.B). The Census Bureau will publish up to three sets of estimates for a geographic area depending on its total population.

- The Census Bureau plans to publish multiyear estimates based on 5 calendar years of sample data for all statistical, legal, and administrative entities, including census tracts, block groups, and small incorporated places, such as cities and towns. These 5-year estimates are based on data collected during the 60 months of the 5 most recent collection years.
- For geographic entities with populations of at least 20,000, the Census Bureau will also publish 3-year estimates based on data collected during the 36 months of the 3 most recent collection years.
- For geographic entities with populations of at least 65,000, the Census Bureau will also publish single-year estimates based on data collected during the 12 months of the most recent calendar year.

When subsequent 3- and 5-year period estimates are produced, data from the most recent year will replace data from the earliest year of the previous estimation period.

The basic estimation approach is a ratio estimation procedure that results in the assignment of two sets of weights: a weight to each sample person record, both household and group quarters (GQ) persons, and a weight to each sample housing unit (HU) record. Ratio estimation is a method that takes advantage of auxiliary information (in this case, population estimates by sex, age, race, and Hispanic origin, and estimates of total HUs) to increase the precision of the estimates as well as correct for differential coverage by geography and demographic detail. This method also produces ACS estimates consistent with the population estimates from the Population Estimates Program (PEP) of the Census Bureau by these characteristics and the estimates of total HUs for each county in the United States.

For any given tabulation area, a characteristic total is estimated by summing the weights assigned to the people, households, families, or HUs possessing the characteristic. Estimates of population characteristics are based on the person weight. Estimates of family, household, and HU characteristics are based on the HU weight. As with most household surveys, weights are used to bring the characteristics of the sample more into agreement with those of the full population by compensating for differences in sampling rates across areas, differences between the full sample and the interviewed sample, and differences between the sample and independent estimates of basic demographic characteristics (Alexander et al., 1997).

Section B describes the 2007 single-year weighting methodology for calculating person weights for the GQ sample records. This weighting for GQ persons is done independently of the weighting for HUs. Sections C, D, E, and F describe the 2007 single-year weighting methodology for calculating HU weights and person weights for the household sample records. The weighting for household persons makes use of the GQ person weights so that the household and GQ person weights

can be combined to produce estimates of the total population. While the methodology for the multiyear weighting is largely the same as the single-year weighting methodology, Section G outlines where the 2005–2007 3-year weighting methodology differs from the 2007 single-year methodology.

2007 ACS Group Quarters Person Weighting

Since the 2006 data collection year, estimates from the ACS have included data from both people living in both HUs and GQs. The weighting of GQ persons is performed in three major steps. The first step calculates the sampling base weights, which includes adjustments for subsampling that occurs at the time of interview. The second step adjusts the interviewed person records for nonresponse. The third step adjusts the person weights so that the weighted estimates conform to estimates from the PEP at the state by major GQ type group level. The basic weighting area used for the GQ weighting is the state.

Sampling Weight

The sampling of GQ persons has two phases—the initial sampling of hits and the subsampling of GQ persons associated with those hits (see Chapter 4 for more details). The initial sampling of GQ persons has a uniform sampling rate of 2.5 percent. Thus, the initial base weight (*BW*) for all GQ persons is equal to the inverse of the sampling rate, 40. This initial weight reflects the sampling probability of the sample hit and the within-GQ sampling probability of the persons if the population of the GQ is equal to the expected value given on the frame. If the observed population is different from the expected value on the frame, then the within-GQ sampling rate will be adjusted to select the same number of sample persons and the weights need to be adjusted accordingly. This adjusted base weight is called the preliminary final base weight (*PFBW*).

The adjustment of the initial base weight (*BW*) for the subsampling that occurs at the time of interview depends on whether the GQ remains in the size stratum that it was initially assigned at the time of sampling based on the new observed population.

GQs in the small size stratum (those whose expected population are 15 or fewer) that remain in the small size stratum based on their observed population will keep their original base weight of 40 since a take-all procedure is used as long as the observed population is 15 or fewer. However, if the small GQ has an observed population of 16 or more, a subsampling procedure is performed to select 10 GQ residents to interview. The base weight in this case is adjusted by the "take every x resident" necessary to select the 10 residents.

GQs in the large size stratum (those whose expected population are 16 or more) will have their base weight adjusted in all situations where the observed population differs from the expected population of the GQ. If the observed size of the large GQ is 10 or more, the base weight is adjusted by the ratio of the observed population to the expected population. If the observed size is fewer than 10 persons, then the base weight is adjusted by the fraction of 10 over the expected size. These adjustments to the initial base weight are summarized in Table 11-1.

Table 11.1 **Calculation of the Preliminary Final Base Weight (*PFBW*)**

Size stratum at time of sampling	Observed population		
	Less than 10 persons	11 to 15 persons	16 or more persons
Small stratum.....	BW	BW	BW * (Observed population) / 10
Large stratum.....	BW * 10 / (Expected population)	BW * (Observed population) / (Expected population)	BW * (Observed population) / (Expected population)

The final step in calculating the sampling weights is a weight-trimming procedure. This procedure caps all preliminary final base weights at 350 and then spreads the excess weight via a ratio adjustment to other GQ person interviews within the same state and major GQ type group. The type groups are defined in Table 11.2. The resulting weights after trimming are then defined as the final base weights (*FBW*) that include all sampling probabilities with the trimming applied.

Table 11.2 **Major GQ Type Groups**

Major GQ type group	Definition	Institutional/Noninstitutional
1	Correctional institutions	Institutional
2	Juvenile detention facilities	Institutional
3	Nursing homes	Institutional
4	Other Llng-term care facilities	Institutional
5	College dormitories	Noninstitutional
6	Military facilities	Noninstitutional
7	Other noninstitutional facilities	Noninstitutional

Calculation of the GQ NonInterview Adjustment Factor

A noninterview adjustment factor is calculated to account for the eligible GQ residents who do not complete an interview. This occurs in a single step where the noninterview adjustment cells are defined, within state, by major GQ type group by county. If a cell contains fewer than 10 interviews and has any number of noninterviews or if the noninterview factor is greater than 2, then cells are collapsed across counties within the same major GQ type group in an attempt to preserve the state by type group weighted totals. If the new collapsed cell still fails one or both of the collapsing criteria, then it is collapsed to a subset of the type groups within the same institutional/noninstitutional class as shown in Table 11.2. If needed, all cells with the same institutional/noninstitutional class are collapsed together across all type groups in the class. If further collapsing is still required, then all cells within the state are collapsed together. In practice, these last two collapsings are rarely, if ever, used. The GQ Noninterview Adjustment Factor (*GQNIF*) for each eligible cell is then calculated:

$GQNIF_i$ = Total final GQ person sample base weights of interviewed GQ persons

÷

Total final GQ person sample base weights of interviewed and noninterviewed GQ persons

$$= \frac{\sum_{j \in \text{Resp}_i} FBW_{ij} + \sum_{j \in \text{NonResp}_i} FBW_{ij}}{\sum_{j \in \text{Resp}_i} FBW_{ij}},$$

where

FBW_{ij} = Final GQ person sample base weight for the *j*th person within the *i*th adjustment cell.

All interviewed GQ persons are adjusted by this noninterview factor. All noninterviews including those persons who were found to be out-of-scope are assigned a factor of 0.0. The computation of the weight after the noninterview adjustment factor is summarized in Table 11.3.

Table 11.3 **Computation of the Weight After the GQ Noninterview Adjustment Factor (*WGQNIF*)**

Interview status	$WGQNIF_{ij}$
Interviewed	FBW_{ij} X $GQNIF_i$
Noninterviewed and out-of-scope	0

Calculation of the GQ Person Post-Stratification Factor

The third and last step in the GQ person weighting process is to apply the GQ Person Post-Stratification Factor (*GQPPSF*). In 2004, a project was undertaken to research an adequate method for applying controls in the single-year weighting of both the household and GQ persons (see Weidman et al., 2007, for more details). The goal of that research was to determine what was the best method to achieve, as a primary goal, accurate estimates for GQ characteristics at the state

level while also achieving, as a secondary goal, reasonable estimates for the total population at the county level. The research compared four alternative options for controlling GQ persons, either separately or in combination with HU persons. The results showed that it is feasible to control the GQ data at the state level by major GQ type group and combine those results with the weighting of the household population by weighting area to produce adequate estimates of the total population for all levels of aggregation. The choice of this methodology is further supported by the nature of the PEP GQ population estimates which are updated and maintained by major GQ type group.

The post-stratification cells are defined by state by major GQ type group and all sample interview persons are placed in their appropriate cells. If a cell contains fewer than 10 GQ persons or the ratio of the PEP population estimate to the ACS estimate calculated using the *WGQNIF* weight is outside of the interval 1/3.5 to 3.5, then the cell is collapsed to a subset of the type groups within the same institutional/noninstitutional class as was done for the noninterview adjustment collapsing. If the new cell fails one or both criteria, then all cells within the same institutional/ noninstitutional class are collapsed together. If further collapsing is required, then all cells within the state are collapsed together. In practice, most cells pass the criterion with either no collapsing or collapsing to a subset of the type groups within the same institutional/noninstitutional class. The GQ Person Post-Stratification Factor (*GQPPSF*) for each eligible cell is then calculated:

$GQPPSF_i$ = PEP GQ population estimate

÷

Total adjusted GQ person weight after the noninterview adjustment for all interviewed persons

$$= \frac{GQPOP_i}{\sum_{j \in \text{(interviewed)}} WGQNIF_{ij}}$$

where

$GQPOP_i$ = PEP GQ population estimate housing unit estimate for the *i*th adjustment cell.

Multiplying the *GQPPSF* by the weighting after the GQ noninterview adjustments, *WGQNIF*, results in the final unrounded GQ person weight, *WGQPPSF*. These weights are then rounded to form the final GQ person weights.

11.2 2007 ACS HOUSING UNIT WEIGHTING—OVERVIEW

Single-year weighting is implemented in three stages. In the first stage, weights are computed to account for differential selection probabilities based on the sampling rates used to select the HU sample. In the second stage, weights of responding HUs are adjusted to account for nonresponding HUs. In the third stage, weights are controlled so that the weighted estimates of HUs and persons by age, sex, race, and Hispanic origin conform to estimates from the PEP of the Census Bureau at a specific point in time. The estimation methodology is implemented by "weighting area," either a county or a group of less populous counties.

11.3 2007 ACS HOUSING UNIT WEIGHTING—PROBABILITY OF SELECTION

The first stage of weighting involves two steps. In the first step, each HU is assigned a basic sampling weight that accounts for the sampling probabilities in both the first and second phases of sample selection. Chapter 4 provides more details on the sampling. In the second step, these sampling weights are adjusted to reduce variability in the monthly weighted totals.

Sampling Weight

The first step is to compute the basic sampling weight for the HU based on the inverse of the probability of selection. This sampling weight is computed as a multiplication of the base weight (*BW*) and a computer-assisted personal interviewing (CAPI) subsampling factor (*SSF*). The *BW* for an HU is calculated as the inverse of the final overall first-phase sampling rate as given in Chapter 4, Table 4.2. HUs sent to CAPI are eligible to be subsampled (second-phase sampling) at one of the rates described in Table 4.4. Those selected for the CAPI subsample, and for which no late mail return is received in the CAPI month, are assigned a CAPI *SSF* equal to the inverse of their (second-phase) subsampling rate. Those not selected for the CAPI subsample receive a factor of 0.0. HUs for which a completed mail return is received, regardless if it was eligible for CAPI, or a computer-assisted telephone interviewing (CATI) interview is completed receive a CAPI *SSF* of 1.0. The CAPI *SSF* is then used to calculate a new weight for every HU, the weight after CAPI subsampling factor (*WSSF*). It is equal to the base weight times the CAPI subsampling factor. After each of the subsequent weighting steps, with one exception that will be noted, a new weight is calculated as the product of the new factor and the weight following the previous step. For additional details about the weighting steps discussed in this and the following section, see Asiala (2004).

Table 11.4 **Computation of the Weight After CAPI Subsampling Factor (*WSSF*)**

Weighting step	Sample disposition				
	Mail respondent	CATI respondent	CAPI sampled units	CAPI nonsampled units	CAPI eligible, but then becomes mail respondent
Base Weight (*BW*)	1 ÷ (overall sampling rate)	1 ÷ overall sampling rate)	1 ÷ (overall sampling rate)	1 ÷ (overall sampling rate)	1 ÷ (overall sampling rate)
CAPI subsampling factor (*SSF*)	1	1	1 ÷ (CAPI subsampling rate)	0	1
Weight after subsampling factor (*WSSF*) = *BW* × *SSF*	1 ÷ (overall sampling rate)	1 ÷ (overall sampling rate)	1 ÷ (overall sampling rate) × 1 ÷ (CAPI subsampling rate)	0	1 ÷ (overall sampling rate)

Note: Table summarizes computation of the *WSSF* by the weighting step and the sample dispostion.

Variation in the Monthly Sample Factor

The goal of ACS estimation is to represent the characteristics of a geographic area across the specified period. For single-year estimates, this period is 12 months, and for 3- and 5-year estimates, it is 36 and 60 months, respectively. The annual sample is allocated into 12 monthly samples. The monthly sample becomes a basis for the operations of the ACS data collection, preparation, and processing, including weighting and estimation.

The data for HUs assigned to any sample month can be collected at any time during a 3-month period. For example, the households in the January sample month can have their data collected in January, February, or March. Each HU in a sample belongs to a tabulation month (the month the interview is completed). This is either the month the processing center checked in the completed mail questionnaire or the month the interview is completed by CATI or CAPI.

Because of seasonal variations in response patterns, the number of HUs in tabulation months may vary, thereby over-representing some months and under-representing other months in the single- and multiyear estimates. For this reason, an even distribution of HU weights by month is desirable. To smooth out the total weight for all sample months, a variation in monthly response factor (*VMS*) is calculated for each month as:

VMS_i = Total sample base weights of all HUs in that sample month

÷

Total adjusted weight after CAPI subsampling factor of all HUs interviewed in that sample month

$$= \frac{\sum_{j \in Month_i} BW_{ij}}{\sum_{j \in Month_i} WSSF_{ij}}$$

where

BW_{ij} = base weight for sampled HU *j* within the *i*th month,

$WSSF_{ij}$ = adjusted HU weight after the CAPI subsampling factor for interviewed HU *j* within the *i*th month.

This adjustment factor is computed within each of the 2,005 ACS weighting areas (either a county or a group of less populous counties). The index for weighting area is suppressed in this and all other formulas for weighting adjustment factors.

Table 11.5 illustrates the computation of the *VMS* adjustment factor within a particular county. In this example, the total base weight (*BW*) for each month is 100 (as shown on line 1 of this table). The total weight (*WSSF*) across modes within each month varies from 90 to 115 (as shown on line 5). The *VMS* factors are then computed by month as the ratio of the total *BW* to the total *WSSF* (as shown on line 6).

Table 11.5 **Example of Computation of *VMS***

Line	Month				
	March	April	May	June	July
Line 1: Total base weight (BW) across released samples	100	100	100	100	100
Total weight after CAPI subsampling (*WSSF*) by mode:					
Line 2: (a) Mail	55 (Mar sample)	45 (Apr sample)	40 (May sample)	45 (Jun sample)	50 (Jul sample)
Line 3: (b) CATI	30 (Feb sample)	25 (Mar sample)	30 (Apr sample)	30 (May sample)	25 (Jun sample)
Line 4: (c) CAPI	30 (Jan sample)	25 (Feb sample)	20 (Mar sample)	25 (Apr sample)	30 (May Sample)
Line 5: Total weight *WSSF* across modes (a+b+c)	115	95	90	100	105
Line 6: *VMS* adjustment factor	100 ÷ 115	100 ÷ 95	100 ÷ 90	100 ÷ 100	100 ÷ 105

The adjusted weights after the variation of monthly response adjustment (*WVMS*) are a product of the weights after CAPI subsampling factor (*WSSF*) and the variation of monthly response factor (*VMS*). When the *VMS* factor is applied, the total *VMS* weights (*WVMS*) across all HUs tabulated in a sample month will be equal to the total base weight of all HUs selected in that month's sample. The result is that each month contributes approximately 1/12 to the total single-year estimates. In other words, the single-year estimates of ACS characteristics are a 12-month average without over- or under-representing any single month due to variation in monthly response. Analogously, each month contributes approximately 1/36 and 1/60 to the 3- and 5-year estimates, respectively.

11.4 2007 ACS HOUSING UNIT WEIGHTING—NONINTERVIEW ADJUSTMENT

The noninterview adjustment uses three factors to account for sample HUs for which an interview is not completed. During data collection, nothing new is learned about the HU or person characteristics of noninterviewed HUs, so only characteristics known at the time of sampling can be

used in adjusting for them. In other surveys and censuses, characteristics that have been shown to be related to HU response include census tract, building type (single- versus multiunit structure), and month of data collection (Weidman et al., 1995). Within counties, if a sufficient number of sample HUs were available to fill the cells of a three-way cross-classification table formed by these variables, a simultaneous adjustment for these three factors could occur. There are more than 65,000 tracts, however, so there would not be enough sample for even the two-way cross-classification of tract by month of data collection. As a result, the noninterview adjustment is carried out in two steps—one based on building type and census tract, and one based on building type and tabulation month. Once these steps are completed and the factors are applied, the sum of the weights of the interviewed HUs will equal the sum of the *VMS* weights of the interviewed plus noninterviewed HUs.

Note that vacant units and ineligible units such as deletes are excluded from the noninterview adjustment.[1] The weight corresponding to these HUs remains unchanged during this stage of the weighting process since it is assumed that all vacant units and deletes are properly identified in the field and therefore are not eligible for the noninterview adjustment. The weighting adjustment is carried out only for the occupied, temporarily occupied (those HUs which are occupied but whose occupants do not meet the ACS residency criteria), and noninterviewed HUs. After completion of the adjustment to the weights of the interviewed HUs, the noninterviewed HUs can be dropped from subsequent weighting steps; their assigned weights will be equal to 0.

The noninterview adjustment steps are applied to all HUs interviewed by any mode—mail, CATI, or CAPI. However, nearly all noninterviewed HUs belong to the CAPI sample, so characteristics of CAPI nonrespondents may be closer to those of CAPI respondents than to mail and CATI respondents. To account for this possible mode-related noninterview bias, a mode noninterview adjustment factor is computed after the two previously mentioned noninterview adjustment steps.

Calculation of the First Noninterview Adjustment Factor

In this step, all HUs are placed into adjustment cells based on the cross-classification of building type (single- versus multiunit structures) and census tract. If a cell contains fewer than 10 interviewed HUs, it is collapsed with an adjoining tract until the collapsed cell meets the minimum size of 10.[2] Cells with no noninterviews are not collapsed, regardless of size, unless they are forced to collapse with a neighboring cell that fails the size criterion. The first noninterview adjustment factor (*NIF1*) for each eligible cell is:

$NIF1_i$ = Total HU weight after variation in monthly response factor of interviewed occupied and temporarily occupied HUs and noninterviewed HUs

÷

Total HU weight after variation in monthly response factor of interviewed occupied and temporarily occupied HUs

$$= \frac{\sum_{j \in \text{Resp}_i} WVMS_{ij} + \sum_{j \in \text{NonResp}_i} WVMS_{ij}}{\sum_{j \in \text{Resp}_i} WVMS_{ij}},$$

[1] Deletes or out-of-scope addresses fall into three categories: (1) addresses of living quarters that have been demolished, condemned, or are uninhabitable because they are open to the elements; (2) addresses that do not exist; (3) addresses that identify commercial establishments, units being used permanently for storage, or living arrangements known as group quarters.

[2] Data are sorted by the weighting area, building type, and tract. Within a building type, a tract that has 10 or more responses is put in its own tract. A tract that has no nonresponses and some responses (even though the total is fewer than 10) is put in its own tract. A tract that has nonresponses and fewer than 10 responses is collapsed with the next tract. If the final tract needs to be collapsed, it is collapsed with the previous tract.

where

$WVMS_{ij}$ = Adjusted HU weight after the variation in monthly response adjustment for the *j*th HU within the *i*th adjustment cell.

All occupied and temporarily occupied interviewed HUs are adjusted by this first noninterview factor. Vacant and deleted HUs are assigned a factor of 1.0, and noninterviews are assigned a factor of 0.0. The computation of the weight after the first noninterview adjustment factor is summarized in Table 11.6.

Table 11.6 **Computation of the Weight After the First Noninterview Adjustment Factor (*WNIF1*)**

Interview status	$WNIF1_{ij}$
Occupied or temporarily occupied HU	$WVMS_{ij} \times NIF1_i$
Vacant or deleted HU	$WVMS_{ij}$
Noninterviewed HU	0

where

$WNIF1_{ij}$ = Adjusted HU weight after the first noninterview adjustment factor for the *j*th HU within the *i*th adjustment cell.

Calculation of the Second Noninterview Adjustment Factor

The next step is the second noninterview adjustment. In this step, all HUs are placed into adjustment cells based on the cross-classification of building type and tabulation month. If a cell contains fewer than 10 interviewed HUs, it is collapsed with an adjoining tabulation month until the collapsed cell has at least 10 interviewed HUs.[3] Cells with no noninterviews are not collapsed, regardless of size, unless they are forced to collapse with a neighboring cell that fails the size criterion. The second noninterview factor (*NIF2*) for each eligible cell is:

$NIF2_i$ = Total HU weight after variation in monthly response factor of interviewed occupied and temporarily occupied HUs and noninterviewed HUs

÷

Total HU weight after first noninterview factor of interviewed occupied and temporarily occupied HUs

$$= \frac{\sum_{j \in \text{Resp}_i} WVMS_{ij} + \sum_{j \in \text{NonResp}_i} WVMS_{ij}}{\sum_{j \in \text{Resp}_i} WNIF_{1\,ij}}.$$

NIF1 weights for all occupied and temporarily occupied interviewed HUs are adjusted by this second noninterview factor. Vacant and deleted HUs are given a factor of 1.0, and noninterviews are assigned a factor of 0.0. The computation of the weight after the second noninterview adjustment factor is summarized in Table 11.7.

[3]Data are sorted by the weighting area, building type, and tabulation month. Within a building type, a tabulation month that has 10 or more responses is put in its own month. A tabulation month that has no nonresponses and some responses (even though the total is fewer than 10) is put in its own month. A tabulation month that has nonresponses and fewer than 10 responses is collapsed with the next month. If the final tabulation month needs to be collapsed, it is collapsed with the previous month.

Table 11.7 **Computation of the Weight After the Second Noninterview Adjustment Factor (*WNIF2*)**

Interview status	$WNIF2_{ij}$
Occupied or temporarily occupied HU	$WNIF1_{ij} \times NIF2_i$
Vacant or deleted HU	$WNIF_{ij}$
Noninterviewed HU	0

where

$WNIF2_{ij}$ = Adjusted HU weight after the second noninterview adjustment for the *j*th HU within the *i*th adjustment cell.

Calculation of the Mode Noninterview Factor and Mode Bias Factor

One element not accounted for by the two noninterview factors above is the systematic differences that exist between characteristics of households that return census mail forms and those that do not (Weidman et al., 1995). The same element has been observed in the ACS across response modes. Virtually all noninterviews occur among the CAPI sample, and people in these HUs may have characteristics that are more similar to CAPI respondents than to mail and CATI respondents. Since the noninterview factors (*NIF1* and *NIF2*) are applied to all HUs interviewed by any mode, compensation may be needed for possible mode-related noninterview bias. The mode bias factor ensures that the total weights in the cells defined by a cross-classification of selected characteristics are the same as if the weight of noninterview HUs had been assigned only to CAPI HUs, but the factor distributes the weight across all respondents (within the cells) to reduce the effect on the variance of the resulting estimates.

The first step in the calculation of the mode bias noninterview factor (*MBF*) is to calculate an intermediate factor, referred to as the mode noninterview factor (*NIFM*). The *NIFM* is not used directly to compute an adjusted weight; instead, it is used as a factor applied to the *WVMS* weight to allow the calculation of the *MBF*. The cross-classification cells are defined for building type by tabulation month. Only HUs interviewed by CAPI and noninterviews are placed in the cells. If a cell contains fewer than 10 interviewed HUs, it is collapsed with an adjoining month. Cells with no noninterviews are never collapsed unless they are forced to collapse with a neighboring cell that fails the size criterion. The *NIFM* for a cell is:

$NIFM_i$ = Total HU weight after variation in monthly response factor of CAPI interviewed occupied and temporarily occupied HUs, and noninterviewed HUs

÷

Total HU weight after variation in monthly response factor of CAPI interviewed occupied and temporarily occupied HUs

$$= \frac{\sum_{j \in \text{CAPIresp}_i} WVMS_{ij} + \sum_{j \in \text{Nonresp}_i} WVMS_{ij}}{\sum_{j \in \text{CAPIresp}_i} WVMS_{ij}}.$$

This mode noninterview factor is assigned to all CAPI-interviewed occupied and temporarily occupied HUs. HUs for which interviews are completed by mail or CATI, vacant HUs, and deleted HUs are given a factor of 1.0. Noninterviews are given a factor of 0.0. The *NIFM* factor is used in the next step only. Note that the *NIFM* adjustment is applied to the *WVMS* weight rather than the HU weight after the first and second noninterview adjustments (*WNIF1* and *WNIF2*). The computation of the weight after the mode noninterview adjustment factor is summarized in Table 11.8.

Table 11.8 **Computation of the Weight After the Mode Noninterview Adjustment Factor (*WNIFM*)**

Interview status	$WNIFM_{ij}$
Occupied or temporarily occupied HU	$WVMS1_{ij} \times NIFM_i$
Vacant or deleted HU	$WVMS_{ij}$
Noninterviewed HU	0

where

$WNIFM_{ij}$ = Adjusted HU weight after the mode noninterview adjustment for the *j*th HU within the *i*th adjustment cell.

Next, a cross-classification table is defined for tenure (three categories: HU owned, rented, or temporarily occupied), tabulation month (12 categories), and marital status of the householder (three categories: married/widowed, single, or unit is temporarily occupied). All occupied and temporarily occupied interviewed HUs are placed in their cells. If a cell has fewer than 10 interviewed HUs, the cells with the same tenure and month are collapsed across all marital statuses. If there are still fewer than 10 interviewed HUs, the cells with the same tenure are collapsed across all months. The mode bias factor (*MBF*) for each cell is then calculated as:

MBF_i = Total weight after mode noninterview factor of interviewed occupied and temporarily occupied HUs

÷

Total weight after second noninterview adjustment factor of interviewed occupied and temporarily occupied HU

$$= \frac{\sum_{j \in Resp_i} WNIFM_{ij}}{\sum_{j \in Resp_i} WNIFM2_{ij}} .$$

All interviewed occupied and temporarily occupied HUs are adjusted by this mode bias factor, and the remaining HUs receive the factor 1.0. These adjustments are applied to the *WNIF2* weights. The computation of the weight after the mode bias factor is summarized in Table 11.9 below.

Table 11.9 **Computation of the Weight After the Mode Bias Factor (*WMBF*)**

Interview status	$WMBF_{ij}$
Occupied or temporarily occupied HU	$WNIF2_{ij} \times MBF_i$
Vacant, deleted, or noninterviewed HU	$WNIF2_{ij}$

where

$WMBF_{ij}$ = Adjusted HU weight after the mode bias factor adjustment for the *j*th HU within the *i*th adjustment cell.

11.5 2007 ACS HOUSING UNIT WEIGHTING—HOUSING UNIT AND POPULATION CONTROLS

This stage of weighting forces the ACS total HU and person weights to conform to estimates from the Census Bureau's PEP. The PEP of the Census Bureau annually produces estimates of population by sex, age, race, and Hispanic origin, and total HUs for each county in the United States as of July 1. The ACS estimates are based on a probability sample, and will vary from their true population values due to sampling and nonsampling error (see Chapters 12 and 14). In addition, we can see

from the formulas for the adjustment factors in the previous two sections that the ACS estimates also will vary based on the combination of interviewed and noninterviewed HUs in each tabulation month. As part of the process of calculating person weights for the ACS, estimates of totals by sex, age, race, and Hispanic origin are controlled to be equal to population estimates by weighting area. There are two reasons for this: (1) to reduce the variability of the ACS HU and person estimates, and (2) to reduce bias due to under-coverage of HUs and the people within them in household surveys. The bias that results from missing these HUs and people is partly corrected by using these controls (Alexander et al., 1997).

The assignment of final weights involves the calculation of three factors based on the HU and population controls. The first adjustment involves the independent HU estimates. A second and separate adjustment relies on the independent population estimates. The final adjustment is implemented to achieve consistency between the ACS estimates of occupied HUs and householders.

Models for PEP estimates of housing units and population

The Census Bureau produces estimates of total HUs for states and counties as of July 1 on an annual basis. The estimates are computed based on a model:

$$\mathbf{HU0X = HU00 + (NC0X + NM0X) - HL0X}$$

where the suffix "X" indicates the year of the HU estimates, and

HU0X = Estimated 200X HUs

HU00 = Geographically updated Census 2000 HUs

NC0X = Estimated residential construction, April 1, 2000, to July 1, 200X

NM0X = Estimated new residential mobile home placements, April 1, 2000, to July 1, 200X

HL0X = Estimated residential housing loss, April 1, 2000, to July 1, 200X.

For more detailed background on the current methodology used for the HU estimates, readers can visit <http://www.census.gov/popest/topics/methodology/> and select "Housing Unit Estimates."

The Census Bureau also produces population estimates as of July 1 on an annual basis. Those estimates are computed based on the following simplified model:

$$\mathbf{P1 = P0 + B - D + NDM + NIM + NMM},$$

where

P1 = population at the end of the period (current estimate year)

P0 = population at the beginning of the period (previous estimate year)

B = births during the period

D = deaths during the period

NDM = net domestic migration during the period

NIM = net international migration during the period

NMM = net military movement during the period.

In practice, the model is considerably more complex to leverage the best information available from multiple sources. For more detailed background on the current methodology used for the population estimates, readers can visit <http://www.census.gov/popest/topics/methodology/> and select "State and County Population Estimates."

Production of the population estimates for Puerto Rico is limited to population totals by *municipio*, and by sex-age distribution at the island level. For this reason, estimates of totals by *municipio*, sex, and age for the PRCS are controlled so as to be equal to the population estimates. Currently, there are no HU controls available for Puerto Rico.

Calculation of Housing Unit Post-Stratification Factor

Note that both HU and population estimates used as controls have a reference date of July 1 which means that the 12-month average of ACS characteristics is controlled to the population with the reference date of July 1. If person weights are controlled to the population estimates as of that date, it is logical that HUs also are controlled to those estimates to achieve a consistent relationship between the two totals.

The HU post-stratification factor is employed to adjust the estimated number of ACS HUs by weighting area to agree with the PEP estimates. For the *i*th weighting area, this factor (*HPF*) is:

HPF_i = PEP HU estimate

÷

Total adjusted HU weight after the mode bias factor of interviewed occupied, interviewed temporarily occupied, and vacant HUs

$$= \frac{HU_i}{\sum_{j \in (\text{interviewed and vacant})} WMBF_{ij}}$$

where

HU_i = PEP housing unit estimate for the *i*th weighting area.

The denominator of the *HPF* formula aggregates the adjusted HU weight after the mode bias factor adjustment (*WMBF*) across 12 months for the interviewed occupied, interviewed temporarily occupied, and vacant HUs. All HUs then are adjusted by this HU post-stratification factor. Therefore, $WHPF = WMBF \times HPF$, where *WHPF* is the adjusted HU weight after the HU post-stratification factor adjustment.

Calculation of Person Weights

The next step in the weighting process is to assign weights to persons via a three-dimensional raking-ratio estimation procedure. This is done so that (1) the combined estimates of spouses and unmarried partners conform to the combined estimate of married-couple and unmarried-partner households; (2) the estimate of householders conforms to the estimate of occupied housing units; and (3) the estimates for certain demographic groups are equal to their population estimates.

Each person in an interviewed occupied HU is assigned an initial person weight equal to the HU weight after the HU post-stratification factor is applied (*WHPF*). Next, there are three steps of ratio adjustment. The first step uses three cells to classify persons by spousal or unmarried partner relationship to the householder. The second step uses two cells to classify persons by householder and nonhouseholder. The third step uses up to 156 cells defined by race/Hispanic origin, sex, and age. The steps are defined as follows:

Step 1: Spouses and Unmarried Partners. All persons are placed into one of three cells:

1. Persons who are the primary person in a two-partner relationship—all householders in a married-couple or unmarried-partner household.
2. Persons who are the secondary person in a two-partner relationship—all spouses or unmarried partners in those same households.
3. Balance of population—all persons not fitting into the first two cells.

The marginals for the first two cells are both equal to the estimate of married-couple plus unmarried-partner households using the *WHPF* weight. The marginal for the third cell is equal to the PEP total population estimate minus the sum of the marginals used for the other two cells. In this manner, the estimate of total population is controlled to the PEP total population estimate.

Step 2: Householders. The second step assigns all persons to one of two cells:

1. Householders
2. Nonhouseholders

The marginal for householders is the estimate of occupied HUs using the *WHPF* weight. The marginal for nonhouseholders is equal to the PEP total population estimate minus the marginal used for the first cell in order to control for total population.

Step 3: Race-Hispanic Origin/Sex/Age. The third step assigns all persons to one of up to 156 cells: six classifications of race-Hispanic origin by sex by 13 age groups. The marginals for these rows at the weighting area level come from the PEP population estimates. Some weighting areas will not have sufficient sample to support all 156 cells and in these cases some collapsing is necessary. This collapsing is done prior to the raking and remains fixed for all iterations of the raking.

Race and Hispanic origin are combined to define six unique race-ethnicity groups consistent with those used in weighting the Census 2000 long form. These groups are created by crossing "Non-Hispanic" with the five major single race groups, plus the group of all Hispanics regardless of race. The race-ethnicity groups are:

1. Non-Hispanic White
2. Non-Hispanic Black
3. Non-Hispanic American Indian and Alaska Native
4. Non-Hispanic Asian
5. Non-Hispanic Native Hawaiian or Pacific Islander
6. Hispanic

The assignment of a single major race to a person can be complicated because people can identify themselves as being of multiple races. People responding either with multiple races or "Other Race" are included in one of the six race-ethnicity groups for estimation purposes only. Subsequent ACS tabulations are based on the full set of responses to the race question.

Initial estimates of population totals are obtained from the ACS sample for each of the weighting race-ethnicity groups. These estimates are calculated based on the initial person weight of *WHPF*. Estimates from the Census Bureau's PEP also are available for each weighting race-ethnicity group. These total population estimates are used to control ACS total population estimates to be equal to the PEP by weighting area.

The initial sample and population estimates for each weighting race-ethnicity group are tested against a set of criteria that require a minimum of 10 sample people and a ratio of the population control to the initial sample estimate that is between (1/3.5) and 3.5. This is done to reduce the effect of large weights on the variance of the estimates. If there are weighting race-ethnicity groups that do not satisfy these requirements, they are collapsed until all groups satisfy the collapsing criteria. Collapsing decisions are made following a specified order in the following way (see Asiala, 2007, for further details):

1. If the requirements are not met when all non-Hispanic race groups are combined, then all weighting race-ethnicity groups are collapsed together and the collapsing is complete.
2. If the requirements are not met for Hispanics, the Hispanics are collapsed with the largest non-Hispanic non-White group.
3. If the requirements are not met for any non-Hispanic non-White group, it is collapsed with the largest (prior to collapsing) non-Hispanic non-White group.
4. If the largest collapsed non-Hispanic non-White group still does not meet the requirements, it is collapsed with the surviving non-Hispanic non-White groups in the following order until the requirements are met: Black, American Indian and Alaska Native, Asian, and Native Hawaiian or Pacific Islander.

5. If all non-Hispanic non-White groups have been collapsed together and the collapsed group still does not meet the requirements, it is collapsed with the non-Hispanic White group.
6. If the requirements are not met for the non-Hispanic White group, then it is collapsed with the largest non-Hispanic non-White group.

Within each collapsed weighting race-ethnicity group, the persons are placed in sex-age cells formed by crossing sex by the following 13 age categories: 0–4, 5–14, 15–17, 18–19, 20–24, 25–29, 30–34, 35–44, 45–49, 50–54, 55–64, 65–74, and 75+ years. If necessary, these cells also are collapsed to meet the requirements of the same sample size and a ratio between (1/3.5) and 3.5. The goals of the collapsing scheme are to keep children age 0–17 together whenever possible by first collapsing across sex within the first three age categories. In addition, the collapsing rules keep men age 18–54, women age 18–54, and seniors 55+ in separate groups by collapsing across age.

The initial sample cell estimates are then scaled and rescaled via iterative proportional fitting, or raking, so that the sum in each row or column consecutively agrees with the row or column household estimate (Steps 1 & 2) or population estimate (Step 3). This procedure is iterated a fixed number of times, and final person weights are assigned by applying an adjustment factor to the initial weights.

The scaling and rescaling between rows and columns is referred to as an iteration of raking. An iteration of raking consists of the following three steps. (The weighting matrix is included to facilitate the discussion below.) The three-step process has been split out into two tables, Table 11.10 and Table 11.11, for clarity.

Table 11.10 **Steps 1 and 2 of the Weighting Matrix**

		Step 2		Step 1 Control
		Householder	Nonhouseholder	
Step 1	Householder in two-partner relationship			Survey estimate of married-couple and unmarried-partner households
	Spouse/unmarried partner in two-partner relationship			Survey estimate of married-couple and unmarried-partner households
	Balance of population			PEP total population estimate minus the sum of the two controls above
Step 2 Control		Survey estimate of occupied housing units	PEP total population estimate minus the control for householders	

Table 11.11 **Steps 2 and 3 of the Weighting Matrix**

			Step 2		Step 3 Control
			Householder	Nonhouseholder	
Step 3	Non-Hispanic White	0–4 Males			PEP population estimate for the collapsed cell by weighting area
		0–4 Females			
		…			
		75+ Females			
	Non-Hispanic AIAN	…			
	Non-Hispanic Asian	…			
	Non-Hispanic NHPI	…			
	Hispanic	…			
	Step 2 Control		Survey estimate of occupied housing units	PEP total population estimate minus the control for householders	

Step 1. At this step, the initial person weights are adjusted to make both the sum of the weights of householders in married-couple or unmarried-partner households and the sum of the weights

of their spouses or unmarried partners equal to the survey estimate of married-couple and unmarried-partner households. This is done using the HU weight after the HU post-stratification factor adjustment. The weights of all other persons are adjusted to make the sum of all weights equal to the PEP total population estimate.

Step 2. The Step 1 adjusted person weights are adjusted again to make the sum of the weights of all householders equal to the survey estimate of occupied HUs using the HU weight after the HU post-stratification factor adjustment. The Step 1 adjusted weights of all other persons are adjusted to make the sum of all weights equal to the total population estimate.

Step 3. The Step 2 adjusted person weights are adjusted a third time by the ratio of the population estimates of race-Hispanic origin/age/sex groups to the sum of the Step 2 weights for sample people in each of the demographic groups described previously.

The three steps of ratio adjustment are repeated in the order given above until the predefined stopping criterion is met. The stopping criterion is a function of the difference between Step 2 and Step 3 weights. The weights obtained from Step 3 of the final iteration are the final person weights.

A single factor, the person post-stratification factor (*PPSF*), is calculated at the person level, which captures the entire adjustment accomplished by the ratio-raking estimation. It is calculated as follows:

PPSF = final person weight ÷ initial person weight.

The factor is calculated and applied to each person, so that their weights become the product of their initial weights and the factor.

ACS single-year estimates are produced for geographic areas with populations of at least 65,000, including incorporated places, for which population estimates also are published annually. Since population controls are applied at the weighting area level, occasionally the ACS estimate of total population for a large place within a weighting area may be far enough from its population estimate to cause confusion among data users. To avoid these anomalies, methodologies are being investigated to control person weights to total population for places with populations of at least 65,000 within weighting areas.

Calculation of Final Housing Unit Factors

Prior to the calculation of person weights, each HU has a single weight which is independent of the characteristics of the persons residing in the HU. After the calculation of person weights, a new HU weight is computed by taking into account the characteristics of the householder in the HU. In each interviewed occupied HU, the householder defined as the reference person (one of the persons who rents or owns the HU) is identified. Adjustment of the HU weight to account for the householder characteristics is done by assigning a householder factor (*HHF*) for an HU equal to the person post-stratification factor (*PPSF*) of the householder.[4] Their *PPSF*s give an indication of under-coverage for households whose householders have the same demographic characteristics. The *HHF* adjustment uses this information to adjust for the resultant bias. Vacant HUs are given an *HHF* of 1.0 because they have no householders.

The adjusted HU weight accounting for householder characteristics is computed as a multiplication of the adjusted HU weight after the HU post-stratification factor adjustment (*WHPF*) with the householder factor (*HHF*). Therefore, $WHHF = WHPF \times HHF$, where *WHHF* is the adjusted HU weight after the householder factor adjustment. The HU weight after the householder factor adjustment becomes the final HU weight.

The ACS weighting procedure results in two separate sets of weights, one for HUs and one for persons residing within HUs. However, since the housing unit weight is equal to the person weight of the householder, the survey will produce logically consistent estimates of occupied housing units,

[4]In the calculation of person weights, the *PPSF* is used to adjust person weight so that the ACS population estimates conform to PEP estimates by demographic characteristics.

households, and householders. With this weighting procedure, the survey estimate of total housing units will differ slightly from the PEP total housing unit estimates. The difference between the ACS estimate the PEP estimate nationally, however, was less than 5,000 in 2006.

11.6 MULTIYEAR ESTIMATION METHODOLOGY

The multiyear estimation methodology involves reweighting the data for each sample address in the 3- or 5-year period and is not just a simple average of the single-year estimates. The weighting methodology for the multiyear estimation is very similar to the methodology used for the single-year weighting. Thus, only the differences between the single- and multiyear weighting are described in this section.

Pooling the data

The data for all sample addresses over the multiyear period are pooled together into one file. The single-year base weights are then adjusted by the reciprocal of the number of years in the period so that each year contributes its proportional share to the multiyear estimates. For example, for the 2005–2007 3-year weighting, the base weights are all divided by three.

The interview month assigned to each address is also recoded so that all the data from the entire period appears as though it came from a 1-year period. For example, in the 2005–2007 3-year weighting, all addresses that were originally assigned an interview month of January 2005, 2006, or 2007 are assigned the common interview month of January. Thus, when the weighting is performed, those records will all be treated as though they come from the same month for the *VMS*, *NIF2*, *NIFM*, and *MBF* adjustments. By pooling the records across years in this manner, the noninterview adjustments, in particular, require less collapsing because of the larger sample in each cell. This, in turn, should better preserve the seasonal trends that may be present in the population as captured by the ACS.

Geography

The geography for all sample addresses in the period are updated into the common geography of the final year. This allows the tabulation of the data to be in a consistent, constant geography that is the most recent and likely most relevant to data users. When tabulating estimates for an area, all interviews from the period that are considered to be inside the boundaries of that area in the final year of the period will be included in the estimates regardless if they were considered to be inside the boundaries for that area at the time of interview. As a by-product of this methodology, the ACS is also able to publish multiyear estimates for newly created places or counties that did not exist when the interviews for the addresses in that place or county were collected.

Derivation of the multiyear controls

Since the multiyear estimate is an estimate for the period, the controls are not those of a particular year but rather they are the average of the annual independent population estimates over the period. The PEP refreshes their entire time series of estimates going back to the previous census each year using the most current data and methodology. Each of these time series are considered a "vintage." In order for the ACS to make use of the best available population estimates as controls, the multiyear weighting uses the population estimates of the most recent vintage for all years in the period in order to derive the multiyear controls.

These derived estimates are created for the HU, GQs population, and total population for use as controls in the multiyear weighting. The derived county-level HU estimates are the simple average across all years in the period. Since the average is typically not an integer, the result is rounded to the final integerized estimate. Likewise, the derived GQ population estimates for state by major type group are the simple average across all years in the period. Those averages are then control rounded so that the rounded state average estimate is within one of the unrounded estimate. Finally, the derived total population estimates by race, ethnicity, age, and sex are averaged across all years in the period and control rounded to form the final derived estimates. This is done prior to the collapsing of the estimates into the 156 cells per weighting area needed for the demographic dimension of the household person weighting as described in the single-year person weighting section.

Model-assisted estimation

Once the data are pooled and put into the geography of the final year, they are weighted using the single-year weighting methodology through the *MBF* adjustment. It is after this adjustment that the only weighting step specific to the multiyear weighting methodology is implemented, the model-assisted estimation procedure. An earlier research project (Starsinic, 2005) compared the variances of ACS tract-level estimates formed from the 1999–2001 ACS to the variances of the Census 2000 long-form estimates. The results of that research showed that the variances of the ACS tract-level estimates were higher in relation to the long form than what we expected based on sample size alone. The primary source of that increased variance was attributed to the lack of ACS subcounty controls at the tract-level or lower as was used for the long form.

Several options were explored on how the ACS estimates of variance for subcounty estimates might be improved. One option considered was to use the ACS sampling frame counts as subcounty controls. Other options explored ways to create subcounty population controls, including tract-level population controls. The final approach, and the one that was chosen, introduces a model-assisted estimation step into the multiyear weighting that makes use of both the sampling frame counts and administrative records to reduce the level of variance in the subcounty estimates (Fay, 2006). An important feature of the model-assisted estimation procedure is that the administrative record data is not used directly to produce ACS estimates. The administrative record data are only used to help reduce the level of variance. The published ACS estimates are still formed from weighted totals of the ACS survey data.

The entire model-assisted estimation process is summarized in these steps:

1. Create frame counts for places and Minor Civil Divisions (MCDs) that contain at least 10,000 in population and at least 300 HU addresses (the 5-year estimation will use tracts simply satisfying the latter criterion).
2. Link the administrative records to the ACS sampling frame (the Master Address File [MAF]) and drop administrative records that cannot be linked.
3. Form unweighted place- and MCD-level totals (tract-level for the 5-year estimates) of the linked administrative record characteristics.
4. Apply the *WMBF* weights at the HU level to the linked administrative records that fall into the ACS sample. The weighted estimates at this step represent (essentially) unbiased estimates of the unweighted totals in Step 2.
5. Using generalized regression estimation, fit a model to calibrate the ACS weights so that the weighted totals from the linked ACS records match the unweighted totals from Step 2 and so that the weighted ACS estimate of HUs match the frame totals in Step 1. The categories of the variables considered in the regression are collapsed or removed as necessary to fit a good model.
6. Proceed with the remaining steps of the ACS weighting starting with the *Housing Unit Poststratification (HPF)* Factor adjustments, including the person weighting using the derived multiyear controls as described in the preceding section.

Frame Counts: The *BWs*, which reflect the sampling probabilities of selection, should sum to the count of records on the sampling frame at the county and, generally, the subcounty level. However, after the noninterview adjustments the weighted subcounty distribution of the interviewed sample cases can deviate from the original frame distribution. This can impact both the subcounty estimates and the variances on those estimates. The use of frame counts as subcounty controls reestablishes the original distribution of HU addresses on the frame in the weighted sample. For the 3-year weighting, these frame counts are calculated at the place- or MCD-level. If the place or MCD has a PEP population estimate of 10,000 or more then the ACS weights are controlled to those frame counts at that subcounty level. For the 5-year weighting, these frame counts will be computed for tracts. This control to the frame counts is the simplest model and is used if a model with administrative record data cannot be estimated. Otherwise, it is one part of the entire calibration performed in this step.

Link Administrative Records to Frame: The administrative record data used for this step is created from linking two primary files maintained by the Data Integration Division at the Census Bureau. The first file includes person characteristics and has been created from a combination of social security and census information. The second file uses administrative records to identify all possible addresses of the persons on the first file. A merged file is then created which contains only the age, sex, race, and Hispanic origin of each person and an identifier that links that person to the best address available in the MAF via a Master Address File ID (MAFID). No other characteristics or publicly identifiable information are present on the file. This file is updated annually to account for new births, death information, and for updated address information.

Administrative Universe Counts: For each MAFID, it is possible to create household demographic totals of people by age/sex and race/ethnicity from the merged administrative records for each address that is matched to the MAF. The age/sex totals are calculated within seven categories:

1. All persons age 0–17
2. All persons age 18–29
3. Males age 30–44
4. Females age 30–44
5. Males age 45–64
6. Females age 45–64
7. All persons age 65 and older

The race/ethnicity totals are calculated within four categories:

1. All Hispanics regardless of race
2. All non-Hispanic Blacks
3. All non-Hispanic Whites
4. All non-Hispanics other races

These household-level totals can then be used to create unweighted place- and MCD-level administrative record universe totals using the geography associated with the address.

Weighted Administrative Sample Counts: The administrative records that match to the sampling frame can also be linked to the actual ACS sample records themselves. Using the *WMBF* weights, the records that match to the ACS sample can then be used to create weighted administrative record totals for the same geographic areas. Since the ACS sample weights should reflect the frame counts, these weighted administrative record totals should be an unbiased estimate of the unweighted universe totals.

Applying GREG Estimation: Using generalized regression estimation (or GREG), the ACS weights are first calibrated so that the weighted administrative record totals match the unweighted universe counts for the seven age/sex categories. Two conditions are checked: is the regression equation solvable and are all of the resulting weights greater than 0.5? If either condition fails then the age/sex categories are collapsed and the regression is attempted again. Two levels of collapsing are attempted:

1. Collapsing across age/sex categories into three categories: all persons age 0–17, all persons age 18–44, and all persons 45 and older.
2. Collapsing all categories into a single cell of total administrative persons.

The alternative: do not make use of the administrative record data.

If the regression passes using at least the single cell of total administrative persons, then an attempt is made to add the race/ethnicity covariates to the model. First, a collapsing procedure is run that tests which race/ethnicity categories can be used. The criteria for including a

race/ethnicity category in the regression is that both the administrative records universe count for the category being tested and the total for all other categories must be greater than 300 persons. This procedure is carried out first for the largest race/ethnicity category not including the non-Hispanic White category, then the next largest such category, and finally the last remaining category other than non-Hispanic White.

As an example, if the largest category other than non-Hispanic White was the Hispanic category, then the first test would be if (1) the Hispanic category had a universe count which was greater than 300 and (2) the other three categories combined had a universe count greater than 300. If it passes, the Hispanic category is flagged for inclusion and the remaining categories are tested. If the next largest category is non-Hispanic Black, it is tested to determine if its universe count is greater than 300 and if the balance, now only the non-Hispanic other races and non-Hispanic White, is greater than 300. If it passes, then the procedure moves on to test the smallest category other than non-Hispanic White. In this example, that is the non-Hispanic other race category. If a similar test on that category fails (or on any previous attempt) then the race collapsing is complete and the covariates for each race/ethnicity category that passed are added to the model. The regression is then attempted including both the age/sex and race/ethnicity covariates. The same conditions used in the age/sex category collapsing are applied to the new attempt. If the regression passes both conditions then the covariate matrix is considered final. If the regression fails either condition, then the smallest race/ethnicity category is not included in the model and the regression is attempted again. This process continues until either the regression passes or all race/ethnicity covariates have been removed.

Apply the GREG Weighting Factor (GWTF): The final result of this step is the creation of the GWTF for each ACS record, which captures the calibration performed in the regression. A summary of the impact of the GWTF is given in Table 11.12.

Table 11.12 **Impact of GREG Weighting Factor Adjustment**

Interview status	and the ACS record is:	Impact of *GWTF*
▪Noninterview ▪CAPI nonsampled	Not applicable	No impact (factor set to 1)
▪Interview (occupied or vacant) ▪Field determined ineligible	In and out-of-scope place/MCD that has either insufficient population or frame counts	No impact (factor set to 1)
	In an in-scope place/MCD but does not match to administrative data or the model using administrative data fails	Adjusts weights to calibrate to frame counts for the area
	In an in-scope place/MCD, matches to the administrative data and the model using administrative data passes	Adjusts weights to calibrate to frame counts and calibrate weighted administrative data to administrative universe counts

This factor is then applied to the WMBF weights to create the weight after the GREG Weighting Factor (WGWTF). The computation of this weight is summarized in Table 11.13.

Table 11.13 **Computation of the Weight After the GREG Weighting Factor**

Interview status	$WGWTF_j$
Interview or field determined ineligible housing unit	$WMBF_j \times GWTF_j$
All others	0

After this step is complete, the multiyear weighting mirrors the single-year weighting, picking up again at the *HPF* step.

Other multiyear estimation steps

In addition to the adjustments to the single-year weighting methodology for weighting the multi-year data, there are other steps involved in the multiyear estimation that are not weighting related. These include standardizing definitions of variables, updating the geography for place of work and migration characteristics, and the adjustment of income, value, and other dollar amounts for inflation over the period. The details of these adjustments are given in Chapter 10.

11.7 REFERENCES

Alexander, C., S. Dahl, and L. Weidman. (1997). "Making Estimates From the American Community Survey." *JSM Proceedings, Social Statistics Section*, Alexandria, VA: American Statistical Association, pp. 88–97. <http://www.census.gov/acs/www/AdvMeth/Papers/ACS/Paper9.htm>.

Asiala, M. (2007). "Specifications for Weighting the ACS 2006 HU Sample (ACS06-W-5)." 2006 American Community Survey Weighting Memorandum Series #ACS06-W-5, June 27, 2007 Draft Census Bureau Memorandum to S. Schechter Bortner from D. Whitford, Washington, DC: U.S. Census Bureau.

Fay, R. (2006). "Using Administrative Records With Model-Assisted Estimation for the American Community Survey." *JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 2995–3001.

Starsinic, M. (2005). "American Community Survey: Improving Reliability for Small Area Estimates." *JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 3592–3599.

U.S. Census Bureau (2007). Methodology: Housing Unit Estimates (2006), Washington, DC: U.S. Census Bureau, <http://www.census.gov/popest/topics/methodology/>.

U.S. Census Bureau (2008). "Methodology: State and Count Population Estimates (2007)," Washington, DC: U.S. Census Bureau. <http://www.census.gov/popest/topics/methodology/>.

Weidman, L., C. Alexander, G. Diffendahl, and S. Love. (1995). "Estimation Issues for the Continuous Measurement Survey." *JSM Proceedings, Survey Research Methods Section*, Alexandria, VA: American Statistical Association, pp. 596–601. <http://www.census.gov/acs/www/AdvMeth/Papers/Acs/Paper5.htm>.

Weidman, L., M. Ikeda, and J. Tsay. (2007). "Comparison of Alternatives for Controlling Group Quarters Person Estimates in the American Community Survey," Statistical Research Division Research Series RRS2007-4, Washington, DC: U.S. Census Bureau.

Chapter 12.
Variance Estimation

12.1 OVERVIEW

Sampling error is the difference between an estimate based on a sample and the corresponding value that would be obtained if the estimate were based on the entire population (as from a census). Note that sample-based estimates will vary depending on the particular sample selected from the population. Measures of the magnitude of sampling error, such as the variance and the standard error (the square root of the variance), reflect the variation in the estimates over all possible samples that could have been selected from the population using the same sampling methodology.

The American Community Survey (ACS) is committed to providing its users with measures of sampling error along with each published estimate. To accomplish this, all published ACS estimates are accompanied either by 90 percent margins of error or confidence intervals both based on ACS direct variance estimates. Due to the complexity of the sampling design and the weighting adjustments performed on the ACS sample, unbiased design-based variance estimators do not exist. As a consequence, the direct variance estimates are computed using a replication method that repeats the estimation procedures independently several times. The variance of the full sample is then estimated by using the variability across the resulting replicate estimates. Although the variance estimates calculated using this procedure are not completely unbiased, the current method produces variances that are accurate enough for analysis of the ACS data.

For Public Use Microdata Sample (PUMS) data users, replicate weights are provided to approximate standard errors for the PUMS-tabulated estimates. Design factors are also provided with the PUMS data, so PUMS data users can compute standard errors of their statistics using either the replication method or the design factor method.

12.2 VARIANCE ESTIMATION FOR ACS HOUSING UNIT AND PERSON ESTIMATES

Unbiased estimates of the variance do not exist because of the systematic sample design, as well as the ratio adjustments used in estimation. As an alternative, ACS implements a replication method for variance estimation. An advantage of this method is that the variance estimates can be computed without consideration of the form of the statistics or the complexity of the sampling or weighting procedures, such as those being used by the ACS.

The ACS employs the same replication method for variance estimates as was used in all of its testing phases—the Successive Differences Replication (SDR) method (Wolter, 1984; Fay and Train, 1995; and Judkins, 1990). The SDR was designed to be used with systematic samples for which the sort order of the sample is informative, as in the case of the ACS's geographic sort. Applications of this method were developed to produce estimates of variances for the Current Population Survey (CPS) (U.S. Census Bureau, 2002) and Census 2000 Long Form estimates (Gbur and Fairchild, 2002).

In the SDR method, the first step in creating a replicate estimate is constructing the replicate factors, from which the replicate weights are calculated by multiplying the base weight for each housing unit (HU) by the replicate factor. The weighting process then is rerun to create a new set of replicate weights. Given these replicate weights, replicate estimates are created by using the same estimation method as the original estimate, but applying each set of replicate weights instead of the original weights. Finally, the replicate and original estimates are used to compute the variance estimate based on the variability between the replicate estimates and the full sample estimate measured across the replicates.

The following steps produce the ACS direct variance estimates:

1. Compute replicate factors.
2. Compute replicate weights.
3. Compute variance estimates.

Replicate Factors

Computation of replicate factors begins with the selection of a Hadamard matrix of order R (a multiple of 4), where R is the number of replicates. A Hadamard matrix **H** is a k-by-k matrix with all entries either 1 or −1, such that $\mathbf{H'H} = k\mathbf{I}$ (that is, the columns are orthogonal). For ACS, the number of replicates is 80 ($R = 80$). Each of the 80 columns represents one replicate.

Next, a pair of rows in the Hadamard matrix is assigned to each record (HU or group quarters (GQ) person). An algorithm is used to assign two rows of an 80×80 Hadamard matrix to each HU. The ACS uses a repeating sequence of 780 pairs of rows in the Hadamard matrix assigned to each record, in sort order (Navarro, 2001a). The assignment of Hadamard matrix rows repeats every 780 records until all records receive a pair of rows from the Hadamard matrix. The first row of the matrix, in which every cell is always equal to one, is not used.

The replicate factor for each record then is determined from these two rows of the 80×80 Hadamard matrix. For record i ($i = 1, \ldots, n$, where n is sample size) and replicate r ($r = 1, \ldots, 80$), the replicate factor is computed as:

$$f_{i,r} = 1 + 2^{-1.5} a_{R1,r} - 2^{-1.5} a_{R2,r}$$

where $R1_i$ and $R2_i$ are respectively the first and second row of the Hadamard matrix assigned to the i-th HU, and $a_{R1_i,r}$ and $a_{R2_i,r}$ are respectively the matrix elements (either 1 or −1) from the Hadamard matrix in rows $R1_i$ and $R2_i$ and column r. Note that the formula for $f_{i,r}$ yields replicate factors that can take one of three approximate values: 1.7, 1.0, or 0.3. That is;

- If $a_{R1_i,r} = +1$ and $a_{R2_i,r} = +1$, the replicate factor is 1.
- If $a_{R1_i,r} = -1$ and $a_{R2_i,r} = -1$, the replicate factor is 1.
- If $a_{R1_i,r} = +1$ and $a_{R2_i,r} = -1$, the replicate factor is approximately 1.7.
- If $a_{R1_i,r} = -1$ and $a_{R2_i,r} = +1$, the replicate factor is approximately 0.3.

The expectation is that 50 percent of replicate factors will be 1, and the other 50 percent will be evenly split between 1.7 and 0.3 (Gunlicks, 1996).

The following example demonstrates the computation of replicate factors for a sample of size five, using a Hadamard matrix of order four:

$$H = \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix}$$

Table 12.1 presents an example of a two-row assignment developed from this matrix, and the values of replicate factors for each sample unit.

Table 12.1 **Example of Two-Row Assignment, Hadamard Matrix Elements, and Replicate Factors**

Case # (i)	Row assignment		Hadamard matrix element								Approximate replicate factor			
			Replicate 1		Replicate 2		Replicate 3		Replicate 4					
	$R1_i$	$R2_i$	$a_{R1i,1}$	$a_{R2i,1}$	$a_{R1i,2}$	$a_{R2i,2}$	$a_{R1i,3}$	$a_{R2i,3}$	$a_{R1i,4}$	$a_{R2i,4}$	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$
1	2	3	+1	+1	−1	+1	+1	−1	−1	−1	1	0.3	1.7	1
2	3	4	+1	+1	+1	−1	−1	−1	−1	+1	1	1.7	1	0.3
3	4	2	+1	+1	−1	−1	−1	+1	+1	−1	1	1	0.3	1.7
4	2	3	+1	+1	−1	+1	+1	−1	−1	−1	1	0.3	1.7	1
5	3	4	+1	+1	+1	−1	−1	−1	−1	+1	1	1.7	1	0.3

Note that row 1 is not used. For the third case ($i = 3$), rows four and two of the Hadamard matrix are to calculate the replicate factors. For the second replicate ($r = 2$), the replicate factor is computed using the values in the second column of rows four (−1) and two (−1) as follows:

$$f_{3,2} = 1 + 2^{-1.5} a_{4,2} - 2^{-1.5} a_{2,2} = 1 + (2^{-1.5} \times -1) - (2^{-1.5} \times -1) = 1$$

Replicate Weights

Replicate weights are produced in a way similar to that used to produce full sample final weights. All of the weighting adjustment processes performed on the full sample final survey weights (such as applying noninterview adjustments and population controls) also are carried out for each replicate weight. However, collapsing patterns are retained from the full sample weighting and are not determined again for each set of replicate weights.

Before applying the weighting steps explained in Chapter 11, the set of replicate sampling weights is computed. With the replicate factor assigned, the replicate sampling weight for replicate r is computed by multiplying the full sample weight after computer-assisted personal interviewing (CAPI) subsampling factor (*WSSF*— see Chapter 11 for the computation of this weight) by the replicate factor $f_{i,r}$; that is, $RWSSF_{i,r} = WSST_i \times f_{i,r}$, where $RWSSF_{i,r}$ is the replicate weight after CAPI subsampling factor for the i-th HU and the r-th replicate ($r = 1, \ldots, 80$).

One can elaborate on the previous example of the replicate construction using five cases and four replicates: Suppose the full sample *WSSF* values are given under the second column of the following table (Table 12.2). Then, the replicate weight after CAPI subsampling factor (*RWSSF*) values are given in columns 7–10.

Table 12.2 **Example of Computation of Replicate Weight After CAPI Subsampling Factor (*RWSSF*)**

Case #	$WSSF_i$	Approximate replicate factor				Replicate weight after CAPI subsampling factor			
		$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	$RWSSF_{i,1}$	$RWSSF_{i,2}$	$RWSSF_{i,3}$	$RWSSF_{i,4}$
1	100	1	0.3	1.7	1	100	29	171	100
2	120	1	1.7	1	0.3	120	205	120	35
3	80	1	1	0.3	1.7	80	80	23	137
4	120	1	0.3	1.7	1	120	35	205	120
5	110	1	1.7	1	0.3	110	188	110	32

The rest of the weighting process (Chapter 11) then is applied to each replicate weight $RWSSF_{i,r}$ (starting from the adjustment for variation in monthly response (*VMS*) and proceeding to the population control adjustment or raking). Basically, the weighting adjustment process is repeated independently 80 times and the $RWSSF_{i,r}$ is used in place of $WSSF_i$ (as in Chapter 11).

By the end of this process, 80 final replicate weights for each HU and person record are produced.

Variance Estimates

Given the replicate weights, the computation of variance for any ACS estimate is straightforward. Suppose that θ is an ACS estimate of any type of statistic, such as mean, total, or proportion. Let $\hat{\theta}_0$ denote the estimate computed based on the full sample weight, and $\theta_1, \theta_2, \ldots, \theta_{80}$, denote the estimates computed based on the replicate weights. The variance of θ_0 $v(\theta_0)$ is estimated as the sum of squared differences between each replicate estimate θ_r ($r = 1, \ldots, 80$) and the full sample estimate $\hat{\theta}_0$. The formula is as follows:[1]

$$v(\hat{\theta}_0) = \frac{4}{80} \sum_{r=1}^{80} (\hat{\theta}_r - \hat{\theta}_0)^2.$$

[1] A general replication-based variance formula can be expressed as $v(\hat{\theta}_0) = \sum_{r=1}^{n} c_r (\hat{\theta}_r - \hat{\theta}_0)^2$, where c_r is the multiplier related to the r-th replicate determined by the replication method. For the SDR method, the value of c_r is 4 / R, where R is the number of replicates (see Fay and Train, 1995).

This equation holds for count estimates as well as any other types of estimates, including percents, ratios, and medians.

There are certain cases, however, where this formula does not apply. The first and most important cases are estimates that are "controlled" to population totals and have their standard errors set to zero. These are estimates that are forced to equal intercensal estimates during the weighting process's raking step—for example, total population and collapsed age, sex, and Hispanic origin estimates for weighting areas. Although race is included in the raking procedure, race group estimates are not controlled; the categories used in the weighting process (see Chapter 11) do not match the published tabulation groups because of multiple race responses and the "Some Other Race" category. Information on the final collapsing of the person post-stratification cells is passed from the weighting to the variance estimation process in order to identify estimates that are controlled. This is done independently for all weighting areas and then is applied to the geographic areas used for tabulation. Standard errors for those estimates are set to zero, and published margins of error are set to "*****" (with an appropriate accompanying footnote).

Another special case deals with zero-estimated counts of people, households, or HUs. A direct application of the replicate variance formula leads to a zero standard error for a zero-estimated count. However, there may be people, households, or HUs with that characteristic in that area that were not selected to be in the ACS sample, but a different sample might have selected them, so a zero standard error is not appropriate. For these cases, the following model-based estimation of standard error was implemented.

For ACS data in a census year, the ACS zero-estimated counts (for characteristics included in the 100 percent census ("short form") count) can be checked against the corresponding census estimates. At least 90 percent of the census counts for the ACS zero-estimated counts should be within a 90 percent confidence interval based on our modeled standard error.[2] Let the variance of the estimate be modeled as some multiple (K) of the average final weight (for a state or the nation). That is:

$$v(0) = K \times \text{(average weight)}$$

Then, set the 90 percent upper bound for the zero estimate equal to the census count:

$$\begin{aligned} &\text{Upper Confidence Bound} \\ &= 0 + 1.645 \times \text{SE}(0) \\ &= 1.645 \times \sqrt{K \times \text{(average weight)}} \\ &= \text{census count} \end{aligned}$$

Solving for K yields:

$$K = \left(\frac{\text{census count}}{1.645}\right)^2 \frac{1}{\text{(average weight)}} .$$

K was computed for all ACS zero-estimated counts from 2000, which matched Census 2000 100 percent counts, and then the 90th percentile of those Ks was determined. Based on the Census 2000 data, we use a value for K of 400 (Navarro, 2001b). As this modeling method requires census counts, the 400 value can next be updated using the 2010 Census and 2010 ACS data.

For publication, the standard error (SE) of the zero count estimate is computed as:

$$SE(0) = \sqrt{400 \times \text{(average weight)}} .$$

The average weights (the maximum of the average housing unit and average person final weights) are calculated at the state and national level for each ACS single-year or multiyear data release. Estimates for geographic areas within a state use that state's average weight, and estimates for geographic areas that cross state boundaries use the national average weight.

[2] This modeling was done only once, in 2001, prior to the publication of the 2000 ACS data.

Finally, a similar method is used to produce an approximate standard error for both ACS zero and 100 percent estimates. We do not produce approximate standard errors for other zero estimates, such as ratios or medians.

Variance Estimation for Multiyear ACS Estimates

The same methodology described above covers both variance estimation for 1-year, 3-year, and 5-year ACS estimates. No changes to the methodology are necessary due to using multiple years of sample data.

12.3 MARGIN OF ERROR AND CONFIDENCE INTERVAL

Once the standard errors have been computed, margins of error and confidence bounds are produced for each estimate. These are the measures of overall sampling error presented along with each published ACS estimate. All published ACS margins of error and the lower and upper bounds of confidence intervals presented in the ACS data products are based on a 90 percent confidence level, which is the Census Bureau's standard (U.S. Census Bureau, 2008a).

A margin of error contains two components: the standard error of the estimate, and a multiplication factor based on a chosen confidence level. For the 90 percent confidence level, the value of the multiplication factor used by the ACS is 1.645. The margin of error of an estimate θ can be computed as:

$$\text{Margin of error } (\theta) = 1.645 \times \text{se}(\theta),$$

where se(θ) is the standard error of the estimate θ. Given this margin of error, the 90 percent confidence interval can be computed as:

$$\theta \pm [\text{margin of error } (\theta)];$$

that is, the lower bound of the confidence interval is [θ − margin of error (θ)], and the upper bound of the confidence interval is [θ + margin of error (θ)]. Roughly speaking, this interval is a range that will contain the "true value" of the estimated characteristic, with a known probability.

Users are cautioned to consider "logical" boundaries when creating confidence bounds from the margins of error. For example, a small population estimate may have a calculated lower bound less than zero. A negative number of people does not make sense, so the lower bound should be set to zero instead. Likewise, bounds for percents should not go below zero percent or above 100 percent. For other characteristics, like income, negative values may be legitimate.

Given the confidence bounds, a margin of error can be computed as the difference between an estimate and its upper or lower confidence bounds:

$$\textit{Margin of Error} = \max(\textit{upper bound} - \textit{estimate}, \textit{estimate} - \textit{lower bound})$$

Using the margin of error (as published or calculated from the bounds), the standard error is obtained as follows:

$$\textit{Standard Error} = \textit{Margin of Error} / 1.645$$

For ranking tables and comparison profiles, the ACS provides an indicator as to whether two estimates are statistically significantly different at the 90 percent confidence level. That determination is made by initially calculating:

$$Z = \frac{Est_1 - Est_2}{\sqrt{SE(Est_1)^2 + SE(Est_2)^2}}.$$

If $Z < -1.645$ or $Z > 1.645$, the difference between the estimates is significant at the 90 percent level. Determinations of statistical significance are made using unrounded values of the standard errors, so users may not be able to achieve the same result using the standard errors derived from the rounded estimates and margins of error as published. Only pairwise tests are used to determine significance in the ranking tables; no multiple comparison methods are used.

12.4 VARIANCE ESTIMATION FOR THE PUMS

The Census Bureau cannot possibly predict all combinations of estimates and geography that may be of interest to data users. Data users can download PUMS files and tabulate the data to create estimates of their own choosing. The ACS PUMS contains a subset of the full ACS sample. Thus, estimates from the ACS PUMS file can be different from the published ACS estimates that are based on the full ACS sample.

Users of the ACS PUMS files can compute the estimated variances of their statistics using one of two options: (1) the replication method using replicate weights released with the PUMS data, and (2) the design factor method described below.

For the replicate method, direct variance estimates based on the SDR formula as described in Section B above can be implemented. Users can simply tabulate 80 replicate estimates in addition to their desired estimate by using the provided 80 replicate weights, and apply the variance formula:

$$v(\hat{\theta}_0) = \frac{4}{80} \sum_{r=1}^{80} (\hat{\theta}_r - \hat{\theta}_0)^2.$$

Similar to methods used to calculate standard errors for PUMS data from Census 2000, the ACS PUMS provides tables of design factors for various topics such as age for persons or tenure for HUs. The 2007 ACS PUMS design factors are published at national and state levels (U.S. Census Bureau, 2008b), and were calculated using 2005 ACS data. PUMS design factors will be updated periodically, but not on an annual basis. The design factor approach was developed based on a model that uses a standard error from a simple random sample as the base, and then inflates it to account for an increase in the variance caused by the complex sample design. Standard errors for almost all counts and proportions of persons, households, and HUs are approximated using design factors. For single-year ACS PUMS files beginning with 2005, use:

$$SE\left(\hat{Y}\right) = DF \times \sqrt{99 \times \hat{Y} \times \left(1 - \frac{\hat{Y}}{N}\right)}$$

for a total, and

$$SE\left(\hat{p}\right) = DF \times \sqrt{\frac{99}{B} \times \hat{p} \times (100 - \hat{p})}$$

for a percent,

where

$\hat{Y}$ = the estimate of total or a count.

$\hat{p}$ = the estimate of a percent.

DF = the appropriate design factor based on the topic of the estimate.

N = the total for the geographic area of interest (if the estimate is of HUs, the number of HUs is used; if the estimate is of families or households, the number of households is used; otherwise the number of persons is used as N).

B = the base (denominator) of a percent.

The factor 99 in the formula is the value of the finite population correction factor for the PUMS, which is computed as $(100 - f) / f$, where f (given as a percent) is the sampling rate for the PUMS data. Since the PUMS is approximately a 1 percent sample of HUs, $(100 - f) / f = (100 - 1) / 1 = 99$.

For 3-year PUMS files beginning with 2005–2007, the 3 years' worth of data represent approximately a 3 percent sample of HUs. Hence, the finite population correction factor for 3-year PUMS is $(100 - f) / f = (100 - 3) / 3 = 97 / 3$. To calculate standard errors from 3-year PUMS data, substitute 97 / 3 for 99 in the above formulas.

The design factor (*DF*) is defined as the ratio of the standard error of an estimated parameter (computed under the replication method described in Section B) to the standard error based on a simple random sample of the same size. The *DF* reflects the effect of the actual sample design and estimation procedures used for the ACS. The *DF* for each topic was computed by modeling the relationship between the standard error under the replication method (*RSE*) with the standard error based on a simple random sample (*SRSSE*); that is, $RSE = DF \times SRSSE$, where the *SRSSE* is computed as follows:

$$SRSSE\left(\hat{Y}\right) = \sqrt{39 \times \hat{Y} \times \left(1 - \frac{\hat{Y}}{N}\right)}$$

The value 39 in the formula above is the finite population correction factor based on an approximate sampling fraction of 2.5 percent in the ACS; that is, 100 − 2.5) / 2.5 = 97.5 / 2.5 = 39.

The value of *DF* is obtained by fitting this (no intercept) regression model $RSE = DF \times SRSSE$ using standard errors (*RSE*, *SRSSE*) for various published table estimates at the national and state levels. The values of *DF*s by topic can be obtained from the "PUMS Accuracy of the Data (2007)" (U.S. Census Bureau, 2008b). The documentation also provides examples on how to use the design factor GVFs to compute standard errors for the estimates of totals, means, medians, proportions or percentages, ratios, sums, and differences.

The topics for the 2007 PUMS design factors are, for the most part, the same ones that were available for the Census 2000 PUMS. We recommend to users that, in using the design factor approach, if the estimate is a combination of two or more characteristics, the largest *DF* for this combination of characteristics is used. The only exceptions to this are items crossed with race or Hispanic origin; for these items, the largest DF is used, excluding race or Hispanic origin *DF*s.

12.5 REFERENCES

Fay, R., and G. Train. (1995). "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties." *Proceedings of the Section on Government Statistics.* Alexandria, VA: American Statistical Association, pp. 154–159, <http://www.census.gov/hhes/www/saipe/asapaper/FayTrain95.pdf>.

Gbur, P., and L. Fairchild. (2002). "Overview of the U.S. Census 2000 Long Form Direct Variance Estimation." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association, pp. 1139–1144.

Gunlicks, C. (1996). "1990 Replicate Variance System (VAR90-20)." Internal U.S. Census Bureau Memorandum for Documentation, June 4, 1996.

Judkins, D. R. (1990). "Fay's Method for Variance Estimation." *Journal of Official Statistics,* Vol. 6, No. 3, 1990, pp. 223–239.

Navarro, A. (2001a). "2000 American Community Survey (ACS) Comparison County Replicate Factors (ACS-V-01)." Internal U.S. Census Bureau Memorandum to C. Alexander, Washington, DC, May 23, 2001.

Navarro, A. (2001b). "Estimating Standard Errors of Zero Estimates." Internal U.S. Census Bureau Draft Memorandum to C. Alexander, Washington, DC, November 6, 2001.

U.S. Census Bureau (2002). "Current Population Survey: Technical Paper 63RV—Design and Methodology." Washington, DC, 2002, <http://www.census.gov/prod/2002pubs/tp63rv.pdf>.

U.S. Census Bureau (2008a). "Census Bureau Standard: Dissemination of Census and Survey Data Products." Washington, DC, 2008, <http://www.census.gov/quality/S17-Ov1.2DisseminationData.pdf>.

U.S. Census Bureau (2008b). "PUMS Accuracy of the Data (2007)." Washington, DC, 2008, <http://www.census.gov/acs/www/Downloads/2007/AccuracyPUMS.pdf>.

Wolter, K. M. (1984). "An Investigation of Some Estimators of Variance for Systematic Sampling." *Journal of the American Statistical Association*, Vol. 79, 1984, pp. 781–790.

Chapter 13.
Preparation and Review of Data Products

13.1 OVERVIEW

This chapter discusses the data products derived from the American Community Survey (ACS). ACS data products include the tables, reports, and files that contain estimates of population and housing characteristics. These products cover geographic areas within the United States and Puerto Rico. Tools such as the Public Use Microdata Sample (PUMS) files, which enable data users to create their own estimates, also are data products.

ACS data products will continue to meet the traditional needs of those who used the decennial census long-form sample estimates. However, as described in Chapter 14, Section 3, the ACS will provide more current data products than those available from the census long form, an especially important advantage toward the end of a decade.

Most surveys of the population provide sufficient samples to support the release of data products only for the nation, the states, and, possibly, a few substate areas. Because the ACS is a very large survey that collects data continuously in every county, products can be released for many types of geographic areas, including many smaller geographic areas such as counties, townships, and census tracts. For this reason, geography is an important topic for all ACS data products.

The first step in the preparation of a data product is defining the topics and characteristics it will cover. Once the initial characteristics are determined, they must be reviewed by the Census Bureau Disclosure Review Board (DRB) to ensure that individual responses will be kept confidential. Based on this review, the specifications of the products may be revised. The DRB also may require that the microdata files be altered in certain ways, and may restrict the population size of the geographic areas for which these estimates are published. These activities are collectively referred to as disclosure avoidance.

The actual processing of the data products cannot begin until all response records for a given year or years are edited and imputed in the data preparation and processing phases, the final weights are determined, and disclosure avoidance techniques are applied. Using the weights, the sample data are tabulated for a wide variety of characteristics according to the predetermined content. These tabulations are done for the geographic areas that have a sample size sufficient to support statistically reliable estimates, with the exception of 5-year period estimates, which will be available for small geographic areas down to the census tract and block group levels. The PUMS data files are created by different processes because the data are a subset of the full sample data.

After the estimates are produced and verified for correctness, Census Bureau subject matter analysts review them. When the estimates have passed the final review, they are released to the public. A similar process of review and public release is followed for PUMS data.

While the 2005 ACS sample was limited to the housing unit (HU) population for the United States and Puerto Rico, starting in sample year 2006, the ACS was expanded to include the group quarters (GQ) population. Therefore, the ACS sample is representative of the entire resident population in the United States and Puerto Rico. In 2007, 1-year period estimates for the total population and subgroups of the total population in both the United States and Puerto Rico were released for sample year 2006. Similarly, in 2008, 1-year period estimates were released for sample year 2007.

In 2008, the Census Bureau will, for the first time, release products based on 3 years of ACS sample, 2005 through 2007. In 2010, the Census Bureau plans to release the first products based on 5 years of consecutive ACS samples, 2005 through 2009. Since several years of samples form the basis of these multiyear products, reliable estimates can be released for much smaller geographic areas than is possible for products based on single-year data.

In addition to data products regularly released to the public, other data products may be requested by government agencies, private organizations and businesses, or individuals. To accommodate such requests, the Census Bureau operates a custom tabulations program for the ACS on a fee basis. These tabulation requests are reviewed by the DRB to assure protection of confidentiality before release.

Chapter 14 describes the dissemination of the data products discussed in this chapter, including display of products on the Census Bureau's Web site and topics related to data file formatting.

13.2 GEOGRAPHY

The Census Bureau strives to provide products for the geographic areas that are most useful to users of those data. For example, ACS data products are already disseminated for many of the nation's legal and administrative entities, including states, American Indian and Alaska Native (AIAN) areas, counties, minor civil divisions (MCDs), incorporated places, congressional districts, as well as data for a variety of other geographic entities. In cooperation with state and local agencies, the Census Bureau identifies and delineates geographic entities referred to as "statistical areas." These include regions, divisions, urban areas (UAs), census county divisions (CCDs), census designated places (CDPs), census tracts, and block groups. Data users then can select the geographic entity or set of entities that most closely represent their geographic areas of interest and needs.

"Geographic summary level" is a term used by the Census Bureau to designate the different geographic levels or types of geographic areas for which data are summarized. Examples include the entities described above, such as states, counties, and places (the Census Bureau's term for entities such as for cities and towns, including unincorporated areas). Information on the types of geographic areas for which the Census Bureau publishes data is available at <http://www.census.gov/geo/www/garm.html>.

Single-year period estimates of ACS data are published annually for recognized legal, administrative, or statistical areas with populations of 65,000 or more (based on the latest Census Bureau population estimates). Three-year period estimates based on 3 successive years of ACS samples are published for areas of 20,000 or more. If a geographic area met the 1-year or 3-year threshold for a previous period but dropped below it for the current period, it will continue to be published as long as the population does not drop more than 5 percent below the threshold. Plans are to publish 5-year period estimates based on 5 successive years of ACS samples starting in 2010 with the 2005–2009 data. Multiyear period estimates based on 5 successive years of ACS samples will be published for all legal, administrative, and statistical areas down to the block-group level, regardless of population size. However, there are rules from the Census Bureau's DRB that must be applied.

The Puerto Rico Community Survey (PRCS) also provides estimates for legal, administrative, and statistical areas in Puerto Rico. The same rules as described above for the 1-year, 3-year, and 5-year period estimates for the U.S resident population apply for the PRCS as well.

The ACS publishes annual estimates for hundreds of substate areas, many of which will undergo boundary changes due to annexations, detachments, or mergers with other areas.[1] Each year, the Census Bureau's Geography Division, working with state and local governments, updates its files to reflect these boundary changes. Minor corrections to the location of boundaries also can occur as a result of the Census Bureau's ongoing Master Address File (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER®) Enhancement Project. The ACS estimates must

[1]The Census Bureau conducts the Boundary and Annexation Survey (BAS) each year. This survey collects information on a voluntary basis from local governments and federally recognized American Indian areas. The information collected includes the correct legal place names, type of government, legal actions that resulted in boundary changes, and up-to-date boundary information. The BAS uses a fixed reference date of January 1 of the BAS year. In years ending in 8, 9, and 0, all incorporated places, all minor civil divisions, and all federally recognized tribal governments are included in the survey. In other years, only governments at or above various population thresholds are contacted. More detailed information on the BAS can be found at <http://www.census .gov/geo/www/bas/bashome.html>.

reflect these legal boundary changes, so all estimates are based on Geography Division files that show the geographic boundaries as they existed on January 1 of the sample year or, in the case of multiyear data products, at the beginning of the final year of data collection.

13.3 DEFINING THE DATA PRODUCTS

For the 1999 through 2002 sample years, the ACS detailed tables were designed to be comparable with Census 2000 Summary File 3 to allow comparisons between data from Census 2000 and the ACS. However, when Census 2000 data users indicated certain changes they wanted in many tables, ACS managers saw the years 2003 and 2004 as opportunities to define ACS products based on users' advice.

Once a preliminary version of the revised suite of products had been developed, the Census Bureau asked for feedback on the planned changes from data users (including other federal agencies) via a *Federal Register* Notice (Fed. Reg. #3510-07-P). The notice requested comments on current and proposed new products, particularly on the basic concept of the product and its usefulness to the data users. Data users provided a wide variety of comments, leading to modifications of planned products.

ACS managers determined the exact form of the new products in time for their use in 2005 for the ACS data release of sample year 2004. This schedule allowed users sufficient time to become familiar with the new products and to provide comments well in advance of the data release for the 2005 sample.

Similarly, a *Federal Register* Notice issued in August 2007 shared with the public plans for the data release schedule and products that would be available beginning in 2008. This notice was the first that described products for multiyear estimates. Improvements will continue when multiyear period estimates are available.

13.4 DESCRIPTION OF AGGREGATED DATA PRODUCTS

ACS data products can be divided into two broad categories: aggregated data products, and the PUMS, which is described in Section 13.5 ("Public Use Microdata Sample").

Data for the ACS are collected from a sample of housing units (HUs), as well as the GQ population, and are used to produce estimates of the actual figures that would have been obtained by interviewing the entire population. The aggregated data products contain the estimates from the survey responses. Each estimate is created using the sample weights from respondent records that meet certain criteria. For example, the 2007 ACS estimate of people under the age of 18 in Chicago is calculated by adding the weights from all respondent records from interviews completed in 2007 in Chicago with residents under 18 years old.

This section provides a description of each aggregated product. Each product described is available as single-year period estimates; unless otherwise indicated, they will be available as 3-year estimates and are planned for the 5-year estimates. Chapter 14 provides more detail on the actual appearance and content of each product.

These data products contain all estimates planned for release each year, including those from multiple years of data, such as the 2005–2007 products. Data release rules will prevent certain single- and 3-year period estimates from being released if they do not meet ACS requirements for statistical reliability.

Detailed Tables

The detailed tables provide basic distributions of characteristics. They are the foundation upon which other data products are built. These tables display estimates and the associated lower and upper bounds of the 90 percent confidence interval. They include demographic, social, economic, and housing characteristics, and provide 1-, 3-, or 5-year period estimates for the nation and the states, as well as for counties, towns, and other small geographic entities, such as census tracts and block groups.

The Census Bureau's goal is to maintain a high degree of comparability between ACS detailed tables and Census 2000 sample-based data products. In addition, characteristics not measured in the Census 2000 tables will be included in the new ACS base tables. The 2007 detailed table products include more than almost 600 tables that cover a wide variety of characteristics, and another 380 race and Hispanic-origin iterations that cover 40 key characteristics. In addition to the tables on characteristics, approximately 80 tables summarize allocation rates from the data edits for many of the characteristics. These provide measures of data quality by showing the extent to which responses to various questionnaire items were complete. Altogether, over 1,300 separate detailed tables are provided.

Data Profiles

Data profiles are high-level reports containing estimates for demographic, social, economic, and housing characteristics. For a given geographic area, the data profiles include distributions for such characteristics as sex, age, type of household, race and Hispanic origin, school enrollment, educational attainment, disability status, veteran status, language spoken at home, ancestry, income, poverty, physical housing characteristics, occupancy and owner/renter status, and housing value. The data profiles include a 90 percent margin of error for each estimate. Beginning with the 2007 ACS, a comparison profile that compares the 2007 sample year's estimates with those of the 2006 ACS also will be published. These profile reports include the results of a statistical significance test for each previous year's estimate, compared to the current year. This test result indicates whether the previous year's estimate is significantly different (at a 90 percent confidence level) from that of the current year.

Narrative Profiles

Narrative profiles cover the current sample year only. These are easy-to-read, computer-produced profiles that describe main topics from the data profiles for the general-purpose user. These are the only ACS products with no standard errors accompanying the estimates.

Subject Tables

These tables are similar to the Census 2000 quick tables, and like them, are derived from detailed tables. Both quick tables and subject tables are predefined, covering frequently requested information on a single topic for a single geographic area. However, subject tables contain more detail than the Census 2000 quick tables or the ACS data profiles. In general, a subject table contains distributions for a few key universes, such as the race groups and people in various age groups, which are relevant to the topic of the table. The estimates for these universes are displayed as whole numbers. The distribution that follows is displayed in percentages. For example, subject table S1501 on educational attainment provides the estimates for two different age groups—18 to 24 years old and 25 years and older, as a whole number. For each age group, these estimates are followed by the percentages of people in different educational attainment categories (high school graduate, college undergraduate degree, etc.). Subject tables also contain other measures, such as medians, and they include the imputation rates for relevant characteristics. More than 40 topic-specific subject tables are released each year.

Ranking Products

Ranking products contain ranked results of many important measures across states. They are produced as 1-year products only, based on the current sample year. The ranked results among the states for each measure are displayed in three ways—charts, tables, and tabular displays that allow for testing statistical significance.

The rankings show approximately 80 selected measures. The data used in ranking products are pulled directly from a detailed table or a data profile for each state.

Geographic Comparison Tables (GCTs)

GCTs contain the same measures that appear in the ranking products. They are produced as both 1-year and multiyear products. GCTs are produced for states as well as for substate entities, such as congressional districts. The results among the geographic entities for each measure are displayed as tables and thematic maps (see next).

Thematic Maps

Thematic maps are similar to ranking tables. They show mapped values for geographic areas at a given geographic summary level. They have the added advantage of visually displaying the geographic variation of key characteristics (referred to as themes). An example of a thematic map would be a map showing the percentage of a population 65 years and older by state.

Selected Population Profiles (SPPs)

SPPs provide certain characteristics from the data profiles for a specific race or ethnic group (e.g., Alaska Natives) or some other selected population group (e.g., people aged 60 years and older). SPPs are provided every year for many of the Census 2000 Summary File 4 iteration groups. SPPs were introduced on a limited basis in the fall of 2005, using the 2004 sample. In 2008 (sample year 2007), this product was significantly expanded. The earlier SPP requirement was that a substate geographic area must have a population of at least 1,000,000 people. This threshold was reduced to 500,000, and congressional districts were added to the list of geographic types that can receive SPPs. Another change to SPPs in 2008 is the addition of many country-of-birth groups.

Groups too small to warrant an SPP for a geographic area based on 1 year of sample data may appear in an SPP based on the 3- or 5-year accumulations of sample data. More details on these profiles can be found in Hillmer (2005), which includes a list of selected race, Hispanic origin, and ancestry populations.

13.5 PUBLIC USE MICRODATA SAMPLE

Microdata are the individual records that contain information collected about each person and HU. PUMS files are extracts from the confidential microdata that avoid disclosure of information about households or individuals. These extracts cover all of the same characteristics contained in the full microdata sample files. Chapter 14 provides information on data and file organization for the PUMS.

The only geography other than state shown on a PUMS file is the Public Use Microdata Area (PUMA). PUMAs are special nonoverlapping areas that partition a state, each containing a population of about 100,000. State governments drew the PUMA boundaries at the time of Census 2000. They were used for the Census 2000 sample PUMS files and are known as the "5 percent PUMAs." (For more information on these geographic areas, go to <http://www.census.gov/prod/cen2000/doc/pums.pdf>.)

The Census Bureau has released a 1-year PUMS file from the ACS since the survey's inception. In addition to the 1-year ACS PUMS file, the Census Bureau plans to create multiyear PUMS files from the ACS sample, starting with the 2005–2007 3-year PUMS file. The multiyear PUMS files combine annual PUMS files to create larger samples in each PUMA, covering a longer period of time. This will allow users to create estimates that are more statistically reliable.

13.6 GENERATION OF DATA PRODUCTS

Following conversations with users of census data, the subject matter analysts in the Census Bureau's Housing and Household Economic Statistics Division and Population Division specify the organization of the ACS data products. These specifications include the logic used to calculate every estimate in each data product and the exact textual description associated with each estimate. Starting with the 2006 ACS data release, only limited changes to these specifications have occurred. Changes to the data product specifications must preserve the ability to compare estimates from one year to another and must be operationally feasible. Changes must be made no later than late winter of each year to ensure that the revised specifications are finalized by the spring of that year and ready for the data releases beginning in the late summer of the year.

After the edited data with the final weights are available (see Chapters 10 and 11), generation of the data products begins with the creation of the detailed tables data products with the 1-year period estimates. The programming teams of the American Community Survey Office (ACSO) generate these estimates. Another staff within ACSO verifies that the estimates comply with the specifications from subject matter analysts. Both the generation and the verification activities are automated.

The 1-year data products are released on a phased schedule starting in the summer. Currently, the Census Bureau plans to release the multiyear data products late each year, after the release of the 1-year products.

One distinguishing feature of the ACS data products system is that standard errors are calculated for all estimates and are released with the latter in tables. Subject matter analysts also use the standard errors in their internal reviews of estimates.

Disclosure Avoidance

Once plans are finalized for the ACS data products, the DRB reviews them to assure that confidentiality of respondents has been protected.

Title 13 of the United States Code (U.S.C.) is the basis for the Census Bureau's policies on disclosure avoidance. Title 13 says, "Neither the Secretary, nor any other officer or employee of the Department of Commerce may make any publication whereby the data furnished by any particular establishment or individual under this title can be identified . . ." The DRB reviews all data products planned for public release to ensure adherence to Title 13 requirements, and may insist on applying disclosure avoidance rules that could result in the suppression of certain measures for small geographic areas. (More information about the DRB and its policies can be found at <http://www.factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_pageId=su5_confidentiality>.

To satisfy Title 13 U.S.C., the Census Bureau uses several statistical methodologies during tabulation and data review to ensure that individually identifiable data will not be released.

Swapping. The main procedure used for protecting Census 2000 tabulations was data swapping. It was applied to both short-form (100 percent) and long-form (sample) data independently. Currently, it also is used to protect ACS tabulations. In each case, a small percentage of household records is swapped. Pairs of households in different geographic regions are swapped. The selection process for deciding which households should be swapped is highly targeted to affect the records with the most disclosure risk. Pairs of households that are swapped match on a minimal set of demographic variables. All data products (tables and microdata) are created from the swapped data files.

For PUMS data the following techniques are employed in addition to swapping:

Top-coding is a method of disclosure avoidance in which all cases in or above a certain percentage of the distribution are placed into a single category.

Geographic population thresholds prohibit the disclosure of data for individuals or HUs for geographic units with population counts below a specified level.

Age perturbation (modifying the age of household members) is required for large households containing 10 people or more due to concerns about confidentiality.

Detail for categorical variables is collapsed if the number of occurrences in each category does not meet a specified national minimum threshold.

For more information on disclosure avoidance techniques, see Section 5, "Current disclosure avoidance practices" at <http://www.census.gov/srd/papers/pdf/rrs2005-06.pdf>.

The DRB also may determine that certain tables are so detailed that other restrictions are required to ensure that there is sufficient sample to avoid revealing information on individual respondents. In such instances, a restriction may be placed on the size of the geographic area for which the table can be published. Current DRB rules require that detailed tables containing more than 100 detailed cells may not be released below the census tract level.

The data products released in the summer of 2006 for the 2005 sample covered the HU population of the United States and Puerto Rico only. In January 2006, data collection began for the population living in GQ facilities. Thus, the data products released in summer 2007 (and each year

thereafter) covered the entire resident population of the United States and Puerto Rico. Most estimates for person characteristics covered in the data products were affected by this expansion. For the most part, the actual characteristics remained the same, and only the description of the population group changed from HU to resident population.

Data Release Rules

Even with the population size thresholds described earlier, in certain geographic areas some very detailed tables might include estimates with unacceptable reliability. Data release rules, based on the statistical reliability of the survey estimates, were first applied in the 2005 ACS. These release rules apply only to the 1- and 3-year data products.

The main data release rule for the ACS tables works as follows. Every detailed table consists of a series of estimates. Each estimate is subject to sampling variability that can be summarized by its standard error. If more than half of the estimates in the table are not statistically different from 0 (at a 90 percent confidence level), then the table fails. Dividing the standard error by the estimate yields the coefficient of variation (CV) for each estimate. (If the estimate is 0, a CV of 100 percent is assigned.) To implement this requirement for each table at a given geographic area, CVs are calculated for each table's estimates, and the median CV value is determined. If the median CV value for the table is less than or equal to 61 percent, the table passes for that geographic area and is published; if it is greater than 61 percent, the table fails and is not published.

Whenever a table fails, a simpler table that collapses some of the detailed lines together can be substituted for the original. If the simpler table passes, it is released. If it fails, none of the estimates for that table and geographic area are released. These release rules are applied to single- and multiyear period estimates based on 3 years of sample data. Current plans are not to apply data release rules to the estimates based on 5 years of sample data.

13.7 DATA REVIEW AND ACCEPTANCE

After the editing, imputation, data products generation, disclosure avoidance, and application of the release rules have been completed, subject matter analysts perform a final review of the ACS data and estimates before release. This final data review and acceptance process helps to ensure that there are no missing values, obvious errors, or other data anomalies.

Each year, the ACS staff and subject matter analysts generate, review, and provide clearance of all ACS estimates. At a minimum, the analysts subject their data to a specific multistep review process before they are cleared and released to the public. Because of the short time available to review such a large amount of data, an automated review tool (ART) has been developed to facilitate the process.

ART is a computer application that enables subject matter analysts to detect statistically significant differences in estimates from one year to the next using several statistical tests. The initial version of ART was used to review 2003 and 2004 data. It featured predesigned reports as well as ad hoc, user-defined queries for hundreds of estimates and for 350 geographic areas. An ART workgroup defined a new version of ART to address several issues that emerged. The improved version has been used by the analysts since June 2005; it is designed to work on much larger data sets and a wider range of capabilities, with faster response time to user commands. A team of programmers, analysts, and statisticians then developed an automated tool to assist analysts in their review of the multiyear estimates. This tool was used in 2008 for the review of the 2005–2007 estimates.

The ACSO staff, together with the subject matter analysts, also have developed two other automated tools to facilitate documentation and clearance for required data review process steps: the edit management and messaging application (EMMA), and the PUMS management and messaging application (PMMA). Both are used to track the progress of analysts' review activities and both enable analysts and managers to see the current status of files under review and determine which review steps can be initiated.

13.8 IMPORTANT NOTES ON MULTIYEAR ESTIMATES

While the types of data products for the multiyear estimates are almost entirely identical to those used for the 1-year estimates, there are several distinctive features of the multiyear estimates that data users must bear in mind.

First, the geographic boundaries that are used for multiyear estimates are always the boundary as of January 1 of the final year of the period. Therefore, if a geographic area has gained or lost territory during the multiyear period, this practice can have a bearing on the user's interpretation of the estimates for that geographic area.

Secondly, for multiyear period estimates based on monetary characteristics (for example, median earnings), inflation factors are applied to the data to create estimates that reflect the dollar values in the final year of the multiyear period.

Finally, although the Census Bureau tries to minimize the changes to the ACS questionnaire, these changes will occur from time to time. Changes to a question can result in the inability to build certain estimates for a multiyear period containing the year in which the question was changed. In addition, if a new question is introduced during the multiyear period, it may be impossible to make estimates of characteristics related to the new question for the multiyear period.

13.9 CUSTOM DATA PRODUCTS

The Census Bureau offers a wide variety of general-purpose data products from the ACS designed to meet the needs of the majority of data users. They contain predefined sets of data for standard census geographic areas. For users whose data needs are not met by the general-purpose products, the Census Bureau offers customized special tabulations on a cost-reimbursable basis through the ACS custom tabulation program. Custom tabulations are created by tabulating data from ACS edited and weighted data files. These projects vary in size, complexity, and cost, depending on the needs of the sponsoring client.

Each custom tabulation request is reviewed in advance by the DRB to ensure that confidentiality is protected. The requestor may be required to modify the original request to meet disclosure avoidance requirements. For more detailed information on the ACS Custom Tabulations program, go to <http://www.census.gov/acs/www/Products/spec_tabs/index.htm>.

Chapter 14.
Data Dissemination

14.1 OVERVIEW

This chapter deals with the 1-year and 3-year data products. Future versions of this document will include a discussion of the 5-year data products. The American Community Survey (ACS) data products and supporting documentation are released in several series and at several Internet locations. The primary Web site for data dissemination is the American FactFinder (AFF); supporting documentation can be found on the ACS Web site and the Census Bureau's File Transfer Protocol (FTP) site.

Since 2000, the ACS has been tabulating and publishing single year estimates for specific areas with populations of 250,000 or more. In 2005, the ACS expanded its sample size to cover all of the United States and the Commonwealth of Puerto Rico. In summer 2006, the ACS started releasing data annually for areas with populations of 65,000 or more. In 2008, the ACS is releasing 3-year period estimates for areas with a population of 20,000 or more on an annual basis. For smaller areas, it will take 5 years to accumulate a large enough sample to produce releasable estimates. Once those data are collected, the Census Bureau will release tabulations annually, based on 5-year period data for areas as small as census tracts and block groups.

Federal agencies distribute billions of dollars among states, tribal governments, and population groups, based on social and economic data. In the past, the statistics that determined services locations and program funding came in large part from the long-form sample of the decennial census. As the ACS continues to grow, its data products will provide updated versions of many of the long-form products from Census 2000. Beginning in 2010, the decennial census no longer will include a long-form sample, and ACS data products will provide high-quality, updated annual statistics for comparisons of the demographic, social, economic, and housing characteristics of areas and population groups. The ACS statistics also will show trends and relative differences between areas and population groups. These data products will continue to meet the needs of those who previously used the decennial census sample statistics, and will provide more current statistics than those available from the census long-form sample, which reflect only one point in time.

14.2 SCHEDULE

Data Release Timetable

By 2010, the information on social, demographic, economic, and housing characteristics previously available only once every 10 years will be available annually through the ACS for all areas. Each year thereafter, these areas will get new estimates based on the 5-year interval ending in the latest completed sample year.

Figure 14.1 summarizes the data products release schedule. In 2006, the first set of 1-year estimates was released for specific areas with populations of 65,000 and more. These areas will continue to receive 1-year estimates annually. In 2008, data collected over a 3-year period (2005–2007) was released for areas with at least 20,000 people. These areas will continue to receive 3-year estimates annually. In 2010, the first 5-year products will be released based on data collected in 2005–2009. These products will be produced for areas down to census tracts and block groups. Once 3- and 5-year products are produced, annual updates will follow, as indicated by Table 14.1.

Table 14.1 **Data Products Release Schedule**

Data product	Population threshold	Year of data release							
		2006	2007	2008	2009	2010	2011	2012	2013
1-year estimates...	65,000+	2005	2006	2007	2008	2009	2010	2011	2012
3-year estimates...	20,000+			2005–2007	2006–2008	2007–2009	2008–2010	2009–2011	2010–2012
5-year estimates...	All areas*					2005–2009	2006–2010	2007–2011	2008–2012

* All legal, administrative, and statistical geographic areas down to the tract and block group level.

14.3 PRESENTATION OF TABLES

American FactFinder

The AFF Web site contains data maps, tables, and reports from a variety of censuses and surveys. AFF lists these data sets by program areas and survey years. AFF contains data for a wide variety of surveys including the Decennial Census, the ACS, the Population Estimates Program, the Economic Census, and the Annual Economic Surveys.

The AFF is the primary Web access tool for ACS data. Data products include detailed tables, data profiles, comparison profiles (1-year data only), narrative profiles, ranking tables and charts (single year data only), geographic comparison tables, thematic maps, subject tables, selected population profiles, and downloadable public use microdata sample (PUMS) files.

ACS Web Site

The ACS Web site contains a wealth of information, documentation, and research papers about ACS. The site contains important metadata, including more than 50 population concept definitions and more than 40 housing concept definitions. The ACS Web site can be found at <http://www.census.gov/acs/www>.

Documentation on the accuracy of the data also is included, and provides information about the sample design, confidentiality, sampling error, nonsampling error, and estimation methodology. The errata section lists updates made to the data. The geography section gives a brief explanation of the Census Bureau's geographic hierarchy, common terms, and specific geographic areas presented.

File Transfer Protocol (FTP) Site

The FTP site is intended for advanced users of census and ACS data. This site provides quick access to users who need to begin their analyses immediately upon data release. The data are downloaded into Excel, PDF, or text files. Users of the FTP site can import the files into the spreadsheet/database software of their choice for data analysis and table presentation. Documentation describing the layout of the site in the README file is available in the main directory on the FTP server. The FTP site can be accessed through the ACS Web site.

Chapter 15.
Improving Data Quality by Reducing Nonsampling Error

15.1 OVERVIEW

As with all surveys, the quality of the American Community Survey (ACS) data reflects how well the data collection procedures address potential sources of nonsampling error, including coverage error, nonresponse and measurement errors, and errors that may arise during data capture and processing. Chapters 4 and 11 provide information regarding the steps the ACS takes to reduce sampling error while still managing costs.

There are four primary sources of nonsampling error (Groves, 1989):

- Coverage Error. The failure to give some units in the target population any chance of selection into the sample, or giving units more than one chance of selection.
- Nonresponse Error. The failure to collect data from all units in the sample.
- Measurement Error. The inaccuracy in responses recorded on survey instruments, arising from:
  - The effects of interviewers on the respondents' answers to survey questions.
  - Respondents' inability to answer questions, lack of requisite effort to obtain the correct answer, or other psychological or cognitive factors.
  - Faulty wording of survey questions.
  - Data collection mode effects.
- Processing Error. Errors introduced after the data are collected, including:
  - Data capture errors.
  - Errors arising during coding and classification of data.
  - Errors arising during editing and item imputation of data.

This chapter identifies the operations and procedures designed to reduce these sources of nonsampling error and thus improve the quality of the data. It also includes information about ACS Quality Measures, which provide data users an indication of the potential for nonsampling error. The ACS releases the survey estimates, as well as the Quality Measures, at the same time each year, so that users can consider data quality in conjunction with the survey estimates. The ACS Quality Measures are available on the American FactFinder (AFF) Web site <http://factfinder.census.gov/home/saff/main.html?lang=en> for ACS data beginning with 2007 (and all multiyear data). The Quality Measures for years 2000 to 2006 are located on the ACS Quality Measures Web site <http://www.census.gov/acs/www/UseData/sse/index.htm>.

15.2 COVERAGE ERROR

All surveys experience some degree of coverage error. It can take the form of under-coverage or over-coverage. Under-coverage occurs when units in the target population do not have a chance of selection into the sample; for example, addresses not listed on the frame, or people erroneously excluded from a household roster. Over-coverage occurs when units or people have multiple chances of selection; for example, addresses listed more than once on the frame, or people included on a household roster at two different sampled addresses. In general, coverage error can affect survey estimates if the characteristics of the individuals or units excluded or included in

error differ from the characteristics of those correctly listed in the frame. Over- and under-coverage sometimes can be adjusted as part of the poststratification process, that is, adjusting weights to independent population control totals. Chapter 11 provides more details regarding the ACS weighting process.

The ACS uses the Master Address File (MAF) as its sampling frame, and includes several procedures for reducing coverage error in the MAF. These procedures are described below. Chapter 3 provides further details.

- Twice a year, the U.S. Census Bureau receives the U.S. Postal Service (USPS) Delivery Sequence File (DSF) that includes the addresses including a house number and street name rather than a rural route or post-office box. This file is used to update the city-style addresses on the MAF.
- The ACS nonresponse follow-up operation provides ongoing address and geography updates.
- The MAF includes address updates from special census operations.
- The Community Address Updating System (CAUS) can provide address updates (as a counterpart to the DSF updates) that cover predominately rural areas where city-style addresses generally are not used for mail delivery. CAUS was put on hold in late 2006 and is expected to be back in 2010. CAUS was put on hold because of the address canvassing operation for the 2010 Census.

The ACS Quality Measures contain housing- and person-level coverage rates (as indicators of the potential for coverage error). The coverage rates are located on the AFF for ACS data for 2007 and beyond (including all multiyear data). Coverage rates for prior years (2000 to 2006) are available on the ACS Quality Measures Web site.

Coverage rates for the total resident population are calculated by sex at the national, state, and Puerto Rico geographies, and at the national level only for Hispanics and non-Hispanics crossed by the five major race categories: White, Black, American Indian and Alaska Native, Asian, and Native Hawaiian and Other Pacific Islander. The total resident population includes persons in both housing units (HUs) and group quarters (GQ). In addition, these measures include a coverage rate specific to the GQ population at the national level. Coverage rates for HUs are calculated at the national and state level, with the exception of Puerto Rico because independent HU estimates are not available.

The coverage rate is the ratio of the ACS population or housing estimate of an area or group to the independent estimate for that area or group, multiplied by 100. The Census Bureau uses independent data on housing, births, deaths, immigration, and other categories to produce official estimates of the population and HUs each year. The base for these independent estimates is the decennial census counts. The numerator in the coverage rates is weighted to reflect the probability of selection into the sample, subsampling for personal visit follow-up, and is adjusted for unit nonresponse. The weight used for this purpose does not include poststratification adjustments (weighting adjustments that make the weighted totals match the independent estimates), since the control totals serve as the basis for comparison for the coverage rates. The ACS corrects for potential over- or under-coverage by controlling to these official estimates on specific demographic characteristics and at specific levels of geography.

As the coverage rate for a particular subgroup drops below 100 percent (indicating under-coverage), the weights of its members are adjusted upward in the final weighting procedure to reach the independent estimate. If the rate is greater than 100 percent (indicating over-coverage), the weights of its members are adjusted downward to match the independent estimates.

15.3 NONRESPONSE ERROR

There are two forms of nonresponse error: unit nonresponse and item nonresponse. Unit nonresponse results from the failure to obtain the minimum required data from an HU in the sample. Item nonresponse occurs when respondents do not report individual data items, or provide data considered invalid or inconsistent with other answers.

Surveys strive to increase both unit and item response to reduce the potential for bias introduced into survey estimates. Bias results from systematic differences between the nonrespondents and the respondents. Without data on the nonrespondents, surveys cannot easily measure differences between the two groups. The ACS reduces the potential for bias by reducing the amount of unit and item nonresponse through procedures and processes listed below.

- Response to the ACS is mandated by law, and information about the mandatory requirement to respond is provided in most materials and reinforced in any communication with respondents in all stages of data collection.
- The ACS survey operations include two stages of nonresponse follow-up: a computer-assisted telephone interview (CATI) follow-up for mail nonrespondents, and a computer-assisted personal interview (CAPI) follow-up for a sample of remaining nonrespondents and unmailable addresses cases.
- The mail operation implements a strategy suggested in research studies for obtaining a high mail response rate (Dillman, 1978): a prenotice letter, a message on the envelope of the questionnaire mailing package stating that the response is "required by law," a postcard reminder, and a second mailing for nonrespondents to the initial mailing.
- The mailing package includes a frequently asked questions (FAQ) motivational brochure explaining the survey, its importance, and its mandatory nature.
- The questionnaire design reflects accepted principles of respondent friendliness and navigation, making it easier for respondents to understand which items apply to them, as well as providing cues for a valid response at an item level (such as showing the format for reporting dates, or using a prefilled '0' to indicate reporting dollar amounts rounded to the nearest whole number). Similarly, the CATI and CAPI instruments direct interviewers to ask the appropriate questions.
- The questionnaire provides a toll-free telephone number for respondents who have questions about the ACS in general or who need help in completing the questionnaire.
- The ACS includes a telephone failed-edit follow-up (FEFU) interview with mail respondents who either failed to respond to specific critical questions, or who indicated a household size of six or more people. (The mail form allows data for only five people, so the FEFU operation collects data for any additional persons.)
- The ACS uses permanent professional interviewers trained in refusal conversion methods for CATI and CAPI.
- Survey operations include providing support in other languages: a Spanish paper questionnaire is available on request, and there is a Spanish CATI/CAPI instrument. Also, there are CATI and CAPI interviewers who speak Spanish or other languages as needed.

Unit Nonresponse

The Census Bureau presents survey response and nonresponse rates as part of the ACS Quality Measures. The survey response rate is the ratio of the units interviewed after data collection to the estimate of all units that were eligible to be interviewed. Data users can find survey response and nonresponse rates on the AFF for ACS data for 2007 and beyond (including multiyear estimates). The same rates for data years 2000 to 2006 are available on the ACS Quality Measures Web site. The ACS Quality Measures provide separate rates for HUs and GQ persons. For the HU response rate, the numerator includes all cases that were interviewed after mail, telephone, and personal visit follow-up. For the GQ person response rate, the numerator includes all interviewed persons after the personal visit. For both rates, the numerator includes completed interviews as well as partial interviews with adequate information for processing.

To accurately measure unit response, the ACS estimates the universe of cases eligible to be interviewed and the survey noninterviews. The estimate of the total number of eligible units becomes the denominator of the unit response rate.

The ACS Quality Measures also include the percentage of cases that did not respond to the survey by the reason for nonresponse. These reasons include refusal, unable to locate the sample unit, no one home during the data collection period, temporarily absent during the interview period, language problem, insufficient data (not enough data collected to consider it a response), and other (such as "sample address not accessible"; "death in the family"; or cases not followed up due to budget constraints, which last occurred in the winter of 2004). For the GQ rates, there are two additional reasons for noninterview: whole GQ refusal, and whole GQ other (such as unable to locate the GQ).

Item Nonresponse

The ACS Quality Measures provide information about item nonresponse. When respondents do not report individual data items, or provide data considered invalid or inconsistent with other answers, the Census Bureau imputes the necessary data. The imputation methods use either rules to determine acceptable answers (referred to as "assignment") or answers from similar people or HUs ("allocation"). Assignment involves logical imputation, in which a response to one question implies the value for a missing response to another question. For example, first name often can be used to assign a value to sex. Allocation involves using statistical procedures to impute for missing values. The ACS Quality Measures include summary allocation rates as a measure of the extent to which item nonresponse required imputation. Starting with the 2007 ACS data (including ACS multiyear data), the Quality Measures include only two item allocation rates: overall HU characteristic imputation rate and overall person characteristic imputation rate. These rates are available on the AFF at the national and state level. However, the ACS releases imputation tables on AFF that allow users to compute allocation rates for all published variables and all published geographies. Allocation rates for all published variables from 2000 to 2006 are available on the ACS Quality Measures Web site at the national and state level.

15.4 MEASUREMENT ERROR

All surveys encounter some form of measurement error, which is defined as the difference between the recorded answer and the true answer. Measurement error may occur in any mode of data collection and can be caused by vague or ambiguous questions easily misinterpreted by respondents; questions that respondents cannot answer or questions where respondents deliberately falsify answers for social desirability reasons (see Tourangeau and Yan (2007) for information on social desirability); or interviewer characteristics or actions such as the tone used in reading questions, the paraphrasing of questions, or leading respondents to certain answers.

The ACS minimizes measurement error in several ways, some of which also help to reduce nonresponse.

- As mandated in the Census Bureau Standard "Pretesting Questionnaires and Related Materials for Surveys and Censuses (Version 1.2)" <http://www.census.gov/quality/S11-0_v1.2_Pretesting.pdf>, ACS pretests new or modified survey questions in all three modes before introducing them into the ACS.
- The ACS uses a questionnaire design that reflects accepted principles of respondent friendliness and navigation.
- The ACS mail questionnaire package includes a questionnaire instruction booklet that provides additional information on how to interpret and respond to specific questions.
- Respondents may call the toll-free telephone questionnaire assistance (TQA) line and speak with trained interviewers for answers to general ACS questions or questions regarding specific items.
- Differences among the mail, CATI, and CAPI questionnaires are reduced through questionnaire and instrument design methods that reflect the strengths and limitations of each mode of collection (for example, less complicated skip patterns on the mail questionnaire, breaking up questions with long or complicated response categories into separate questions for telephone administration, and including respondent flash cards for personal visit interviews).

- The CATI/CAPI instruments automate or direct skips, and show the interviewer only those questions appropriate for the person being interviewed.
- The CATI/CAPI instruments include functionality that facilitates valid responses. For example, the instruments check for values outside of the expected range to ensure that the reported answer reflects an appropriate response.
- Training for the permanent CATI and CAPI interviewing staff includes instruction on reading the questions as worded and answering respondent questions, and encompasses extensive role-playing opportunities. All interviewers receive a manual that explains each question in detail and provides detailed responses to questions often asked by respondents.
- Telephone interview supervisors and specially-trained staff monitor CATI interviews and provide feedback regarding verbatim reading of questions, recording of responses, interaction with respondents, and other issues.
- Field supervisors and specially-trained staff implement a quality reinterview program with CAPI respondents to minimize falsification of data.
- The CATI/CAPI instruments include a Spanish version, and bilingual interviewers provide language support in other languages.

Note that many of these methods are the same as those used to minimize nonresponse error. Methods that make it easier for the respondent to understand the questions also increase the chances that the individual will respond to the questionnaire.

15.5 PROCESSING ERROR

The final component of nonsampling error is processing error—error introduced in the postdata collection process of turning the responses into published data. For example, a processing error may occur in keying the data from the mail questionnaires. The miscoding of write-in responses, either clerically or by automated methods, is another example. The degree to which imputed data differ from the truth also reflects processing error—specifically imputation error. A number of practices are in place to control processing error (more details are discussed in Chapters 7 and 10). For example:

- Data capture of mail questionnaires includes a quality control procedure designed to ensure the accuracy of the final keyed data.
- Clerical coding includes a quality control procedure involving double-coding of a sample of the cases and adjudication by a third keyer.
- By design, automated coding systems rely on manual coding by clerical staff to address the most difficult or complicated responses.
- Procedures for selecting one interview or return from multiple returns for an address rely on a review of the quality of data derived from each response and the selection of the return with the most complete data.
- After completion of all three phases of data collection (mail, CATI, and CAPI), questionnaires with insufficient data do not continue in the survey processing, but instead receive a noninterview code and are accounted for in the weighting process.
- Edit and imputation rules reflect the combined efforts and knowledge of subject matter experts, as well as experts in processing, and include evaluation and subsequent improvements as the survey continues to progress.
- Subject matter and survey experts complete an extensive review of the data and tables, comparing results with previous years' data and other data sources.

15.6 REFERENCES

Biemer, P., and L. Lyberg. (2003) *Introduction to Survey Quality*, Hoboken, NJ: John Wiley and Sons.

Dillman, D. (1978) *Mail and Telephone Surveys: The Total Design Method*, New York: John Wiley and Sons.

Groves, R. M. (1989) *Survey Errors and Survey Costs*, New York: John Wiley and Sons.

Groves, R. M., M. P. Couper, F. J. Fowler, J. M. Lepkowski, E. Singer, and R. Tourangeau. (2004) *Survey Methodology*, Hoboken, NJ: John Wiley and Sons.

Tourangeau, R., and T. Yan. (2007) "Sensitive questions in surveys," *Psychological Bulletin*, 133(5): 859–883.

U.S. Census Bureau (2002) "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: May 2002, Report 2: Demonstrating Survey Quality," Washington, DC.

U.S. Census Bureau (2004) "Meeting 21st Century Demographic Data Needs—Implementing the American Community Survey: Report 7: Comparing Quality Measures: The American Community Survey's Three-Year Averages and Census 2000's Long Form Sample Estimates," Washington, DC.

U.S. Census Bureau (2006) "Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses Version 1.2," Washington, DC.

Acronyms

AIANHH	American Indian Area/Alaska Native Area/Hawaiian Homeland
AIANSA	American Indian/Alaska Native Village Statistical Area
ACS	American Community Survey
ACSO	American Community Survey Office
AFF	American FactFinder
ALMI	Automated Listing and Mapping Instrument
ART	Automated Review Tool
ART II	Revised Automated Review Tool
ASA	American Statistical Association
BAS	Boundary and Annexation Survey
BOP	Bureau of Prisons
C2SS	Census 2000 Supplementary Survey
CAPI	Computer-Assisted Personal Interviewing
CATI	Computer-Assisted Telephone Interviewing
CAUS	Community Address Updating System
CCD	Census County Division
CDP	Census Designated Place
CIC	Census Information Center
CM	Continuous Measurements
CPI	Consumer Price Index
CPS	Current Population Survey
CQR	Count Question Resolution
CV	Coefficient of Variation
DAAL	Demographic Area Address Listing
DCF	Data Capture File
DOC	Department of Commerce
DOT	Department of Transportation
DRB	Disclosure Review Board

DSCMO	Decennial Systems and Contract Management Office
DSF	Delivery Sequence File
EEOC	Equal Employment Opportunity Commission
EMMA	Edit Management and Messaging Application
FAIP	Federal Agency Information Program
FEFU	Failed Edit Follow-Up
FR	Field Representative
FSCPE	Federal-State Cooperative Program for Population Estimates
FTP	File Transfer Protocol
GAO	Government Accountability Office
GPO	Government Printing Office
GQ	Group Quarters
GQFQ	Group Quarters Facility Questionnaire
GQMOS	Group Quarters Measure of Size
GQNF	Group Quarters Noninterview Factor
GUMOS	Government Unit Measure of Size
HU	Housing Unit
HUD	Department of Housing and Urban Development
IPE	Intercensal Population Estimates
IVR	Interactive Voice Recognition
KFI	Key-From-Image
KFP	Key-From-Paper
LUCA	Local Update of Census Addresses
MAF	Master Address File
MAFGOR	Master Address File Geocoding Office Resolution
MBF	Mode Bias Factor
MCD	Minor Civil Division
MOS	Measure of Size
NAS	National Academy of Sciences
NHIS	National Health Interview Survey
NPC	National Processing Center
OMB	Office of Management and Budget

OMR	Optical Mark Recognition
PAPI	Paper and Pencil Interviewing
PDF	Portable Document Format
PIO	Public Information Office
PMMA	PUMS Management and Messaging Application
POP	Population Division
PRA	Paperwork Reduction Act
PRCS	Puerto Rico Community Survey
PUMA	Public Use Microdata Area
PUMS	Public Use Microdata Sample
RA	Remote Alaska
RO	Regional Offices
SDC	State Data Center
SDR	Successive Differences Replication
SIPP	Survey of Income and Program Participation
SP	Special Place
SSI	Supplemental Security Income
SSS	Special Sworn Status
TDD	Telephone Device for the Deaf
TEFU	Telephone Edit Follow-Up
TIGER®	Topologically Integrated Geographic Encoding and Referencing
TQA	Telephone Questionnaire Assistance
UA	Urbanized Area
UAA	Undeliverable As Addressed
UNECE	United Nations Economic Commission for Europe
US	United States
USC	United States Code
USDA	United States Department of Agriculture
USPS	United States Postal Service
WDS	Web Data Server

Glossary of Terms

100 Percent Data. A term used in 2000 to describe the data that were asked of "100 percent" of the population in Census 2000. That is, questions that were collected for all people on both the census short-form and long-form questionnaires. In 2000, this included sex, relationship, age/date of birth, Hispanic origin, race, and tenure.

Accessibility. One of four key dimensions of survey quality, accessibility refers to the ability of the data users to readily obtain and use survey products.

Acceptability Index. The average number of basic ACS items reported per person, including sex, age (counted double), relationship, marital status, Hispanic origin, and race. A questionnaire for an occupied unit must have an acceptability index of 2.5 or greater to be considered an interview.

Accuracy. One of four key dimensions of survey quality. Accuracy refers to the difference between the survey estimate and the true (unknown) value. Attributes are measured in terms of sources of error (for example, coverage, sampling, nonresponse, measurement, and processing).

Address Control File. The residential address list used in the 1990 census to label questionnaires, control the mail response check-in operation, and determine the nonresponse follow-up workload.

Address Corrections from Rural Directories. A post-Census 2000 Master Address File (MAF) improvement operation where Census Bureau staff reviewed commercial directories for 300 rural counties in 10 Midwestern states to obtain new city-style addresses for MAF records that did not contain a city-style address. Conducted in 2002, over 15,000 city-style addresses were associated with MAF records that previously lacked a city-style address.

Address Listing. A Census 2000 field operation to develop the address list in areas with predominantly non-city-style mailing addresses. A lister captured the address and/or a physical/location description for each living quarters within a specified assignment area. The lister marked the location of each residential structure on a block map by placing a spot on the map indicating its location and assigning a map spot number. The lister also updated and corrected features on the map if necessary. This activity was called "prelist" in the 1990 census.

Administrative Entities. Geographic areas, usually with legally defined boundaries but often without elected officials, created to administer elections and other governmental functions. Administrative areas include school districts, voting districts, ZIP codes, and nonfunctioning Minor Civil Divisions (MCDs) such as election precincts, election districts, and assessment districts.

Allocation. Imputation method required when values for missing or inconsistent items cannot be derived from the existing response record. In these cases, the imputation must be based on other techniques such as using answers from other people in the household, other responding housing units, or people believed to have similar characteristics. Such donors are reflected in a table referred to as an allocation matrix.

American Community Survey (ACS) Alert. This periodic electronic newsletter informs data users and other interested parties about news, events, data releases, congressional actions, and other developments associated with the ACS.

American Community Survey Demonstration Program. The full set of testing, research, and development program activities that started in 1994 and continued until the ACS was fully implemented in 2005.

American Community Survey Full Implementation. The period beginning in January 2005 during which the ACS interviewing of its housing unit sample was conducted in every county and Puerto Rico *municipio* as well as all American Indian and Alaska Native Areas and Hawaiian Homelands. The full implementation initial sample size is approximately 3 million addresses each year, and includes group quarters (GQ) facilities which were added beginning in January 2006.

American Community Survey Test Sites. The ACS demonstration program expanded from an initial four test counties in 1996 to 36 test counties in 1999. When the term ACS test site is used, it refers to data from these 36 counties.

American FactFinder (AFF). An electronic system for access and dissemination of Census Bureau data on the Internet. The system offers prepackaged data products and user-selected data tables and maps from Census 2000, the 1990 Census of Population and Housing, the 1997 and 2002 Economic Censuses, the Population Estimates Program, annual economic surveys, and the ACS.

American Indian Area, Alaska Native Area, Hawaiian Homeland (AIANAHH). A Census Bureau term referring to the following types of areas: federal and state American Indian reservations, American Indian off-reservation trust land areas (individual or tribal), Oklahoma tribal statistical areas (in 1990 tribal jurisdictional statistical area), tribal designated statistical areas, state designated American Indian statistical areas, Alaska Native Regional Corporations, Alaska Native village statistical areas, and Hawaiian Homelands.

Assignment. Imputation method in which values for a missing or inconsistent item can be derived from other responses from the sample housing unit or person. For example, a first name can be used to determine and assign the sex of a person.

Automated Address Unduplication. An ongoing MAF improvement activity completed twice a year (coinciding with the delivery sequence file (DSF) refresh of the MAF) where, through automated means, pairs of city-style addresses are identified as identical based on house number, street name, five-digit ZIP code, and within structure identifier (if one exists). These addresses are linked for future operations to control duplication.

Automated Clerical Review. The ACS program run on raw mail return data to determine whether or not a case goes to failed-edit follow-up. The name reflects the fact that it was originally done clerically. The operator checks for missing content and for large households (more than five members) and for coverage inconsistencies.

Automated Editing. Editing that is accomplished using software, as opposed to being done clerically.

Automated Listing and Mapping Instrument (ALMI). Software used primarily by Census Bureau field representatives for the purpose of locating an address or conducting an address listing operation. The ALMI combines data from the MAF and the Topologically Integrated Geographic Encoding and Referencing (TIGER®) System database to provide users with electronic maps and associated addresses. ALMI functionality allows users to edit, add, delete, and verify addresses, streets, and other map features, view a list of addresses associated with a selected level of geography, and view and denote the location of housing units on the electronic map.

Automated Review Tool (ART). A Web-based computer application designed to help subject matter analysts quickly review and approve ACS estimates.

Automated Review Tool II (ART II). The next generation of the ART. It is aimed at providing analysts with reports at a more detailed level than the previous version.

Base Tables. Tables that provide the most detailed estimates on all topics and geographic areas from the ACS. Base tables also include totals and subtotals. These tables form the data source for the "Derived Products." Base tables are also known as detailed tables.

Base Weight. The base weight for an address is equal to the inverse of the probability with which the address was selected for the sample as determined by the sample design. Since these weights are based only on the initial probability of selection, they are known as *a priori* to the data collection phase. This is the weight for a housing unit before any adjustments are made. The base weight is also known as the unbiased weight.

Be Counted Enumeration and Be Counted Questionnaire. The Be Counted program provided a means for people who believed they were not counted to be included in Census 2000. The Census Bureau placed Be Counted questionnaires at selected sites that were easily accessible to and frequented by large numbers of people. The questionnaires also were distributed by the Questionnaire Assistance Centers and in response to requests received through Telephone Questionnaire Assistance.

Blaise. An authoring application that produces an instrument used to collect data using computer-assisted telephone interviewing (CATI) or computer-assisted personal interviewing (CAPI).

Block. A subdivision of a census tract (or, prior to 2000, a block numbering area), a block is the smallest geographic entity for which the Census Bureau tabulates decennial census data. Many blocks correspond to individual city blocks bounded by streets, but blocks—especially in rural areas—may include many square miles and may have some boundaries that are not streets. The Census Bureau established blocks covering the entire nation for the first time in 1990. Previous censuses back to 1940 had blocks established only for part of the nation. Over 8 million blocks were identified for Census 2000.

Block Canvassing. A Census 2000 field operation to ensure the currency and completeness of the MAF within the mailout/mailback area. Listers traveled in their assignment areas to collect and verify information to ensure that their address listing pages (derived from the MAF) contained a mailing address for every living quarters. They especially looked for hidden housing units (such as attics, basements, or garages converted into housing units) and houses that appeared to be one unit but actually contained multiple housing units. They also updated and corrected their Census Bureau maps.

Block Group. A subdivision of a census tract (or, prior to 2000, a block numbering area), a block group is a cluster of blocks having the same first digit of their four-digit identifying number within a census tract.

Boundary and Annexation Survey (BAS). An annual survey of all counties and statistically equivalent entities, all or selected incorporated places and minor civil divisions, all or selected federally recognized American Indian reservations and off-reservation trust lands, and Alaska Native Regional Corporations, to determine the location of legal limits and related information as of January 1 of the survey year.

Case Management. A tool used by field representatives that allows them to manage their interview assignments on their laptops.

Census 2000 Supplementary Survey (C2SS). The C2SS was an operational test conducted as part of the research program in Census 2000, and used the ACS questionnaire and methods to collect demographic, social, economic, and housing data from a national sample. This evaluation study gave the Census Bureau essential information about the operational feasibility of converting from the census long-form sample to the ACS.

Census County Division (CCD). A subdivision of a county that is a relatively permanent statistical area established cooperatively by the Census Bureau and state and local government authorities. Used for presenting decennial census statistics in those states that do not have well-defined and stable minor civil divisions that serve as local governments.

Census Designated Place (CDP). A statistical entity that serves as a statistical counterpart of an incorporated place for the purpose of presenting census data for a concentration of population, housing, and commercial structures that is identifiable by name, but is not within an incorporated place. CDPs usually are delineated cooperatively with state, Puerto Rico, Island Area, local, and tribal government officials, based on the Census Bureau guidelines. For Census 2000, CDPs did not have to meet a population threshold to quality for the tabulation of census data.

Census Geography. A collective term referring to the types of geographic areas used by the Census Bureau in its data collection and tabulation operations, including their structure, designations, and relationships to one another.

Census Information Center (CIC). The CIC program is a cooperative activity between the Census Bureau and the national nonprofit organizations representing interests of racial and ethnic communities. The program objective is to make census information and data available to the participating organizations for analysis, policy planning, and for further dissemination through a network of regional and local affiliates.

Census Sample Data. Population and housing information collected only on the census long form for a sample of households.

Census Tract. A small, relatively permanent statistical subdivision of a county delineated by a local committee of census data users for the purpose of presenting data. Census tract boundaries normally follow visible features, but may follow governmental unit boundaries and other nonvisible features; they always nest within counties. Designed to be relatively homogeneous units with respect to population characteristics, economic status, and living conditions at the time of establishment, census tracts average about 4,000 inhabitants.

City-Style Address. An address that consists of a house number and street or road name; for example, 201 Main Street. The address may or may not be used for the delivery of mail, and may include apartment numbers/designations or similar identifiers.

Coding. The process of associating numeric codes with write-in strings. For example, the write-in associated with Place of Birth is turned into a three-digit code.

Coefficient of Variation (CV). The ratio of the standard error (square root of the variance) to the value being estimated, usually expressed in terms of a percentage (also known as the relative standard deviation). The lower the CV, the higher the relative reliability of the estimate.

Cold Deck Values. The values used to initialize matrices used for hot-deck allocation.

Collapsing. Reducing the amount of detail shown in a base table to comply with data release rules.

Community Address Updating System (CAUS). A post-Census 2000 MAF improvement program that provides a systematic methodology for enhancement and update of address and feature information. Designed to provide a rural counterpart to the update of the city-style addresses received from the U.S. Postal Service's Delivery Sequence File, CAUS identifies and conducts listing operations in selected geographic areas suspected of experiencing growth that is either not available from or appears to be incomplete in the U.S. Postal Service's Delivery Sequence File. Address and feature updates collected for CAUS are added to the MAF and the TIGER® System.

Comparison Profile. Comparison profiles are available from the ACS for 1-year estimates beginning in 2007. These tables are available for the United States, the 50 states, the District of Columbia, and geographic areas with a population of more than 65,000.

Complete Interview. The ACS interview is classified as complete when all applicable questions have been answered on the mail form, or during a CATI or CAPI interview. The interview may include responses of "Don't Know" and "Refused" to specific questions.

Computer-Assisted Personal Interviewing (CAPI). A method of data collection in which the interviewer asks questions displayed on a laptop computer screen and enters the answers directly into a computer.

Computer-Assisted Telephone Interviewing (CATI). A method of data collection using telephone interviews in which the questions to be asked are displayed on a computer screen and responses are entered directly into a computer.

Confidence Interval. The sample estimate and its standard error permit the construction of a confidence interval that represents the degree of uncertainty about the estimate. Each ACS estimate is accompanied by the upper and lower bounds of the 90 percent confidence interval, or the 90 percent margin of error, from which a confidence interval can be constructed. A 90 percent confidence interval can be interpreted roughly as providing 90 percent certainty that the interval defined by the upper and lower bounds contains the true value of the characteristic.

Confidentiality. The guarantee made by law (Title 13, United States Code) to individuals who provide census information, regarding nondisclosure of that information to others.

Congressional Tool Kit. A collection of documents developed for members of Congress that explain how and why the ACS is conducted, its benefits, and how to obtain additional information. The Tool Kit originally was distributed as hard copies in 3-ring binders and is now available as a series of online portable document format (PDF) files.

Consumer Price Index (CPI). The CPI program of the Bureau of Labor Statistics produces monthly data on changes in the prices paid by urban consumers for a representative basket of goods and services.

Control File. A file which represents the current status of any case in sample in the ACS.

Controlled. During the ACS weighting process, the intercensal population and housing estimates are used as survey controls. Weights are adjusted so that ACS estimates conform to these controls.

Count Question Resolution (CQR). A process followed in Census 2000 whereby state, local, and tribal government officials could ask the Census Bureau to verify the accuracy of the legal boundaries used for Census 2000, the allocation of living quarters and their residents in relation to those boundaries, and the count of people recorded by the Census Bureau for specific living quarters.

Cross Tabulation. The joint distribution of two or more data characteristics, where each of the categories of one characteristic is repeated for each of the categories of the other characteristic(s). A cross-tabulation in a base table is denoted where "BY" is used as the conjunction between characteristics; for example, "AGE BY SEX" or "AGE BY SEX BY RACE."

Current Population Survey (CPS). Monthly sample survey of the U.S. population that provides employment and unemployment estimates as well as current data about other social and economic characteristics of the population. Collected for the Bureau of Labor Statistics by the Census Bureau.

Current Residence. The concept used in the ACS to determine who should be considered a resident of a sample address. Everyone who is currently living or staying at a sample address is considered a resident of that address, except people staying there for two months or less. People who have established residence at the sample address and are away for only a short period of time are also considered to be current residents.

Custom Tabulations. The Census Bureau offers a wide variety of general purpose data products from the ACS. These products are designed to meet the needs of the majority of data users and contain predefined sets of data for standard census geograpic areas, including both political and statistical geography. These products are available on the American FactFinder and the ACS Web site.

For users with data needs not met through the general purpose products, the Census Bureau offers "custom" tabulations on a cost-reimbursable basis, with the ACS Custom Tabulation program. Custom tabulations are created by tabulating data from ACS microdata files. They vary in size, complexity, and cost depending on the needs of the sponsoring client.

Data Capture File. The repository for all data captured from mail return forms and by CATI and CAPI Blaise instruments.

Data Collection Mode. One of three ACS methods (mail, telephone, personal visit) of data collection.

Data Profiles. Data products containing estimates of key demographic, social, economic, and housing characteristics. Data swapping is done by editing the source data or exchanging records for a sample of cases. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics. Because the swap often occurs within a neighboring area, there is usually no effect on the marginal totals for the area or for totals that include data from multiple areas.

De Facto Residence Rules. De facto means "in fact." A de facto residence rule would define survey residents as all people living or staying at the sample address at the time of the interview without considering other factors such as the amount of time they will be staying there. Such a rule would exclude people away from a regular residence even if they were away for only that one day. The ACS is using a de facto residence rule when determining the residents of GQ facilities eligible to be sampled and interviewed for the survey.

Delivery Sequence File (DSF). A U.S. Postal Service (USPS) computer file containing all mailing addresses serviced by the USPS. The USPS continuously updates the DSF as its letter carriers identify addresses for new delivery points and changes in the status of existing addresses. The Census Bureau uses the DSF as a source for maintaining and updating its MAF.

Demographic Area Address Listing (DAAL). A post-Census 2000 program associated with coverage improvement operations, address list development, and automated listing for the CAUS and demographic household surveys. The program uses automated listing methods to update the inventory of living quarters, and also updates the street network in selected blocks.

Derived Products. Derived products are informational products based largely on estimates from the base tables.

Detailed Tables. See **Base Tables**.

Disclosure Avoidance (DA). Statistical methods used in the tabulation of data prior to releasing data products to ensure the confidentiality of responses. See **Confidentiality**.

Disclosure Review Board (DRB). A board comprised of Census Bureau staff who review and must approve all data products based on disclosure avoidance rules before they can be released to the public.

Edit. To subject data to program logic to check for missing data and inconsistencies.

Edit Management and Messaging Application (EMMA). An Internet application used by ACS subject-matter analysts to show the status of edit review and to relay analyst's relevant comments.

Estimates. Numerical values obtained from a statistical sample and assigned to a population parameter. Data produced from the ACS interviews are collected from samples of housing units. These data are used to produce estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology.

Evaluation Studies. Research and evaluation conducted by Census Bureau staff and external experts to assess a broad set of topics including the feasibility and the quality of the data products produced by the ACS.

Failed Edit Follow-Up (FEFU). Data collection activity of mail response records designed to collect missing information. Mail returns failing the automated clerical review edit are contacted by telephone.

Federal Agency Information Program (FAIP). A long-term program of information and technical partnership with federal agencies. The FAIP is designed to establish a relationship with each agency that will identify the unique opportunities and challenges it faces in using ACS data. The program targets assistance based on the needs and resources of each federal agency in order to help the agency make a smooth transition to using ACS data.

Federal Government Unit (FGU). Any of a variety of civil divisions; places and is used for sampling.

Federal Register Notice. Published by the Office of the Federal Register, National Archives and Records Administration (NARA), the *Federal Register* is the official daily publication for rules, proposed rules, and notices of federal agencies and organizations, as well as executive orders and

other presidential documents. Information describing proposed data collection must be posted on the *Federal Register* for public review and comment for a 30-day period and must take place before the Office of Management and Budget (OMB) can provide final clearance for the data collection.

Federal-State Cooperative Program for Population Estimates (FSCPE). FSCPEs are state-level organizations, designated by their respective governors, to work cooperatively with the Census Bureau's Population Estimates Program in the production of subnational population estimates and in making data broadly available to the public.

Field Representative (FR). A Census Bureau employee who interviews people to obtain information for a census or survey.

File Transfer Protocol (FTP). A process that allows a user to download large files and datasets from American FactFinder.

Final Outcome Code. A code assigned to a CATI or CAPI case at the conclusion of the data collection which characterizes the status of the case, such as "completed occupied interview" or "respondent refusal noninterview."

First Stage Sample. ACS first stage sampling maintains five 20 percent partitions of the MAF by determining which addresses were in the first stage sample 4 years prior and excluding them. This ensures that no address is in sample more than once in any 5-year period. The first phase sample is the universe from which the second phase sample is selected.

Five-Year Estimates. Estimates based on 5 years of ACS data. These estimates are meant to reflect the characteristics of a geographic area over the entire 60-month period and will be published for all geographic areas down to the census block group level.

Functioning Governmental Unit (FGU). A general purpose government that has the legal capacity to elect or appoint officials, raise revenues, provide surveys, and enter into contracts.

General Coding. The process whereby write-in answers to Hispanic origin, race, ancestry, and language are categorized into codes. This is accomplished using an automated system approach, relying on a set of growing dictionaries of write-ins against which responses are computer matched. Responses that are not found in the dictionaries are sent to subject matter experts who code them. These new responses are added to the computer dictionaries for subsequent use.

Geocoding. The assignment of an address, structure, key geographic location, or business name to a location that is identified by one or more geographic codes. For living quarters, geocoding usually requires identification of a specific census block.

Geographic Summary Level. A geographic summary level specifies the content and the hierarchical relationships of the geographic elements that are required to tabulate and summarize data. For example, the county summary level specifies the state-county hierarchy. Thus, both the state code and the county code are required to uniquely identify a county in the United States or Puerto Rico.

Government Printing Office (GPO). A federal agency responsible for producing, procuring, and disseminating printed and electronic publications of the Congress as well as the executive departments and establishments of the federal government.

Governmental Unit Measure of Size (GUMOS). The smallest measure of size associated with a given block. It is used in the sample selection operation to determine the initial sampling rate at the block level.

Group Quarters (GQ) Facilities. A GQ facility is a place where people live or stay that is normally owned or managed by an entity or organization providing housing and/or services for the residents. These services may include custodial or medical care, as well as other types of assistance. Residency is commonly restricted to those receiving these services. People living in GQ facilities are usually not related to each other. The ACS collects data from people living in both housing units and GQ facilities.

Group Quarters Facilities Questionnaire (GQFQ). A Blaise-based automated survey instrument that field representatives (FRs) use to collect new or updated information about a GQ facility. Questions in this survey include facility name, mailing and physical address, telephone number, GQ contact name and telephone number, special place name, and the GQ facility's maximum occupancy and current number of people staying in the GQ facility.

Group Quarters Geocoding Correction Operation. A post-Census 2000 MAF improvement operation implemented to correct errors (mostly census block geocodes) associated with college dormitories in MAF and TIGER®. Conducted by Census Bureau staff, source materials for over 20,000 dormitories were reviewed and used to identify and correct MAF/TIGER® errors.

Group Quarters Listing Sheet. This form is preprinted with information such as GQ name and control number for sample GQ facilities. It is used by FRs when the GQ administrator is unable to provide a list of names or occupied bed locations for person-level sample selection.

Group Quarters Measure of Size (GQMOS). The expected population of a given GQ facility divided by 10. It is used in the sample selection operation to determine the universe of sample units to be sampled. A sample unit is a cluster or group of 10 people.

Hot Deck Imputation. An approach for filling in missing answers with information from like households or persons, with donors determined by geographic location or specific characteristics reported. Hot deck imputation continually updates matrices with data from donors with acceptable data and then provides values from such matrices to recipients who need data.

Household. A household includes all the people who occupy a housing unit that meet all the residence rules of a survey or census.

Housing Unit (HU). A house, apartment, mobile home or trailer, a group of rooms, or a single room occupied as separate living quarters, or if vacant, intended for occupancy as separate living quarters. Separate living quarters are those in which the occupants live separately from any other individuals in the building and have direct access from outside the building or through a common hall. For vacant units, the criteria of separateness and direct access are applied to the intended occupants whenever possible.

Imputation. When information is missing or inconsistent, the Census Bureau uses imputation methods to assign or allocate values. Imputation relies on the statistical principle of "homogeneity," or the tendency of households within a small geographic area to be similar in most characteristics.

Interactive Voice Recognition (IVR). An automated telephone application which allows the caller to hear prerecorded responses to frequently asked questions. The caller may proceed through the application by entering numbers from the telephone key pad or by speaking responses to select which messages he/she wants to hear. The caller may also elect to speak to an interviewer instead of listening to the recorded responses.

Intercensal Estimates. Official Census Bureau estimates of the population of the United States, states, metropolitan areas, cities and towns, and counties; also official Census Bureau estimates of housing units (HUs).

Interim Codes. These are codes assigned to a sample GQ assignment in the GQFQ system by a field representative when scheduling a personal visit to a sample ACS GQ facility, when additional research is needed to locate the GQ facility, or when a return visit to the GQ facility is needed to obtain additional survey information.

Interpolation. Interpolation is frequently used in calculating medians or quartiles based on interval data and in approximating standard errors from tables. Linear interpolation is used to estimate values of a function between two known values. Pareto interpolation is an alternative to linear interpolation. In Pareto interpolation, the median is derived by interpolating between the logarithms of the upper and lower income limits of the median category.

Interview Monitoring. A process in which CATI supervisors, for quality control purposes, listen to interviewers while they are conducting interviews with respondents to assure that the interviewer is following all interviewing procedures correctly. The interviewer is not told when the supervisor is listening, but is given feedback on his/her performance after the monitoring.

Item Nonresponse. The failure to obtain valid responses or responses consistent with other answers for individual data items.

Iterations. Subgroups of the original tabulation universe, especially by race, Hispanic origin, ancestry, and tribal groups. For example, many ACS base tables are iterated by 9 race and Hispanic origin groups.

Joint Economic Edit. An edit which looks at the combination of multiple variables related to a person's employment and income, thereby maximizing the information used for filling any missing related variables.

Key-From-Image (KFI). An operation in which keyers use a software program to capture questionnaire responses by typing responses directly into the scanned image of a questionnaire displayed on their work station screen.

Key-From-Paper (KFP). An operation in which keyers use a software program to capture questionnaire responses from a hard-copy of the questionnaire.

Legal Entity. A geographic entity whose origin, boundary, name, and description result from charters, laws, treaties, or other administrative or governmental action, such as the United States, states, the District of Columbia, Puerto Rico, the Island Areas, counties, cities, boroughs, towns, villages, townships, American Indian reservations, Alaska Native villages, congressional districts, and school districts. The legal entities and their boundaries that the Census Bureau recognizes are those in existence on January 1 of each calendar year.

List/Enumerate. A method of decennial census data collection in some of the more remote, sparsely populated areas of the United States and the Island Areas, where many of the households do not have mail delivery to city-style addresses. Enumerators list the residential addresses within their assignment areas on blank address register pages, map spot the location of the residential structures on Census Bureau maps, and conduct an interview for each household.

Local Update of Census Addresses (LUCA). A Census 2000 program, established in response to requirements of Public Law 103-430, that provided an opportunity for local and tribal governments to review and update individual address information or block-by-block address counts from the MAF and associated geographic information in the TIGER® database. The goal was to improve the completeness and accuracy of both computer files. Individuals working with the addresses had to sign a confidentiality agreement before a government could participate. Also called the Address List Review Program.

Long Form. The decennial census long-form questionnaire was used to survey a sample of the U.S. population. It contained the questions on the census short form and additional detailed questions relating to the social, economic, and housing characteristics of each individual and household.

Lower Bound. Represents the low end of the 90 percent confidence interval of an estimate from a sample survey. A 90 percent confidence interval can be interpreted roughly as providing 90 percent certainty that the true number falls between the upper and lower bounds.

Mailing Address. The address used by a living quarters, special place, business establishment, and the like for mail delivery by the USPS. It can be a house number and street or road name, which may be followed by an apartment, unit, or trailer lot designation; a building or apartment complex name and apartment designation; a trailer park name and lot number; a special place/GQ facility name; a post office box or drawer; a rural route or highway contract route, which may include a box number; or general delivery. A mailing address includes a post office name, state abbreviation, and ZIP Code. A mailing address may serve more than one living quarters, establishment, and so on.

Mailout-Mailback. A method of data collection in which the USPS delivers addressed questionnaires to housing units. Residents are asked to complete and mail the questionnaire to a specified data capture center.

Main Phase Sample. The annual ACS sample is chosen in two phases. During the first phase, referred to as the main phase, approximately 98 percent of the total ACS sample is chosen. The main phase sample addresses are allocated to the 12 months of the sample year. The second phase, referred to as supplemental sample selection, is implemented to represent new construction.

Master Address File (MAF). The Census Bureau's official inventory on known living quarters (housing units and GQ facilities) and selected nonresidential units (public, private, and commercial) in the United States. The file contains mailing and location address information, geocodes, and other attribute information about each living quarters. The Census Bureau continues to update the MAF using the USPS DSF and various automated, computer-assisted, clerical, and field operations.

Master Address File Geocoding Office Resolution (MAFGOR). An operation in which census staff try to find the location of addresses from the USPS that did not match to the records in the TIGER® database. Staff use atlases, maps, city directories, and the like to locate these addresses and add their streets and address ranges to the TIGER® database.

Master Address File/TIGER® Reconciliation. A post-Census 2000 MAF improvement activity where census staff reviewed and corrected map spot inconsistencies in over 1,800 counties. Over 75,000 MAF records in nonmailout/mailback blocks were corrected. The most common types of MAF corrections were the assignment of map spots to MAF records such that they are consistent with the TIGER® database, and the identification and linkage of duplicate MAF records.

Margin of Error (MOE). Some ACS products provide an MOE instead of confidence intervals. An MOE is the difference between an estimate and its upper or lower confidence bounds. Confidence bounds can be created by adding the MOE to the estimate (for the upper bound) and subtracting the MOE from the estimate (for the lower bound). All published ACS MOEs are based on a 90 percent confidence level.

Measure of Size (MOS). A generic term used to refer to the estimated size of a specific administrative or statistical area. It is used in the sample selection operation to determine the initial sampling rate at the block level.

Measurement Error. Also referred to as "response error," measurement error occurs when the response received differs from the "true" value as a result of the respondent, the interviewer, the questionnaire, the mode of collection, the respondent's record-keeping system(s) or other similar factors.

Median. This measurement represents the middle value (if n is odd) or the average of the two middle values (if n is even) in an ordered list of data values. The median divides the total frequency distribution into two equal parts: one-half of the cases fall below the median and one-half of the cases exceed the median. Medians in the ACS are estimated using interpolation methods.

Metadata. Information about the content, quality, condition, and other characteristics of data. Metadata related to tables presented in American FactFinder can be found by clicking on column headings or by clicking "Help" and then "Census Data Information."

Minor Civil Division (MCD). A primary governmental and/or administrative subdivision of a county, such as a township, precinct, or magisterial district. MCDs exist in 28 states and the District of Columbia. In 20 states, all or many MCDs are general-purpose governmental units: Connecticut, Illinois, Indiana, Kansas, Maine, Massachusetts, Michigan, Minnesota, Missouri, Nebraska, New Hampshire, New Jersey, New York, North Dakota, Ohio, Pennsylvania, Rhode Island, South Dakota, Vermont, and Wisconsin. Most of these MCDs are legally designated as towns or townships.

Multiyear Estimates. Three- and five-year estimates based on multiple years of ACS data. Three-year estimates will be published for geographic areas with a population of 20,000 or more. Five-year estimates will be published for all geographic areas down to the census block group level.

Municipio. Primary legal divisions of Puerto Rico. These are treated as county equivalents.

Narrative Profile. A data product that includes easy-to-read descriptions for a particular geography.

National Processing Center (NPC). The permanent Census Bureau processing facility in Jeffersonville, Indiana. Until 1998, it was called the Data Preparation Division.

Non-City-Style Address. A mailing address that does not use a house number and street or road name. This includes rural routes and highway contract routes, which may include a box number; post office boxes and drawers; and general delivery.

Noninterview/Nonresponse. A sample address which was eligible for an interview, but from which no survey data was obtained.

Nonresponse Error. Error caused by survey failure to get a response to one or possibly all of the questions. Nonresponse error is measured in the ACS by survey response rates and item nonresponse rates.

Nonresponse Follow-Up. An operation whose objective is to obtain complete survey information from housing units for which the Census Bureau did not receive a completed questionnaire by mail. In the ACS, telephone and personal visit methods are used for nonresponse follow-up.

Nonsampling Error. Total survey error can be classified into two categories—sampling error and nonsampling error. Errors that occur during data collection (for example, nonresponse error, response error, and interviewer error) or data capture fall under the category of nonsampling error.

Office of Management and Budget (OMB). OMB assists the President in the development and execution of policies and programs. OMB has a hand in the development and resolution of all budget, policy, legislative, regulatory, procurement, e-government, and management issues on behalf of the President. OMB is composed of divisions organized either by agency and program area or by functional responsibilities. However, the work of OMB often requires a broad exposure to issues and programs outside of the direct area of assigned responsibility. In accordance with the Paperwork Reduction Act of 1995, the Census Bureau submits survey subjects, questions, and information related to sampling, data collection methods, and tabulation of survey data to OMB for approval and clearance.

Operational Response Rates. Response rates for data collection operations conducted in the ACS—Mail, CATI, CAPI, and FEFU operations.

Optical Mark Recognition (OMR). Technology that uses a digital image of a completed questionnaire and computer software to read and interpret the marking of a response category and to convert that mark into an electronic response to the survey question.

Overcoverage. Extent to which a frame includes units from the target population more than once, giving the unit multiple chances of selection, as well as the extent to which the frame includes units that are not members of the target population.

Period Estimates. An estimate based on information collected over a period of time. For ACS the period is either 1 year, 3 years, or 5 years.

Point-in-Time Estimates. An estimate based on one point in time. The decennial census long-form estimates for Census 2000 were based on information collected as of April 1, 2000.

Population Controls. Intercensal estimates used in weighting ACS sample counts to ensure that ACS estimates of total population and occupied housing units agree with official Census Bureau estimates.

Primary Sampling Unit (PSU). The PSU for the housing unit sample selection is the address. For the GQ sample selection it is groups of ten expected interviews. For the small GQ sample selection operation it is the GQ facility. All residents of small GQ facilities in sample are included in the person sample.

Processing Error. Error introduced in the postdata collection process of taking the responses from the questionnaire or instrument and turning those responses into published data. Thus, processing error occurs during data capture, coding, editing, imputation, and tabulation.

Public Use Microdata Area (PUMA). An area that defines the extent of territory for which the Census Bureau releases Public Use Microdata Sample (PUMS) records.

Public Use Microdata Sample (PUMS) Files. Computerized files that contain a sample of individual records, with identifying information removed, showing the population and housing characteristics of the units and people included on those forms.

Public Use Microdata Sample (PUMS) Management and Messaging Application (PMMA). This system is the PUMS version of EMMA, and is used by analysts to communicate with the data processing team about their review of the PUMS files.

Puerto Rico Community Survey (PRCS). The counterpart to the ACS that is conducted in Puerto Rico.

Quality Assurance (QA). The systematic approach to building accuracy and completeness into a process.

Quality Control (QC). Various statistical methods that validate that products or operations meet specified standards.

Quality Index. A measure of the quality of a particular return which is used when there are multiple returns for a particular sample unit.

Quality Measures. Statistics that provide information about the quality of the ACS data. The ACS releases four different quality measures with the annual data release: 1) initial sample size and final interviews; 2) coverage rates; 3) response rates, and; 4) item allocation rates for all collected variables.

Raking. An iterative procedure whereby a series of ratio adjustments are performed and then repeated. Each ratio adjustment corresponds to a dimension of the raking matrix. The goal of the procedure is to achieve a high degree of consistency between the weighted marginal totals and the control totals used in the ratio adjustment. The raking ratio estimator is also known as iterative proportional fitting.

Ranking Table. Ranking tables are tables and related graphics that show the rank order of a key statistic or derived measure across various geographic areas, currently states, counties, and places.

Recodes. Variables on data files that are the result of combining values from more than one variable.

Reference Period. Time interval to which survey responses refer. For example, many ACS questions refer to the day of the interview; others refer to “the past 12 months” or “last week.”

Regional Office (RO). One of 12 permanent Census Bureau offices established for the management of all census and survey operations in specified areas.

Relevance. One of four key dimensions of survey quality. Relevance is a qualitative assessment of the value contributed by the data. Value is characterized by the degree to which the data serve to address the purposes for which they are produced and sought by users (including mandate of the agency, legislated requirements, and so on.)

Remote Alaska. Rural areas in Alaska which are difficult to access. In these areas, all ACS sample cases are interviewed using the personal visit mode. Field representatives attempt to conduct interviews for all cases in specific areas of remote Alaska during a single visit. All sample cases in remote Alaska are interviewed in either January through April or September through December.

Residence Rules. The series of rules that define who (if anyone) is considered to be a resident of a sample address for purposes of the survey or census.

Respondent. The person supplying survey or census information about his or her living quarters and its occupants.

Respondent Errors. The respondent's failure to provide the correct answer to a survey question for any reason, such as poor comprehension of the question meaning, low motivation to answer the question, inability to retrieve the necessary information, or an unwillingness to answer the question truthfully.

Response Categories. The response options for a particular survey question shown on the paper questionnaire, read to the respondent in a CATI interview or read or presented on a flashcard to the respondent in a CAPI interview.

Response Errors. Also referred to as measurement error, response error is any error that occurs during the data collection stage of a survey resulting in a deviation from the true value for a given survey question or questions. Errors made by respondents, interviewer errors such as misreading a question or guiding the response to a particular category, and poorly designed data collection instruments or questionnaires all contribute to response error.

Rolling Sample. A rolling sample design jointly selects k nonoverlapping probability samples, each of which constitutes $1/F$ of the entire population. One sample is interviewed each time period until all of the sample has been interviewed after k periods.

Sample Month. The first month of a sample's 3-month interview period.

Sampling Entity. Geographic and statistical entities eligible to be used in determining the sampling strata assignment.

Sampling Error. Errors that occur because only part of the population is directly contacted. With any sample, differences are likely to exist between the characteristics of the sampled population and the larger group from which the sample was chosen.

Sampling Frame. Any list or device that, for purposes of sampling, delimits, identifies, and allows access to the sampling units, which contain elements of the sampled population. The frame may be a listing of persons, housing units, businesses, records, land segments, and so on. One sampling frame or a combination of frames may be used to cover the entire sampled population.

Sampling Rate. Proportion of the addresses in a geographical area, or residents of a GQ facility, who are selected for interview in a particular time period.

Sampling Variability. Variation that occurs by chance because a sample is surveyed rather than the entire population.

Second Stage Sample. The set of addresses selected from the first phase sample using a systematic sampling procedure. This procedure employs seven distinct sampling rates.

Selected Population Profiles (SPPs). An ACS data product that provides certain characteristics for a specific race or ethnic group (for example, Alaska Natives) or other population subgroup (for example, people aged 60 years and over). SPPS are produced directly from the sample microdata (that is, not a derived product).

Short Form. The decennial census short-form questionnaire includes questions on sex, age/date of birth, relationship, Hispanic origin, race, and tenure.

Single-Year Estimates. Estimates based on the set of ACS interviews conducted from January through December of a given calendar year. These estimates will be published for geographic areas with a population of 65,000 or more.

Size Thresholds. Population sizes of geographical areas that determine when data products will first be released for that area; for example, areas with 65,000 or greater populations will get single-year profiles in 2006 and every year thereafter; areas with 20,000 or greater populations will receive 3-year data products in 2008 and every year thereafter. There are no population size thresholds applied to the 5-year data products other than those imposed by the DRB.

Small Area Income and Poverty Estimates (SAIPE). Census Bureau program that prepares mathematical model-based estimates of selected characteristics of the United States, states, and school districts.

Special Census. A federal census conducted at the request and expense of a local governmental agency to obtain a population count between decennial censuses.

Special Place (SP). A special place is an entity that owns and/or manages one or more GQ facilities. A special place can be in the same building or location as the GQ facility or it can be at a different location than the GQ facility it manages or oversees.

Special Sworn Status (SSS) or Special Sworn Status (SSS) Individual. Individuals with SSS are defined as non-Census Bureau personnel who require access to census information or confidential data. An SSS individual is bound by Census Bureau confidentiality requirements, as authorized by Title 13, United States Code.

Standard Error. The standard error is a measure of the deviation of a sample estimate from the average of all possible samples.

State Data Center (SDC). A state agency or university facility identified by the governor of each state and state equivalent to participate in the Census Bureau's cooperative network for the dissemination of census data.

Statistical Areas. Defined and intended to provide nationally consistent definitions for collecting, tabulating, and publishing federal statistics for a set of geographic areas.

Statistical Significance. The determination of whether the difference between two estimates is not likely to be from random chance (sampling error) alone. This determination is based on both the estimates themselves and their standard errors. For ACS data, two estimates are "significantly different at the 90 percent level" if their difference is large enough to infer that there was a less than 10 percent chance that the difference came entirely from random variation.

Strata. See **Stratum.**

Stratum. A grouping or classification that has a similar set of characteristics.

Subsampling. Refers to the sampling of a sample. The cases that are not completed by mail or through a telephone interview become eligible for CAPI interviewing. This winnowing of the sample is referred to as subsampling.

Subject Tables. Data products organized by subject area that present an overview of the information that analysts most often receive requests for from data users.

Successive Differences Replication (SDR). A variance estimation methodology to be used for surveys with a systematic sample. The initial sampling weights are multiplied by sets of 80 predetermined factors, and then reprocessed through the weighting system to produce 80 new sets to replicate weights. The 80 replicate weights and the final production weights are used to estimate the variance of ACS estimates.

Sufficient Partial Interview. A sufficient partial interview means that the Census Bureau accepts an interview as final even if the respondent did not provide a valid response for all applicable items.

Summary File 3 (SF 3). This file presents base tables on population and housing characteristics from Census 2000 sample topics, such as income and education. It also includes population estimates for ancestry groups and selected characteristics for a limited number of race and Hispanic or Latino categories.

Summary File 4 (SF 4). This file presents data similar to the information included on Summary File 3. The data from Census 2000 are shown down to the census tract level for 336 race, Hispanic or Latino, American Indian and Alaska Native, and ancestry categories.

Supplemental Sample. The sample that is selected from new addresses (primarily new construction) and allocated to the last 9 months of the sample year. This is done in January of the sample year.

Survey. A data collection for a sample of a population. Surveys are normally less expensive to conduct than censuses, hence, they may be taken more frequently and can provide an information update between censuses.

Survey of Income and Program Participation (SIPP). A longitudinal survey conducted by the Census Bureau that collects data periodically from the same respondents over the course of several years. The SIPP produces data on income, taxes, assets, liabilities, and participation in government transfer programs.

Survey Quality. The four key elements of survey quality include relevance, accuracy, timeliness, and accessibility.

Survey Response Rates. A measure of total response across all three modes of data collection, calculated as the ratio of the estimate of the interviewed units to the estimate of all units that should have been interviewed. The ACS weights the survey response rate to reflect the sample design, including the subsampling for the CAPI.

Swapping. See **Data Swapping.**

Systematic Errors. Errors or inaccuracies occurring in data consistently in one direction, which can distort survey results. By definition, any systematic error in a survey will occur in all implementations of that same survey design.

Tabulation Month. The month associated with a sample case which is used in producing estimates. Also known as the Interview Month, it reflects the response month, which may or may not be the same as the sample month.

Tabulation Universe. The specific category of people, households, or housing units on which estimates are based; for example, people aged 25 and over or occupied housing units.

Targeting. In the context of the ACS language program, this refers to the identification of geographic areas warranting specific language tools.

Telephone Questionnaire Assistance (TQA). A process which allows respondents to call a toll-free telephone number to receive help when completing the survey questionnaire. This process also allows respondents to complete the survey over the telephone with an interviewer.

Thematic Maps. Data products that show the geographic patterns in statistical data. Thematic maps are a complement to the ranking tables, and are a tool to visually display on a map the geographic variability of a key summary or derived measure.

Three-Year Estimates. Estimates based on 3 years of ACS data. These estimates are meant to reflect the characteristics of a geographic area over the entire 36-month period. These estimates will be published for geographic areas with a population of 20,000 or more.

Timeliness. One of four key dimensions of survey quality. Timeliness refers to both the length of time between data collection and the first availability of a product and to the frequency of the data collection.

Title 13 (U.S. Code). The law under which the Census Bureau operates and that guarantees the confidentiality of census information and establishes penalties for disclosing this information.

Topcoding. A disclosure avoidance practice whereby extremely low or high values are masked by replacing them with a value that represents everything above or below a certain value.

Topologically Integrated Geographic Encoding and Referencing (TIGER®) System or Database. A digital (computer-readable) geographic database that automates the mapping and related geographic activities required to support the Census Bureau's census and survey programs.

Tract. See **Census Tract.**

Undeliverable-As-Addressed (UAA). A USPS notification that a mailing piece could not be delivered to the designated address.

Undercoverage. The extent to which the sampling frame does not include members of the target population thus preventing those members from having any chance of selection into the sample.

Unit Nonresponse. The failure to obtain the minimum required data from a unit in the sample.

Unmailable. A sample address that is inadequate for delivery by the USPS.

Update/Leave (U/L). A method of data collection used in Census 2000 and other censuses, whereby enumerators canvassed assignment areas and delivered a census questionnaire to each housing unit. At the same time, enumerators updated the address listing pages and Census Bureau maps. The household was asked to complete and return the questionnaire by mail. This method was used primarily in areas where many homes do not receive mail at a city-style address; that is, the majority of United States households not included in mailout/mailback areas. U/L was used for all of Puerto Rico in Census 2000.

Upper Bound. Represents the high end of the 90 percent confidence interval of an estimate from a sample survey. A 90 percent confidence interval can be interpreted roughly as providing 90 percent certainty that the true number falls between the upper and lower bounds.

Urbanizacion. An area, sector, or residential development, such as a neighborhood, within a geographic area in Puerto Rico.

Urbanized Area (UA). A densely settled territory that contains 50,000 or more people. The Census Bureau delineates UAs to provide a better separation of urban and rural territory, population, and housing in the vicinity of large places.

Usual Residence. The concept used to define residence in the decennial census. The place where a person lives and sleeps most of the time.

Voluntary Methods Test. A special test conducted at the request of Congress in 2002 to measure the impact on the ACS of changing the data collection authority from mandatory to voluntary.

WebCATI. A control system which is used to track and assign cases to individual telephone interviewers. WebCATI evaluates the characteristics of each case (for example, the date and time of the previous call) and the skills needed for each case (for example, the need for the case to be interviewed in Spanish), and delivers the case to the next available interviewer who possesses the matching skill.

Web Data Server (WDS). A research tool for reporters, SDCs, CICs, ROs, and internal Census Bureau analysts. WDS features a user-friendly interface that allows users to quickly access, visualize, and manipulate ACS base tables.

Weighting. A series of survey adjustments. Survey data are traditionally weighted to adjust for the sample design, the effects of nonresponse, and to correct for survey undercoverage error.