American Community Survey Design and Methodology (January 2014)

United States™ Census Bureau

Version 2.0
January 30, 2014

[This page intentionally left blank]

Document Log	
Document Title:	American Community Survey Design and Methodology (January 2014)
Approval Authority	James B. Treat, Chief, ACSO
Document Approval Date	January 30, 2014
Document Author(s)	Nancy Torrieri, ACSO, DSSD, and SEHSD Program Staff
Critical Reviewer(s)	ACS Program Senior Staff

Sensitivity Assessment

This document does not contain any:

- Title 5, Title 13, Title 26, or Title 42 protected information
- Procurement information
- Budgetary information
- Personally identifiable information

Acknowledgements

The development of the 2014 edition of *ACS Design and Methodology* took place under the direction of James B. Treat, Chief, American Community Survey Office (ACSO). Nancy K. Torrieri, Special Assistant to the Chief, ACSO, provided overall management and coordination. The American Community Survey Program is under the direction of Frank A. Vitrano, Associate Director for the 2020 Census and Lisa M. Blumerman, Assistant Director for the American Community Survey and the Decennial Census.

Staff of the ACSO, the Decennial Statistical Studies Division, the Social, Economic and Housing Statistics Division, and the Population Division contributed to this report as primary authors, secondary authors, developers of graphics, tables, or other content, and as reviewers. These contributors include Mark E. Asiala, Lawrence M. Bates, Judy G. Belton, Cheryl V. Chambers, Melissa C. Chiu, Grace L. Clemons, Donna M. Daily, Kenneth B. Dawson, Gail M. Denby, Tomas E. Encarnacion, Sirius C. Fuller, Deborah H. Griffin, Steven P. Hefter, Liza L. Hill, Todd R. Hughes, Karen Humes, Michelle E. Jiles, Karen E. King, Kenneth C. Kowalk, Terrina L. Long, Ana J. Montalvo, David A. Raglin, Dameka M. Reese, Kanin L. Reese, Monique L. Rhames, Daniel Sommers, Nicholas M. Spanos, Jennifer G. Tancreto, Anthony G. Tersine, Nancy K. Torrieri, Kai T. Wu, and Matthew A. Zimolzak.

MITRE Corporation staff, led by Jeb S. Gaul, reviewed an early draft of the report.

Catherine M. Rosol of ERIMAX, Inc. formatted the report and prepared it for release.

.

Table of Contents

Figures

Tables

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. This 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 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 is used to administer many kinds of government programs and to distribute more than $400 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. The sample was increased to about 800,000 addresses in 2000 to test the feasibility of conducting the ACS concurrent with conducting a decennial census. The demonstration period continued through 2004, and the Census Bureau collected sufficient information to produce data for the nation, states, and most geographic areas with 250,000 or more population. Evaluations and comparisons with the results from the Census 2000 long form data collection demonstrated the quality of ACS 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. The ACS program 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 residents.

ACS data are now available for all areas. Currently, the ACS publishes single-year and multi-year estimates for all areas, including those with populations of less than 20,000. All estimates are updated annually, with data published for the largest areas with populations of 65,000 or

more in 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 2014 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. Since then, we have issued two revised editions of the Report, and provided revisions to several chapters that describe key design changes in the ACS program. This edition includes information on changes to the ACS program since 2009 and through December 2013. This period covers several key recent developments in the ACS program. These include the initiation of a program review in 2011, and the addition of an internet response mode in 2013.

We hope that data users find this report helpful and that it will aid in improving an understanding of the ACS statistical design and the methods.

Dedicated Staff and A Cooperative Public Are Essential to Success. The ACS is the largest household survey conducted by the federal government. The ACS program has been successful in large part because of the innovation and dedication of many people who have worked hard to achieve the program's goals, and the willingness of the public to participate as survey respondents.

All of the primary survey activities are designed and managed by the staff at Census Bureau headquarters in Suitland, MD. These staff continually strive to improve the accuracy of the ACS estimates, streamline ACS operations, analyze ACS data, and conduct important research and evaluation to achieve greater efficiencies and program effectiveness. They also serve as educational resources and experts for the countless data users who come to the Census Bureau in need of technical assistance and help to use ACS data.

In addition, the Census Bureau's field partners in the six Field Regional Offices, thousands of field representatives across the country who collect ACS data, and survey managers and other staff at the Census Bureau's National Processing Center (NPC) in Jeffersonville, IN, and at the Census Bureau telephone call centers in Jeffersonville, IN; Hagerstown, MD; and Tucson, AZ make it possible to achieve a smooth and efficient running of a very complex and demanding survey operation.

At the most fundamental level, the ACS program's achievements are based on the willingness of the public to provide information that make it possible for the Census Bureau to release summarized data for the nation, states, local and tribal governments, and many other data users. Millions of Americans willingly provide the data that are collected each year by the ACS. The Census Bureau thanks each and every respondent who takes the time and effort to participate in the ACS.

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 estimates for the same small areas (census tracts and block groups) formerly surveyed via the decennial census long-form sample. Initially, five years of samples were required to produce these small-area data. Once the Census Bureau, released its first 5-year estimates in December 2010; new small-area statistics now are 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 2013 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 16 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 2013 data collection year, which now includes an Internet option.

- 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 estimates. 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.

- Chapter 16 describes the ACS research and evaluation program.

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 five 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. Once the first five year ACS estimates were released in 2010, what might be called an enhancement stage began. Currently underway, this stage has included a fundamental reexamination of the systems and processes that underlie the ACS and an exploration of new methods, techniques, and approaches designed to improve the ACS program and the Census Bureau's relationships with stakeholders and data users.

2.2 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.

2.3 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.

2.4 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.

2.5 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. 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, were 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, became available late in 2010 for areas down to the smallest block groups, census tracts, and small local governments. Beginning in 2010, the nation had a 5-year period estimate, available as an alternative to the decennial census long-form sample, for nearly all geographic areas recognized by the Census Bureau, including census tracts and block groups.

2.6 The ACS Program Review

With the publication of the first five-year estimates, the Census Bureau met its goal of replacing the decennial census long form with the ACS since those estimates were designed to be comparable to the long form estimates produced following each decennial census. This benchmark event was followed by planning for a detailed review of the systems and processes that underlie the ACS program. An initial goal of this review, which began in 2011, was to identify possible opportunities for improvements. By 2012, the review was well underway, and had expanded to include other aspects of the ACS program. In 2013, as part of what was by then called the ACS Program Review, managers in several divisions that contributed to the ACS participated in a series of meetings and off-site events designed to envision the ACS program of the future. Other developments associated with the program review include:

1) the organization of a set of review teams (the highest level of which was comprised of division chiefs) to function as a set of program management boards. These boards

provide oversight to the review of technical decisions for the program (for example, whether to develop a new data tabulation);

2) the incorporation of a risk review and the formal acknowledgement of the total resources required to implement planned improvements to the program;

3) the use of off-site events and organizational management techniques to more effectively solicit the views of staff in the day-to-day management of the program;

4) the adoption of more effective documentation processes;

5) the recognition of the need for greater involvement with corporate infrastructure solutions such as adaptive design; and,

6) the creation of an ACS Data Users Group to more effectively solicit information on what stakeholders need to use ACS estimates, and what usability issues arise among stakeholders with similar interests in ACS data applications.

2.7 ACS Stakeholders and External Engagement

The ACS program depends heavily on engaging stakeholders in the development of the program, and seeking stakeholder input as much as possible in decisions affecting ACS data products. 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.

As a formal ACS testing period was under development, ACS managers started forming plans to ensure local communities were aware of their inclusion as test sites for the ACS. ACS testing was launched in four sites in 1995 as described earlier in this chapter. From March 1996 to November 1999, 31 town hall-style meetings were held throughout the country, with more than 600 community representatives attending the meetings. Similar meetings took place in the years to follow. A series of three regional outreach meetings, in Dallas, TX; Grand Rapids, MI; and Seattle, WA, were held in mid-2004, with an overall attendance of more than 200 individuals representing data users, academicians, the media, and local governments. Other early stakeholder engagement efforts included the development of special-purpose advisory panels in partnership with the Committee on National Statistics of the National Academies of Science and a Rural Data Users Conference held in May 1998 in Alexandria, Virginia, to discuss issues of concern to representatives of small areas and populations. Annual meetings of individual State Data Center representatives and affiliate organizations have frequently featured presentations to update members on the latest ACS program developments and data products.

Changes based on stakeholder input were important in shaping the design and development of the ACS and continue to influence its form as the ACS program moves forward. 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. From 2003 to 2005, the Census Bureau invited federal agencies to participate in an ACS Federal Agency Information Program designed to arrange meetings at federal agencies where specific questions by federal agency representatives on the ACS design, methods, and data products could be addressed by Census Bureau technical experts.

In 2007, the Committee on National Statistics (CNSTAT) 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. In 2013, the Census Bureau requested that CNSTAT convene a workshop on the benefits of the ACS to a broad array of non-federal data users. A summary of the workshop is described in the CNSTAT publication, “Benefits, Burdens, and Prospects of the American Community Survey: Summary of a Workshop.”

Meetings with the Decennial Census Advisory Committee, the Census Advisory Committee of Professional Associations, and the Race and Ethnic Advisory Committees have provided opportunities for ACS staff to receive specific advice on the ACS design, survey methods, and data products. The Census Bureau’s Field Division Partnership and Data Services Staff and regional directors all have played prominent role in communicating the importance of participating in ACS data collection to state and local government representatives, and circulate pamphlets and similar publications to explain the ACS program and its benefits to communities. The latest example of such a publication is the ACS Information Guide, available at:

http://www.census.gov/acs/www/about_the_survey/american_community_survey/

The ACS staff regularly brief 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 brief Members of Congress regularly on multiple aspects of the ACS, including data collection.

The number of scope of groups and organizations that represent ACS stakeholders has expanded dramatically since the survey was implemented. The chart below lists representative ACS stakeholder organizations. Some of these organizations represent broad areas of interest, such as statistical methodology, public opinion research, demography, regional science, sociology, and geography. Others advocate for specific interests, such as housing, transportation, and education; some represent population groups such as the elderly, veterans, and American Indians and Alaska Natives; or professional groups that represent a specific occupation, such as librarian.

Table 2-1: Representative Stakeholder Organizations for the ACS

American Association of Public Opinion Research	Council for Community and Economic Research
Association of Public Data Users	Joint Center for Political and Economic Studies
American Library Association	National Association of Towns and Townships
American Marketing Association	National Council of La Raza
American Sociological Association	National Conference of State Legislatures
American Statistical Association	National Congress of American Indians
Association of American Geographers	National Urban League
Asian American Federation	Population Association of America
Brookings Institution	Population Reference Bureau
Children's Defense Fund	Rural Sociological Society
Council of Professional Associations on Federal Statistics	Urban Institute

Efforts have been made in the past to promote the international sharing of the Census Bureau's experiences with the development and implementation of the ACS. Presentations have been given to many international visitors who have come to the Census Bureau to learn about surveys and censuses, including, in November 2013, the Office of National Statistics of the United Kingdom. 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.

2.8 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: Nonresponse 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 Master Address File (MAF) is the Census Bureau's official inventory of known housing units (HUs), group quarters (GQs), and selected non-residential units (public, private, and commercial) in the United States and Puerto Rico. It serves as the source of addresses for the American Community Survey (ACS), other Census Bureau demographic surveys, and the decennial census. 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 resulting database is called the MAF/TIGER database (MTdb).

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 programs. The MAF was used as the initial frame for the ACS, in its state of existence at the conclusion of Census 2000. Updates from nationwide 2010 Census operations were incorporated into the MTdb and were included in the ACS sampling frame in the middle of 2010. 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 3.2 provides basic information about the MAF and its content. Sections 3.3 and 3.4 describe the MAF development and update activities for HUs in the United States and Puerto Rico. Section 3.5 describes the MAF development and update activities for GQs. Finally, Section 3.6 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 characteristics and/or location description of the unit or any relationships to other units, residential or commercial status, latitude and longitude coordinates, and source and history information indicating the operation(s) that added/updated the record (see Section 3.3). ACS

obtains this information from the MAF in files called MAF extracts (see Section 3.6) and uses it for sampling, data collection, and data tabulation activities.

The geographic codes in the MTdb identify a variety of areas, including states, counties, county subdivisions, places, American Indian areas, Alaska Native areas, Hawaiian Homelands, census tracts, block groups, and blocks. Two important geographic code sets are the 2010 Census tabulation geography set, based on the January 1, 2010 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 2012 reflect legal boundaries as of January 1, 2012). Each MAF record contains geographic codes from 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 frequently 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 or a post office (PO) box format. Examples of these types of addresses are RR 2, Box 9999, Anytown, ST 99988 and PO Box 123, Anytown, ST 99988.

In the United States, city-style addresses are most prevalent in urban and suburban areas, and accounted for 98.2 percent of all residential addresses in the MAF at the conclusion of the 2010 Census. 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 (U.S. Census Bureau 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 0.3 percent of all residential addresses in the MAF at the conclusion of the 2010 Census. 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. Post Office Box addresses cannot be located in the field because they are associated with a post office location, not a structure location.

To help field staff locate non-city-style addresses in the field, the MAF often contains a location description[1] of the unit and/or its latitude and longitude coordinates. 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 non-residential. A residential address represents a housing unit in which a person or persons live or could live. A non-residential address represents a structure, or a unit within a structure, that is used for a purpose other than residence. While the MAF includes many non-residential addresses, it is not a comprehensive source of such addresses (U.S. Census Bureau 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 1.5 percent of the United States residential addresses in the MAF at the conclusion of the 2010 Census.

For more information on the MAF, including a description of its content and structure, see U.S. Census Bureau (2000b).

3.3 Master Address File Development and Updating for the U.S. Housing Unit Inventory

MAF Development in the United States

For the 1990 and earlier decennial censuses, the Census Bureau compiled address lists 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.

The Census Bureau created the MAF by merging city-style addresses from the 1990 Address Control File;[2] field listing operations;[3] the USPS's DSF; and addresses supplied by local governments through partnership programs, such as the Local Update of Census Addresses

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

[2] 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 (U.S. Census Bureau 2000a, p. XVII-1).

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

(LUCA)[4] and other Census 2000 activities, including the Be Counted Campaign.[5] At the conclusion of Census 2000, the MAF contained a complete inventory of known HUs nationwide.

For details on the address list development for Census 2000, see U.S. Census Bureau (2000a).

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 non-city 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 and 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, the Census Bureau implemented these operations to improve the addresses recognized in Census 2000 and their associated characteristics. CQR was implemented again after the 2010 Census.

The 2010 Census operations improved the coverage and quality of the MAF. The operations included several nationwide field canvassing and enumeration operations. In preparation for the 2010 Census, the Census Bureau implemented a nationwide address canvassing field operation (with the exception of remote areas in Alaska and rural Maine) to update the housing unit inventory in the MAF. Other field operations to support the 2010 Census enumeration identified HU and GQ corrections, additions, and deletions and updated the MAF with those results. Additionally, the Census Bureau repeated the same partnership and count coverage programs used for Census 2000 for the 2010 Census, including the LUCA[6] and the Be Counted programs. The Census Bureau determined the final 2010 Census status of each HU record in the MAF in late 2010. These operations improved the MAF extracts used for the ACS sample selection. ACS and the 2010 Census planners worked together closely to assess the impact of the decennial operations on the ACS. For details on the 2010 Census operations, see U.S. Census Bureau (2011).

[4] 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 Census 2000 address list for all blocks in which the participating governments questioned the number of living quarter addresses.

[5] 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).

[6] The Census Bureau redesigned the LUCA program for the 2010 Census, allowing participants a choice of several methods of reviewing the census address or housing unit inventories in their jurisdictions. Participant feedback was included in the Address Canvassing field operation for verification.

Some ongoing improvement operations are designed to deal with errors remaining from the 2010 Census, while others aim to keep pace with post-2010 Census address development. In the remainder of this section, we discuss several ongoing operations, including DSF updates, ACS nonresponse follow-up updates, the Geographic Support System Initiative, and Demographic Area Address Listing (DAAL) updates. We also discuss the Community Address Updating System (CAUS), which the Census Bureau employs in rural areas. Table 3-1 summarizes the development and improvement activities.

Table 3-1: Master Address File Development and Improvement

Initial Input (2000 and earlier)	Improvements (Post-2000)
1990 Decennial Census address control file	DSF updates
USPS Delivery Sequence File (DSF)	ACS personal visit
Local government updates	Community Address Updating System (CAUS)
Other Census 2000 activities	Demographic Area Address Listing (DAAL) Operations
	2010 Census field operations
	Other 2010 Census activities
	Geographic Support System Initiative

Delivery Sequence File (DSF)

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. The DSF is the primary source of new city-style-addresses used to update the MAF between decennial censuses. 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.

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 housing units, 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).

Address Updates from ACS Personal Visit

Field representatives (FRs) can obtain address updates or corrections for each HU visited during the personal visit phase of the ACS. The ACS conducts this follow-up for a sample of addresses. The Census Bureau updates the MAF 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).

Demographic Area Address Listing (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 American Housing Survey (AHS), 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. The Census Bureau conducts listings for DAAL on laptop computers using the Automated Listing and Mapping Instrument (ALMI) software. The ALMI uses extracts from the current MTdb 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. In October 2011, Global Positioning System (GPS) functionality was enabled in the ALMI. This functionality allowed the FRs to collect latitude and longitude coordinates for the structure. 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 the MTdb. Starting in 2013, the demographic surveys are only listing in the following 13 states: Alabama, Alaska, Arkansas, Kentucky, Maine, Mississippi, Montana, New Hampshire, New Mexico, Oklahoma, Vermont, West Virginia, and Wyoming (Kennel, et al. 2011). For details on DAAL, see Perrone (2005).

Community Address Updating System (CAUS)

The Census Bureau designed the CAUS program 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. An automated field data collection operation, CAUS was designed to provide a rural counterpart to the update of city-style addresses received from the DSF. It 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 were not used to provide adequate coverage. The Census Bureau selected CAUS blocks 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. The CAUS program was suspended from October 2007 to March 2010 until the 2010 Census Address Canvassing and field follow-up activities were completed.

The CAUS program resumed listing activities again in April 2010. Approximately 30,000 blocks were listed from October 2010 through September 2012. Beginning in October 2012, and subject to available resources, the Census Bureau plans for the CAUS program to list approximately 1,500 blocks per year.

For details on the CAUS program and its block selection methodology, see Hartman (2009, 2011) and Schar (2012a, 2012b).

Geographic Support System Initiative

The Geography Division of the U.S. Census Bureau has already begun preparations for the 2020 Census and future surveys by initiating a broad-based geographic support system initiative. The initiative covers many aspects of geographic support for these programs, including investigating various partnering opportunities with local governments, and pursuing commercial resources and crowdsourcing, to maintain the MTdb throughout the decade.

All of these MAF improvement activities and operations contribute to the overall update of the MTdb.

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 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 Census 2000 for Puerto Rico, see U.S. Census Bureau (2004b).

The Census Bureau designed Census 2000 procedures and processing systems to capture, process, transfer, and store information for the conventional three-line mailing address. Mailing addresses in Puerto Rico generally incorporate the urbanization name (a geographic area roughly equivalent to a neighborhood), 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 U.S. 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 to fully implement the Puerto Rico Community Survey (PRCS) starting in 2005. 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 were added to the MAF, 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 for the 2005 ACS began in November 2004 with the best address information available given these shortcomings. 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. Future versions of this document will include discussions of these operations and MAF development and updating in Puerto Rico.

As part of the MAF/TIGER redesign effort in the middle of the last decade, the Census Bureau redesigned the MAF to accommodate the Puerto Rico specific address components that were lacking previously. The MAF now accommodates these specific address components, allowing the potential to update the MAF in Puerto Rico by census field operations and other methods.

In preparation for the 2010 Census, the Census Bureau conducted address canvassing in Puerto Rico as it was in the United States, updating the inventory of housing units in the MAF for Puerto Rico prior to the 2010 Census. Results from the 2010 Census Update/Leave and follow-up operations also updated the MAF addresses in Puerto Rico. The Census Bureau determined the final 2010 Census status of each HU record in the MAF in Puerto Rico in late 2010.

3.5 Master Address File Development and Updating For Group Quarters in the United States and Puerto Rico

MAF Development for 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 Census Bureau incorporated the address and identification information for each GQ into the MAF.

MAF Improvement Activities and Operations for 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.

The Census Bureau collects information on the new GQ facilities and updated address information for existing GQ facilities 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.

Prior to 2010 Census operations, the Census Bureau developed a process to supplement these activities to create an updated GQ universe from which to select the ACS sample. The Census Bureau constructed the ACS GQ universe 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 (BoP) and a file from the Department of Defense (DoD) containing military bases and vessels. The Census Bureau also conducted research to identify new migrant worker locations, new state prisons, and state prisons that had closed.

ACS FRs, while conducting the Group Quarters Facility Questionnaire (GQFQ), collect updated address and geographic location information. Updates collected via the GQFQ were used 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-Level Phase) in Section 8.2 of Chapter 8.

The Address Canvassing operation for the 2010 Census identified records as “other living quarters” or OLQs. All OLQs and GQs were then visited in the Group Quarters Validation operation where their final status as a HU or GQ was determined. GQs were then enumerated in the GQ Enumeration operation. The Census Bureau applied updates from all of these operations to the MAF. The Census Bureau determined the final 2010 Census status for each GQ in late 2010.

The final Census universe of GQs is the basis of the ACS GQ frame for 2012 and beyond. ACS also includes GQs that were identified as having no population on Census Day as those GQs may contain people if visited at another time of the year. New GQs from ongoing operations, such as DAAL and CQR, are also included in the ACS GQ frame. The Census Bureau updates the ACS

GQ frame with results from ACS GQ data collection operations as well as results of state prison research using the individual state Department of Corrections websites. ACS continues to partner with the BoP to ensure the most accurate GQ frame for federal prisons.

For more information on the post-2010 Census ACS GQ frame, see Bates (2011) and Aubuchon (2011).

3.6 American Community Survey Extracts from the Master Address File

Data from the MTdb are provided for use with the 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 3.2).

The MAF, as an inventory of living quarters (HUs and GQs) and some non-residential 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. Each Census Bureau program that relies on the MAF defines the set of valid addresses for their individual program.

Since the ACS frame must be as complete as possible, the Census Bureau applies filtering rules during the creation of the ACS extracts to minimize both overcoverage and undercoverage and to obtain an inclusive listing of addresses. For example, the ACS filter rules include 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 has not been linked to a census tract and block yet. 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 mail to the address. In this regard, the ACS filtering rules differ from those for the 2010 Census. For the 2010 Census, EDS records were included on the list of addresses to be updated in Address Canvassing, but ungeocoded records were excluded. Ungeocoded records and EDS records added to the MAF after Address Canvassing were excluded from all post-Address Canvassing operations.

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, it must meet the conditions set forth in the filter.

Filtering rules change, and with them, the ACS frame. The most significant recent change to the ACS filter was the incorporation of results from 2010 Census operations. Prior to Address Canvassing, the largest source of HUs on the ACS frame was HUs tabulated in Census 2000. Address Canvassing results were incorporated into the MAF in time to be included the ACS frame by mid-2010. Once the Census Bureau established the final 2010 Census HU universe, the basis of the ACS HU frame became the list of HUs tabulated in the 2010 Census. The post-2010

ACS frame consists of 2010 Census addresses plus any new records added to the MAF after the 2010 Census, including post-Census DSF adds, new or validated records from DAAL, CQR, special censuses, and Census tests, and 2010 Census deletes that persist on the DSF.

As discussed above, the ACS attempts to create a sampling frame that is as accurate as possible by minimizing both overcoverage and undercoverage[7]. 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 1.9 to 5.2 percent for the years 2002- 2009. See Kephart (2010) for a discussion of this issue.

For details on the ACS requirements for MAF extracts, see Zimolzak (2012). 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 Chapters 7 and 8. For details on the MAF, its coverage, and the implications of extract rules on the ACS frame, see Shapiro and Waksberg (1999) and Kephart (2010).

[7] Definitions of the terms “overcoverage” and “undercoverage” are provided in the Glossary.

3.7 References

Aubuchon, Dianne (2011). ‘‘Creating the Group Quarters Universe for the American Community Survey for Sample Year 2012.’’ Internal U.S. Census Bureau Memorandum From D. Whitford to J. Treat, Draft, Washington, DC, September 26, 2011.

Bates, Lawrence M. (2011). ‘‘Creating the Group Quarters Universe for the American Community Survey for Sample Year 2011.’’ Internal U.S. Census Bureau Memorandum From D. Whitford to J. Treat, Washington, DC, April 26, 2011.

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.

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.

Hartman, James (2009). “Overview of the Community Address Updating System’s Processing System.” Internal U.S. Census Bureau Memorandum From D. Whitford for Documentation, Washington, DC, December 14, 2009.

Hartman, James (2011). “Community Address Updating System Program Overview.” Internal U.S. Census Bureau Memorandum From D. Whitford for The Record, Washington, DC, April 11, 2011.

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.

Kennel, Timothy, Wneck, Jeff, White, Chengee, Tekansik, Sarah, Rottach, Reid, Rawlings, Andreea, Parmer, Randy, Nguyen, T. Trang, Lubich, Antoinette, and Farber, James (2011). “Recommendations for Coverage Improvement Frame Creation and Sampling Methodology for the 2010 Demographic Surveys Sample Redesign (Doc. #2010-3.8-R-5, Version 1.0,” Internal U.S. Census Bureau Memorandum From Within PSU MAF Sampling Team (WG 3.8) to Distribution List, Washington, DC, October 17, 2011.

Kephart, Kathleen (2010). ‘‘National Estimate of Coverage of the MAF for 2009,’’ Internal U.S. Census Bureau Memorandum From D. Whitford to T. Trainor, Washington, DC, November 30, 2010.

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.

Schar, Bryan (2012a). “An Investigation into Expanding the Community Address Updating System Universe.” U.S. Census Bureau Memorandum From D. Whitford to ACS Research and Evaluation Steering Committee, Washington, DC, May 24, 2012.

Schar, Bryan (2012b). “Refreshing the Community Address Updating System Block Score for 2011 Selection.” U.S. Census Bureau Memorandum From A. Tersine to ACS Research and Evaluation Steering Committee, Washington, DC, September 27, 2012.

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 (2000a). ‘‘Census 2000 Operational Plan.’’ Washington, DC, December 2000.

U.S. Census Bureau (2000b). ‘‘MAF Basics.’’ Washington, DC, 2000.

U.S. Census Bureau (2004). ‘‘Census 2000 Topic Report No. 14: Puerto Rico.’’ Washington, DC, 2004.

U.S. Census Bureau (2011). “2010 Census Operational Plan and Accompanying Operational Requirements Document.” Washington, DC, June 29, 2011.

Zimolzak, Matthew and Bates, Lawrence (2012). ‘‘Customer Requirements Document for American Community Survey Geographic Products V2.0.” Washington, DC, March 1, 2012.

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 residents of group quarters (GQ) facilities. As described in Chapter 3, we derive the sampling frames from which we draw these samples 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 (U.S.) and Puerto Rico.

We select independent HU address samples for each of the 3,143 counties and county equivalents in the U.S., including the District of Columbia, as well as for each of the 78 municipalities in Puerto Rico. In 2004, we selected samples of HU addresses for every county and county equivalent for field data collection in 2005.[8] Each year from 2005–2010, we selected approximately 2.9 million HU addresses in the U.S. and 36,000 HU addresses in Puerto Rico. Beginning in 2011, we implemented the following changes to the ACS sample designs:

- We increased the HU sample in June 2011, bringing the size of the sample selected to 3.54 million addresses per year.
- We added several new HU sampling rates that better control the allocation of the sample and improve estimate reliability for small areas.
- We increased the follow-up sample to 100 percent in select geographic areas.

In addition, starting in 2013, we restricted the assignment of the GQ sample for college dorms to the non-summer months (January–April, September–December).

Full-implementation samples of GQ facilities and persons are selected independently within each state, including the District of Columbia and Puerto Rico. This began in 2006. In 2006 and 2007, the ACS and the PRCS included approximately 2.5 percent of the expected number of residents in GQ facilities. Beginning in 2008, we increased the sampling rates in 16 states with small GQ populations to meet publication thresholds. See Chapters 7 and 8 for details of the data collection methods.

This chapter presents details on the selection of the HU address and GQ samples. The final section describes the differences in sampling and data collection methodology for some hard to reach areas in Alaska (referred to as Remote Alaska). The section on Remote Alaska also details recently modified sampling and data collection procedures for these areas.

[8] In the remainder of this chapter, the term "county" refers to counties, county equivalents, and municipalities.

4.2 Housing Unit Sample Selection

There are two phases of HU address sampling for each county.[9] During first-phase sampling, we assign blocks to sampling strata, calculate sampling rates, and select the sample. During the second phase of sampling, we select a sample of nonresponding addresses for Computer Assisted Personal Interviewing (CAPI). This is the CAPI sample.

First-phase sampling produces the annual ACS initial sample of addresses and includes two processes—main and supplemental sampling. The main and supplemental samples in the first-phase sampling include two stages. The first stage sample selection systematically assigns new addresses to sub-frames and identifies the appropriate sub-frame associated with a specific year's sample. The second stage sample selection systematically selects the sample from the selected sub-frame.

Figure 4-1 provides a visual overview of the housing unit address sampling process.

[9]Throughout this chapter, "addresses" refers to valid ACS addresses that have met the filter criteria (Bates, Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey, 2013).

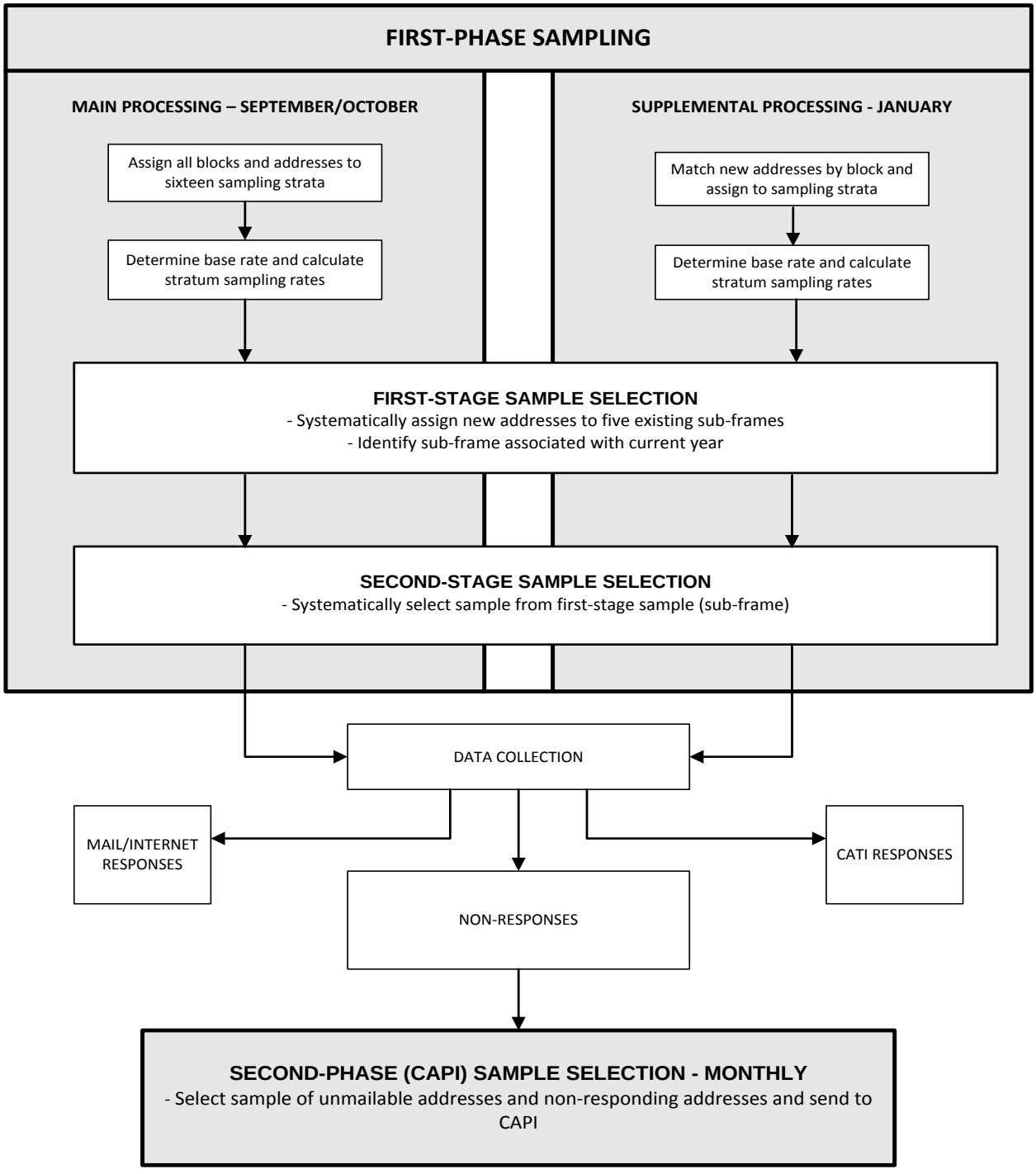


Figure 4-1: Assignment of Blocks (and their addresses) to Second-stage Sampling

4.3 First Phase Sample

The first phase of sampling is comprised of two separate stages. The first stage of first phase sampling maintains five distinct partitions, or sub-frames, of the addresses on the sampling frame within each county. Each county sub-frame is a representative sample of addresses in the county. We assign these sub-frames to specific years and rotate them annually. The sub-frames maintain their annual designation over time. First stage sampling systematically sorts and assigns addresses that are new to the frame to one of the five sub-frames.[10] First stage sampling also determines the sampling rates for each stratum for the current sample year.

The second stage of first phase sampling selects a sample of the addresses from the current year's sub-frame and allocates this sample to the twelve months of the year for data collection.

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

One of the ACS design requirements is that no HU address be in sample more than once in any five-year period. To accommodate this restriction, the addresses in the frame are assigned systematically to five sub-frames, each containing roughly 20 percent of the frame, and each being a representative sample. Addresses from only one of these sub-frames are eligible to be in the ACS sample each year and each sub-frame is used every fifth year. For example, 2014 will have the same addresses in its sub-frame as did 2009 with the addition of all new addresses that we assigned to that sub-frame during the 2010–2014 time period. As a result, we must perform both the main and supplemental sample selection in two stages. The first stage partitions the sampling frame into the five sub-frames and determines the sub-frame for the current year. The second stage, described in more detail below, selects addresses to be included in the ACS from the sub-frame eligible for the sample year.

Prior to the 2005 sample selection, there was a one-time allocation of all addresses then present on the ACS frame to the five sub-frames. In subsequent years, we must systematically allocate addresses new to the frame to these five sub-frames. We accomplish this by sorting the addresses in each county by stratum and geographic order including tract, block, street name, and house number. We then assign addresses sequentially to each of the five existing sub-frames. 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 assignment to a sub-frame. Similarly, during supplemental sampling, only addresses new to the MAF since main sampling go through first-stage sampling.

[10] All existing addresses retain their previous assignment to one of the five sub-frames. The five sub-frames are maintained to meet the requirement that no address be in sample more than once in a five-year period.

The ACS and PRCS reflect two separate sampling operations carried out at different times of the year: (1) main sampling, which occurs in September and October of the year preceding the sample year, and (2) supplemental sampling, which occurs in January of the sample year. This allows an opportunity for new addresses to have a chance of selection into the sample. 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 sample year. The MAF available at the time of main sampling, obtained in the July preceding the sample year, reflects address updates 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 through September of the year preceding the sample year. During supplemental sampling, we assign addresses new to the frame systematically to the five sub-frames using the same process for new addresses as in the main sample.

First Phase, First-Stage Sample: Determining the Sampling Rates

Each year, we must determine the specific set of sampling rates for each of the thirteen non-fixed rate sampling strata defined in Table 4-1. Before we can do this, we must perform the following two steps. The first step is to calculate a base rate (BR) for the current year. Thirteen of the sixteen sampling rates are a function of a base rate. The three fixed rate strata are 15 percent, 10 percent, and 7 percent. Column 3 of Table 4-1 shows the relationship between the base rate and the sixteen sampling rates. Beginning in 2009, the number of new addresses differed from what was expected by enough to warrant the calculation of a separate set of sampling rates for supplemental sample selection. This led to separate supplemental sampling rates beginning with the 2010 sample selection.

The distribution of addresses by sampling stratum, coupled with the target sample size of 3.54 million, allows us to set up and solve a simple algebraic equation for the BR.

The second step is the calculation of the sampling rates using the value of BR and the equations in Table 4-1. Beginning in June, 2011 we increased the sample size to a monthly level corresponding to an annual 3.54 million sample (approximately 295,000 per month). Between January of 2005 and May of 2011, the monthly sample corresponded to an annual sample of approximately 2.9 million (roughly 242,000 per month).

First-Phase, First-Stage Sample: First-Phase Sampling Rates

Columns 2 and 3 of Table 4-1 provide the sampling rates for the 2013 ACS for the U.S. and Puerto Rico, respectively (Sommers, 2012b).

Table 4-1: 2013 ACS/PRCS Main Sampling Rates

Stratum	Sampling Rates[1]	
	United States	Puerto Rico
Blocks in smallest sampling entities (0<SEMOS≤200)	15.0	(NA)
Blocks in small sampling entities (200<SEMOS≤400)	10.0	(NA)
Blocks in medium sampling entities (400<SEMOS≤800)	7.0	7.0
Blocks in large sampling entities (800<SEMOS≤1,200)	4.4	(NA)
Blocks in large sampling entities (SEMOS>1,200) and smallest tracts (0<TMOS≤400) with predicted levels of completed interviews prior to CAPI sampling≤60 percent	5.5	4.9
Blocks in large sampling entities (SEMOS>1,200) and smallest tracts (0<TMOS≤400) with predicted levels of completed interviews prior to CAPI sampling>60 percent	5.1	
Blocks in large sampling entities (SEMOS>1,200) and small tracts (400<TMOS≤1,000) with predicted levels of completed interviews prior to CAPI sampling≤60 percent	4.4	3.9
Blocks in large sampling entities (SEMOS>1,200) and small tracts (400<TMOS≤1,000) with predicted levels of completed interviews prior to CAPI sampling>60 percent	4.0	
Blocks in large sampling entities (SEMOS>1,200) and medium tracts (1,000<TMOS≤2,000) with predicted levels of completed interviews prior to CAPI sampling≤60 percent	2.7	2.4
Blocks in large sampling entities (SEMOS>1,200) and medium tracts (1,000<TMOS≤2,000) with predicted levels of completed interviews prior to CAPI sampling>60 percent	2.5	
Blocks in large sampling entities (SEMOS>1,200) and large tracts (2,000<TMOS≤4,000) with predicted levels of completed interviews prior to CAPI sampling≤60 percent	1.6	1.4
Blocks in large sampling entities (SEMOS>1,200) and large tracts (2,000<TMOS≤4,000) with predicted levels of completed interviews prior to CAPI sampling>60 percent	1.4	
Blocks in large sampling entities (SEMOS>1,200) and larger tracts (4,000<TMOS≤6,000) with predicted levels of completed interviews prior to CAPI sampling≤60 percent	0.9	0.8
Blocks in large sampling entities (SEMOS>1,200) and larger tracts (4,000<TMOS≤6,000) with predicted levels of completed interviews prior to CAPI sampling>60 percent	0.9	
Blocks in large sampling entities (SEMOS>1,200) and largest tracts (6,000>TMOS) with predicted levels of completed interviews prior to CAPI sampling≤60 percent	0.5	(NA)
Blocks in large sampling entities (SEMOS>1,200) and largest tracts (6,000>TMOS) with predicted levels of completed interviews prior to CAPI sampling>60 percent	0.5	

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

NA Not applicable.

[1] In percent.

Since the design of the ACS calls for a target annual address sample of approximately 3.54 million in the U.S. and 36,000 in Puerto Rico, we reduce the sampling rates for all but the smallest sampling entity strata (SEMOS≤800) each year as the number of addresses in the U.S. and Puerto Rico increases. However, as shown in Table 4-1, among the strata where the rates are decreasing, the relationship of the sampling rates will remain proportionally constant. The sampling rates for the smallest sampling entities will remain at 15 percent, 10 percent, and 7 percent.

The sampling rates that we use to select the sample include strata for blocks in certain census tracts in the U.S. These tracts are projected to have the highest rates of completed questionnaires by mail and by the telephone follow-up operation, called 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. Note that the initial identification of these tracts, performed in 2004 was used in the 2005 sample selection and was revised in 2007 based on more recent data and has been used since the 2008 sample selection.

Specifically, we multiply the sampling rates by 0.92 (reduced by 8 percent) for blocks in the U.S. in the six strata in which the SEMOS was greater than 1,200. We make this adjustment for blocks in tracts that we predict will have a level of completed mail and CATI interviews of at least 60 percent, and at least 75 percent mailable addresses.

Because of this adjustment, there are sixteen sampling rates used in the U.S., and ten in Puerto Rico, as shown in columns 2 and 3 of Table 4-1. See the research report (Asiala, 2005) for a full description of the relationship between this reduction and the CAPI sampling rates. This reduction does not occur in Puerto Rico, therefore there are ten sampling strata eligible to be used in Puerto Rico. Only six strata in Puerto Rico contain valid addresses on the 2013 main sampling frame, so for 2013, we only used the six sampling rates shown in Table 4-1.

First-Phase, Second-Stage Sampling: Selection of Addresses

As noted earlier, the second stage of first phase sampling selects a sample of the addresses from the current year's sub-frame. We partition this sub-frame by county and select the addresses from the sub-frame in each county. Second stage sampling allocates this sample to the twelve months of the year for data collection. This process results in the creation of the initial annual ACS sample.

We sort the addresses in each county by stratum and the first-stage order of selection. After sorting, we select systematic samples of addresses using a sampling rate approximately equal to the final sampling rate divided by 20 percent.[11]

First-Phase, Second-Stage Sampling: Assigning Addresses to the Second-Stage Sampling Strata

Each year, the main sampling operation assigns each block to one of the sixteen sampling strata, and consequently, assigns each block one of sixteen sampling rates.[12] 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, we must sample areas with smaller populations at higher rates relative to those areas with larger populations. We base the stratum assignment for a block 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:

- Counties,
- Places with active and functioning governments,[13]
- 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,[14] or
- Census Designated Places (CDPs) in Hawaii only.

We base the sampling stratum for most blocks 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, we derive block-level counts of addresses from the main MAF. This count is converted to an estimated number of occupied HUs by multiplying it by the proportion of occupied HUs in the block in the 2010 Census.

[11]The second-stage rate is approximately equal to the sampling rate divided by 20 percent since the first-stage sampling rate is approximately 20 percent, and the first-stage rate times the second-stage rate equals the overall sampling rate. An adjustment is made to account for uneven distributions of addresses in the county level sub-frames.

[12] From 2005 – 2010 five sampling strata were used.

[13] Functioning governments have elected officials who can provide services and raise revenue.

[14] 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.

For American Indian and Alaska Native Statistical Areas (AIANSA[15]) and Tribal Subdivisions, we multiply the estimated number of occupied HUs by the proportion of its population that responded as American Indian or Alaska Native (either alone or in combination) in the 2010 Census.[16] For each sampling entity, we sum the estimate across all blocks in the entity to create the MOS for the entity. In AIANSAs, if the sum of these estimates across all blocks is non-zero, 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. For detail, see the computer specifications for calculating the MOS for the ACS (Sommers, 2012a). We assign each block the smallest MOS of all the sampling entities in which the block is contained and we refer to it as the Smallest Entity Measure of Size, or SEMOS.

If the SEMOS is greater than 1,200, we base the stratum assignment for the block on the MOS for the census tract that contains it. The sum of the estimated number of occupied HUs across all of its blocks.is the MOS for each tract (TMOS). Using SEMOS and TMOS, we can assign blocks to the sixteen strata defined in columns 1 and 2 in Table 4-2 below.

[15] AIANSA 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 Areas Geographic Program for Census 2000, see the publication in the *Federal Register* **Invalid source specified.**.

[16] 2010 Census information was used for the first time to define the measures of size in the 2012 sample selection.

Table 4-2: Sampling Strata Thresholds and Relationship between the Base Rate and the Sampling Rates

Stratum	Smallest Entity Measure of Size (SEMOS) and Tract Measure of Size (TMOS)	Sampling Rates
Blocks in smallest sampling entities	0 < SEMOS ≤ 200	15% (fixed)
Blocks in small sampling entities	200 < SEMOS ≤ 400	10% (fixed)
Blocks in medium sampling entities	400 < SEMOS ≤ 800	7% (fixed)
Blocks in large sampling entities	800 < SEMOS ≤ 1,200	2.8×BR
Blocks in large sampling entities (SEMOS > 1,200) and smallest tracts with predicted levels of completed interviews prior to CAPI sampling ≤ 60 percent	0 < TMOS ≤ 400	3.5×BR
Blocks in large sampling entities (SEMOS > 1,200) and smallest tracts with predicted levels of completed interviews prior to CAPI sampling > 60 percent		0.92×3.5×BR
Blocks in large sampling entities (SEMOS > 1,200) and small tracts with predicted levels of completed interviews prior to CAPI sampling ≤ 60 percent	400 < TMOS ≤ 1,000	2.8×BR
Blocks in large sampling entities (SEMOS > 1,200) and small tracts with predicted levels of completed interviews prior to CAPI sampling > 60 percent		0.92×2.8×BR
Blocks in large sampling entities (SEMOS > 1,200) and medium tracts with predicted levels of completed interviews prior to CAPI sampling ≤ 60 percent	1,000 < TMOS ≤ 2,000	1.7×BR
Blocks in large sampling entities (SEMOS > 1,200) and medium tracts with predicted levels of completed interviews prior to CAPI sampling > 60 percent		0.92×1.7×BR
Blocks in large sampling entities (SEMOS > 1,200) and large tracts with predicted levels of completed interviews prior to CAPI sampling ≤ 60 percent	2,000 < TMOS ≤ 4,000	BR
Blocks in large sampling entities (SEMOS > 1,200) and large tracts with predicted levels of completed interviews prior to CAPI sampling > 60 percent		0.92×BR
Blocks in large sampling entities (SEMOS > 1,200) and larger tracts with predicted levels of completed interviews prior to CAPI sampling ≤ 60 percent	4,000 < TMOS ≤ 6,000	0.6×BR
Blocks in large sampling entities (SEMOS > 1,200) and larger tracts with predicted levels of completed interviews prior to CAPI sampling > 60 percent		0.92×0.6×BR
Blocks in large sampling entities (SEMOS > 1,200) and largest tracts with predicted levels of completed interviews prior to CAPI sampling ≤ 60 percent	6,000 < TMOS	0.35×BR
Blocks in large sampling entities (SEMOS > 1,200) and largest tracts with predicted levels of completed interviews prior to CAPI sampling > 60 percent		0.92×0.35×BR

Figure 4-2 shows a Census Block that is in City A and 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

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 $400 < \text{SEMOS} \leq 800$ 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 twelve strata with SEMOS > 1,200 depending on the size of the census tract (TMOS - not shown in the diagram) and the predicted level of completed interviews prior to CAPI sampling in the tract. In this example, suppose the TMOS is 1,800, and the predicted level of completed interviews prior to CAPI sampling is ≤ 60 percent, then the Census Block will be placed in the $1{,}000 < \text{TMOS} \leq 2{,}000$ stratum with a predicted level of completed interviews prior to CAPI sampling ≤ 60 percent.

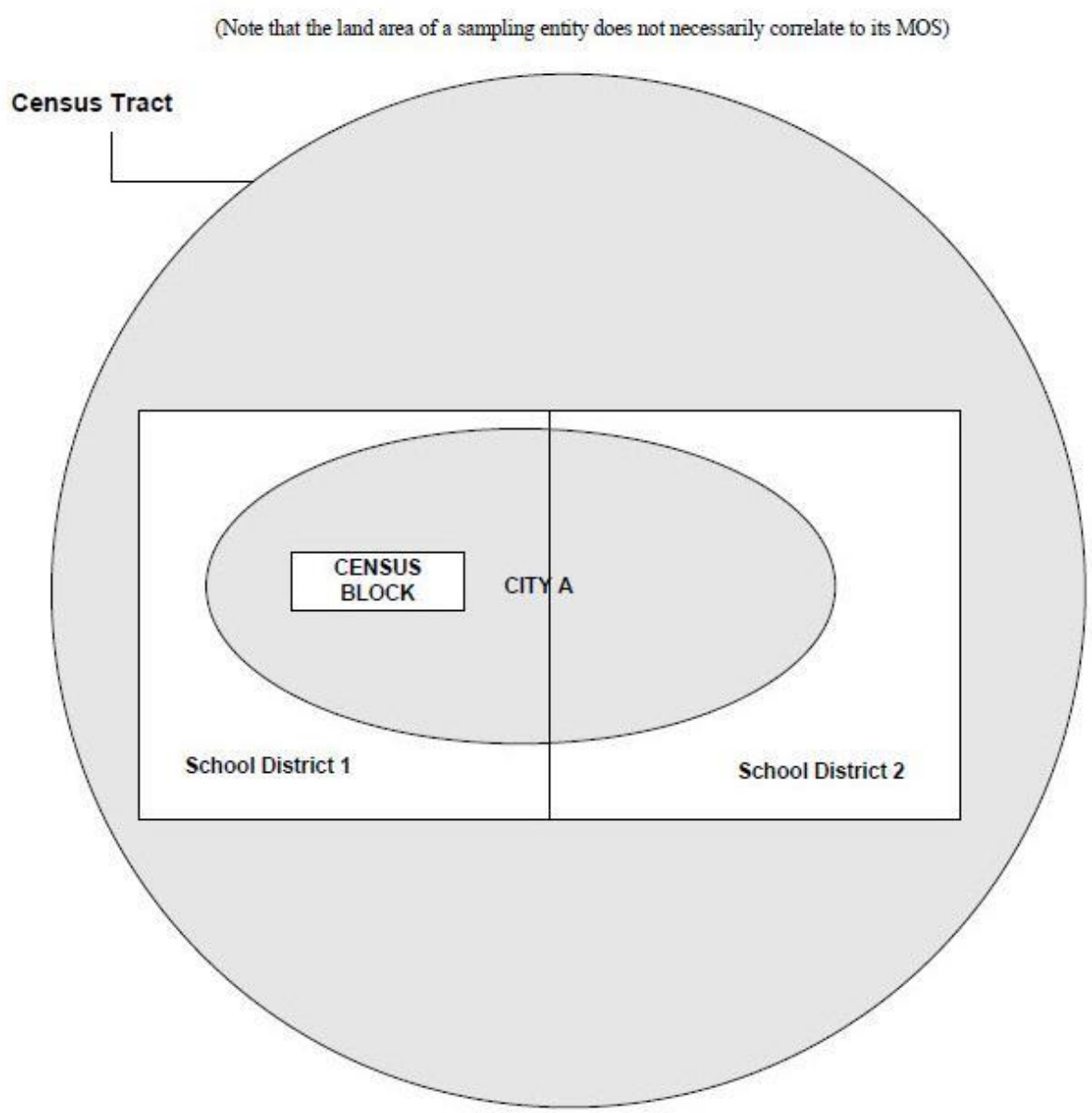


Figure 4-2: Assignment of Blocks (and their addresses) to Second-stage Sampling

First-Phase, Second-Stage Sampling: Sample Month Assignment for Address Samples

We must assign each sample address for a particular year to a specific data collection month. The set of all addresses assigned to a specific month is the month's sample or panel. We sort addresses selected during main sampling by stratum and geography and assign them systematically to the 12 months of the year. However, we assign addresses that one of several Census Bureau household surveys have also selected, to an ACS data collection month based on the interview month(s) for these other household surveys.[17] 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.

We sort the supplemental sample by stratum and geography and systematically assign this sample to the months of July through December. Since this sample is only approximately one percent of the total ACS sample, very few addresses are also in one of the other household surveys in the specified months. Therefore, we chose not to implement the procedure described above to move the ACS data collection month for cases in common with the current surveys during supplemental first-phase sampling.

4.4 Second-Phase Sampling for CAPI follow-up

The ACS uses four modes of data collection—Internet, mail, telephone, and personal visit. (See Chapter 7 for more information on data collection.) Mailable sample addresses are eligible to complete the survey during the entire three-month time period. We send all mailable addresses with available telephone numbers for which we receive no Internet or mail response during the first data collection month to CATI for follow-up. We conduct CATI follow-up for these cases during the second month. Cases without a completed Internet or mail questionnaire or a completed CATI interview are eligible for CAPI in the third month, as are the unmailable addresses. An address is unmailable if the address is incomplete or directs mail to only a post office box. Table 4-3 summarizes the eligibility of addresses for CAPI sampling.

Table 4-3: Addresses Eligible for CAPI Sampling

Mailable Address	Responds to Mailing (Internet or questionnaire)	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

[17]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 CAPI sample selects a systematic sample of these addresses for CAPI data collection each month using the rates shown in Table 4-4. The selection is made after sorting within county by CAPI sampling rate, mailable versus unmailable, and geographic order within the address frame (Keathley, 2010).

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 the subsampling, CAPI cases generally have larger weights than cases completed by Internet, mail or CATI. The variance of the estimates for an area will tend to increase as the proportion of Internet, 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 U.S. that are predicted to have low levels of interviews completed by Internet, 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 interviews and mail/Internet/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 we initially used the sampling rates of 1-in-3 for mailable and 2-in-3 for unmailable addresses. On the basis of early response results observed during the first months of the PRCS, we changed the CAPI sampling rate for mailable addresses in all Puerto Rico tracts to 1-in-2 beginning in June 2005.

We made several enhancements to the CAPI sampling beginning with the 2011 sample, to increase the reliability of the ACS estimates for populations in certain well-defined geographic areas. Beginning in January of 2011, we send all Remote Alaska sample addresses to CAPI where previously they had been sampled at the rate assigned to unmailable cases (2-in-3). In addition, we send all unmailable addresses and all addresses that did not respond via Internet, mail, or CATI to CAPI in the following areas:

- Hawaiian Homelands
- Alaska Native Village Statistical Areas
- All American Indian areas with at least ten percent of the population responding to the 2010 Census as American Indian or Alaska Native (alone or in combination).

Table 4-4 summarizes the CAPI sampling rates that are used for addresses of each particular type.

Table 4-4: CAPI Sampling Rates

Address and Tract Characteristics	CAPI Sampling rate (percent)
United States	
Addresses in Remote Alaska, mailable and unmailable addresses in American Indian areas with 10 percent or more American Indian population (alone or in combination) in the 2010 Census, mailable and unmailable addresses in Hawaiian Homelands, mailable and unmailable addresses in Alaska Native Village Statistical Areas	100.0
Other unmailable addresses	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

4.5 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. We classify each GQ facility according to its GQ type. (For more information on GQ facilities, see Chapter 8.) As noted previously, the 2005 ACS did not include GQ facilities, but we have included GQs since 2006. We select the GQ sample for a given year during a single operation carried out in September and October of the previous year. The most recently available updated MAF as well as lists from other sources and operations define the sampling frame of GQ facilities and their locations. The ultimate sampling units for the GQ sample are the GQ residents, not the facilities. The GQ samples are independent state-level samples.

The ACS sampling and data collection operations exclude certain GQ types , including domestic violence shelters, soup kitchens, regularly scheduled mobile food vans, targeted non-sheltered outdoor locations, 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, we control ACS estimates of the total population to be consistent with the Population Estimates Program estimate of the GQ resident population from all GQs, even those excluded from the ACS.

We classify all GQ facilities into one of two groups: (1) small GQ facilities (having 15 or fewer people according to 2010 Census or updated information); (2) large GQ facilities (with an expected population of more than 15 people). There are approximately 94,000 small GQ facilities and 68,000 large GQ facilities on the 2013 GQ sampling frame. We create two sampling strata to sample the GQ facilities. The first stratum includes both small GQ facilities and those with no available population count. The second stratum 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. We compute a GQ measure of size (GQMOS) for use in sampling the large GQ facilities. The GQMOS for each large GQ is the expected population count divided by 10.

The sampling procedures differ for these two strata. We sample GQs in the small GQ stratum like addresses in the HU sample, and collect data 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 five-year period. People are the ultimate sampling unit for GQs in the large stratum, where groups of 10 people ("hits") 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 and small GQs, large GQ facilities are eligible for sampling each year. For more detail, see the computer specifications for the GQ sampling (Cyffka, 2012).

4.6 Small Group Quarters Stratum Sample

For the small GQ stratum, a two-phase, two-stage sampling procedure is used. In the first phase, we select a GQ facility sample 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 1 systematically assigns small GQ facilities to a sub-frame associated with a specific year. During the second stage, we select a systematic sample of the small GQ facilities. In the second phase of sampling, we interview all people in the facility as long as there are 15 or fewer at the time of interview. Otherwise, we select and interview a sub-sample of 10 people.

First Phase of Small GQ Sampling—Stage One: Random Assignment of GQ Facilities to Sub-frames

The sampling procedure for 2006 assigned all of the GQ facilities in the small stratum to one of five 20 percent sub-frames. We sort the GQ facilities within each state 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, we assigned new GQ facilities systematically to the five sub-frames. The sub-frame for the 2013 GQ sample selection contains the facilities previously designated to the sub-frame for calendar year 2013 and 20 percent of new small GQ facilities added since the 2012 sampling activates. The small GQ facilities in the 2013 sub-frame will not be eligible for sampling again until 2018, since the once-in-five-years 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 systematic sample of the GQ facilities from the assigned sub-frame within each state. The GQs are sorted by new versus previously existing addresses and the order in which they were selected during stage one sampling. Regardless of their actual size, all of these small GQ facilities have the same probability of selection. The second-stage sampling rate combined with the 1-in-5 first-stage sampling rate yields an overall first-phase-sampling rate equal to the sampling rate for each state. As an example, if the sampling rate for the state is 2.5 percent, then the second-stage sampling rate would be 1-in-8 so that overall the GQ sampling would be (1-in-5) × (1-in-8) = 1-in-40 = 2.5 percent. Table 4-5 shows the 2012 state level sampling rates.

Table 4-5: 2012 Group Quarters State-level Sampling Rates

State	Sampling Rate (percent)	State	Sampling Rate (percent)
Alabama	2.17	Montana	3.96
Alaska	4.19	Nebraska	2.46
Arizona	2.05	Nevada	3.63
Arkansas	2.21	New Hampshire	2.90
California	2.49	New Jersey	2.72
Colorado	2.33	New Mexico	2.77
Connecticut	2.37	New York	2.29
Delaware	5.00	North Carolina	2.34
District of Columbia	2.77	North Dakota	4.49
Florida	2.34	Ohio	2.39
Georgia	2.39	Oklahoma	2.39
Hawaii	3.00	Oregon	2.50
Idaho	4.13	Pennsylvania	2.53
Illinois	2.21	Rhode Island	2.63
Indiana	2.35	South Carolina	2.26
Iowa	2.40	South Dakota	3.51
Kansas	2.39	Tennessee	2.30
Kentucky	2.38	Texas	2.12
Louisiana	2.60	Utah	3.00
Maine	3.09	Vermont	4.39
Maryland	2.39	Virginia	2.20
Massachusetts	2.22	Washington	2.45
Michigan	2.79	West Virginia	2.31
Minnesota	2.47	Wisconsin	2.47
Mississippi	2.32	Wyoming	6.97
Missouri	2.25	Puerto Rico	2.50

Second Phase 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, interviewers perform a field sub-sampling operation to reduce the total number of sampled people to 10, similar to the groups of ten selected in the large GQ stratum.

4.7 Large Group Quarters Stratum Sample

Unlike the HU address and small GQ samples, we do not divide the large GQ facilities into five sub-frames. The ultimate sampling units for large GQ facilities are people, not the facility itself, and we conduct interviews in groups of ten. We use a two-phase sampling procedure. The first phase indirectly selects the GQ facilities by selecting groups of ten within the facilities. The second phase 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 the 2010 Census, its GQMOS is 55 and there are 55 groups of ten that are eligible for selection in the sample.

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

We sort all of the large GQ facilities in a state by GQ type and geographical order in the large GQ frame, and select a systematic sample of groups of ten. For this reason, in states with a 2.5 percent sampling rate, 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 in a state with a 2.5 percent sampling rate and between 40 and 80 groups will have at least one group selected with certainty. If the GQ facility has between 80 and 120 groups, it 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 large 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 randomly selects the 10 people to be included, one from each group of ten being interviewed. The instrument is pre-loaded 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 10, 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 pre-loaded 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 10; if the observed population is less than 10, the instrument selects everyone in the GQ.

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, we try 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, we could interview someone in a very large GQ facility in consecutive months. All GQ facilities in this stratum are eligible for selection every year, regardless of their sample status in previous years.

Sample Month Assignment for Small and Large Group Quarter Samples

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

Due to operational and budgeting constraints, we assign the same month to all sample groups of ten within certain types of correctional GQs or military barracks. For example, we assign all samples in federal prisons to September, and data collection may take up to 4.5 months, an exception to the six weeks allowed for all other GQ types. For the samples in non-federal 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.

Beginning with the 2013 GQ sample, we no longer assign college dorms to the months of May-August. This is in response to the relatively low interview rates at these GQs during the summer months. In 2013, we only assign college dorm GQs to January-April and September-December.

4.8 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. There are approximately 30,000 HU addresses and 500 GQs in Remote Alaska. 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 county, place, AIANSA, or block group (in that order) in which it is contained. We assign all designated addresses located in each of these geographical entities to either January or September in such a way as to balance workloads between the months and to keep groups of cases together geographically. We sort the addresses for each month by county and geographical order in the address frame, and beginning in 2011, all sample addresses are sent directly to CAPI (bypassing mail, Internet, and CATI for the HU sample) in the appropriate month. We assign the GQ sample

in Remote Alaska to January or September using the same procedure and allow up to four months to complete the HU and GQ data collection for each of the two data collection periods.

4.9 References

Asiala, M. (2005). American Comunity Survey Research Report: Differential Sub-Sampling in the Computer Assisted Personal Interview Sample Selection in Areas of Low Cooperation Rates. DSSD 2005 American Community Survey Documentation Memorandum Series #ACS05-DOC-2. Washington, DC: U.S. Census Bureau.

Bates, L. (2013). Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey. DSSD 2013 American Community Survey Universe Creation Memorandum Series #ACS13-UC-1. Washington, DC: U.S. Census Bureau.

Cyffka, K. (2012). Specifications for Selecting the American Community Survey Group Quarters Sample. DSSD 2013 American Community Survey Sampling Memoradum Series #ACS13-S-6. Washington, DC: U.S. Census Bureau.

Keathley, D. (2010). American Community Survey: Specifications for Selecting the Computer Assisted Personal Interview Samples. DSSD 2010 American Community Survey Sampling Memorandum Series #ACS10-S-45. Washington, DC: U.S. Census Bureau.

Sommers, D. (2012a). Creating the Governmental Unit Measure of Size (GUMOS) Datasets for the American Community Survey and the Puerto Rico Community Survey. DSSD 2013 American Community Survey Sampling Memorandum Series #ACS13-S-1. Washington, DC: U.S. Census Bureau.

Sommers, D. (2012b). Specifications for Selecting the Main and Supplemental Housing Unit Address Samples for the American Community Survey. DSSD 2013 American Community Survey Sampling Memoradum Series #ACS13-S-3. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau. (2000). American Indian and Alaska Native Areas Geographic Program for Census 2000; Notice. Federal Register , 65 (121), 39062-39069.

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, whose role is to provide the Census Bureau with guidelines for pretesting, field testing, and implementing new content and changes to existing ACS content. This chapter provides detail on the history of content development for the ACS, current survey content, and the content determination process and policy. Especially noteworthy for this Design and Methodology Report is the creation of a new group to provide additional advice and counsel to the OMB and the Director of the Census Bureau on how the ACS can best fulfill its role in the portfolio of Federal household surveys and provide the most useful information with the least amount of burden.

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.[18] Beginning in 2005, the ACS collects this detailed information on an ongoing basis, thereby providing more accurate and timely data than was possible previously. Since 2010, the decennial census only includes a short form that collects basic information for a total count of the nation's population.[19]

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 case law requirements imposed by the U.S. federal court system (required) needed the data.
- The data were necessary for Census Bureau operational needs and there was no explicit

[18] 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.

[19] In addition to counting each person in every household, the basic information included on the 2010 Census short form included a very select set of key demographic characteristics needed for voting rights and other legislative requirements, including tenure at residence, sex, age, relationship, Hispanic origin, and race.

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. The Interagency Committee for the ACS 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 members examined each question and provided for their agency 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 with which the data are needed. They cited the text of statutes and other legislative documentation and classified 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 by committee 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. Since 2002, the process of examining the justifications for each question has been repeated several times. This updating process occurred most recently in 2012. Under the leadership of OMB, a request was sent to Federal agencies to link ACS content to Federal Agency requirements to ensure that federal needs for ACS data are clearly authorized.

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

5.3 Initial ACS/PRCS Content – 2003-2007 Content

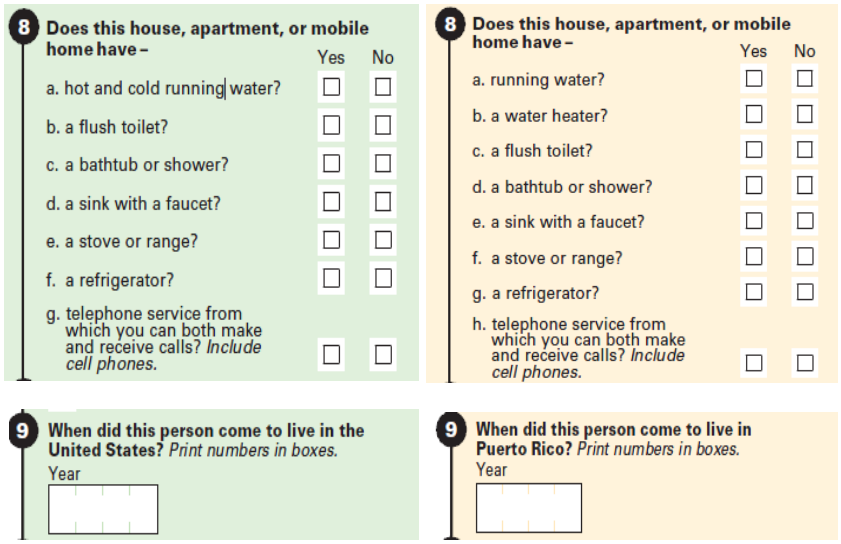
In 2003-2007, the ACS consisted of 25 housing and 42 population questions (six basic and 36 detailed population questions). (See Table 5-1 for a complete list of ACS topics.) The ACS GQ questionnaire consisted of every population question in the population column of Table 5-1, with the exception of the relationship to householder question. The ACS GQ questionnaire also includes one housing question, the food stamp benefit question.

Table 5-1: 2014 ACS Topics Listed by Type of Characteristic and Question Number

Housing	Population
H1 Units in Structure	P1 Name
H2 Year Structure Built	P2 Relationship to Householder
H3 Year Householder Moved into Unit	P3 Sex
H4 Acreage	P4 Date of Birth
H5 Agriculture Sales	P5 Hispanic Origin
H6 Business on Property	P6 Race
H7 Rooms and Bedrooms	P7 Place of Birth
H8 Plumbing and Kitchen Facilities, Telephone Service	P8 Citizenship
H9 Computer Use	P9 Year of Entry
H10 Internet Accessibility	P10 Type of School and School Enrollment
H11 Internet Subscription	P11 Educational Attainment
H12 Vehicles Available	P12 Field of Degree
H13 House Heating Fuel	P13 Ancestry
H14 Cost of Utilities	P14 Language Spoken at Home, Ability to Speak English
H15 Food Stamp Benefit	P15 Residence 1 Year Ago (Migration)
H16 Condominium Status and Fee	P16 Health Insurance
H17 Tenure	P17 Disability: Sensory, Physical
H18 Monthly Rent	P18 Disability: Mental, Self-Care
H19 Value of Property	P19 Disability: Going out Alone, Ability to Work
H20 Real Estate Taxes	P20 Marital Status
H21 Insurance for Fire, Hazard, and Flood	P21 Marital History
H22 Mortgage Status, Payment, Real Estate Taxes	P22 Number of Times Married
H23 Second or Junior Mortgage Payment or Home Equity Loan	P23 Last Year Married
H24 Mobile Home Costs	P24 Fertility
	P25 Grandparents as Caregivers
	P26 Veteran Status
	P27 Years of Military Service
	P28 Veterans Disability
	P29 Worked Last Week
	P30 Place of Work
	P31 Means of Transportation
	P32 Private Vehicle Occupancy
	P33 Time Leaving Home to Go to Work
	P34 Travel Time to Work
	P35 Layoff, Temporarily Absent, Informed of Recall or Return
	P36 Looking for Work
	P37 Available for Work
	P38 When Last Worked
	P39 Weeks Worked
	P40 Usual Hours Worked Per Week
	P41 Class of Worker
	P42 Name of Employer
	P43 Type of Business
	P44 Business Classification
	P45 Occupation
	P46 Primary Job Activity
	P47 Income in the Past 12 Months (by type of income)
	P48 Total Income

Puerto Rico Community Survey (PRCS) Content

The content for the PRCS is identical to that used in the United States (ACS) with the exception of six questions worded differently to accommodate cultural and geographic differences between the two areas. (See Figure 5-1 for an example of ACS questions modified for the PRCS.)



ACS (2013) **PRCS (2013)**

Figure 5-1: Examples of two ACS questions modified for the PRCS

5.4 Content Policy and Content Change Process

In 2006, the ACS content development and change process hinged on the status of a question as ‘‘mandatory,’’ ‘‘required,’’ or “programmatic,” consistent with the constraints of the traditional long form.

In 2006, the OMB, in consultation with Congress and the Census Bureau, adopted a more flexible approach to content determinations for the ACS. In the new content determination process, 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, the OMB will only approve necessary questions for inclusion on the ACS. The OMB’s responsibility under the Paperwork Reduction Act requires that new questions demonstrate the practical utility of the data and that the respondent burden be minimized (especially for the mandatory ACS collections).

The Census Bureau's ACS Content Policy is used as a basic guideline for handling 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.

One vision for the ACS that emerged as the early years of the ACS program was that of a national resource to address emerging policy questions in the public sector. The idea that the ACS could be flexible, providing more frequent opportunities to add new content, via new questions on the ACS itself, or though follow-on or supplementary modules, was put forward in discussions with federal agencies about the program's future. Because response to the survey is required by law, the ACS has generally attained very high participation rates relative to other Federal government surveys, so for users concerned about statistical bias due to nonparticipation, the ACS estimates are attractive. Also, the continuing efforts to improve the Census Bureau' Master Address File that is used for the decennial census, ACS, and other surveys the Census Bureau conducts make the ACS attractive in terms of its coverage rates.

At the same time, and serving to counterbalance this vision, the OMB and the Census Bureau recognized the need to develop priorities for including questions on the ACS and ensure that common decision criteria were used to add or delete a question from the ACS, to use the ACS as a frame for follow-on surveys, or include a module of questions for a subsample of ACS cases. This recognition, and guidance from the OMB that the respondent burden for completing the ACS (measured as the number of minutes each respondent requires to complete the ACS form), would remain fixed, led to the creation of a group tasked with advising the OMB and Census Bureau on issues related to content practices and policies.

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 OMB 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). The Census Bureau will add new pretested questions to the ACS only after the OMB gives approval.

In 2012, the Interagency Council on Statistical Policy Subcommittee on the ACS (ICSP-SACS) was established and tasked with assisting the OMB and Census Bureau through annual and ad hoc activities to review the justifications for ACS questions and propose priorities for including questions on the ACS.

The Charter of the Interagency Council on Statistical Policy Subcommittee on the American Community Survey[20] describes the governing authority for the ICSP-SAC. This document will be updated from time to time as the responsibilities of the subcommittee evolve. Membership on the ICSP-SAC includes representatives from federal statistical agencies and major statistical programs, some of whom serve on rotating basis.

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 are 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 is incorporated into the ACS only after pretesting, potentially including a field test, has been completed, and the OMB has provided final approval.

[20] *Charter* (2012)

- The requesting federal agency assists with the development of a draft question(s), works with the Census Bureau and other agencies to develop or revise the question, and submits 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 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 Content Testing and the ACS Methods Panel

The Census Bureau uses the term "content tests" in describing the testing, research, and evaluation processes used to determine the best wording, format, and placement of proposed new questions or revisions to existing questions on the ACS. Content tests are one of several kinds of tests the Census Bureau conducts as part of the ACS Methods Panel. The ACS Methods Panel tests, in addition to content, proposed improvements to ACS data collection methods and techniques. Such improvements include, for example, the addition to the initial ACS mailing package of a brochure assisting speakers of languages other than English; the use of a new format to organize questions and guidance provided on the ACS questionnaire; or the development of new instructions used by ACS interviewers for the CAPI phase of data collection. This chapter will focus on content testing alone. Descriptions of other kinds of methods panel testing are included in other chapters of this report.

The methodology used for content testing is designed to be similar to ACS data collection in a regular production cycle. The Census Bureau collects data on the quality of the responses obtained in the test. Response variance, gross difference rates, item nonresponse rates, and measures of distributional changes from the regular production cases serve as indicators of the quality of the test questions relative to current ACS questions. Content testing and analysis takes place over approximately two years, so that the results are not implemented until at least two years following the date associated with the content test. The following sections describe Content Tests conducted in 2006, 2007, and 2010, and the implementation of testing results in the content of the 2008, 2009, and 2013 ACS, starting with the 2006 ACS Content Test and concluding with the implementation of the 2013 ACS.

5.6 Content Testing, 2006 to 2009

In 2004, planning began for the 2006 ACS Content Test, so Census could field-test the content changes in the ACS before it finalized the 2008 ACS questionnaire. Consistent with procedures described above, 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.

In 2006, the National Center for Education Statistics (NCES) began work with the U.S. Census Bureau, as well as two groups of academic researchers, to develop and test alternative formats of a Field of Degree question that asked the field of degree for a person having a bachelor's degree or higher level of educational attainment. Based on the preliminary research, Census developed and tested two alternative formats of the question in the Census Bureau's 2007 ACS Methods Test, completed in fall 2007. The result of the Field of Degree Content Test was to use an open ended question for field of degree.

In 2008, no content or other kinds of Methods Panel testing took place because of the unavailability of funding. In 2009, ACS Methods Panel testing did not include any content testing.

5.7 2008-2009 ACS

Reflecting the results of the 2006 Content test, the 2008 ACS included new questions on health insurance coverage, marital history, and Veterans Administration (VA) service-connected disability. Revisions to existing questions were also implemented in the 2008 ACS.

In 2009, the Census Bureau added a field of degree question with an open-ended response category to the 2009 ACS as Question 12. Other changes from the 2008 ACS to the 2009 ACS were relatively few and minor compared to comparable changes from 2007 to 2008. Changes to the ACS questionnaire from 2005 to 2009 are described in detail at: http://www.census.gov/acs/www/Downloads/questionnaires/SQuestChanges05to09.pdf

5.8 2010-2012 Content Testing

In response to federal agencies' requests for new and revised ACS questions, the Census Bureau conducted the 2010 ACS Content Test. The Interagency Committee for the ACS helped identify possible changes to ACS content and additional new content that would be the subject of testing. The primary objective was to test whether changes to question wording, response categories, and the redefinition of underlying constructs could improve the quality of the data collected. The Census Bureau proposed to evaluate changes to the questions, or, for new questions, to compare the performance of question versions to each other as well as to other well-known sources of such information. The proposed topics for content testing were new questions on computer and Internet usage and parental place of birth, and revisions to veteran's identification and period of

service, cash public assistance income, wages and property income, and the Food Stamp program name.

5.9 2010-2013 ACS

Figure 5-2 identifies changes to the ACS paper questionnaire form between 2010 and 2013.

Questions	2010	2011	2013
Housing Questions			
Computer Use			N
Internet Accessibility			N
Internet Subscription			N
Food Stamp Benefit	R		
Population Questions			
Veteran Status			R
Period of Military Service			R
Income in the Past 12 Months by Type of Income			R
Total Income			R
Administrative Pages			
Revised page flow instructions		R	
Cover Page/Front Page			R

R = Revised
N =New

Description of change

2010

- Food Stamp Benefit: Revised the Food Stamp Benefit Question to include the official name for the Food Stamp program (The Supplemental Nutritional Assistance Program) "SNAP" and included additional respondent's instructions about "WIC", the School Lunch Program and the Food Banks

2011

- Revised the instructions located at the end of each detailed person questions' page

2013

- Revised the cover page to include the URL for the Internet Data Collection Mode of the survey https://respond.census.gov/acs
- Added a new question about the type of personal computer the respondent owns and additional instructions were included on which computer devices to exclude from the answer
- Added a new question about internet accessibility
- Added a new question about internet subscription
- Veteran Status: The answer categories were reduced from five answer choices to four answer choices
- Period of Military Services: Reduced the number of answer categories from eleven to nine by merging four categories into two. Merged *May 1975 to August 1980* and *September 1980 to July 1990* into one category, *September 1980 to July 1990*. Merged *February 1955 to February 1961* and *March 1961 to July 1964* into one category *February 1955 to July 1964*
- Income in the Past 12 Months by Type of Income: Increased write-in field length by one for 47a, 47b, and 47c and 48

Figure 5-2: Changes to the ACS Paper Questionnaire Form between 2010 and 2013

A full description of the overall 2010 ACS Content Test and topic-specific research objectives, methodology, and empirical results is available at:

http://www.census.gov/acs/www/library/by_series/content_test_evaluation_reports/

Based on results of the 2010 Content Test, one new question topic, computer ownership and internet usage, was added. In addition, OMB approved the modification of one housing question and four population questions for the 2013 ACS: veterans status and period of service; wages, interest/dividends income; public assistance income; and food stamps.

Computer and Internet Usage

As authorized by the Broadband Data Improvement Act of 2008, the Federal Communications Commission sponsored the computer and Internet usage topic. The Broadband Data Improvement Act requires that the Secretary of Commerce, in consultation with the Federal Communications Commission, expand the American Community Survey to elicit information from residential households, including those located on native lands, to determine whether persons at such households own or use computers at their address, whether persons subscribe to Internet service and, if so, whether they subscribe to dial-up or broadband Internet service at that address. The additions to the questionnaire consist of three questions with a mix of fixed choice and open-ended responses.

Modified Questions

At the request of the Food and Nutrition Service, Census revised one housing question on food stamps to incorporate the program name change to the Supplemental Nutrition Assistance Program (SNAP). The Census Bureau revised the property income and wage questions to improve response by breaking up these questions into shorter pieces to improve comprehension when an interviewer asked the questions. Census incorporated this change into the interviewer-administered modes only. At the request of the Department of Veteran Affairs, Census revised the veteran status and period of service questions to simplify the reporting categories. The new version is for all collection modes.

5.10 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.

Charter of the Interagency Council on Statistical Policy Subcommittee on the American Community Survey, available at http://www.census.gov/acs/www/Downloads/operations_admin/ICSP_Charter.pdf

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. Since 2006, the ACS has 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 6.2 and 6.3 in this chapter detail the interview and residence rules. Section 6.4 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 or Internet responses, 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.[21] 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 two consecutive months or less (often described as the 2-month rule). Under this rule, anyone who has been or will be living for two 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 two months or less, not that they have been staying in the sample unit for two months or less. In general, people who are away from the sample unit for two 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 two 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 two 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 two months. There are three exceptions:

- Children (below college age) who are away at boarding school or summer camp for more than two 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 two months, is considered to be a current resident. A person whose length of stay at the sample HU is for two months or less and has another place where he or she stays for periods of more than two months is not considered a current resident.

[21] 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.

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 includes 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 four data collection modes (Internet, mail, telephone, and in-person interviews). A paper questionnaire and automated Internet instrument are used in the self-response 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 Internet, 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 follow-up 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.
- **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 2013 ACS questionnaire includes 24 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 that extends out at least six inches, 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 and Kitchen Facilities.** Answers to this question are used to estimate the number of HUs that do not have complete plumbing facilities or do not have complete kitchen 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. 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. The house or apartment has telephone service available if cellular telephones are used by household members.

- **Computer Usage.** This question measures usage or ownership of desktop, laptop, netbook or notebook computers, handheld computers, smart mobile phones, other handheld wireless computers, and any other type of computers (which have applications that allow them to function like a desktop or laptop computer). GPS devices, digital music players, and devices with only limited computing capabilities (such as household appliances) are excluded.

- **Internet Access.** This question determines whether any member of the household has access to the Internet at the unit. If yes, it also determines whether that access is provided with or without a subscription service.

- **Internet Subscription Type.** For respondents that indicate that they do access the Internet with a subscription service, this question categorizes the type of subscription used. "Dial-up service" uses a regular telephone line to connect to the Internet. "DSL service" is a broadband Internet service that uses a regular telephone line and, unlike dial-up, allows users to be online and use the phone at the same time. "Cable modem service" is a broadband Internet service that uses a cable TV line. "Fiber-optic service" is a broadband Internet service that uses a fiber-optic line. "Mobile broadband plan for a computer or a cell phone" include wireless broadband Internet service that can be accessed through a portable modem or cell phone. "Satellite Internet service" is a broadband Internet service that uses a satellite dish.

- **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 one 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.

- **Supplemental Nutrition Assistance Program Benefit.** The Food and Nutrition Service of the U.S. Department of Agriculture (USDA) administers the Supplemental Nutrition Assistance (Food Stamp) Program through state and local welfare offices. This 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 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 three 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 three 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.

- **Field of Degree**. Persons with a bachelor's degree or higher are asked to provide the specific major of this person's bachelor's degree. If this person has more than one bachelor's degree, or more than one major, this question requests names of all majors for all bachelor's degrees.

- **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.

- **Health Insurance.** This question measures the insured and uninsured by asking about coverage through an employer, direct purchase from an insurance company, Medicare, Medicaid or other government-assistance health plans, military health care, VA health care, Indian Health Service, or other types of health insurance or coverage plans. Plans that cover only one type of health care (such as dental plans) or plans that only cover a person in case of an accident or disability are not included.

- **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.

- **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.'' "Now married" includes married persons regardless of whether his or her spouse is living in the household, unless they are separated. If the person's only marriage was annulled, the person should be classified as "never married." The "divorced" category should be selected

only if the person has a divorce decree. Couples who live together (unmarried people, people in common-law marriages) report the marital status they consider the most appropriate.

- **Marital History.** Data are collected on whether the person got married, widowed or divorced in the past 12 months, the total number of times the person has been married, and year in which the person last got married. A person is considered divorced in the past 12 months only if the person has received a divorce decree in the past 12 months. Marriages ending in annulment should not be included in the count of total marriages.

- **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 four to six months of initial training or yearly summer camps. All other civilians aged 18 and older are classified as nonveterans.

- **Service-Connected Disability Rating.** This question determines whether the person has a VA service-connected disability rating, and if yes, identifies the numeric rating. The "0 percent" category should only be selected if the person has a service-connected disability rating of zero, and should not be used to indicate no rating.

- **Work Status.** People aged 16 and older who have worked one 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 four 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 GQ questionnaire includes all of the population items included on the HU questionnaire, except for relationship. One housing question, Supplemental Nutrition Assistance Program 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

The data collection operation for housing units (HUs) consists of four modes: Internet, mail, telephone, and personal visit. For most HUs, the first phase includes a mailed request to respond via Internet, followed later by an option to complete a paper questionnaire and return it by mail. If no response is received by mail or Internet, 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).

The ACS includes 12 monthly independent samples. Data collection for each sample lasts for three months, with mail and Internet 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/Internet phase, one is in the CATI phase, and one is in the CAPI phase.

ACS sample panel	Month of data collection					
	2013					
	January	February	March	April	May	June
November 2012	Personal visit					
December 2012	Phone	Personal visit				
January 2013	Mail/ Internet	Phone	Personal visit			
February 2013		Mail/ Internet	Phone	Personal visit		
March 2013			Mail/ Internet	Phone	Personal visit	
April 2013				Mail/ Internet	Phone	Personal visit
May 2013					Mail/ Internet	Phone
June 2013						Mail/ Internet

Figure 7-1: ACS Data Collection Consists of Three Overlapping Phases

Figure 7-2 summarizes the distribution of interviews and noninterviews for the 20012 ACS. Among the ACS sample addresses eligible for interviewing in the United States, approximately 48 percent were interviewed by mail, seven percent by CATI, and 42 percent were represented by CAPI interviews. Three percent were noninterviews.

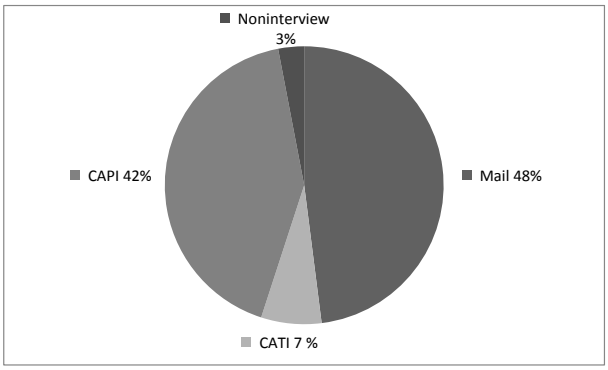


Figure 7-2: Distribution of ACS Interviews and Noninterviews – Source: 2012 ACS Sample

7.2 Mail and Internet Phase

Self-response by mail or Internet is the least expensive method of data collection, and the success of the program depends on high levels of self-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 97 percent of the 2012 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 64 percent of the addresses in Puerto Rico were considered mailable in 2012.

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/Internet and telephone phases.

Mail and Internet Strategy

Because a high level of self-response is critical, the ACS employs multiple mailings to encourage respondents to complete the survey via the Internet or to return a paper questionnaire. 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 and additional reminder postcard are sent to sample addresses when there is no response two weeks after mailing the initial mail package. Households that have not responded by mail, but are not eligible for telephone follow-up are sent yet another postcard at the start of the following month. (Details of each are provided below.)

Prenotice Letter. The first mailing consists of a prenotice letter, signed by the Census Bureau's director, alerting residents that they will receive instructions on how to complete the survey in a few days and encouraging them to do so 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 two days of the month, in which case the emailout schedule begins one 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. In addition to the prenotice letter, a multi-lingual brochure is included in this mailing. This brochure provides general survey information in English, Spanish, Russian, Chinese, Korean and Vietnamese, and also provides a toll-free number for respondents to receive telephone questionnaire assistance in each language.

Initial Mail Package. The next mailing is the initial mail package. On the front of the envelope is a boxed message 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, an instruction card for responding via the Internet, and a brochure.

- **Cover Letter.** The cover letter is from the Census Bureau's director. It reminds householders that they received the prenotice letter a few days earlier and encourages them to go online to complete the survey as soon as possible. The letter then explains the purpose of the ACS and how the data are used, as well as informs the respondent that if they do not have access to the Internet, a paper questionnaire will automatically be sent to them. Finally, a toll-free telephone number is included for respondents if they have questions or need help completing the questionnaire.

- **Instruction Card.** This card directs respondents to the website where respondents will complete the survey via Internet, and informs them they will need information pre-printed on the card in order to log into the survey. It also provides a toll-free number they may call if they have questions or need help. The card provides this information in English on one side, and in Spanish on the other.

- **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.

First Reminder Postcard. The third mailing is a postcard, printed on white cardstock and signed by the director of the Census Bureau. The postcard is mailed on Thursdays, three days after the initial mail package, and reminds respondents to complete the survey via the Internet. The reminder postcard also is printed in-house, using print-on-demand technology to merge text and addresses.

Replacement Mail Package. The fourth mailing is sent only to those sample addresses from which the initial questionnaire has not been returned. It is mailed on Thursdays, about 2½ weeks after the initial mail package. The contents are similar 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. Additionally, an ACS questionnaire and postage-paid return envelope is included.

ACS Questionnaire. The 2013 ACS questionnaires are 28-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, and information on how to respond via the Internet. 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.

Second Reminder Postcard. The fifth mailing is a postcard, printed on white cardstock signed by the director of the Census Bureau. The postcard is mailed on Mondays, three days after the replacement mail package, and reminds respondents to return their questionnaires or respond via Internet. This postcard also is printed in-house, using print-on-demand technology to merge text and addresses.

Additional Postcard. The final mailing is sent at the start of the second month to only those households that did not respond via mail or Internet, and for whom we have not obtained a phone number to contact them during the telephone phase. This card is printed on green cardstock and

larger than the other postcards sent to the household, and explains that we may contact them in person if they do not complete the survey.

In Puerto Rico, a slightly different set of mailings are used since there is not an Internet response option available for Puerto Rico households. In Puerto Rico, the initial mail package includes a paper questionnaire rather than an instruction card for responding via Internet, the replacement packages does not include an instruction card for responding via Internet and there is no second reminder postcard. The timing of the replacement mail package is approximately 3 ½ weeks after the initial mail package for Puerto Rico households.

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 three days before the mailing of the replacement mail package. This file excludes addresses from which a response was received by mail or Internet during the first two 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. An additional updated address file for the additional postcard is provided to the NPC three weeks after replacement mailings which excludes addresses from which a response was received via mail or Internet as well as addresses sent to the telephone phase of follow-up.

Most mail and Internet responses are received within five weeks after the initial mail package is sent, but the Census Bureau will continue to accept mail or Internet responses for three months from the start of each monthly sample. After a specified cutoff date, late returns will not be included in the data set.

Check-In of Paper Questionnaires

The United States Postal Service (USPS) returns all completed ACS paper 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 three days to check in a form, although usually they check in all the forms they receive within one 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 are ineligible for the replacement mail packages. UAAs are eligible for the CATI and CAPI operations.

Telephone Questionnaire Assistance (TQA)

Respondents that call the toll-free TQA number reach an interactive voice recognition (IVR) telephone system that provides answers to questions about completing the questionnaire, or assists respondents in requesting a questionnaire 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 completed survey 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 gives the respondent the option to complete the questionnaire over the telephone. Agents use an automated survey instrument to capture the respondent’s answers. Respondents may also contact TQA staff in order to reset their Internet questionnaire due to losing the PIN number provided when the respondent first accessed their questionnaire online.

Household members from approximately six percent of the mailable addresses called the toll-free number for assistance in 2006 and 2007. For less than one percent of the mailable addresses in 2011 and 2012, 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 of Paper Questionnaires

After the paper 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 three 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 from paper or the data is returned from the Internet, the data files are processed in batches through a computerized edit to check coverage consistency. 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 completed surveys.

Cases failing the edit for coverage consistency can take three forms. First, since the ACS paper 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. Second, Internet responses that indicate the sample address is vacant or a business are also treated as coverage failures. Third, the amount of data-defined fields is less than expected and the case data will not be retained for downstream processes. 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 three 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 or Internet 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. 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 five 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 responds to the survey via mail or Internet at any time during the CATI operation, that case is removed from the CATI sample and is considered a mail/Internet 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 key match the answers provided by the respondent.

Approximately 850 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 2012about 110,000 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/Internet and CATI operations have been completed, a CAPI subsample is selected from two categories of cases. Mailable addresses with neither a mail/Internet response 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. All eligible addresses in Hawaiian Homelands, Alaska Native Village Statistical areas and a subset of American Indian areas are sent to CAPI without subsampling. (U.S Census Bureau 2012).

The CAPI operation is conducted by Census Bureau field representatives (FRs) operating from the Census Bureau's six regional offices (ROs). The sampled cases are distributed among the six ROs based on their geographic boundaries. The New York 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 two months.

The FRs are trained to remain polite but persistent when attempting to obtain responses. They also are trained on how to handle almost any situation, from responding to a household that claims to have returned its questionnaire by mail, or responded by Internet, 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 averaged approximately 58,000 cases each month in 2012. 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. Remote Alaska respondents are also not eligible to respond to the ACS via the Internet. 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, and we do not subsample for CAPI in Remote Alaska as we do elsewhere.

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 one month to four 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, respond using the Internet, 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. 2012. ‘‘Accuracy of the Data (2012).’’ Washington, DC, 2012, http://www.census.gov/acs/www/Downloads/data_documentation/Accuracy/ACS_Accuracy_of_Data_2011ccuracy/ACS_Accuracy.

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/university student housing, residential treatment centers, skilled nursing facilities, group homes, military barracks, correctional facilities, workers' group living quarters and Job Corps centers, and emergency and transitional shelters. 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. Data collection for each monthly sample lasts six 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 the administrator of the selected GQ (referred to as the GQ level interview), and second, the FR conducts interviews with a sample of individuals from the facility (referred to as the person- or resident-level interview).

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 to draw a random sample of residents 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 uses a Computer-Assisted Personal Interview (CAPI) automated instrument to collect detailed information for each sampled resident. FRs also have the option to distribute a bilingual (English/Spanish) questionnaire to residents for self-response if unable to complete a CAPI interview. 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:
http://www.census.gov/acs/www/about_the_survey/forms_and_instructions/#ACSGQ

Preparation

Outside vendors print most GQ data collection materials, such as the questionnaires, the questionnaire information guide booklet, brochures, the information card booklet, and Privacy Act notices. 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. On a monthly basis, the Census Bureau headquarters provides label/address files for GQ materials to the NPC. The NPC receives the files approximately eight weeks prior to the sample months, and is responsible for using these files to assemble GQ and resident-level packages. Each GQ level package contains questionnaire labels, the FAQ brochure, the Survey Package Control List for Special Sworn Status (SSS) Individuals, an Instruction Manual for SSS Individuals, a listing sheet, Thank You letter, and a copy of the Introductory Letter mailed to the GQ. Each resident-level package includes a questionnaire, resident level Introductory Letters, FAQ Brochure, and a Thank You letter. The NPC delivers both packages and other materials to the ROs two weeks before the start of each survey monthly panel.

8.2 Group Quarters (Facility-Level Phase)

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

During the GQ-level interviews, FRs verify sample GQ information such as the name, address, and GQ type. FRs also obtain a roster of residents currently living at the GQ. The GQFQ randomly selects residents for person-level interviews. The information obtained from GQ 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 two 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 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 the NPC using

print-on-demand technology, which merges the letter text and the sample GQ name and address. There are special letters for administrators at college/university student housing and health care facilities because these are some of the most challenging GQ types to interview.

GQ Frequently Asked Questions (FAQ) Brochure. This brochure contains FAQs about the ACS and GQ facilities, and is mailed to the sample GQ facility along 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.

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 try to gain the GQ contact’s 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. The ACS Group Quarters Branch may also contact the GQ to explain the nature of the survey and try to gain cooperation.

Visiting the GQ Facility

Upon arrival at the facility, the FR updates or verifies the 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. Depending on the size of the facility, either a sample or all of the residents will be interviewed.

After determining that the GQ facility is in scope for GQ data collection, the FR asks for a roster 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 roster 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 GQFQ 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, but does not conduct person-level interviews.

The sample of GQ residents to be interviewed is generated from the GQFQ instrument through a systematic sample selection. (See Section 8.3 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 the line number corresponding to the register, a place for name, the sample person location description, a telephone number, 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.

An interim or final outcome is assigned to each GQ-level interview, and reasons for GQ refusals or noninterviews are also specified. The GQFQ assigns 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.

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 up to 15 sample residents at each assigned GQ facility using a Computer-Assisted Personal Interviewing (CAPI) automated instrument.

Person-Level Survey Instruments and 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.

Computer Assisted Personal Interviewing (CAPI) Questionnaire Instrument (QI). The CAPI QI is the preferred method of data collection. FRs use the CAPI QI to conduct face-to-face interviews with sample GQ residents. Interviews can be conducted in both English and Spanish. The GQ QI instrument is designed to record detailed population information for one person. It does not include housing questions except for the food stamp benefit question. The QI contains skip patterns based on GQ type. For example, sample residents living in nursing facilities and correctional facilities are not asked the journey-to-work questions.

ACS GQ Paper Questionnaire. The FR distributes a paper GQ questionnaire to residents for self-response when a face-to-face CAPI interview cannot be conducted. This questionnaire is a bilingual, 16 page, two-color, flip-style booklet. Eight blue pages make up the English language GQ questionnaire and, when flipped over, eight green pages make up the Spanish language version. Like the QI, the GQ paper questionnaire is designed to record detailed population information for one person. It does not include housing questions except for the food stamp benefit question. The paper questionnaire does not have skip patterns based on GQ type.

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 Frequently Asked Question Brochure. Every sample GQ resident is given a FAQ brochure. This brochure provides answers to questions about the ACS GQ program.

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

Completing the GQ CAPI Automated Questionnaire Instrument (QI) or Paper Questionnaire

There are several ways for an FR to obtain a completed interview. The preferred method is for the FR to conduct a face-to-face interview with the sampled resident using the CAPI Questionnaire Instrument (QI); however, other data collection methods may be necessary. The FR may conduct a telephone interview with the sample resident using CAPI QI, conduct a face-to-face CAPI proxy interview with a relative, guardian, or GQ contact; or leave a paper 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 two 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 Special Sworn Status (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. Relatives or 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

When a CAPI interview is conducted, the QI contains automated edit checks within the instrument to ensure the quality of the interview and to determine the final outcome of the interview (completed, sufficient partial, or insufficient partial interview). When paper questionnaires are used for self-response, the edit screen in the CAPI instrument is used by FRs to verify 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 FRs also determine whether the self-response questionnaire is a completed interview, a sufficient partial, or an incomplete interview. The FR records the final outcome code for each self-response paper questionnaire on the Census Use Only page on the questionnaire.

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 conducts a GQ-level assignment review. This review is necessary to ensure that all CAPI interviews have been conducted, and that all self-response questionnaires dropped off have been accounted for.

FRs ship paper 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. CAPI interviews are transmitted from laptops to ACSO processing staff on a nightly basis.

8.4 Check-In and Data Capture

CAPI QI interview data is transmitted nightly via automated procedures to Census Bureau headquarters. Based on the final outcome code recorded for each paper questionnaire, the RO separates blank questionnaires from those with data. Only questionnaires that contain completed, or sufficient partial data 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 three days to check in a form, although they usually check in all the forms they receive within one day. The check-in process results in batches of 50 questionnaires for data capture.

The NPC accepts completed questionnaires shipped from the RO on a weekly basis, for a period of six weeks from the start of the sample month. Each RO closes out the sample month GQ assignments and accounts for all questionnaires. The NPC completes the sample month check-in within seven 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-Image (KFI) technology. NPC clerical staff key the data from the questionnaires and transmit data files to Census Bureau headquarters each night. Keyers have approximately two months to complete the keying for a given interview panel.

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.

Since 2007, there have been concerns about the coverage/data from college dorms during the summer months, identified as May-August. Fewer students live in dorms during the summer months compared with other months of the year, which causes a large increase in the number of noninterviews during this period. Beginning January 2013, the ACS collected data at college dorms in only the nonsummer months of January-April, and from September-December of each data collection year. By reallocating dorms from the summer months into nonsummer months we expect that we will improve GQ data in several ways. For example, we expect to have better data from which to impute responses. In addition, this change will reduce data collection costs because fewer FRs will be sent to college dorms that are largely vacant in summer months.

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 7.4 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 one 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 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.

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. The ACS program provides language assistance in many forms, including translated instruments and other survey materials, bilingual interviewers, and multiple language support by telephone. 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. Language support can help individuals with limited English skills to understand the survey questions, their rights as respondents, and the importance of the ACS.

The ACS language assistance program includes the use of several tools to support each mode of data collection—mail, Internet, telephone, and personal visit. Staff developed these tools 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, the ACS successfully interviewed non-English speakers 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 language program includes assistance in many languages during the telephone and personal visit nonresponse follow-up stages, as well as some assistance in other languages during the mail and Internet phases.

This chapter provides 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 each of the four data collection modes. The chapter closes with a discussion of associated research and evaluation activities.

9.2 Background

The 2010 Decennial Census Program 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 worked 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 2011 ACS found that 60.6 million people (20.8 percent of the population five years and over) spoke a language other than English at home with about 41.8 percent speaking English less than ‘‘very well.’’ The population five years and older speaking a

language other than English at home in 2011 represents a 158.2 percent increase since 1980. (Ryan, 2013).

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 surveys and censuses follow specific guidelines when they translate data collection instruments, respondent letters, and other respondent materials.

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 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 2004). The ACS utilizes these Census Bureau guidelines in the preparation of data collection instruments, advance letters, and other respondent communications.

9.4 Mail and Internet Data Collection

Beginning in January 2013, the ACS added an Internet option to complete the survey online. The mailing requesting response by Internet and the Internet instrument are available in both English and Spanish. The Census Bureau currently mails ACS questionnaires to each nonresponding address in a single language. In the United States, households receive English language forms, while in Puerto Rico, they receive Spanish forms. The cover of the English and Spanish questionnaires contain 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 2012, the Census Bureau received requests for Spanish questionnaires from less than 0.01 percent of the mailout sample, approximately 200 forms requests per panel (Fish, 2013). In 2011, the Census Bureau added to the pre-notice letter a multi-lingual brochure tested in 2009 and providing information in English, Spanish, Russian, Chinese, Korean, and Vietnamese (Joshipura, 2010). In 2012, the Census Bureau began making available Chinese and Korean language assistance guides when requested by the respondent.

Language assistance guides include a full translation of the questionnaire for use as reference by both respondents and interviewers.

The ACS provides telephone questionnaire assistance in English, Spanish, Chinese, Russian, Korean, and Vietnamese. A call to the toll-free Spanish, Chinese, Russian, Korean, and Vietnamese help numbers reaches an in-language 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. Fish (2010a) and Fish (2010b) estimate the language needs in the 2006 - 2008 ACS CAPI and CATI operations. She found that the language workloads in the regional offices were stable over time and that the regional offices successfully met the language needs of the population in their regions by hiring field representatives with necessary language skills. She also found that the call centers successfully support at least 10 of the top 14 critical language needs encountered during CATI.

The regional offices train CAPI interviewers 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 language preferred by a household respondent. 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, Vietnamese, Polish, Portuguese, French, Haitian-Creole, or Arabic 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.

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 has created language assistance guides in Chinese and Korean for interviewers to use while interviewing. These language assistance guides contain the preferred translation in Chinese and Vietnamese. The ACS developed special procedures and an interviewer training module dealing with the collection of data from respondents who do not speak English. The standard classroom interviewer training for all ACS interviewers includes this language assistance 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.

Bilingual interviews currently provide support in more than 30 languages. Interviewer language capabilities include English, Spanish, Portuguese, Chinese, Russian, French, Polish, Korean, Vietnamese, German, Japanese, Arabic, Haitian Creole, Italian, Navajo, Tagalog, Greek, and Urdu.

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 interviewers complete survey interviews. The instruments 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 eight 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 managers to identify CATI and CAPI cases that the call centers and the regional offices did not complete due to language barriers. A profile of this information by language highlights those languages needing greater support. For example, Fish (2010a) found that over the period 2006 to 2008, some regional offices' total language CAPI workloads experienced moderate changes, while others' total language workloads remained stable. These changes were driven mostly by an increase or decrease in the regional offices' English and/or Spanish language workloads. This research also demonstrated that estimated language workloads and the estimated linguistically isolated language workloads aligned well with the available language assistance resources. Regional offices have hired field representatives with the necessary language skills to accommodate their unique linguistically isolated language workloads.

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, the GQ data collection currently uses this form. 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, the ACS established early research and development priorities for the language assistance program. Of critical importance was a benchmarking of the effectiveness of current methods. McGovern (2004) and Griffin and Broadwater (2005) assessed the potential for nonresponse bias due to language barriers. 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. McGovern (2004) also assessed item level nonresponse. She found that the mail returns received from non-English speakers were nearly as complete as those from English speakers and that the interviews conducted by telephone and personal visit with non-English speakers were 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 developed and tested translated language guides for use by respondents and telephone and personal visit interviewers who conduct interviews in Korean and Chinese to reduce the potential for translation errors. The Census Bureau has completed a complete assessment of the Spanish instrument to improve the quality of data collected from Spanish-speaking households.

To improve response in languages other than English and Spanish, the ACS tested inserting a multi-lingual brochure into the mailings. That brochure includes translations of key messages, encouraging respondents to call a toll-free number for assistance. As noted earlier, the ACS added this brochure in 2011. For details of this testing, see Joshipura (2010.) ACS managers plan research and development of additional language assistance materials for the mail and Internet modes. Increasing levels of participation by mail and Internet can reduce survey costs and improve the quality of final ACS data.

9.8 References

Fish, Samantha. (2010b). “Assessment of Language Needs and Language Assistance Resources in the 2006-2008 Computer Assisted Telephone Operation.” Final Report. Washington, DC: U.S. Census Bureau, 2010.

Fish, Samantha. (2013). “Percent of Spanish Questionnaire Requests Out of Mailout Sample. “2013 American Community Survey Office Special Studies Staff Memorandum Series #SSS13-3.” December 17, 2013.

Fish, Samantha. (2010a). “Assessment of Language Needs and Language Assistance Resources in the 2006-2008 Computer Assisted Personal Interviewing Operation.” Final Report. Washington, DC: U.S. Census Bureau, 2010.

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.

Joshipura, Megha. (2010), “Evaluating the Effects of a Multilingual Brochure in the American Community Survey.” Final Report. Washington, DC: U.S. Census Bureau, 2010.

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.

Ryan, Camille. 2013. Language Use in the United States: 2011, American Community Survey Reports, ACS-22. U.S. Census Bureau, Washington, DC.

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

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 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.

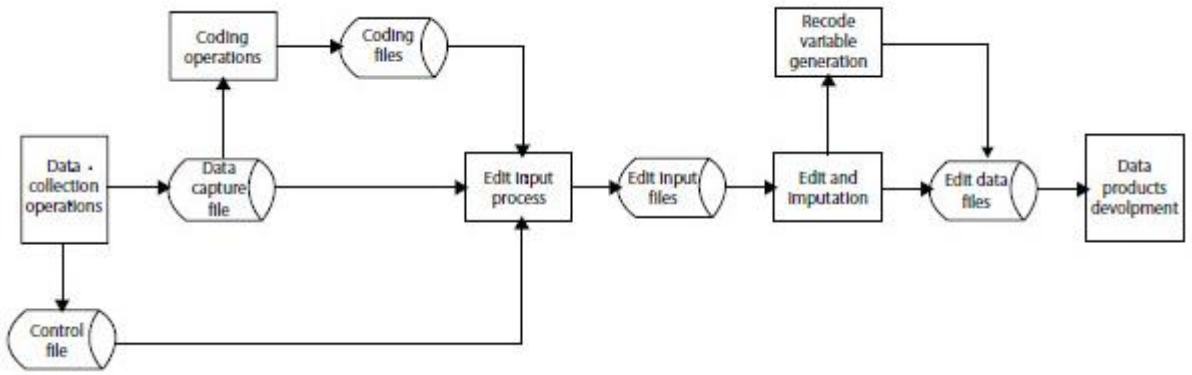


Figure 10-1: American Community Survey (ACS) Data Preparation and Processing

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/internet operations, it documents the receipt and check-in date of questionnaires returned by mail or the date completed on the internet. The status of data capture for 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/internet, 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/internet, CATI, and CAPI. Sampled households first are mailed a request to complete their form on the internet and then are subsequently mailed the ACS questionnaire. Households that do not complete their form by mail/internet self-response and for which a phone number is available receive telephone follow-up. As discussed in Chapter 7, a sample of the non-completed 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/internet are processed daily and are added to a Data Capture File (DCF) on a monthly basis. Data received by mail/internet 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/internet 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.

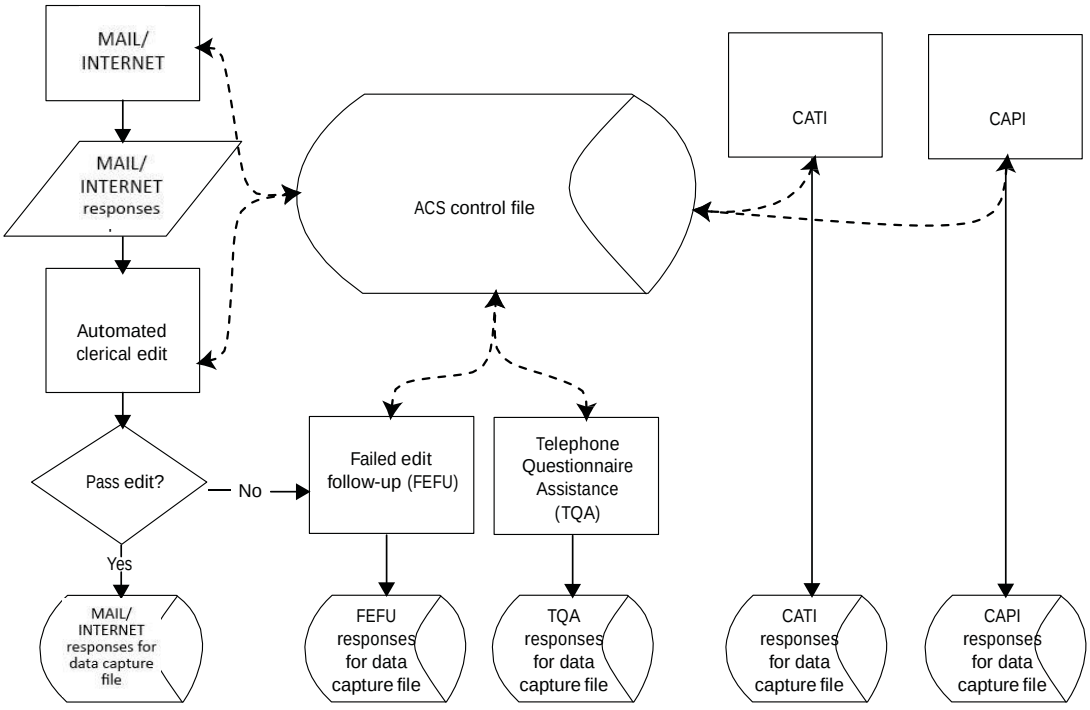


Figure 10-2: Daily Processing of Housing Unit Data

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 or from a CAPI collection operation.

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/internet self response, 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. The same rules are applied to internet responses through an automated set of business rules and are grouped in with the mail data in the capture files. Once the data capture files have gone through this initial data cleaning, the next step is processing the HU questions that require coding.

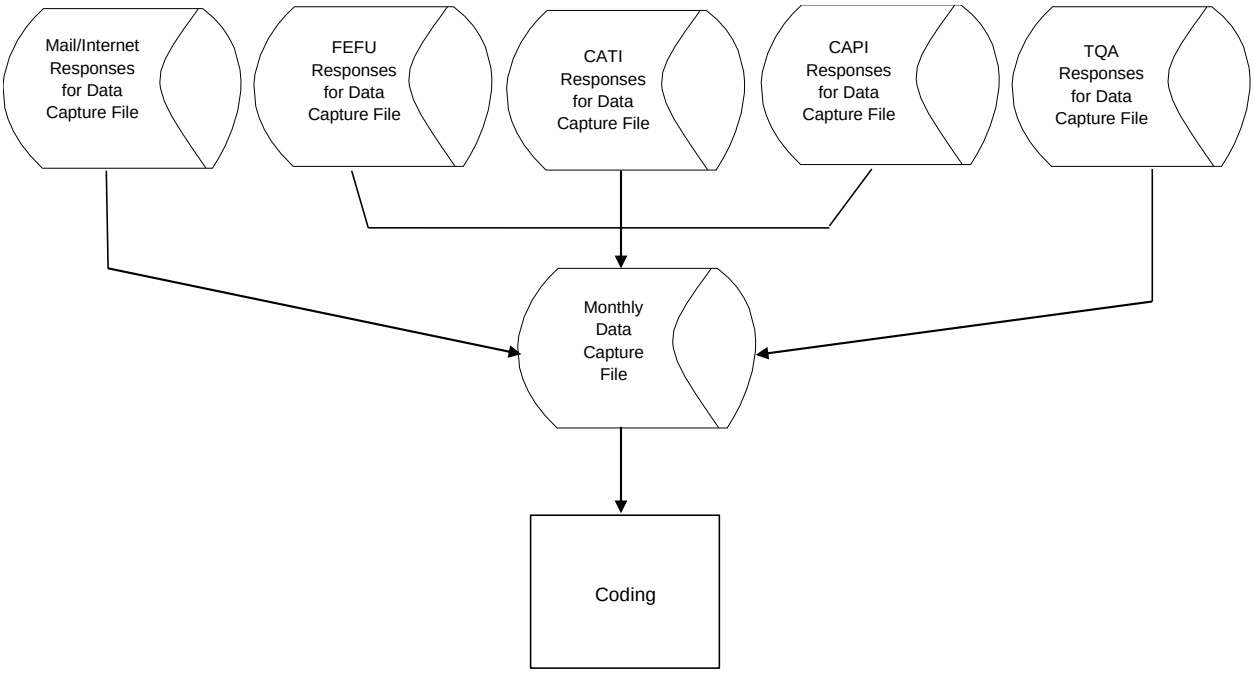


Figure 10-3: Monthly Data Capture File Creation

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.

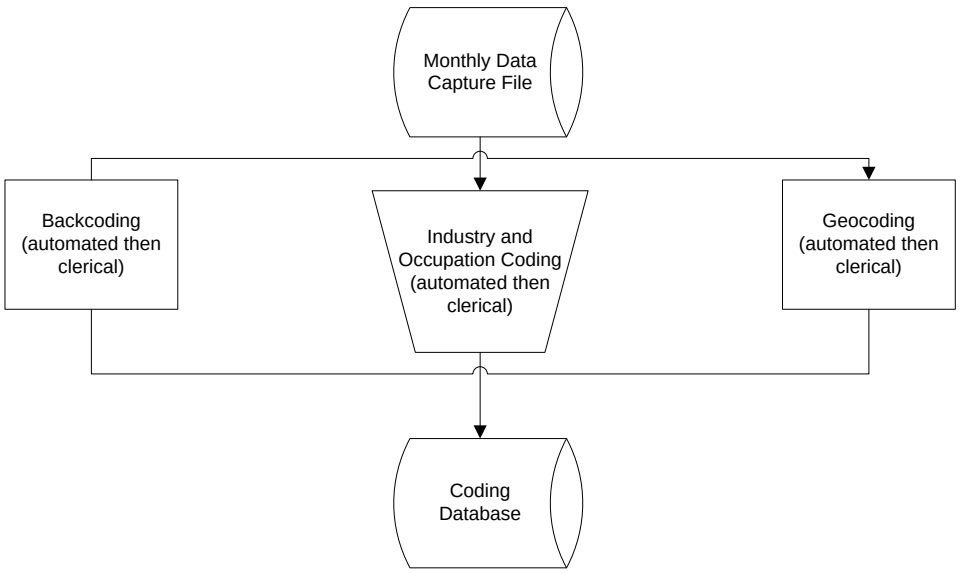


Figure 10-4: American Community Survey (ACS) Coding

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: Type and Method of Coding

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
Health Insurance	Backcoding	Automated with clerical follow-up
Field of Degree	Backcoding	Automated with clerical follow-up
Computer Types	Backcoding	Automated with clerical follow-up
Internet Service	Backcoding	Automated with clerical follow-up
Industry	Industry	Automated with clerical follow-up
Occupation	Occupation	Automated with clerical follow-up
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

In 2013, the ACS added an internet option for response. The autocoding and clerk clerical coding processes have remained the same as for all modes of response. However, the Census Bureau anticipates the rules to change for converting internet response text to data files or for keying mail response text to data files. Rule changes may affect the treatment of special characters (e.g., colons), resulting in a decrease in the match rate of incoming responses to the autocoder data dictionaries, which are based on the current rules. If rule changes occur, then the autocoder may add a process to adjust the incoming text data to match the data dictionaries. For example, the process could remove special characters allowed under the new rules. For records that still require clerk coding after autocoding, the new special characters would be retained and provided to the clerks. If changes occur that affect any part of the industry and occupation coding processes, future editions of this report will discuss the processing adjustments in detail.

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 processes.

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, a percentage of the backcoded items are sent through the quality assurance (QA) process. The algorithm is specified according to the number of returns in a batch. 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.

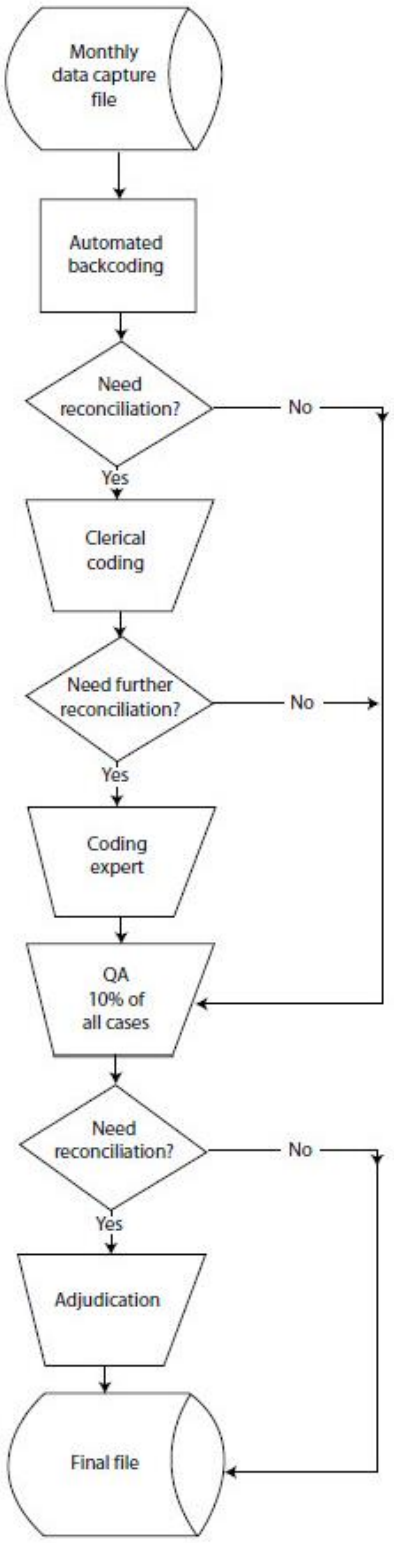


Figure 10-5: Backcoding

Industry and Occupation Coding

The second type of coding is industry and occupation coding, which is slightly different from backcoding. 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 the resident population is 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 write-in 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.

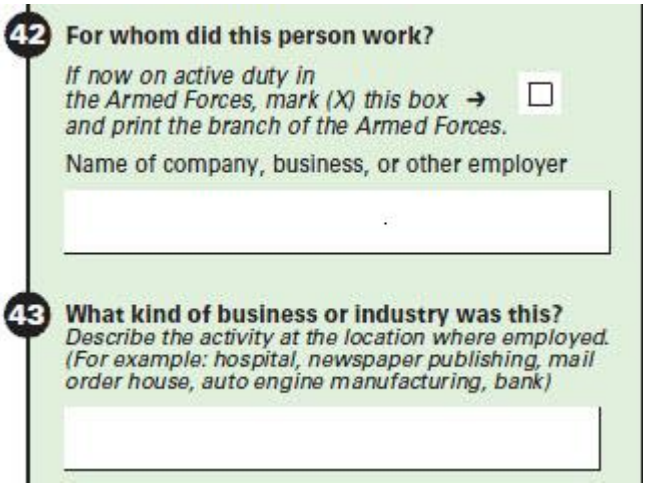


Figure 10-6: ACS Industry Questions

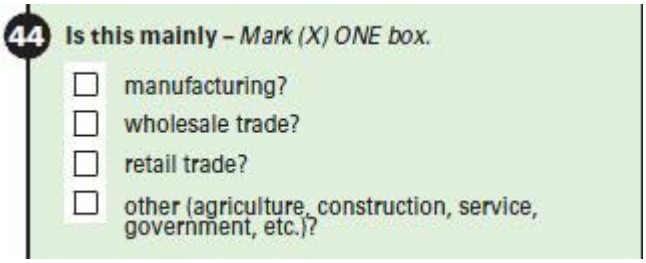


Figure 10-7: ACS Industry Type Question

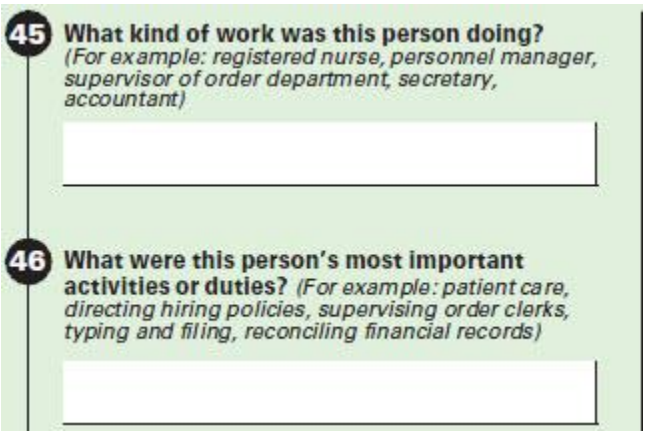


Figure 10-8: ACS Occupation Questions

From these questions, monthly processing converts industry and occupation write-in responses to a code category. Prior to 2012, specialized industry and occupation coders assigned all codes. Industry and occupation items did not go through an automated assignment process. Beginning with the 2012 data collection, industry and occupation coding incorporated automated assignment as a first step in coding. This industry and occupation autocoder is a set of logistic regression models, data dictionaries, and consistency edits ("hardcodes") developed from around two million clerk-coded records, including group quarters and Spanish records (Thompson, et al., 2012). The autocoder assigns an industry or occupation code if the quality score, based on agreement with clerk-coded records, is sufficiently high. If one or both of industry or occupation remain unassigned, these residual records are then assigned a code by specialized industry and occupation clerical coders. When clerical coders are unable to assign a code, the case is sent to an expert, or coding referralist, for a decision. Once these cases are assigned both an industry and an occupation code, they are placed in the general pool of completed responses. Both the autocoder and the clerical coding have independent quality check processes. Figure 10-9 illustrates the industry and occupation coding process beginning with 2012 data.

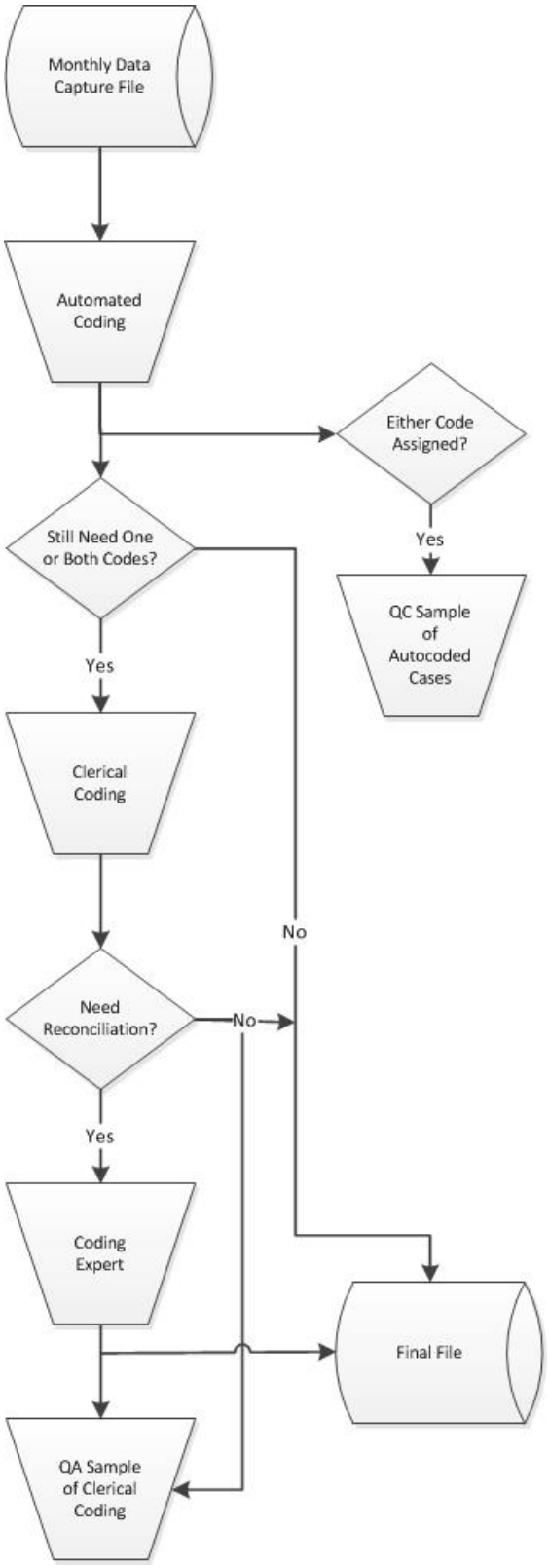


Figure 10-9: Industry and Occupation (I/O) Coding

Both the automated and clerical industry and occupation coding use the Census Classification System to code responses. This system is based on the North American Industry Classification System (NAICS) for industry and the Standard Occupational Classification (SOC) for occupation. The Census Classification System's 4-digit code categories can be bridged directly to the NAICS and SOC for comparisons. However, the degree of correspondence between the systems may vary, depending on the level of specificity of responses collected on the ACS. For instance, some 4-digit Census industry codes correspond to a 2-digit NAICS industry sector, while others correspond to a 6-digit NAICS U.S. industry. Similarly, some 4-digit Census occupation codes correspond to a 3-digit SOC minor occupation group, while others correspond to a 6-digit SOC detailed occupation.

Standardized procedures and additional resources are maintained to aid in the assigning of industry and occupation codes. Both the autocoder and clerical coders are given access to additional responses, such as education level, age, and geographic location. Clerks may also use an alphabetical index of industries and occupations. If the name of the company or business for which a person works is available, clerical coders can look up the name on a reference listing of employers and their industries. The Census Bureau has used many versions of this reference list, often referred to as the Employer Name List (ENL). Some ENLs have been developed from public publications while others have used previously coded records. There are also dedicated clerks who code group quarters records and Spanish records. Finally, the coding referralists have access to even more resources, including access to state registries and use of the Internet for finding more information about the response.

Both the autocoder and clerical coders are subjected to regular quality checks. The autocoder quality control (QC) process begins with referralists coding a sample of autocoded records without seeing the autocoder's assigned code. The two codes are compared and when they disagree, a third code is assigned by a different referralist who sees both the autocoder's and first referralist's codes. This third code is then used in final comparisons as the "correct" code. If the autocoder agrees with the third code, it is considered to be correct, otherwise, the autocoder is considered to be in error. Analysts then review industry or occupation categories with high error rates, looking for patterns in word combinations that yield incorrect autocodes. These "wordbits" are then recoded by referralists, and analysts test these new codes in the data dictionaries and autocoding computer programs, updating them if appropriate.

The clerk quality assurance (QA) process begins with a sample of each coder's records being independently assigned by a different clerk who does not see the assigned code. Prior to 2012, this sample was a fixed percentage of each clerk's coded cases. Since June 2012, the sampling rate is dynamic, based on the number of records a clerk codes in a month. The sample must also meet a minimum sample size. After the samples are re-coded, the codes then are reconciled to determine which is correct. Coders are required to maintain a high monthly agreement rate and a minimum production rate to remain qualified to code. A coding supervisor oversees the QA 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).[22]

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.

[22] The following sections dealing with geocoding rely heavily on Boertlein (2007b).

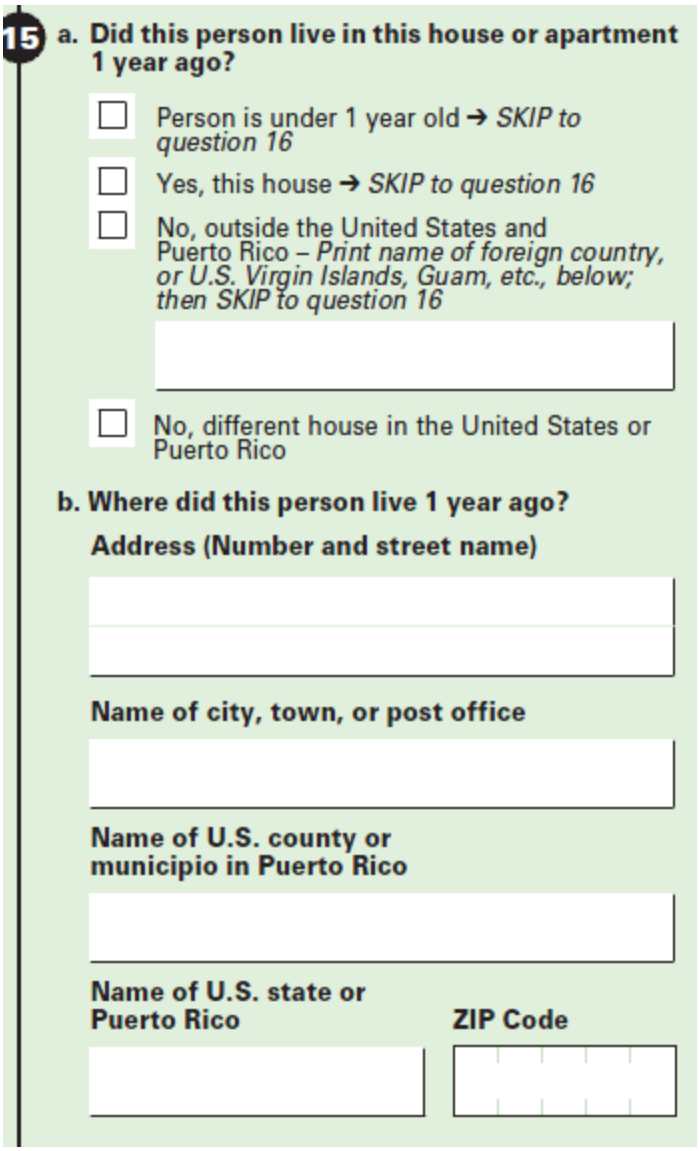


Figure 10-10: ACS Migration Questions

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.

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.

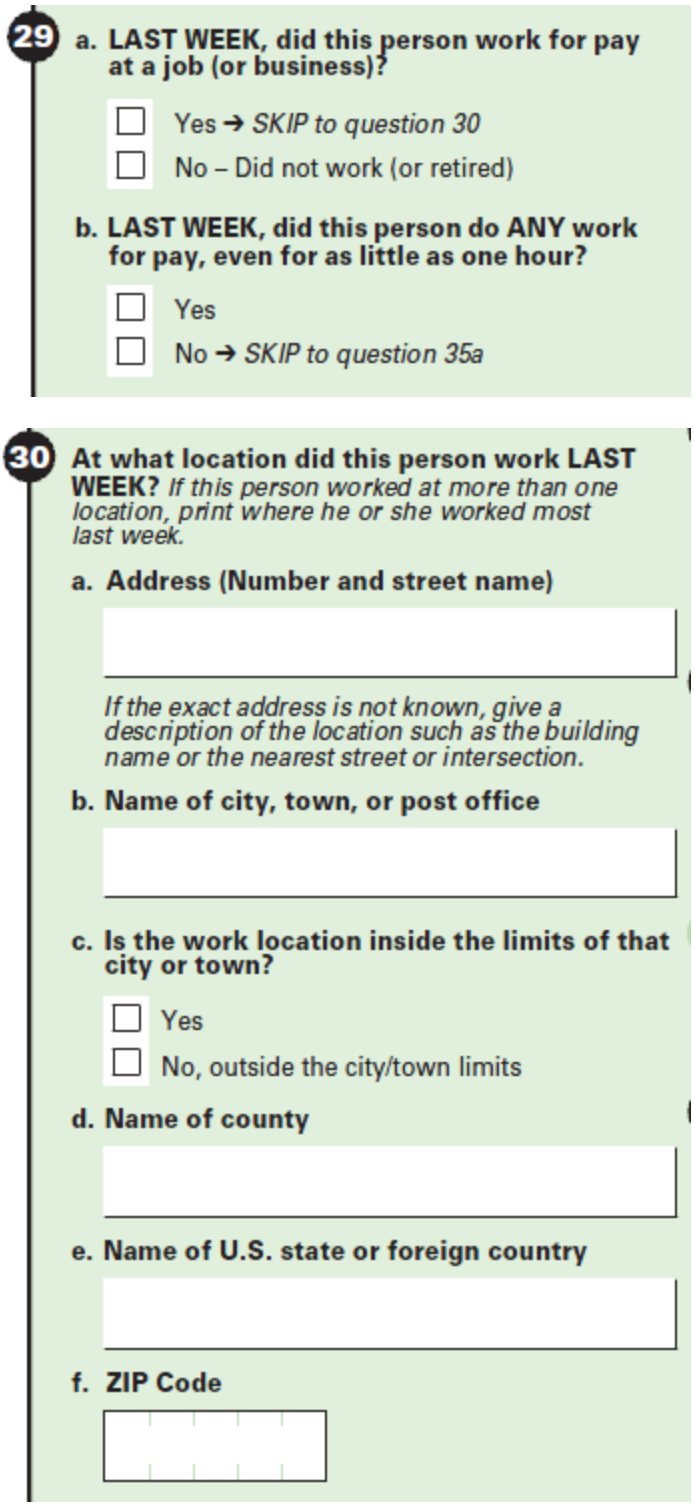


Figure 10-11: ACS Place-of-Work Questions

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.

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 in producing matches for given responses. Table 10-3 summarizes the properties of the geocoding workload 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 Percent
Migration	98 Percent
Place of Work	55 Percent

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 two out of three coders agree on a match, the third coder is assigned an error for the case. Coders must maintain an error rate of less than five percent 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 an uncodeable rate of 35 percent or less. 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 automatic geocoding.

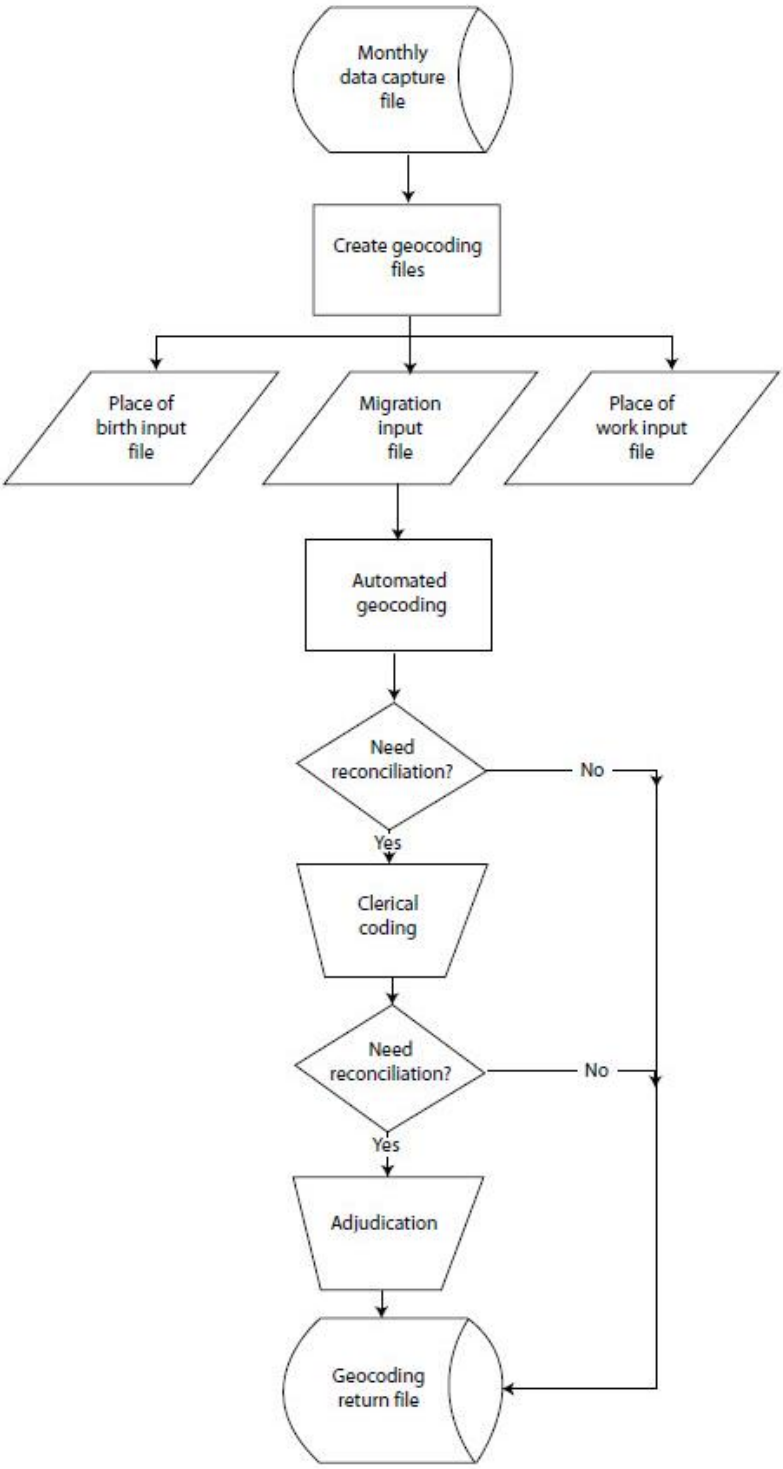


Figure 10-12: Geocoding

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 the 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 [2012].)

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 described in Chapter 11.

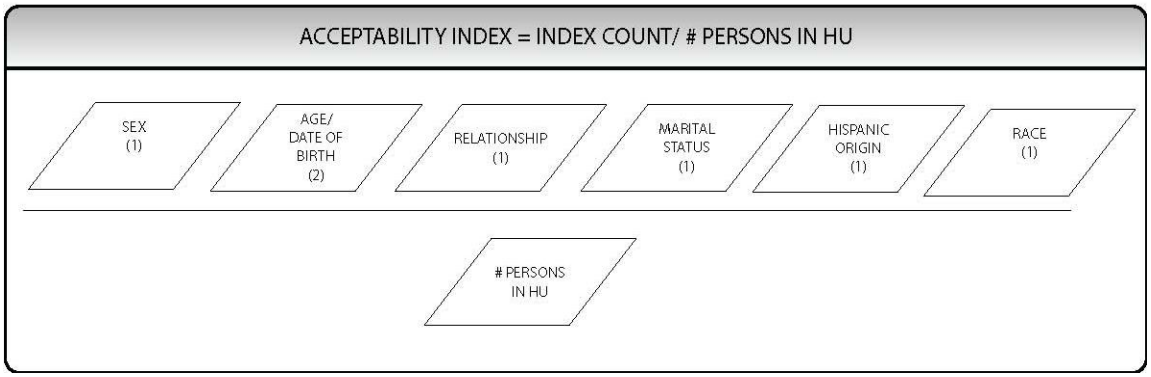


Figure 10-13: Acceptability Index

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

If the Acceptability Index is less 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/internet forms, one in response to the request to complete by internet, and a second in response to the replacement mailing. Depending on the timing, a household might return an internet response or mailed form, but also be interviewed in CATI or CAPI before the internet or 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/internet returns. For the more complete set of rules, see Powers (2012).

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 2013 were included in the 2012 ACS data products released in 2013 because the data collected were associated with the 2012 ACS data collection. Similarly, data collected in December 2013 as part of the 2014 ACS will be included with the 2014 ACS data products that are released in 2015 because the data collected are associated with the 2014 ACS data collection.

10.4 Creating the Select Files and Edit Input File

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,[23] returns that were too incomplete to meet the acceptability requirements. In addition, while the ‘‘current year’’ 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 DCF household data with the codes for computer and internet access. This file is then merged with 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, field of degree, place of work, health insurance, 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,

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

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 internet returns or 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 eleven 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 2013 are not directly comparable to those of December 2013 because the value of the dollar changed during this period. Consumer Price Index (CPI) indexes are used to adjust these income components for inflation. For example, a household interviewed in March 2013 reports their income for the preceding 12 months—March 2012 through February 2013. This reported income is adjusted to the reference year by multiplying it by the 2013 (January–December 2013) CPI and dividing by the average CPI for March 2012–2013.

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 a person's educational attainment, 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 geo-

graphic, 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. 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 those of previous years to check for notable differences. A review is conducted by variable, and results on unweighted data are compared across years to see if there are substantial differences. The initial review takes place with national data, and another final review compares data from smaller geographic areas, such as counties (Jiles, 2007). Analysts also examine unusual individual cases that were changed during editing to ensure accuracy.

These processes also are carried out after weighting and swapping data (discussed in Chapter 12).

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 Weighting

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 weighting 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 were 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) were used as the basis for producing the 3-year ACS tabulated estimates for the multiyear period. This discussion will focus on the 2011-2013 3-year and 2009-2013 5-year files and describe the steps to implement multiyear data processing.

A number of steps must be applied to the previous year's final edited data to make them consistent for multiyear processing. The first step is to update the current residence geography for 2011 and 2012 data to 2013 geography. The most complex 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 (2013). This step is required because the 2011 edited data for these variables and recodes are in 2011 vintage geography, and in 2012 vintage geography for the 2012 edited data. The geocodes in these variables and recodes from prior years need to be converted in some way to current geography. This transformation is accomplished using a matching process to multiyear geographic bridge files to update these variables to 2013 geography (Boertlein, 2008). Inflation adjustments also must be applied to monetary income and housing variables and recodes to inflate them up to a constant reference year of 2013 for the 2011–2013 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 2011–2013 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 2011 and 2012 data using the 2013 algorithm. When all of these modifications have been applied to the prior year's data, these data are combined with the 2013 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.

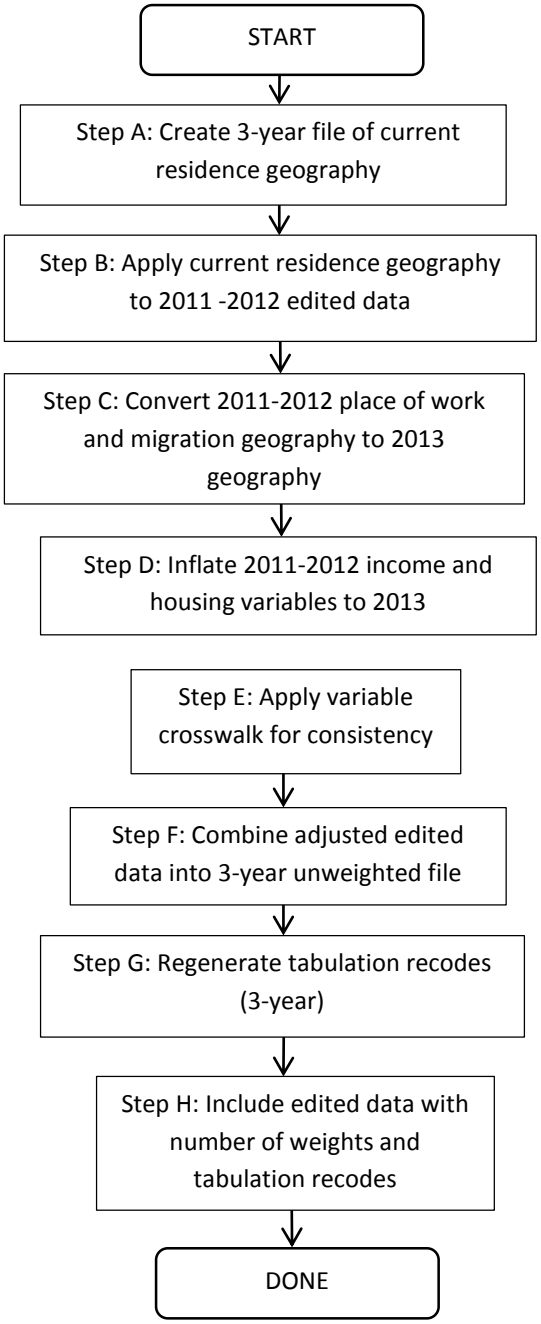


Figure 10-14: Multiyear Edited Data Process

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.

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.

Powers, J. (2012). U.S. Census Bureau Memorandum, ‘‘Specification for Creating the Edit Input and Select Files, 2012 (ACS12-W-6).’’ Washington, DC. 2012.

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.

Thompson, Matthew, Michael Kornbau, and Julie Vesely. (2012). "Creating an Automated Industry and Occupation Coding Process for the American Community Survey." In JSM Proceedings, Section on Statistical Learning and Data Mining. San Diego, CA: American Statistical Association.

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

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 14, "Data Dissemination"). The Census Bureau will publish up to three sets of estimates for a geographic area depending on its total population.

- For all statistical, legal, and administrative entities, including census tracts, block groups, and small incorporated places, such as cities and towns, the Census Bureau publishes 5-year estimates based on data collected during the 60 months of the five most recent calendar 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 three most recent calendar 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.

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. 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, Dahl, & Weidman, 1997).

In particular, the ACS uses ratio estimation to take advantage of independent population estimates by sex, age, race, and Hispanic origin, and estimates of total HUs produced by the Population Estimates Program (PEP) of the Census Bureau. This results in an increase in the precision of the estimates and corrects for under-/overcoverage by geography and demographic detail. This method also produces ACS estimates consistent with the population estimates by these characteristics and the estimates of total HUs for each county in the United States.

For any given geographic 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.

Sections 11.2–11.6 describe the single-year weighting and estimation methodology for calculating person weights for the GQ person records as implemented for the 2011 ACS forward. This weighting for GQ persons is done independently of the weighting for HUs. Sections 11.7–0 describe the single-year weighting methodology for calculating HU weights and person weights for the household sample records for the 2009 ACS forward. 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 11.11 outlines where the multiyear (3- and 5-year) weighting methodology differs from the single-year methodology.

11.2 ACS Group quarters person weighting

Since the 2006 data collection year, estimates from the ACS have included data from both people living in HUs and GQs. The weighting and estimation methodology for GQs significantly changed for the 2011 data year going forward. Readers who are interested in the methodology used prior to 2011 should reference the 12/2010 revision of this chapter posted on the ACS web site. The new methodology was designed to address a significant limitation of the current sample design and the previous weighting methodology. Due to constraints on both sample size and budget, the sample design was optimized at the state level rather than the small area level as is the case for the HU sample. In addition, the lack of independent GQ population estimates at the substate level led to the decision to optimize the weighting at the state level as well to support the GQ products that are released at that level. The trade-off, however, was increased substate variation in both the estimate of total GQ population and the characteristics of that population. As a result of this variation, there were many counties and tracts that did not have GQ representation even with the five-year estimates (Asiala, Beaghen, & Navarro, Using Imputation Methods to Improve the American Community Survey Estimates of the Group Quarters Population for Small Geographies, 2011). This variation was substantial enough to impact the estimates of the characteristics of the total resident population for the substate areas, including counties (Beaghen & Stern, 2009).

To address this limitation, a new GQ estimation methodology was developed and implemented with the 2011 data products. At its core is a mass imputation procedure whereby whole person records taken from the interviewed sample are copied (i.e., imputed) into not-in-sample GQs. By doing so, the GQ estimates better reflect the substate distribution of the GQs present on the sampling frame and reduce the variability in the substate estimates.

This estimation methodology has four basic components:

- Construct enhanced GQ imputation frame
- Select donors for whole person record imputation into select not-in-sample GQs
- Weighting
- Construct the post-imputation microdata

Each component is described in detail in the subsequent sections.

11.3 Construct Enhanced GQ Imputation Frame

The goal of the enhanced GQ imputation frame is to start with the sampling frame for the given year (see Chapter 3 for more details) and update that frame with all information regarding the frame that is collected during the year. Most updates that are available come from sample cases that were fielded after the creation of the sampling frame. These updates include: number of persons residing in the GQ, GQ type, and identification of nonexistent or out-of-scope GQ facilities.

If only the size of the sampled facilities were updated on the enhanced frame then the imputation into the not-in-sample facilities would not reflect the trends observed in the in-sample facilities. For example, if GQs that were in sample for a particular major type are tending to be larger than expected the same trend is expected to occur in the not-in-sample GQs of the same major type. For this reason, the expected populations of the not-in-sample GQs are adjusted using the observed relationship between the observed and expected population of the in-sample GQs. This adjustment is calculated within cells defined by major GQ type (see Table 11-1) by size class (less than 16, 16 to 399, 400 or greater).

Table 11-1: Major GQ Type

Major GQ type	Definition	Institutional/Noninstitutional
0	Correctional Institutions – Federal Prisons	Institutional
1	Correctional institutions - Other	Institutional
2	Juvenile Detention facilities	Institutional
3	Nursing homes	Institutional
4	Other Long-Term Care facilities	Institutional
5	College Dormitories	Noninstitutional
6	Military facilities	Noninstitutional
7	Other Noninstitutional facilities	Noninstitutional

To improve the imputation, a flag is set on the enhanced frame to identify single-sex facilities. A facility is designated as a single sex facility if either the federal Bureau of Prisons demographics file, both the most recent census and historical ACS sample interview data, or the most recent census for facilities with no historical ACS sample interview data reflect a sex distribution that is either at least 90% male or female. GQs identified as a single sex GQ will only have persons of

that sex imputed into that facility. All other GQs will not take sex into account when imputing records into the facility. For more information on creating the enhanced frame, see the detailed computer specifications (Castro, 2012b).

11.4 Select Donors for Imputation

The overarching goal of the imputation procedure is for the substate GQ estimates to better reflect the distribution present on the frame. To accomplish this, this goal is separated into two objectives:

- To establish representation of county by major type in the tabulations for each combination that exists on the frame for the 1-, 3-, and 5-year data.
- To establish representation of tract by major type in the tabulations for each combination that exists on the frame for the 5-year data.

To accomplish these two objectives, while providing some limits on the degree of imputation required, the imputation is targeted towards two groups:

- All not-in-sample GQs that have an expected population of greater than 15 persons will be selected to receive imputed whole person records.
- A subset of the not-in-sample GQs that have an expected population of 15 or fewer persons will likewise be selected as necessary in order to achieve the two objectives stated above.

The larger GQs are selected with certainty to ensure a base distribution of the GQ estimates in the broadest set of geographic areas. Since these GQs contain the largest proportion of the GQ population, targeting these GQs to receive imputed records will have the greatest visibility and impact on the estimates. The smaller GQs are selected only as needed to achieve the stated objectives. Thus, if there is a tract by major type combination that exists on the enhanced frame that is comprised of entirely small GQs, then one small GQ will be selected at random to represent the set of small GQs that exist for that combination.
Once the GQs are selected for imputation, the number of imputed person records to allocate to each GQ is determined. For the larger GQs, the number of imputed GQ person records is calculated as the larger of 2.5% of the expected population or one. For the smaller GQs, the number of imputed person records is the larger of 20% of the expected population or one.

Once the subset of not-in-sample GQs has been selected and the number of GQ imputed records to be assigned to the GQ has been computed, donors from the interviewed sample are selected. The selection process is implemented through an expanding search algorithm that first searches for a donor within county of the same specific GQ type. The specific types are a more detailed breakdown of the seven major types into more than 30 specific types. For example, the major type for correctional institutions is further classified into federal prisons, state prisons, jails, and half-way houses. If a donor is not found, the search expands to within county but of the same

major GQ type. If a donor is still not found, the geographic region is expanded and the process repeats until a donor is found. The levels of search are as follows:

- Within a geographic level, the search is first within the same specific type and then within the same major type
- Geographic levels expand as necessary in the following order: county, state, division, region, nation

In order to guard against the excessive reuse of donors, a particular donor is limited to being used three times within a single tract and five times within a single county. For more information on selecting donors, see the detailed computer specifications (Castro, 2012d).

11.5 GQ Weighting

The GQ weighting makes no distinction between the sampled and imputed GQ person records. The weighting has three basic steps: assigning an initial weight that reflects the observed combined sampled/imputed rate, an adjustment of those weights to match substate totals from the enhanced frame, and a coverage adjustment at the state level.

Base Weights

The base weights (*BW*) for GQ persons are defined so that the sum of the base weights is equal to the domain that they represent. That domain differs depending on whether the GQ is small or large. Large GQs are self-representing and thus the sum of the base weights for the persons in that GQ is equal to the actual or adjusted expected population of the GQ. The base weights for all persons in the GQ are defined to be equal and hence, for the *i*th person in the GQ, *BW* is defined as follows:

BW_i = Actual or adjusted expected population, N_P, of the GQ
÷
Total number of sampled or imputed GQ person records, n_p

$$= \frac{N_P}{n_p}$$

For the small GQs, the domain that the sum of the base weights is to represent is the total GQ population residing in small GQs for the tract by major type combination. Thus the definition of *BW* is adjusted to account for the potential random selection of the small GQ with sampled or imputed data from the set of all small GQs in the tract by major type combination:

BW_i = (Number of small GQs, N_{GQ}, on frame for the tract by major type combination
÷
Number of small GQs, n_{gq}, with either sampled or imputed GQ person records)
×
(Actual or adjusted expected population, N_P, of the GQ
÷
Total number of sampled or imputed GQ person records, n_p)

$$= \frac{N_{GQ}}{n_{gq}} \times \frac{N_P}{n_p}$$

Note that, as defined, the base weights also account for nonresponse within the GQ and within the tract (for small GQs).

Tract-level Constraint

The next steps are a series of constraints to ensure that the weighted totals of the sample and imputed records match the frame totals of adjusted population. One reason why the sum of the initial weights may not match the frame totals is the fact that the base weights of the small GQs reflect the equal probability selection of the small GQs within a tract (for the imputed GQs). While in expectation, the sum of the base weights may match the frame totals at the tract level, there may be a small deviance between the two because the first factor in the base weight calculation does not account for the population totals of the small GQs.

The tract-level constraint is thus defined as follows:

$TRCON_{tg}$ = Sum of adjusted GQ population, *ADJEXPOP*, for all GQs on the enhanced frame within the tract *t* and major type *g*
÷
Sum of base weights for all GQ person records sampled or imputed in tract *t* and major type *g*

$$= \frac{\sum_{\text{GQ } j \in \text{Tract } t, \text{ Major Type } g} ADJEXPOP_j}{\sum_{\text{Person } i \in \text{Tract } t, \text{ Major Type } g} BW_i}$$

The weight after the tract-level constraint, *WTRCON*, is achieved by multiplying the constraint factor by the base weight:

$$WTRCON_i = BW_i \times TRCON_{t(i)g(i)}$$

County-level Constraint

A second source of deviance between the weighted totals and the frame counts are ungeocoded GQs on the frame. These GQs do not have the census block codes required for tabulation but do have a county code assigned to them. For this reason, ungeocoded GQs are ineligible for imputation (they are still eligible for sampling, however, where they can be geocoded during data collection). To maintain consistency with the frame, the population total of all ungeocoded GQs on the frame are distributed to the geocoded GQs within county and major type via the county-level constraint. Note that in 2011, the issue of ungeocoded records is relatively small because of the robustness of the sampling frame that is based on the 2010 Census. In future years, new

updates to the frame that cannot be geocoded through automated means may make this constraint more important.

The county-level constraint is defined as follows:

$CTYCON_{cg}$ = Sum of adjusted GQ population for all GQs on the enhanced frame within the county c and major type g
÷
Sum of the weight after the tract-level constraint for all GQ person records sampled or imputed in county c and major type g

$$= \frac{\sum_{\text{GQ } j \in \text{County c, Major Type } g} ADJEXPOP_j}{\sum_{\text{Person } i \in \text{County c, Major Type } g} WTRCON_i}$$

The weight after the county-level constraint, *WCTYCON*, is achieved by multiplying the constraint factor by the weight after the tract-level constraint:

$$WCTYCON_i = WTRCON_i \times CTYCON_{c(i)g(i)}$$

State-level Constraint

The last constraint is designed to be a safety net in case there exists an ungeocoded GQ in a county where there are no geocoded GQs of the same major type. In that case, the population of that GQ is spread over all GQs of the same major type within the state. In practice, this is a relatively rare situation and the constraint is very close to one.

The state-level constraint is defined as follows:

$STCON_t$ = Sum of adjusted GQ population for all GQs on the enhanced frame within the state s and major type g
÷
Sum of weight after the county-level constraint for all GQ person records sampled or imputed in state s and major type g

$$= \frac{\sum_{\text{GQ } j \in \text{State s, Major Type } g} ADJEXPOP_j}{\sum_{\text{Person } i \in \text{State s, Major Type } g} WCTYCON_i}$$

The weight after the state-level constraint, *WSTCON*, is achieved by multiplying the constraint factor by the weight after the county-level constraint:

$$WSTCON_i = WCTYCON_i \times STCON_{s(i)g(i)}$$

GQ Post-stratification Adjustment to Controls

The final step in the GQ person weighting process is to apply the GQ Person Post-Stratification Factor (*GQPPSF*). The post-stratification cells are defined within state by GQ major type. This is consistent with the nature of the PEP GQ population estimates that are updated and maintained by major type. Using state as the level of geography for the post-stratification allows the GQ distribution on the frame to drive the substate distribution of the estimates.

All sample interviewed and imputed persons are placed in their appropriate cells. The *GQPPSF* for each cell is then calculated:

$GQPPSF_{sg}$ = PEP GQ population estimate for state s and major type g
÷
Sum of weight after the state-level constraint for GQ person records that are either interviewed sample or imputed in state s and major type g

$$= \frac{GQPOP_{sg}}{\sum_{\text{Person } j \in \text{State } s, \text{ Major Type } g} WSTCON_j}$$

where

$GQPOP_{sg}$ = PEP GQ population estimate for state s and major type g.

The weight after post-stratification, *WGQPPSF*, is achieved by multiplying the post-stratification factor by the weight after the GQ state constraint adjustment:

$$WGQPPSF_i = WSTCON_i \times GQPPSF_{s(i)g(i)}$$

These weights are then rounded to form the final GQ person weights. For more information on creating the GQ person weights, see the detailed computer specifications (Castro, 2012a).

11.6 Construct GQ Post-imputation Microdata

The final person-level microdata are assembled by concatenating the sample interview microdata with the imputed records. The microdata for each imputed record is created by joining the geographic information of the GQ selected for imputation with the edited response information from the donor. For geographically-tied characteristics, some adjustments are necessary in order to preserve certain data relationships. For example, if the donor listed the same county for their residence one year ago as their current county of residence, the microdata for the imputed record is adjusted so that the same relationship is true for the done record as was true for the donor record. Similar procedures are performed to preserve analogous relationships for place of work and journey to work. These steps help maintain the integrity of these characteristics for the imputed person records so that the estimates formed from the sampled and imputed records are not adversely impacted. For more information on creating the post-imputation microdata, see the detailed computer specifications (Castro, 2012c).

11.7 ACS Housing Unit Weighting—Overview

The 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. Note that this section reflects the methodology as implemented for the 2011 data prior to the introduction of the internet mode of data collection. It is expected that very little change will occur in the HU weighting with the addition of the internet mode and that all self-response modes, i.e., mail and internet, will be treated equally in the weighting.

11.8 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 Interview (CAPI) subsampling factor (*SSF*). The *BW* for an HU is calculated as the inverse of the final overall first-phase sampling rate which, for 2011, ranges from approximately 0.6 percent to 15 percent. HUs sent to CAPI are eligible to be subsampled (second-phase sampling) at rates generally ranging from 1-in-3 to 2-in-3 except for areas in remote Alaska and select American Indian areas which have a 100 percent CAPI sampling rate (see Chapter 4 for further details). 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 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 *BW* times the *SSF*. 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. Table 11-2 summarizes the computation of the *WSSF* by weighting step and the sample disposition of HUs. Additional information can be found in the detailed computer specifications for the HU weighting (Albright, 2012).

Table 11-2: Computation of the Weight after CAPI Subsampling Factor (*WSSF*)

Weighting step	Sample Disposition				
	Mail respondent	CATI respondent	CAPI sampled units	CAPI non-sampled units	CAPI eligible, but then becomes a 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 sub-sampling rate)	0	1
Weight after subsampling factor (*WSSF*)= *BW* × *SSF*	1 ÷ (overall sampling rate)	1 ÷ (overall sampling rate)	1 ÷ (overall sampling rate) × 1 ÷ (CAPI sub-sampling rate)	0	1 ÷ (overall sampling rate)

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 base weights of all HUs in that sample month
÷
Total weight after CAPI subsampling adjustment factor of all HUs interviewed in that sample month

$$= \frac{\sum_{j \in \text{Month}_i} BW_{ij}}{\sum_{j \in \text{Month}_i} WSSF_{ij}}$$

where

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

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

This adjustment factor is computed within each of the 2,005 ACS single-year 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 11-3 illustrates the computation of the *VMS* adjustment factor within a particular county. In this example, the total *BW* for each month is 100 (as shown on line 1 of this table). The total *WSSF* weight 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 in line 6).

Table 11-3: Example of Computation of *VMS*

	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 weight after the variation of monthly response adjustment (*WVMS*) is the product of the weight after CAPI subsampling factor (*WSSF*) and the variation of monthly response factor

(*VMS*). When the *VMS* factor is applied, the total weight 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.

11.9 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 multi-unit structure), and month of data collection (Weidman, Alexander, Diffendal, & Love, 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, then simultaneous adjustments for these three factors could be made. 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.[24] 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

[24] 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; and (3) addresses that identify commercial establishments, units being used permanently for storage, or living arrangements known as group quarters.

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 multi-unit 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.[25] 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 adjustment factor of interviewed occupied and temporarily occupied HUs and noninterviewed HUs
÷
Total HU weight after variation in monthly response adjustment factor of interviewed occupied and temporarily occupied HUs

$$= \frac{\sum_{j \in \text{Interviews}_i} WVMS_{ij} + \sum_{j \in \text{Noninterviews}_i} WVMS_{ij}}{\sum_{j \in \text{Interviews}_i} WVMS_{ij}}$$

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 below.

Table 11-4: Computation of the Weight after the first Noninterview Adjustment (*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

[25] 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.

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.[26] 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

$$= \frac{\sum_{j \in \text{Interviews}_i} WVMS_{ij} + \sum_{j \in \text{Noninterviews}_i} WVMS_{ij}}{\sum_{j \in \text{Interviews}_i} WNIF1_{ij}}$$

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 11-5.

Table 11-5: 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	$WNIF1_{ij}$
Noninterviewed HU	0

where

$WNIF2_{ij}$ = Adjusted HU weight after the second noninterview adjustment factor 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

[26] 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.

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*). *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 mode noninterview factor (*NIFM*) for a cell is:

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

$$= \frac{\sum_{j \in \text{CAPI Interviews}_i} WVMS_{ij} + \sum_{j \in \text{Noninterviews}_i} WVMS_{ij}}{\sum_{j \in \text{CAPI Interviews}_i} WVMS_{ij}}$$

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 11-6 below.

Table 11-6: Computation of the Weight After the Mode Noninterview Adjustment Factor (*WNIFM*)

Interview Status	$WNIFM_i$
Occupied or temporarily occupied HU interviewed via CAPI	$WVMS_{ij} \times NIFM_i$
Occupied or temporarily occupied HU interviewed via mail or CATI	$WVMS_{ij}$
Vacant or deleted HU	$WVMS_{ij}$
Noninterviewed HU	0

where

$WNIFM_i$ = Adjusted HU weight after the mode noninterview adjustment factor 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 (twelve categories), and marital status of the householder (three categories: married/widowed, single, or the unit is temporarily occupied and thus the marital status is unknown). 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 adjustment factor of interviewed occupied and temporarily occupied HUs
÷
Total weight after second noninterview adjustment factor of interviewed occupied and temporarily occupied HU

$$= \frac{\sum_{j \in \text{Resp}_i} WNIFM_{ij}}{\sum_{j \in \text{Resp}_i} WNIF2_{ij}}$$

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 11-7 below.

Table 11-7: Computation of the Weight after the Mode Bias Adjustment 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 adjustment factor for the *j*th HU within the *i*th adjustment cell.

11.10 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 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. They also produce annually updated estimates of total population for incorporated places and minor civil divisions (MCDs) 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, it can be seen 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 undercoverage of HUs and the people within them in household surveys. The bias that results from missing these HUs and people is partially corrected by using these controls (Alexander, Dahl, & Weidman, 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 HUs 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:

$$HU1X = HU10 + (NC1X + NM1X) - HL1X$$

where the suffix "X" indicates the year of the housing unit estimates, and

HU1X = Estimated 201X HUs

HU10 = Geographically updated 2010 Census HUs

NC1X = Estimated residential construction, April 1, 2010 to July 1, 201X

NM1X = Estimated new residential mobile home placements, April 1, 2010 to July 1, 201X

HL1X = Estimated residential housing loss, April 1, 2010 to July 1, 201X.

More detailed background on the current methodology used for the HU estimates can be found on the Census Bureau's website (U.S. Census Bureau, 2010a).

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:

$$P1 = P0 + B - D + NDM + NIM + NMM,$$

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. More detailed background on the current methodology used for the HU estimates can be found on the Census Bureau's website (U.S. Census Bureau, 2010b).

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.

Creation of the Subcounty Control Areas

The subcounty control areas are formed to give both MCDs and incorporated places the benefit of using subcounty controls. In order to achieve this balance, the basic units for forming the

subcounty areas are the county/MCD/place intersections or parts where the "balance of county" is also considered as another fundamental subcounty area. Note that outside of the strong and weak MCD states (U.S. Census Bureau, 2010c) for which the PEP produce total population estimates this defaults to simply the county/place parts. These subcounty areas are then combined until all subcounty areas within a county have a total population of 24,000 or greater. If it is not possible to partition a county into two or more subcounty areas of this size then the subcounty area is simply coexistent with the county.

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

HPF_i = PEP HU estimate
÷
Total HU weight after the mode bias factor of interviewed occupied, interviewed temporarily occupied and vacant HUs

$$= \frac{HU_i}{\sum_{j \in \text{Occupied and Vacant}_i} WMBF_{ij}}$$

where

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

Note that if the PEP HU subcounty estimates are summed across all subcounty areas within a county, the total is consistent with the PEP county-level HU estimates. 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 × 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 estimate of total population for the subcounty areas conform to the population estimates; (2) the combined estimates of spouses and unmarried partners conform to the combined estimate of married-couple and unmarried-partner

households and the estimate of householders conforms to the estimate of occupied HUs; and (3) the estimates for certain demographic groups are equal to their population estimates.
The population estimates used for the household person weighting are derived from the PEP estimates of total resident population by subtracting from the PEP total the corresponding ACS GQ estimate for that same population. For example, the control total used for county household population is derived by subtracting the ACS GQ estimate of total GQ population from the PEP estimate of total resident population. By doing so, the ACS estimate of total resident population (formed by summing the household and GQ population) conforms to the PEP estimate for the same population. This procedure is also used to derive the controls for subcounty areas and demographics as well.

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 one cell per subcounty control area defined within the weighting area. The second step uses four cells to classify persons by spousal relationship, householder and non-householder. The third step uses up to 156 cells defined by race/Hispanic origin, sex, and age. The steps are defined as follows:

Step 1: Subcounty Population Controls. All persons are assigned to one subcounty area within the weighting area. The marginal totals (i.e., the single-dimension control totals for a raking matrix) are simply equal to the derived household population control totals for the subcounty area as described above.

Step 2: Spouse / Unmarried Partner and Householders. All persons are placed into one of four 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, or
3. Persons who are a householder but do not fit into the first cell, or
4. Balance of population—all persons not fitting into the first three cells.

The marginals for the first two columns of cells are both equal to the estimate of married-couple plus unmarried-partner households using the *WHPF* weight. The marginal for the third column is the estimate of occupied HUs using the *WHPF* weight minus the marginal for the first column. In this manner, the estimate of households, equal to first column plus the third column of cells, is controlled to the estimate of occupied HUs. The marginal for the fourth column is equal to the derived household population estimate minus the sum of the marginals used for the other three columns of cells. In this manner, the estimate of total household population is controlled to the derived population estimates.

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 Alaskan Native (AIAN)
4. Non-Hispanic Asian
5. Non-Hispanic Native Hawaiian or Pacific Islander (NHPI)
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 area/race-ethnicity groups. These estimates are calculated based on the initial person weight of *WHPF*. Estimates from the Census Bureau's PEP also are summarized into estimates for each weighting area/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.

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 Alaskan Native, Asian, and Native Hawaiian or Pacific Islander.
5. If all non-Hispanic non-White groups have been collapsed together 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 11-8 and Table 11-9, for clarity.

Table 11-8: Steps 1 and 2 of the Weighting Matrix

		Step 2				Step 1 Control
		Householder in two-partner relationship	Spouse / unmarried partner in two-partner relationship	Householder not in two-partner relationship	Balance of population	
Step 1	Subcounty Area #1					Derived household population estimate
	…					
	Subcounty Area #n					
Step 2 Control		Survey estimate of married-couple and unmarried-partner households	Survey estimate of married-couple and unmarried-partner households	Survey estimate of all other single-headed households	Derived population estimate minus the sum of the other three controls	

Step 1. At this step, the initial person weights are adjusted to make the sum of the weights of all household persons equal to the derived household population controls for the defined subcounty control area.

Step 2. The Step 1 adjusted 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. In addition, the weights are adjusted so that the sum of the weights of householders not in a two partner relationship equal to the survey estimate of other single-headed households. For both of these constraints, the survey estimate is calculated using the HU weight after the HU post-stratification factor adjustment. Lastly, the weights of all other persons are adjusted to make the sum of all person weights equal to the derived household population estimates.

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 (*WHPF*)

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

Table 11-9: Steps 2 and 3 of the Weighting Matrix

			Step 2			Step 3 Control
			Householder in two-partner relationship	…	Balance of population	
Step 3	Non-Hispanic White	0–4 Males				Derived household population estimate
		0–4 Females				
		…				
		75+ Females				
	Non-Hispanic Black	…				
	Non-Hispanic AIAN	…				
	Non-Hispanic Asian	…				
	Non-Hispanic NHPI	…				
	Hispanic	…				
Step 2 Control			Survey estimate of married-couple and unmarried-partner households	…	Derived population estimate minus the sum of the other three controls	

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. Their *PPSF*s give an indication of undercoverage 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, households, and householders. With this weighting procedure, the survey estimate of total HUs will differ slightly from the PEP total housing unit estimates but is typically within a tenth of a percent at the county level.

11.11 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 one-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 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 one-year period. For example, in the 2007–2009 3-year weighting, all addresses that were originally assigned an interview month of January 2007, 2008 or 2009 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 non-interview 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 is 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 Population Estimates Program 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 housing unit, group quarters population, and total population for use as controls in the multiyear weighting. The derived county-level housing unit 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 group quarters 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 1 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.

The weighting areas used for the multiyear estimation are generally smaller than those used for the single-year estimation. They are still formed by complete counties or aggregations of counties and they must meet a threshold of 400 unweighted person interviews at the time of their formation. In addition, for the five-year estimation, the weighting area must have a minimum population of 2,500. For the three-year estimation, this generally results in most published counties being defined as their own weighting area as is the case for the one-year estimation. However, since there is no publication threshold for the five-year data product, there will be counties which are not their own weighting area and therefore greater differences between the

ACS and PEP estimates of total population may exist. For the formation of the subcounty control areas, the three-year threshold is 8,000 in total population and the five-year threshold is 2,500.

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 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 might improve our estimates of variance for subcounty estimates. 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 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, Using Administrative Records with Model-Assisted Estimation for the American Community Survey, 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 model-assisted estimation step is calculated at the same geographic areas as the subcounty controls for the ACS 3-year data and is calculated at the tract level for the ACS 5-year data. The entire model-assisted estimation process is summarized in these steps.

1. Create frame counts for geographic areas described above that contain at least 300 housing unit addresses.
2. Link the administrative records to the ACS sampling frame (the Master Address File or MAF) dropping administrative records that cannot be linked.
3. Form unweighted geographic totals of the linked administrative record characteristics.
4. Apply the *WMBF* weights at the housing-unit 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 3.
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 3 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 *HPF* adjustments, including the person weighting using the derived multiyear controls as described in the preceding section.

Frame Counts: The base weights (*BW*), 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 the frame counts reestablishes the original subcounty distribution of housing unit addresses on the frame in the weighted sample. For the 3-year weighting, these frame counts are calculated at the same county-place-MCD areas as the areas used for the subcounty controls. 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 tract-, 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. Collapse all categories into a single cell of total administrative persons.

If the condition still fails after the second level of collapsing, then the administrative record data is not used.

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: The final result of this step is the creation of the GREG Weighting Factor (*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 11-10.

Table 11-10: Impact of GREG Weighting Factor Adjustment

Interview Status	and the ACS record is:	Impact of *GWTF*
Non-Interview or	Not Applicable	No impact (factor set to 1)
CAPI Non-Sampled Interview (occupied or vacant) or Field determined ineligible housing unit	In an 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 11-11.

Table 11-11: Computation of the Weight After the GREG Weighting Factor (*WGWTF*)

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 multiyear 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.

11.12 References

Albright, K. (2012). Specifications for Weighting the 2011 1-year, 3-year, and 5-year American Community Survey Housing Unit Samples. DSSD 2011 American Community Survey Memorandum Series #ACS11-W-10. Washington DC: US Census Bureau.

Alexander, C., Dahl, S., & Weidman, L. (1997). Making Estimates from the American Community Survey. JSM Proceedings, Social Statistics Section (pp. 88-97). Alexandria, VA: American Statistical Association.

Asiala, M., Beaghen, M., & Navarro, A. (2011). Using Imputation Methods to Improve the American Community Survey Estimates of the Group Quarters Population for Small Geographies. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Beaghen, M., & Stern, S. (2009). Usability of the American Community Survey Estimates of the Group Quarters Population for Substate Geographies. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Society.

Castro, E. (2012a). Specifications for Calculating the Weights for the 2011 1-Year, 2009–2011 3-Year, and 2007–2011 5-Year American Community Survey GQ Sample. DSSD 2011 American Community Survey Memorandum Series #ACS11-W-9. Washington, DC: US Census Bureau.

Castro, E. (2012b). Specifications for Creating the 2007–2011 Group Quarters Imputation Frames. DSSD 2011 American Community Survey Memorandum Series #ACS11-W-16. Washington, DC: US Census Bureau.

Castro, E. (2012c). Specifications for Creating the Single-Year Post-Imputation Group Quarters Edited Microdata for 2007–2011. DSSD 2011 American Community Survey Memorandum Series #ACS11-W-18. Washington, DC: US Census Bureau.

Castro, E. (2012d). Specifications for Imputing Group Quarters Persons for 2007–2011. DSSD 2011 American Community Survey Memorandum Series #ACS11-W-17. Washington, DC: US Census Bureau.

Fay, R. (2006). Using Administrative Records with Model-Assisted Estimation for the American Community Survey. JSM Proceedings, Survey Research Methods Section (pp. 2995-3001). Alexandria, VA: American Statistical Association.

Starsinic, M. (2005). American Community Survey: Improving Reliability for Small Area Estimates. JSM Proceedings, Survey Research Methods Section (pp. 3592-3599). Alexandria, VA: American Statistical Association.

U.S. Census Bureau. (2010a). Methodology for State and County Total Housing Unit Estimates (Vintage 2009). Retrieved November 17, 2010, from U.S. Census Bureau: http://www.census.gov/popest/topics/methodology/2009-hu-meth.pdf

U.S. Census Bureau. (2010b). Methodology for the State and County Total Resident Population Estimates (Vintage 2009). Retrieved November 17, 2010, from U.S. Census Bureau: http://www.census.gov/popest/topics/methodology/2009-st-co-meth.pdf

U.S. Census Bureau. (2010c). Population Estimates: Geographic Terms and Definitions. Retrieved November 17, 2010, from U.S. Census Bureau: http://www.census.gov/popest/geographic/estimates_geography.html

Weidman, L., Alexander, C., Diffendal, G., & Love, S. (1995). Estimation Issues for the Continuous Measurement Survey. JSM Proceedings, Survey Research Methods Section (pp. 596-601). Alexandria, VA: American Statistical Association.

Chapter 12: Variance Estimation

12.1 Overview

Sampling error is the uncertainty associated with an estimate that is based on data gathered from a sample of the population rather than the full population. 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 Housing Unit and Person Estimates

Unbiased estimates of variances for ACS estimates 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 Successive Differences Replication (SDR) method (Wolter, 1984; Fay & Train, 1995; Judkins, 1990) to produce variance estimates. It has been the method used to calculate ACS estimates of variances since the start of the survey. 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 (U.S. Census Bureau, 2006) and Census 2000 Long Form estimates (Gbur & Fairchild, 2002).

In the SDR method, the first step in creating a variance estimate is constructing the replicate factors. Replicate base weights are then calculated by multiplying the base weight for each housing unit (HU) by the factors. The weighting process then is rerun, using each set of replicate base weights in turn, to create final 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.

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 to assign rows 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,…,n, where n is sample size) and replicate r (r = 1,…,80), the replicate factor is computed as:

$$f_{i,r} = 1 + 2^{-1.5}a_{R1i,r} - 2^{-1.5}a_{R2i,r}$$

where $R1i$ and $R2i$ are respectively the first and second row of the Hadamard matrix assigned to the i-th HU, and $a_{R1i,r}$ and $a_{R2i,r}$ are respectively the matrix elements (either 1 or −1) from the Hadamard matrix in rows $R1i$ and $R2i$ 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 $a_{R1i,r} = +1$ and $a_{R2i,r} = +1$, the replicate factor is 1.
- If $a_{R1i,r} = -1$ and $a_{R2i,r} = -1$, the replicate factor is 1.
- If $a_{R1i,r} = +1$ and $a_{R2i,r} = -1$, the replicate factor is approximately 1.7.
- If $a_{R1i,r} = -1$ and $a_{R2i,r} = +1$, 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		Hadamard matrix element								Approximate replicate			
	$R1_i$	$R2_i$	Replicate 1		Replicate 2		Replicate 3		Replicate 4		$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$
			$a_{R1i,1}$	$a_{R2i,1}$	$a_{R1i,2}$	$a_{R2i,2}$	$a_{R1i,3}$	$a_{R2i,3}$	$a_{R1i,4}$	$a_{R2i,4}$				
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 replicate base weight (RBW) for replicate r is computed by multiplying the full sample base weight (BW— see Chapter 11 for the computation of this weight) by the replicate factor $f_{i,r}$; that is, $RBW_{i,r} = BW_i \times f_{i,r}$, where $RBW_{i,r}$ is the replicate base weight for the i-th HU and the r-th replicate (r = 1, …, 80).

One can elaborate on the previous example of the replicate construction using five cases and four replicates: Suppose the full sample BW values are given under the second column of the following table (Table 12-2). Then, the replicate base weight values are given in columns 7−10.

Table 12-2: Example of Computation of Replicate Base Weight Factor (RBW)

Case #	BW_i	Approximate Replicate Factor				Replicate Base Weight			
		$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	$RBW_{i,1}$	$RBW_{i,2}$	$RBW_{i,3}$	$RBW_{i,4}$
1	100	0.3	1	1	1.7	29	100	100	171
2	120	1.7	1	0.3	1	205	120	35	120
3	80	1	1	1	0.3	80	80	80	23
4	120	0.3	1	1	1.7	35	120	120	205
5	110	1.7	1	0.3	1	188	110	32	110

The rest of the weighting process (Chapter 11) then is applied to each replicate weight $RBW_{i,r}$ (starting from the adjustment for CAPI subsampling) and proceeding to the population control adjustment or raking). Basically, the weighting adjustment process is repeated independently 80 times and the $RBW_{i,r}$ is used in place of BW_i (as in Chapter 11).

By the end of this process, 80 final replicate weights for each HU and person record are produced.

Beginning with the ACS 2011 data products, a new imputation-based methodology was incorporated into processing (see the description in Chapter 11). An adjustment was added to the production replicate weight variance methodology to account for the non-negligible amount of additional variation being introduced by the new technique. For more information regarding this issue, see Asiala and Castro (2012).

Variance Estimates

Given the replicate weights, the computation of variance for any ACS estimate is straightforward. Suppose that $\hat{\theta}$ 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 $\hat{\theta}_1$, $\hat{\theta}_2$, …, $\hat{\theta}_{80}$ denote the estimates computed based on the replicate weights. The variance of $\hat{\theta}_0$, $v(\hat{\theta}_0)$, is estimated as the sum of squared differences between each replicate estimate $\hat{\theta}_r$ (r = 1, …, 80) and the full sample estimate $\hat{\theta}_0$. The formula is as follows[27]:

$$v(\hat{\theta}_0) = \frac{4}{80}\sum_{r=1}^{80}(\hat{\theta}_r - \hat{\theta}_0)^2$$

[27] 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 (Fay & 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 identification 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.

The model-based method requires census counts and ACS data. For ACS data collected 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. 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:

$$\text{Upper Confidence Bound} = 0 + 1.645 \times SE(0) = 1.645 \times \sqrt{K \times (\text{average weight})} = \text{census count}$$

Solving for K yields:

$$K = \left(\frac{\text{census count}}{1.645}\right)^2 \frac{1}{(\text{average weight})}$$

The first *K* was computed for all ACS zero-estimated counts from 2000 which were matched to Census 2000 100 percent counts, and then the 90th percentile of those *K*s was determined. Based

on the Census 2000 data, we used a value for *K* of 400 (Navarro, 2001b) from 2001 through the 2010 data products.

When the results of the 2010 Census became available in 2011, work was done to update the *K* value. The new value for K was determined to be 223 and was used for the first time with the 2011 data products. In 2012, additional research was conducted. As a result, for the 1-year and 3-year data products, the k value continues to be 223, but for the 5-year data products there is a change. One of six possible values (4, 8, 10, 14, 18, or 22) is assigned based on the population size of the geographic area.

For publication, the standard error (*SE*) of the zero count estimate is computed as:

$$SE(0) = \sqrt{K \times (average\ weight)}$$

Where:

K = 400 for ACS data products from 2001 to 2010
223 for ACS data products for 2011
4, 8, 10, 14, 18, 22 or 223 for ACS data products starting with 2012

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.

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 – Finite Population Correction Factor

Through the 2008 and 2006-2008 data products, the same variance estimation methodology described above was implemented for both 1-year and 3-year. No changes to the methodology were necessary due to using multiple years of sample data. However, beginning with the 2007-2009 and 2005-2009 data products, the ACS incorporated a finite population correction (FPC) factor into the 3-year and 5-year variance estimation procedures.

The Census 2000 long form, as noted above, used the same SDR variance estimation methodology as the ACS currently does. The long form methodology also included an FPC factor in its calculation. One-year ACS samples are not large enough for an FPC to have much impact on variances. However, with 5-year ACS estimates, up to 50 percent of housing units in certain blocks may have been in sample over the 5-year period. Applying an FPC factor to multi-year ACS replicate estimates will enable a more accurate estimate of the variance, particularly for small areas. It was decided to apply the FPC adjustment to 3-year and 5-year ACS products, but not to 1-year products.

The ACS FPC factor is applied in the creation of the replicate factors:

$$f_{i,r}^{*} = 1 + \left(2^{-1.5}a_{R1i,r} - 2^{-1.5}a_{R2i,r}\right) \times \sqrt{1 - n/N}$$

where $\sqrt{1 - n/N}$ is the FPC factor. Generically, n is the unweighted sample size, and N is the unweighted universe size. The ACS uses two separate FPC factors: one for HUs responding by mail or telephone, and a second for HUs responding via personal visit follow-up.

The FPC is typically applied as a multiplicative factor "outside" the variance formula. However, under certain simplifying assumptions, the variance using the replicate factors after applying the FPC factor is equal to the original variance multiplied by the FPC factor. This method allows a direct application of the FPC to each housing unit's or person's set of replicate weights, and a seamless incorporation into the ACS's current variance production methodology, rather than having to keep track of multiplicative factors when tabulating across areas of different sampling rates.

The adjusted replicate factors are used to create replicate base weights, and ultimately final replicate weights. It is expected that the improvement in the variance estimate will carry through the weighting, and will be seen when the final weights are used.

The ACS FPC factor could be applied at any geographic level. Since the ACS sampling rates are determined at the small area level (mainly census tracts and governmental units), a low level of geography was desirable. At higher levels, the high sampling rates in specific blocks would likely be masked by the lower rates in surrounding blocks. For that reason, the factors are applied at the census tract level.

Group quarters persons do not have an FPC factor applied to their replicate factors.

12.3 Margin of Error and Confidence Interval

Once the standard errors have been computed, margins of error and/or 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, 2010). 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 $\hat{\theta}$ can be computed as:

$$\text{Margin of error}(\hat{\theta}) = 1.645 \times SE(\hat{\theta})$$

where SE($\hat{\theta}$) is the standard error of the estimate $\hat{\theta}$. Given this margin of error, the 90 percent confidence interval can be computed as:

$$\hat{\theta} \pm \text{Margin of error}(\hat{\theta})$$

That is, the lower bound of the confidence interval is [$\hat{\theta}$ − margin of error ($\hat{\theta}$)], and the upper bound of the confidence interval is [$\hat{\theta}$ + margin of error ($\hat{\theta}$)]. Roughly speaking, this interval is a range that will contain the ‘‘full population 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:

$$\text{Margin of error} = max(\text{upper bound} - \text{estimate}, \text{estimate} - \text{lower bound})$$

Using the margin of error (as published or calculated from the bounds), the standard error is obtained as follows:

$$\text{Standard error}(\hat{\theta}) = \text{Margin of error}(\hat{\theta}) \div 1.645$$

For ranking tables and comparison profiles, the ACS provides an indicator as to whether two estimates, Est_1 and Est_2, 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. Because the ACS PUMS contains only a subset of the full ACS sample, estimates from the ACS PUMS file will often 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.

PUMS Replicate Variances

For the replicate method, direct variance estimates based on the SDR formula as described in Section 12.2 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 then apply the variance formula:

$$v(\hat{\theta}_0) = \frac{4}{80}\sum_{r=1}^{80}(\hat{\theta}_r - \hat{\theta}_0)^2$$

PUMS Design Factor Variances

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. For example, the 2009 ACS PUMS design factors are published at national and state levels (U.S. Census Bureau, 2010a), and were calculated using 2009 ACS data. PUMS design factors are updated periodically, but not necessarily 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 1-year ACS PUMS files beginning with 2005, use:

$$SE(\hat{Y}) \doteq DF \times \sqrt{99 \times \hat{Y} \times \left(1 - \frac{\hat{Y}}{N}\right)}$$

for a total, and

$$SE(\hat{p}) \doteq DF \times \sqrt{\frac{99}{B} \times \hat{p} \times (100 - \hat{p})}$$

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 denominator (base) of the percent.

The value 99 in the formula is the value of the 1-year PUMS FPC factor, 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 3-year PUMS FPC factor 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.

Similarly, 5-year PUMS files, beginning with 2005-2009, represent approximately a 5 percent sample of HUs. So, the 5-year PUMS FPC is 95 / 5 = 19, which can be substituted 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 12.2) 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* × *SRSSE*, where the *SRSSE* is computed as follows:

$$SRSSE(\hat{Y}) \doteq DF \times \sqrt{39 \times \hat{Y} \times \left(1 - \frac{\hat{Y}}{N}\right)}$$

The value 39 in the formula above is the FPC 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 the 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 DFs by topic can be obtained from the "PUMS Accuracy of the Data" statement that is published with each PUMS file. For example, 2009 1-year PUMS DFs can be found in the PUMS Accuracy of the Data (2009) (U.S. Census Bureau (2010b)). The documentation also provides examples on how to use the design factors to compute standard errors for the estimates of totals, means, medians, proportions or percentages, ratios, sums, and differences.

The topics for the 2009 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, after removing the race or Hispanic origin DFs from consideration.

12.5 References

Asiala, M. & Castro, E. (2012). Developing Replicate Weight-Based Methods to Account for Imputation Variance in a Mass Imputation Application. Joint Statistical Meetings: Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association.

Fay, R., & Train, G. (1995). Aspects of Survey and Model Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties. Joint Statistical Meetings: Proceedings of the Section on Government Statistics (pp. 154-159). Alexandria, VA: American Statistical Association.

Gbur, P., & Fairchild, L. (2002). Overview of the U.S. Census 2000 Long Form Direct Variance Estimation. Joint Statistical Meetings: Proceedings of the Section on Survey Research Methods (pp. 1139-1144). Alexandria, VA: American Statistical Association.

Gunlicks, C. (1996). 1990 Replicate Variance System (VAR90-20). Washington, DC: U.S. Census Bureau.

Judkins, D. R. (1990). Fay's Method for Variance Estimation. Journal of Official Statistics, 6(3), 223-239.

Navarro, A. (2001a). 2000 American Community Survey Comparison County Replicate Factors. American Community Survey Variance Memorandum Series #ACS-V-01. Washington, DC: U.S. Census Bureau.

Navarro, A. (2001b). Estimating Standard Errors of Zero Estimates. Washington, DC: U.S. Census Bureau.

U.S. Census Bureau. (2006). Current Population Survey: Technical Paper 66—Design and Methodology. Retrieved from U.S. Census Bureau: http://www.census.gov/prod/2006pubs/tp-66.pdf

U.S. Census Bureau. (2010). Statistical Quality Standard E2: Reporting Results. Washington, DC: U.S. Census Bureau.

Wolter, K. M. (1984). An Investigation of Some Estimators of Variance for Systematic Sampling. Journal of the American Statistical Association, 79, 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 demographic, social, economic and housing characteristics. These products cover geographic areas within the United States and Puerto Rico. The Public Use Microdata Sample (PUMS) files, which enable data users to create their own estimates, are also data products.

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 nationwide survey that collects data continuously every year, products can be released for many types of geographic areas, including many smaller geographic areas such as places, townships, and census tracts.

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 impose restrictions on the publication of estimates for certain geographic areas based on the size of the ACS sample or the population size of such areas. 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 are 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. PUMS estimates are reviewed in a separate process.

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. This step made the ACS sample 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, for the first time, released products based on three years of ACS sample data collected from 2005 through 2007. In 2010, the Census Bureau released the first products based on five years of consecutive ACS samples, 2005 through 2009. Since several years of sample 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, the Census Bureau releases other data products developed through a fee-based special tabulations program. Government agencies, private organizations and businesses, or individuals can request special tabulations of data. As is the case for regular data products, special tabulation requests are reviewed before release by the Census Bureau's Disclosure Review Board (DRB) to assure protection of the confidentiality of individual responses. Section 13.6 provides information on methods used by the DRB to protect census data, including data from the ACS.

Chapter 14 describes the dissemination of the data products discussed in this chapter, the ACS data release schedule, and locations of various products and supporting documents on the Census Bureau's website.

13.2 Geography

The Census Bureau strives to provide ACS estimates for the geographic areas that are most important to data users. For example, data products reflecting ACS estimates are disseminated for many of the nation's legal and administrative areas, including states, American Indian and Alaska Native (AIAN) areas, counties, minor civil divisions (MCDs), incorporated places, congressional districts, as well as a variety of other geographic entities. Some geographic areas for which ACS estimates are produced, such as block groups, census tracts and census designated places, are defined and delineated in cooperation with state and local agencies. Others, such as urban areas, are defined and delineated by the Census Bureau, based on criteria developed by the Census Bureau and reviewed by federal, state, local, and tribal government data users.

Information on the types of geographic areas for which the Census Bureau publishes ACS estimates is available at:
https://www.census.gov/acs/www/data_documentation/areas_published/.

The publication of 1-, 3-, and 5-year ACS estimates is subject to limitations imposed by confidentiality restrictions. One-and 3-year estimates are also subject to data quality filtering, a process designed to limit the publication of low-reliability estimates. 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 five percent below the specific period threshold. The topics of confidentiality restrictions and data quality filtering are covered in more detail in Section 3.6.

The Puerto Rico Community Survey (PRCS) also provides estimates for legal, administrative, and statistical areas in Puerto Rico. The data release and threshold rules described above for the publication of 1-year, 3-year, and 5-year ACS period estimates for the U.S resident population apply for the publication of data for the PRCS.

Many areas for which ACS estimates are produced undergo periodic boundary changes due to annexations, detachments, or mergers with other areas. Each year, the Census Bureau's Geography Division, working with state and local governments, conducts the Boundary and Annexation Survey (BAS) to collect updated boundary information.[28] 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 reflect these legal area boundary changes, so all estimates are based on Geography Division resources that depict the geographic boundaries for legal areas as they existed on January 1 of the sample year. In the case of multiyear estimates, the boundaries of the entity for which the estimate is produced are those effective on January 1 of the final year of data collection for that multiyear estimate. For example, the boundaries of an entity for which a 2011-2013 estimate is produced are those in effect on January 1, 2013.

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 suggested certain changes to many tables, the Census Bureau implemented these preferences for the ACS products.

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 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 data users. Data users provided a wide variety of comments, leading to modifications of planned products.

ACS managers determined the format of the new products for use in the 2005 ACS data release of data collected from the 2004 ACS. This made it possible for data users to become familiar with the new products and to provide comments well in advance of the release of data for the 2005 ACS.

[28]The BAS is described in more detail in the Glossary.

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. Again in 2009, a *Federal Register* Notice was issued to gather user feedback on the first ACS 5-year products.

Since 2010, an ACS Data Products Planning Working Group (DPPWG) and ACS Portfolio Management Governance Board (PMGB) have had a critical role in establishing more efficient processes for considering, reviewing, and approving proposals for new and modified data products. These groups have helped to allocate resources appropriately to develop new products, to consider innovative product design approaches suggested by both internal and external groups, and to revise existing products to make them more useful. They also help to ensure that technical constraints are fully assessed against technical requirements, and that potential product innovations align with future goals.

The ACS Data Users Group (ACS DUG) was formed in early 2013 to provide regional planners, demographers, community leaders and other data users with an externally managed forum for exchanging information on understanding and using ACS data. More information is available on ACS DUG activities at: http://www.acsdatausers.org/

13.4 Description of Aggregated Data Products

ACS data products can be divided into two broad categories: aggregated data products, and products representing extracts of the Public Use Microdata Sample (PUMS), described in Section 13.5.

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 2012 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 2012 in Chicago with residents under 18 years old.

This section provides a description of each aggregated product. Each product described is available as 1-year period estimates; unless otherwise indicated, they are also available as 3-year and 5-year estimates. The data products described below contain all estimates planned for release each year, including those from multiple years of data. Data release rules described in Section 3.6 will prevent certain 1-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 states, as well as for counties, towns, and other small geographic entities such as census tracts and block groups, the latter of which are only available as five year estimates.

The Census Bureau's initial goal in defining ACS data products was 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 were included in the ACS base tables. For example, the ACS 2012 data products include more than 1,470 detailed tables that cover a wide variety of characteristics and over 400 race and Hispanic origin iterations.

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, the Census Bureau published a comparison profile that compares the sample year's estimates with those of each of the four previous years. For example, the 2012 comparison profile compared 2012 estimates with those of the 2011 ACS, 2010 ACS, 2009 ACS and 2008 ACS. 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

Sourced from the data profiles, the narrative profiles allow users to explore a narrative and graphical presentation of the statistics for their own communities. This descriptive report describes a geographic area using custom text and graphics across fifteen topics. This product is available using the 5-year data and can be generated for the Nation, States (including Puerto Rico), Counties, Places, Metro/Micro Areas, Zip Code Tabulation Areas (ZCTAs), Tracts, and American Indian/Alaska Native (AIAN) Areas. The 5-year narrative profiles were released beginning in 2012 and are available at:
http://www.census.gov/acs/www/data_documentation/2012_narrative_profiles.

Subject Tables

Subject 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. There are about 70 topic-specific subject tables 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 tables and tabular displays that allow for testing statistical significance.

The rankings show approximately 90 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

Geographic Comparison Tables (GCTs) contain the same measures that appear in the ranking products, plus additional 100 demographic measures that cannot be produced as ranking tables. 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.

Selected Population Profiles

Selected Population Profiles (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 1- and 3-year estimates. In 2008, the requirement that SPPs could only be produced for sub-state areas of at least 1,000,000 persons was modified to allow for SPPs with minimum population sizes of 500,000. Another change to SPPs in 2008 was the expansion in the number of groups for which SPPs are produced to include additional country-of-birth groups. Groups too small to warrant an SPP for

a geographic area based on one year of sample data may appear in an SPP based on the 3-year accumulations of sample data.

Selected Population Tables and American Indians and Alaska Native Tables

The ACS Selected Population Tables and American Indians and Alaska Native (AIAN) Tables include a set of data products comparable to the Census 2000 Summary File 4/AIAN products. Representing a collection of detailed tables, data profiles, and subject tables, this product provides data iterated by detailed race, ethnicity, and tribal group. This product was first released in May 2012 and included estimates for the period 2006 to 2010. The Census Bureau plans to produce the Selected Population Tables and American Indians and Alaska Native Tables every five years.

Report Production Series and Analytical Applications

Subject matter analysts at the Census Bureau create analytical data products based on recent ACS data. The analytical products include reports, infographics, and online mapping applications. These products provide data users with insightful analysis on the current state of the nation. Some subject matter areas produce special tabulations of the data for their topics. The Census Bureau also produces methodological analysis products based on ACS collection and processing research. Information on these various analytic products can be found at the American Community Survey website at: http://www.census.gov/acs/www/library/by_year/2013/.

13.5 Public Use Microdata Sample (PUMS)

The PUMS comprises individual records that contain information collected about each person and housing unit (HU). PUMS files are extracts from the microdata for which disclosure avoidance technique are applied to protect confidential information about households or individuals. These extracts cover all of the same characteristics contained in the full microdata sample files. The only geography other than state shown on a PUMS file is the Public Use Microdata Area (PUMA). PUMAs are special non-overlapping areas that partition a state; each PUMA contains a population of about 100,000 or more. State governments drew the PUMA boundaries at the time of the Census. PDF-format maps of PUMA boundaries are available from the Census Bureau's website at: http://www.census.gov/geo/maps-data/maps/reference.html.

The Census Bureau releases 1-year, 3-year, and 5-year PUMS files. 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

The subject matter analysts in the Census Bureau's Social, Economic and Housing Statistics Division and Population Division provide the specifications for 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. Since 2006, limited changes to these specifications have occurred. The ACS Data Products Planning Workgroup, which serves as a Change Control Board, must approve specification changes. Once the specifications are validated and verified, tabulation starts when the weighted and edited microdata become available (see Chapters 10 and 11). ACS data products are generated by various automated tabulation systems, beginning with the 1-year period estimates for the detailed tables data 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. This practice differs from that followed for decennial census long form products released for Census 2000 and earlier decennial censuses, for which only the long form estimates appeared in data tables. Users of long form data for these decennial censuses were provided with a description and explanation of the concept of statistical uncertainty, and guidance for calculating the standard errors of individual estimates. However, the specific standard errors applying to each estimate were not calculated for data users and did not appear alongside the estimates in the data tables that included long form estimates.

Subject matter analysts use the standard errors in their internal reviews of estimates. Data users can also use them to determine whether two ACS estimates are statistically different.

13.7 Data Release Rules

Various kinds of restrictions are applied to ACS data to limit the disclosure of information about individual respondents and to limit the production of ACS estimates with unacceptable statistical reliability.

Population Thresholds for ACS Data

Population thresholds restrict the availability of data according to the type of ACS estimate. Areas or groups of 65,000 or more are eligible for 1-, 3-, and 5-year estimates. Areas or groups of 20,000 or more are eligible for 3- and 5-year estimates. Areas or groups of 20,000 or fewer are eligible for 5-year estimates.

Other population threshold rules apply in cases where a drop in population size changes the qualification of an area for types of ACS estimates for which it was previously qualified. Consider, for example, the case of a geographic area that met the 1-year threshold for a previous period, but dropped below it for the current period. The Census Bureau would like to provide the same kind of estimates for such areas every year to preserve the continuity of data for such areas. To give such an area a reasonable chance to retain access to 1-year estimates, the Census Bureau applies a data release rule that provides for the continued production of 1-year estimates unless

the population drops more than five percent below the specific period threshold for 1 year estimates, 65,000. A comparable rule applies to areas that normally qualify for 3-year estimates.

Data Quality Filtering

Another kind of data release rule, data quality filtering, applies to ACS 1-year and 3-year estimates. 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 to meet the rule's requirements and is restricted from publication. 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 3- year estimates, but are not applied to the 5- year estimates.

Disclosure Review Board Rules

Beyond the population thresholds applied to ACS estimates and data quality filtering applied to ACS tables, the Disclosure Review Board (DRB) establishes additional rules that specify what ACS data are released. These rules describe, for example, how medians or other quantiles must be calculated, what requirements apply to means or aggregates, the thresholds for tables involving specific kinds of geographic areas or tables with more than 100 detail cells, and thresholds for tables of unweighted counts of people and housing units.

The DRB uses data swapping as a disclosure limitation technique for PUMS files. Data swapping involves the swap, or interchange, of a small percentage of household records between pairs of households in different geographic regions. 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, and the household pairs in most cases are located in the same census tract. 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.

13.8 Data Review and Acceptance

After the editing, imputation, data products generation, disclosure avoidance, and application of the release rules are complete, 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 subject matter analysts review the ACS estimates following a multistep review process and provide clearance before estimates are released to the public. Because of the short time available to review such a large amount of data, an automated review tool (ART) and various multiyear review tools have been developed to facilitate the process. These tools enable 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 and included functions allowing for ad hoc, user-defined queries for hundreds of single-year estimates covering 350 geographic areas. The improved version has been used by the analysts since June 2005. It was designed to work on much larger data sets and has a wider range of capabilities, with faster response time to user commands. In 2008, a team of programmers, analysts, and statisticians developed an automated multiyear review tool to assist analysts in their review of the multiyear estimates and this tool has been used for the review of 3-year and 5-year estimates.

The ACSO staff, together with the subject matter analysts, also have developed an automated tool to facilitate documentation and clearance for the edit review process. The edit management and messaging application (EMMA) is used to track the progress of analysts' review activities. EMMA reports are readily available to analysts, facilitating collaboration among them and contributing to a more efficient review process.

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 in the middle of a multiyear period, estimates of characteristics related to the new question will not be available until they are included in the entire 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.

In some cases, requests for ACS custom tabulations from federal agencies with statutory requirements for information have led to innovative approaches in providing ACS data. For example, the Census Bureau met a Department of Justice tabulation request for ACS estimates to meet Section 203 requirements of the Voting Rights Act by providing modeled ACS estimates.

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/data_documentation/custom_tabulations/.

Chapter 14: Data Dissemination

14.1 Overview

This chapter provides a description of the release schedule and dissemination process for ACS estimates. As in the past, the primary data dissemination system for ACS tabulated data products is the American FactFinder (AFF): http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml. ACS Summary Files and Public Use Microdata Sample (PUMS) files are available through the Census Bureau's File Transfer Protocol (FTP) site that can be accessed through links on AFF and through the ACS website, www.census.gov/acs. Also, as was the case in the past, the ACS 5-year Summary Files and PUMS are also available via Data Ferrett: http://dataferrett.census.gov/index.html.

New developments in the Census Bureau's data dissemination processes have expanded opportunities for using ACS and other census data by improving access. In the sections that follow, the current ACS data dissemination system is described, and the new developments are highlighted to provide an updated account of ACS data dissemination as of February 2013.

14.2 Schedule

Data Release Timetable

Starting in 2010, the information on social, demographic, economic, and housing characteristics previously available only once a decade is now available annually through the ACS for all areas, as shown by Table 14-1.

The Census Bureau has released 1-year ACS estimates in September; 3-year ACS estimates in October; and 5-year ACS estimates in December. PUMS files have been released on a slightly delayed schedule. Table 14-2 shows the general ACS data release schedule.

Table 14-1: ACS Data Availability by Type of Estimate, 2006-2013 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.

Table 14-2: ACS Data Release Schedule

Planned Release	Data Products	Lowest Level Geography
September	1-Year Data Release on American FactFinder (AFF): • Data Profiles • Comparison Profiles • Narrative Profiles • Selected Population Profiles • Ranking Tables • Subject Tables • Detailed Tables • Geographic Comparison Tables 1-Year Summary File Release on FTP and DataFerrett.	Places, County Subdivisions (where available) Geographies of 65,000+ population Exception: Ranking Tables the lowest level is States
October	3-Year Data Release on AFF: • Data Profiles • Comparison Profiles • Narrative Profiles • Selected Population Profiles • Subject Tables • Detailed Tables • Geographic Comparison Tables 3-Year Summary File Release on FTP and DataFerrett.	Places, County Subdivisions (where available) Geographies of 20,000+ population
	1-Year Public Use Microdata Sample (PUMS) File Release on FTP and DataFerrett.	Public Use Microdata Area (PUMA)
December	5-Year Data Release on AFF: No products below were released via DataFerrett	Census Tracts Exceptions: Geographic Comparison Tables the lowest level is Places/County Subdivisions
	• Data Profiles • Subject Tables • Narrative Profiles • Detailed Tables • Geographic Comparison Tables	
	5-Year Summary File Release on FTP and DataFerrett.	Census Block Groups
	3-Year PUMS File Release on FTP and DataFerrett.	PUMA
January (of the following year)	1-Year, 3-Year, & 5-Year Spanish Version of Puerto Rico Community Survey Data Release on AFF. Narrative Profiles will be discontinued on AFF starting 2013.	Census Tracts Exceptions: Geographic Comparison Tables the lowest level is Places/County Subdivisions
	5-Year PUMS File Release on FTP and DataFerrett.	PUMA
Once every five years	Selected Population Tables and American Indian and Alaska Native Tables	Census Tracts

Notes:
1.The release schedule for the 2013 ACS PUMS products was revised to reflect their planned release in 2014.
2.The 2008-2012 Five-Year Narrative Profiles were released on DataWeb as a HotReport.

14.3 Accessing ACS Data and Supporting Documents

American FactFinder (AFF)

The AFF website contains data maps, tables, and reports from a variety of censuses and surveys including the Decennial Census, the ACS, the Population Estimates Program, the Economic Census, and the Annual Economic Surveys.

The AFF is currently the primary Web access tool for ACS data products, which include detailed tables, data profiles, comparison profiles (1- and 3-year data), ranking tables (1-year data only), and geographic comparison tables, selected population profiles, summary files, and downloadable Public Use Microdata Sample (PUMS) files. As the Census Bureau continues its efforts to make ACS data available through new dissemination vehicles, other forms of ACS data products have emerged. Some of these are discussed later in this chapter.

ACS Website

The ACS website contains a wealth of information, documentation, and research papers about the ACS program. The site also hosts many supporting documents related to data products such as subject definitions, code lists, geographic information, comparison guidance, and notes on new and notables for each data release. Documentation on the accuracy of the data also is included, providing information about the sample design, confidentiality, sampling error, nonsampling error, and estimation methodology. The User Notes and 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.

The ACS website can be found at: http://www.census.gov/acs/www.

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 analysis immediately upon data release. The data can be 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.

The ACS Summary Files and PUMS files are located on the FTP site at http://www2.census.gov/ or they are assembled via AFF link and the ACS website at: http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml or http://www.census.gov/acs/www.

Documentation describing the layout of the site in the README file is available in the main directory on the FTP server.

Data Ferrett and the Research Data Centers

In addition to AFF, the Census Bureau supports Data Ferret, a unique data analysis and extraction tool with recoding capabilities to use ACS and other census data sets with federal, state, and local data. Data Ferrett represents a second source of ACS PUMS files as well as of ACS Summary Files. Access to PUMS files on Data Ferrett does not require knowledge of a programming language, making the files popular with many data users who do not have training in the use of computer program languages such as SPSS and SAS.

Qualified researchers can access ACS microdata through the Census Bureau-sponsored Research Data Centers (RDCs). More information about the RDCs is available at: http://www.census.gov/ces/rdcresearch/index.html#.

New Developments in ACS Data Access and Dissemination

A series of new developments in the access, dissemination, and presentation of ACS and other census data sets has been underway at the Census Bureau over the last few years. For example, the Census Bureau has developed its first mobile app to provide updated statistics to its smartphone and tablet data users for the purpose of highlighting major data releases. The first mobile app highlighted the Economic Census; future mobile apps will highlight ACS and other data sets. The Census Bureau updated its popular Quick Facts site in December 2013 to reflect the latest ACS data; regular updates are planned for the future. The launch of a new tool, "Easy Stats," allows users to build their own tables by selecting a desired topic and geography providing users easy access to facts about people, business, and geography based on 5-year ACS estimates.

New developments in data dissemination also include interactive, online mapping applications designed for faster utilization, display, and extraction of data. Two examples of such applications are the Language Mapper and the Census Flows Mapper. The Language Mapper allows a data user to pinpoint the wide array of languages spoken in homes for communities across the nation, along with a detailed report on rates of English proficiency and the growing number of speakers of other languages. Similarly, the Census Flows Mapper allows for a quick and visual extraction of county-to-county migration flow data. The Census Flows Mapper allows the user to select a county and view its inbound, outbound, and net migration flows. Flows can also be filtered by the number of movers and by characteristics such as sex, age, race and Hispanic origin using the 2006-2010 and 2007-2011 5-year data. In the future, the mapper will include data from additional 5-year files and flows will be filtered by additional characteristics.

The Language Mapper can be accessed at:
http://www.census.gov/hhes/socdemo/language/data/language_map.html

To access the Census Flows Mapper go to:
http://flowsmapper.geo.census.gov/flowsmapper/flowsmapper.html

Other developments that could affect ACS data users in the years to come are a federal digital strategy and new initiatives in data dissemination. The federal digital government strategy addresses the challenges and oportunities associated with the proliferation of mobile devices, new technologies such as cloud computing, and expectations that the Federal Government should be able to deliver and receive digital information and services any time, anywhere, and on any device.[29] This intiative has led the Census Bureau to reexamine IT architecture and services, data

[29] Digital Government, pp. 1-12

products, and data dissemination systems, in an effort to align them with the digital government strategy guidelines for accessibility, usability, and efficiency.

An example of a new data dissemination initiative that the Census Bureau is pursuing is the Microdata Analysis System (MAS). The MAS is an automated system for public access to microdata that meet disclosure restrictions. The MAS permits advanced query and analysis via direct access to tabulations of data that would normally require substantial time to develop, review, and clear. While its adoption would not eliminate the need for custom tabulations, it has the potential to help relieve staff of a significant part of the burden of producing such tabulations, and provide data users with more rapid access to custom tabulations that would normally require weeks of staff time to develop.

Finally, the Census Bureau has developed an Application Programming Interface (API) to improve access to and the usability of census data, including ACS estimates. The API empowers web users and developers to come up with new ways to access and display subsets of statistics they choose to meet their customer's needs. The API is the base upon which diverse methods of disseminating Census data to the public can be developed without issues of data storage and data access. Apps developed from the API include the Census Bureau's Easy Stats web app and the Sunlight Foundation's Sitegeist mobile app. Easy Stats provides theme-based navigation to ACS tables. Sitegeist highlights selected ACS data at the tract level and links the app user back to the census.gov webpage. Additional apps that will allow users to explore ACS data in engaging ways are under development.

14.4 References

White House. "Digital Government: Building a 21st Century Platform to Better Serve the American People. May 23, 2012." Available at: http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government-strategy.pdf.

Chapter 15: Improving Data Coverage by Reducing Non-Sampling Error

15.1 Overview

Total survey error includes two components: sampling error and nonsampling error. Chapters 4 and 11 provide information about the steps the Census Bureau takes to reduce American Community Survey (ACS) sampling error. As with all surveys, the quality of ACS data reflects how well the data collection and processing procedures address potential sources of nonsampling error, including coverage, nonresponse, and measurement errors, as well as errors that may arise during data capture and processing. Groves (1989) identified four primary sources of nonsampling error:

- 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 complete 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 non-sampling error and thus improve the quality of ACS estimates. It also includes information about ACS Quality Measures that provide an indication of the potential for some types of nonsampling error. Finally, it describes the annual review process used to demonstrate that ACS data meet the Census Bureau's Statistical Quality Standards. 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 Quality Measures for years 2000 to currently released data are located on the ACS Quality Measures Website: <http://www.census.gov/acs/www/methodology/sample_size_and_data_quality/>.

Additional data products describing data quality are available through the American FactFinder (AFF): <http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml>.

15.2 Coverage Error

All surveys experience some degree of coverage error. It can take the form of undercoverage or overcoverage. Undercoverage occurs when units in the target population do not have a chance of selection into the sample, for example, when our address frame is missing certain housing units or when respondents erroneously exclude some people from a household roster. Overcoverage occurs when units or people have multiple chances of selection, for example, when our address frame lists a housing unit more than once. In general, coverage error will affect survey estimates if the characteristics of the individuals or units excluded or included in error differ from the characteristics of those correctly listed.

The ACS Quality Measures contain housing-level and person-level coverage rates as indicators of the potential for coverage error. The ACS calculates coverage rates for the total resident population 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. We calculate coverage rates for HUs 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. We weight the numerator in the coverage rates to reflect the probability of selection into the sample, subsampling for personal visit follow-up, and adjustments 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 used in production for this purpose serve as the basis for comparison for the coverage rates.

Over-and undercoverage can be partially adjusted as part of the poststratification process, that is, adjusting weights to independent population control totals. The ACS corrects for potential overcoverage or undercoverage 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 undercoverage), 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 overcoverage), the weights of its members are adjusted downward to match the independent estimates. 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, described below. Chapter 3 provides further details.

- Twice a year, the Census Bureau receives the U.S. Postal Service (USPS) Delivery Sequence File (DSF) that consists of the addresses including a house number and street name rather than a rural route or post-office box. Geography Division uses this file to update the city-style addresses on the MAF.
- The ACS nonresponse follow-up operation provides ongoing address and geography updates.
- 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. The Census Bureau chose to put the CAUS program on hold in late 2006 due to the address canvassing operation for the 2010 Census and restarted the program in 2010. The CAUS program in 2013 plans to list approximately 2,000 blocks.

The ACS conducts a telephone follow-up operation on mail responses with conflicting information about the total number of household members and on all households with six or more persons. This follow up confirms that the ACS includes the appropriate persons, reducing the possible impact of coverage error. The Internet instrument includes an error message to a respondent when they provide conflicting information on total persons, and allows up to 20 household members.

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 a unit in 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.

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. The ACS Quality Measures provide separate rates for HUs and GQ persons. For the HU response rate, the numerator is a weighted estimate of the number of cases that were interviewed after all modes of data collection.

To accurately measure unit response, the ACS estimates the universe of cases eligible to be interviewed, which becomes the denominator of the unit response rate.

The ACS Quality Measures also include the weighted estimates of nonresponse rates broken down 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).

For the GQ person response rate, the numerator is a weighted estimate of all interviewed GQ persons after personal visit. The denominator is the weighted estimate of the total number of persons eligible to be interviewed in GQs. For the GQ rates, there are two additional reasons for 202 noninterviews: 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. Beginning with the 2007 ACS data (including multiyear data), two allocation rates—overall HU characteristic and overall person characteristic allocation rates—are available on the AFF at the nation and state level (plus the county level for five-year periods). Allocation rates of many individual characteristics are available on the ACS Quality Measures Web site at the national and state level for 2000 to the present. In addition, the ACS releases imputation tables on AFF that allow users to compute allocation rates for published variables and all published geographies.

The ACS reduces the potential for nonresponse 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 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 Internet and mail nonrespondents, and

a computer-assisted personal interview (CAPI) follow-up for a sample of remaining nonrespondents and unmailable addresses cases.

- The ACS mailing protocol is based on a strategy shown in research studies to obtain a high response rate (Dillman, 1978; Tancreto et al., 2012, Matthews et al., 2012): a prenotice letter, an invitation to respond online with a message on the envelope stating that response is ‘‘required by law,’’ a postcard reminder, a paper questionnaire mailing for nonrespondents to the initial mailing, and, for those same nonrespondents, a second reminder postcard.
- If we cannot find a telephone number for an address that did not respond by Internet or mail at the point when operations switch to CATI nonresponse follow-up, a third postcard is sent urging response by mail, Internet or by calling the toll-free ACS Telephone Questionnaire Assistance (TQA) help line.
- The mailing package includes a frequently asked questions (FAQ) motivational brochure explaining the survey, its importance, and its mandatory nature.
- Both the online survey and paper questionnaire use designs that reflect accepted principles of respondent friendliness and navigation. The intent of the design is to make it easier for respondents to navigate the survey, as well as provide cues for a valid response at an item level (such as providing examples of the type of response desired, or using a prefilled ‘0’ to indicate reporting monetary amounts rounded to the nearest dollar). Similarly, the Internet, CATI and CAPI instruments direct respondents and interviewers to ask the appropriate questions based on the respondent’s answers.
- The online survey and questionnaire provide a toll-free telephone number for respondents who have questions about the ACS in general or who need help in completing the survey.
- The ACS includes a telephone failed-edit follow-up (FEFU) interview with mail and Internet respondents whose responses indicate a discrepancy between the reported household size and the number of people for whom data are provided, and those 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.) In addition, addresses from the Internet identified as a business or vacant are also sent to FEFU.
- The ACS uses permanent professional interviewers trained in refusal conversion methods for CATI and CAPI.
- Survey operations include providing support in other languages: the online survey is available in Spanish, a Spanish paper questionnaire is available on request, and there is a Spanish CATI/CAPI instrument. There are CATI and CAPI interviewers who speak Spanish and other languages as needed. Furthermore, we send a brochure in six languages (English, Spanish, Chinese, Korean, Russian, and Vietnamese) that provides a toll-free number for respondents to receive telephone questionnaire assistance in each language.

15.4 Measurement Error

Measurement error 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 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. In 2012, the Census Bureau conducted a reinterview with a subsample of ACS respondents to attempt to quantify the amount of inconsistency in reporting between the original interview and reinterview. This study, which will be available in early 2014, found minimal evidence of measurement error for most ACS questions.

The ACS minimizes measurement error in several ways, some of which also help to reduce nonresponse.

- ACS pretests new or modified survey questions in all modes before introducing them into the ACS as mandated in the Census Bureau Standard ‘‘Pretesting Questionnaires and Related Materials for Surveys and Censuses (Version 1.2),’’ available at: < http://www.census.gov/srd/pretest-standards.html >.
- The ACS uses a questionnaire design on the Internet and paper form that reflects accepted principles of respondent friendliness and navigation.[30]
- The ACS online survey provides help via a hyperlink for respondents who need additional information on how to interpret and respond to specific questions. Households that receive the paper ACS questionnaire also receive an instruction booklet that provides the same information.
- Respondents may call the toll-free TQA line and speak with trained interviewers for answers to general ACS questions or questions regarding specific items.
- The presentation of questions in the Internet, mail, CATI, and CAPI data collection modes reflect the strengths and limitations of each mode. For example, the online survey provides the ability to review/edit answers at the end of the survey, 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 Internet and CATI/CAPI instruments automatically navigate the questionnaire, showing only those questions appropriate for the interviewee based on their reported characteristics.

[30] http://www.census.gov/acs/www/Downloads/library/2012/2012_Tancreto_02.pdf

- The Internet and CATI/CAPI instruments include functionality that helps achieve valid responses for some questions. 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 Internet and CATI/CAPI instruments include a Spanish version, and bilingual CATI/CAPI interviewers provide language support in other languages.

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 post-data 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 four phases of data collection (Internet, 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 Census Bureau Statistical Quality Standards

Beginning in 2012, the ACS program conducted a new, formal assessment to demonstrate that its data products meet the Census Bureau's recently released quality standards. Because the ACS covers a very broad set of topics, the Census Bureau chose a selection of seventeen important core 5-year estimates reflecting social, demographic, economic, and housing characteristics to gauge the quality of the survey's products. This assessment uses several of the ACS Quality Measures associated with nonsampling error including coverage rates, unit response rates, and item response rates (Sections 15.2 and 15.3 of this document provide an explanation about most of these types of nonsampling errors). We derived a combined nonsampling error rate from these three measures. Given the ACS goal to produce estimates for small areas, counties were chosen to assess quality for smaller geographies that are large enough to result in stable quality measures. The assessment calculated these four rates at the county level, summarizing those results as medians. To meet the quality standard: the median county-level unit response rate must equal or exceed 60 percent, the median county-level coverage rates must equal or exceed 70 percent and the median county-level item response rates for each one of the 17 estimates must equal or exceed 70 percent. The median county-level combined measure of nonsampling error (the product of a coverage rate, the unit response rate and the item response rate) must equal or exceed 50 percent for each of the 17 estimates.

The assessment also includes a measure of sampling error that does not come from the ACS Quality Measures. Sampling error in survey estimates arises due to the use of probability sampling. The ACS was designed to produce reliable tract-level estimates based on 5-year data. To determine a measure of sampling error, the assessment includes the median tract-level 5-year coefficient of variation (CV) for each of the 17 estimates. The CV is equal to the standard error (SE) of a weighted estimate divided by the weighted estimate itself. The majority of the estimates must have a CV less than or equal to 30 percent to meet the quality standard.

For both 2012 and 2013, the ACS data products have met the Census Bureau's Statistical Quality Standards. There is every indication that ACS will continue to meet them.

15.7 References

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

ACS Design and Methodology Improving Data Quality by Reducing Nonsampling Error 15−5

U.S. Census Bureau

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.

Matthews, B., Davis, M, Tancreto, J.G., Zelenak, M.F, and Ruiter, M (2012) "2011 American Community Survey Internet Tests: Results from Second test in November 2011", U.S. Census Bureau Report.
http://www.census.gov/acs/www/Downloads/library/2012/2012_Matthews_01.pdf

Tourangeau, R., and T. Yan. (2007) ''Sensitive questions in surveys,'' Psychological Bulletin, 133(5): 859−883.

Tancreto, J.G., Zelenak, M.F., Davis, M, Ruiter, M., and Matthews, B. (2012) "2011 American Community Survey Internet Tests: Results from First Test in April 2011", U.S. Census Bureau Report. http://www.census.gov/acs/www/Downloads/library/2012/2012_Tancreto_01.pdf

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

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.

Census Bureau (2006) ''Census Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses Version 1.2,'' Washington, DC.

Chapter 16: Research and Evaluation

16.1 Overview

Short and long-range planning of the ACS includes research and evaluation (R&E) activities on all aspects of the ACS program, including survey methodology and analytical research. This chapter will describe the process for identifying, prioritizing, and implementing R&E activities that comprise topics such as: experiments (quantitative and qualitative studies); simulations (e.g. alternative weighting methodologies, alternative filtering methods, variance estimation, etc.); feasibility and cost/benefit studies; and quality assessments and measurements of error. R&E projects span research on both housing units and group quarters. Every fiscal year, the ACS reprioritizes R&E projects based on new directives or changes in budgets.

The R&E program includes a specific subset of projects known as ACS Methods Panel tests. These are identified, prioritized and approved through the same R&E process. ACS Methods Panel tests are aimed at improving overall ACS data quality; achieving survey efficiencies; and developing and improving ACS questionnaire content and related data collection materials.

Annually, the R&E Working Group (WG), made up of ACS mid-level and senior-level managers, brainstorms R&E projects that are aligned with the ACS program's strategic objectives. In particular, for fiscal year 2013 (FY2013), which covers the period October 2013 to September 2014, all R&E projects align with these two strategic objectives:

- Accurate demographic, social, economic, and housing data products are published at all geographic levels or strategic objective
- Efficient, effective, and adaptable survey data collection methods

After the identification of a list of potential R&E projects, the R&E WG prioritizes the projects based on specific criteria for the fiscal year. For example, the criteria for FY 2013 are based on how closely the project aligns with these seven criteria:

1. The degree of strategic alignment
2. The benefits to stakeholders and/or data users
3. If it has a significant follow on benefit or is a dependency for future projects or programs
4. The urgency of the project
5. If it is mandatory
6. The technical feasibility of the project
7. The management feasibility of the project

Each project is ranked as high, medium, or low on each of the seven criteria by each member of the R&E WG; the rank is then converted into a utility score. Projects with a low utility score are deferred or denied, and an initial list of approved projects is identified.

After the initial list of approved R&E projects is determined, Census Bureau senior-level managers recommend appropriate high-level methodology and metrics for the research projects. These managers have a broad perspective on what research is in progress across the Census Bureau and may know of related external research that is relevant to consider. As a result of their input, other R&E projects may be identified for the ACS program. These new projects may be incorporated into the list. A final round of reviews takes place that includes the consideration of available resources and results in the approval of a list of R&E projects for the fiscal year.

The R&E projects are determined annually; however, additional R&E projects can be added anytime throughout the year. These projects go through a similar review process. The R&E WG reviews potential projects monthly and prioritizes the projects against the current approved projects and available resources. If the project is approved it will be incorporated into the ACS R&E project list for the fiscal year. Current projects may also be deferred based on available resources.

Researchers conduct the R&E projects with oversight and guidance from ACS senior-level managers. The researchers provide monthly status updates and request approval from ACS senior managers if the scope, time, or resources needed for projects change. Before research begins, the researchers develop a Research and Evaluation Analysis Plan (REAP). This plan outlines the history, motivations, and reasons for doing the project. It also explains, in detail, the research questions and the methodology and metrics used. Design assumptions and limitations are also discussed. Subject matter experts, known as critical reviewers, are assigned to review each REAP and to provide feedback and guidance on the methodology and research questions. After the incorporation of all feedback, the critical reviewers and the researcher's supervisors approve the REAP and research begins.

Once the research is complete, preparation of a written report begins to present the findings. After the report is drafted, it under goes review and is briefed to the ACS R&E WG and ACS senior-level managers prior to being finalized. The final R&E reports are put into the ACS R&E memoranda series and made available to the public on the ACS web page located at http://www.census.gov/acs/www/library/by_series/acs_research_evaluation_program.

Appendix: Glossary

Term/Acronym	Description
ACS	American Community Survey
Address Canvassing	A process for updating addresses and map databases to ensure they are accurate and current. Address canvassing operations require that address listing staff (address listers) travel to every census block to which they are assigned; identify every place where people live, stay, or could live or stay; and verify, correct, add, and delete address records as appropriate while also updating map features on an electronic map provided on a hand-held computer.
AFF	American FactFinder
AIANHH	American Indian Areas/Alaska Native Areas/Hawaiian Home Lands
AIANSA	American Indian and Alaska Native Statistical Areas
ALMI	Automated Listing and Mapping Instrument
Automated Coding	A process of coding write-in responses using digital means. Written-in responses are keyed into digital data and then matched to a data dictionary. The data dictionary(ies) contain(s) a list of the most common responses, with a code attached to each. The coding program(s) attempt(s) to match the keyed response to an entry in the dictionary to assign a code.
BLAISE	An open-source scripting software language
BLAISE Instrument	A computer-assisted interviewing system and survey processing tool that supports interactive data review, editing, and other survey processing tasks
BoP	Bureau of Prisons
BR	Base Rate
BW	Base Weight
CAPI	Computer Assisted Personal Interviewing
CATI	Computer Assisted Telephone Interviewing
CAUS	Community Address Updating System
CDP	Census Designated Places
Clerical Coding	A process of coding write-in responses by trained human coders.
Coverage Error	**See Overcoverage** and **Undercoverage**.
CPS	Current Population Survey
CQR	Count Question Resolution
DAAL	Demographic Area Address Listing
DoD	Department of Defense
DSF	Delivery Sequence File
EDS	Excluded from Delivery Statistics
FAQs	Frequently Asked Questions

Term/Acronym	Description
FEFU	Failed-edit Follow-up
FR	Field Representative
FTP	File Transfer Protocol
GPS	Global Positioning System
GQ	Group Quarters
GQFQ	Group Quarters Facility Questionnaire
GQMOS	Group Quarters Measure of Size
GQPPSF	GQ Person Post-Stratification Factor
GREG	Generalized Regression Estimation
GWTF	GREG Weighting Factor
HHF	Householder Factor
HPF	Housing Unit Post-Stratification Factor
HU	Housing Unit
IVR	Interactive Voice Recognition
KFI	Key-from-Image
LACS	Locatable Address Conversion Service
LUCA	Local Update of Census Addresses
MAF	Master Address File
MBF	Mode Bias Factor
MCD	Minor Civil Divisions
Methods Panel	An ongoing research program designed to improve the quality of the estimates, response to the survey, and contain data collection costs of the ACS. It is the vehicle for testing changes before we implement new methods in ACS production. The experimental testing includes proposed new or revised content and proposed enhancements to various data collection methods.
MOS	Measure of Size
MTdb	MAF/TIGER database
NHIS	National Health Interview Survey
NIF1	HU First Noninterview Adjustment Factor
NIF2	HU Second Noninterview Adjustment Factor
NIFM	HU Mode Noninterview Adjustment Factor
NPC	National Processing Center
OMB	Office of Management and Budget
OMR	Optical Mark Recognition
Overcoverage	Errors that occur due to the inclusion in a sample of elements (for example, addresses) that should not be included
PPSF	HU Person Post-Stratification Factor
PRCS	Puerto Rico Community Survey
PUMS	Public Use Microdata Sample
RO	Regional Office
SEMOS	Smallest Entity Measure of Size
SIPP	Survey of Income and Program Participation
SSF	CAPI Subsampling Factor

Term/Acronym	Description
TDD	Telephone Device for the Deaf
ACS Test Counties/ Test Sites	Testing to demonstrate the feasibility of conducting the ACS took place from 1996 to 2001. The testing included 36 counties organized into 31 test sites. Most of these sites were single-county sites; others consisted of multiple, contiguous counties (such as the ACS test site of Vilas and Oneida Counties, WI).
TIGER	Topologically Integrated Geographic Encoding and Referencing
TMOS	Tract Measure of Size
TQA	Telephone Questionnaire Assistance
UAA	Undeliverable as Addressed''
Undercoverage	Inadequate representation of some elements (for example, members of a population group) in a sample
USDA	U.S. Department of Agriculture
USPS	United States Postal Service
VMS	HU Variation in Monthly Sample Factor
WCTYCON	GQ Weight after the County-level Constraint
WGQPPSF	GQ Weight after the GQ Person Post-Stratification Factor
WGWTF	HU Weight after the GREG Weighting Factor
WHHF	HU Weight after the Householder Factor
WHPF	HU Weight after the Housing Unit Post-Stratification Factor
WMBF	HU Weight after the Mode Bias Factor
WNIF1	HU Weight after the First Noninterview Adjustment
WNIF2	HU Weight after the Second Noninterview Adjustment
WNIFM	HU Weight after the Mode Noninterview Adjustment
WPPSF	HU Person Weight after the Person Post-Stratification Factor
WSSF	GQ Weight after the CAPI Subsampling Factor
WSTCON	GQ Weight after the State-level Constraint
WTRCON	GQ Weight after the Tract-level Constraint
WVMS	Weight after the Variation in Monthly Sample Factor