Accuracy of the Data

INTRODUCTION

The data contained in this product are based on the Census 2000 sample. The data are estimates of the actual figures that would have been obtained from a complete count. Estimates derived from a sample are expected to be different from the 100-percent figures because they are subject to sampling and nonsampling errors. Sampling error in data arises from the selection of people and housing units included in the sample. Nonsampling error affects both sample and 100-percent data and is introduced as a result of errors that may occur during the data collection and processing phases of the census. This Appendix provides a detailed discussion of both types of errors and a description of the estimation procedures.

MASTER ADDRESS FILE DEVELOPMENT

The majority of addresses in the country are in what is known for census purposes as Mailout/Mailback areas, which generally consist of city-style addresses. The original source of addresses on the Master Address File (MAF) for the Mailout/Mailback areas was the 1990 Census Address Control File (ACF). The first update to the ACF addresses is a United States Postal Service (USPS) Delivery Sequence File (DSF) of addresses. The November 1997, September 1998, November 1999, and April 2000 DSFs were incorporated into the MAF.

Until shortly before the census, the ACF addresses and the November 1997 and September 1998 residential DSF addresses constituted the MAF. These addresses were tested against Census Bureau geographic information to determine their location at the census block level. The geographic information is maintained in the Census Bureau's Topologically Integrated Geographic Encoding Referencing (TIGER) system. When an address on the MAF can be uniquely matched to the address range in TIGER for a street segment that forms one of the boundaries of a particular block, the address is said to be *geocoded* to that block. Valid and geocoded addresses appeared on each address list used for a field operation.

The Block Canvass operation was the next major address list operation in the Mailout/Mailback areas for Census 2000. Between January and May 1999, there was a 100-percent canvass of every block in these areas. Every geocoded address was printed in a block-by-block address register. Block Canvassing listers identified each address as one of the following: a verified housing unit; a unit with corrections to the street name or directional; a delete; a duplicate, implying the unit exists elsewhere on the list with a different, unmatchable designation, such as a different street name or building name; uninhabitable; or nonresidential. Also, units that were deleted from one block and matched an added unit in another block were called a move.

A cooperative address list check with local governmental units throughout the country, called Local Update of Census Addresses (LUCA) 98, occurred in approximately the same time frame as Block Canvassing. In LUCA 98, the participating governmental units received an address list and were asked for input mostly on added units but also on deleted units and corrected street names or directionals. The outcome of this operation was similar to that of Block Canvassing; units were added to and deleted from blocks, and address corrections were made.

The Decennial Master Address File (DMAF), created in July 1999, was the file used for the main printing of the Census 2000 questionnaires. In Mailout/Mailback areas, the operations that had yielded housing units and their status before this initial printing stage were the ACF, the November 1997 DSF, the September 1998 DSF, LUCA 98, and Block Canvassing.

Updates to the DMAF followed the creation of the initial DMAF. Addresses were added by the November 1999, February 2000, and April 2000 DSFs. The LUCA 98 field verification and appeal processes were address update operations that occurred subsequent to the creation of the initial

DMAF. Units receiving a conflicting status from Block Canvassing and the LUCA 98 operation were sent for field verification by the Census Bureau; the results of the field verification were sent to the governmental units. The governmental unit could appeal the Census Bureau's findings for particular units at this stage. At an appeal, the Census Bureau and the governmental unit submit their evidence of the status of a housing unit for independent review. The Census Address List Appeals Office, a temporary Federal office established outside the Department of Commerce, administered the appeal process. The Director of the Appeals Office (or their designee) was responsible for issuing a written determination that was considered final. Both the field verification and the appeal process had the potential to change the status of a housing unit.

The New Construction operation was another cooperative effort with participating governmental units that added addresses before Census Day. This was a final operation in Mailout/Mailback areas that used governmental units' local knowledge to identify new housing units in February and March of 2000.

After Mailout/Mailback, the second most common method of questionnaire delivery was Update/Leave. Rather than obtaining addresses from the ACF and DSF, the address list for Update/Leave areas was constructed during a Census Bureau field operation called Address Listing. This was due to the fact that addresses in Update/Leave areas were primarily noncity-style. Census employees were sent to the field with maps of their assignment areas and were instructed to record the city-style address, noncity-style address or location description, or possibly some combination of the above, for every housing unit. In addition, the location of the unit was noted on the census map with what is known as a *map spot*. This operation took place in the fall of 1998.

After processing the Address Listing data, the Census Bureau could tabulate the number of housing units in each block. Because the housing units in these areas may have nonstandard mailing addresses and may be recorded in census files solely with a location description, the governmental units participating in the local review operation in these areas were sent lists of housing unit counts by block. This operation was called LUCA 99. When a LUCA 99 participant disagreed with a Census block count, the contested block was sent out for LUCA 99 recanvassing. Census employees were redeployed to make updates to the address list. In addition, there was a LUCA 99 appeal process for settling housing unit status discrepancies that could potentially add units to the address list. The LUCA 99 recanvassing and LUCA 99 appeal process took place at various times during the DMAF updating process. Although most of the LUCA 99 entities had their recanvassing results processed before creation of the initial DMAF, many did not. There were DMAF updates designed specifically for obtaining late recanvassing and appeal results. These updates to the census files occurred in time for USPS delivery of a questionnaire.

The last address list-building operation in the Update/Leave areas was the Update/Leave operation itself. This operation was responsible for having a census questionnaire hand-delivered at every housing unit. The MAF and the maps were updated during this process.

In the most remote regions of the country, housing units were listed at the same time people within them were enumerated for Census 2000. These operations, called List/Enumerate and Remote Alaska enumeration, were the only source of addresses in these regions. All housing units were map spotted at the time of enumeration.

In some other regions of the country where an address list had already been created, the Census Bureau determined that direct enumeration of the population would be more successful than mailback of the forms. This operation was called Update/Enumerate. There were two types of Update/Enumerate areas – urban areas that were formerly Mailout/Mailback and rural areas that were formerly Update/Leave. The urban areas had passed through all the Mailout/Mailback operations up through the point of the creation of the initial DMAF, and the rural areas had passed through Address Listing, and sometimes LUCA 99, by the time of the creation of the initial DMAF. Because of these distinct paths, it was necessary to distinguish between the urban and rural Update/Enumerate areas.

Urban Update/Leave is another special enumeration that took place in areas where mail delivery was considered to be problematic. The addresses had passed through all the operations of the

Mailout/Mailback areas up through the creation of the initial DMAF, but enumerators visited the area during the census. As a result, additions, deletions and corrections to the address list were made.

People who do not receive a questionnaire at their house could submit a Be Counted Form, or they could call Telephone Questionnaire Assistance and have their information collected over the telephone. Addresses from these operations that did not match those already on the DMAF and that were geocoded to a census collection block in an area where census enumeration did not take place were visited in a Field Verification operation to determine if they existed. Verified addresses were added to the address list.

Follow-up operations provided additional information about housing units listed on the DMAF. In Nonresponse Followup (NRFU), enumerators followed up on units that had not returned a preaddressed census form. These units could be enumerated, deemed vacant, or possibly deleted. At the same time, units that did not appear on the address list could be added and enumerated concurrently. Coverage Improvement Follow Up was designated for enumeration at addresses added by New Construction and the later Delivery Sequence Files, as well as a second check on NRFU vacant and deleted units. Adds were also possible. These operations occurred in the Mailout/Mailback, Update/Leave, and Urban Update/Leave areas.

SERVICE-BASED ENUMERATION

Service Based Enumeration was designed to account for people without a usual residence who use service facilities (i.e., shelters, soup kitchens and mobile food vans). Only people using the service facility on the interview day were enumerated. In addition, people enumerated in Targeted Non-Shelter Outdoor Locations (TNSOLS) and people without a usual residence that filed Be Counted Forms (BCF) augmented the count. **This component of the enumeration should *not* be interpreted as a complete count of the population without a usual residence.**

SAMPLE DESIGN

Every person and housing unit in the United States was asked basic demographic and housing questions (for example, race, age, and relationship to householder). A sample of these people and housing units was asked more detailed questions about items, such as income, occupation, and housing costs. The sampling unit for Census 2000 was the housing unit, including all occupants. There were four different housing unit sampling rates: 1-in-8, 1-in-6, 1-in-4, and 1-in-2 (designed for an overall average of about 1-in-6). The Census Bureau assigned these varying rates based on precensus occupied housing unit estimates of various geographic and statistical entities, such as incorporated places and interim census tracts. For people living in group quarters or enumerated at long form eligible service sites (shelters and soup kitchens), the sampling unit was the person and the sampling rate was 1-in-6.

The sample designation method for housing units depended on the data collection procedures. Approximately 95 percent of the population was enumerated by the mailback procedure. In these areas, the Census Bureau used the Decennial Master Address File (DMAF) to select electronically a probability sample. The questionnaires were either mailed or hand-delivered to selected addresses with instructions to complete and mail back the form.

The housing unit sampling rate varied by census block. Long Form Sampling Entities (LFSEs) were used to determine sampling rates in Census 2000 similarly to the way governmental units were used in the 1990 census sample design. LFSEs were:

- Counties and county equivalents (such as parishes in Louisiana).
- Cities.
- Incorporated places (including consolidated cities).
- Census designated places in Hawaii only.

- Minor civil divisions in certain states only (Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Wisconsin).
- School districts (based on the 1995-1996 school year).
- American Indian reservations.
- Tribal jurisdiction statistical areas.
- Alaska Native village statistical areas.

Size estimates for LFSEs were based on housing unit counts from the DMAF and occupancy rates from the 1990 census. If the smallest LFSE that included all or any part of a block had an estimated housing unit count of less than 800, the housing units in the block were sampled at a 1-in-2 rate. If the smallest LFSE that included all or any part of a block had an estimated housing unit count of 800 or more but less than 1,200, housing units in the block were sampled at a 1-in-4 rate. If a block was not in either of the two previous sampling rate categories, and was part of an interim census tract with 2,000 or more estimated housing units, the housing units in the block were sampled at a 1-in-8 rate. Housing units in all remaining blocks (those not assigned to 1-in-2, 1-in-4, or 1-in-8 rates) were sampled at a 1-in-6 rate.

In List/Enumerate areas (accounting for less than 0.5 percent of the housing units), each enumerator was given a blank address register with designated sample lines. Beginning about Census Day, the enumerator systematically canvassed an Assignment Area (AA) and listed all housing units in the address register in the order they were encountered. Completed questionnaires, including sample information for any housing unit listed on a designated sample line, were collected. If an AA contained any blocks that would qualify as above for a 1-in-2 or 1-in-4 rate, all households in the AA were sampled at 1-in-2. Housing units in all other AAs were sampled at 1-in-6.

Housing units in American Indian reservations, tribal jurisdiction statistical areas, and Alaska Native villages were sampled according to the same criteria as other LFSEs, except the sampling rates were based on the size of the American Indian and Alaska Native population in those areas as measured in the 1990 census. Trust lands were sampled at the highest rate of any part of their associated American Indian reservations. If the associated American Indian reservation was entirely outside the state containing the trust land, then the trust land was sampled at a 1-in-2 rate. All Remote Alaska assignment areas were sampled at a rate of 1-in-2. Housing units in Puerto Rico were sampled at a constant 1-in-6 rate in all blocks.

Variable sampling rates provide relatively more reliable estimates for small areas and decrease respondent burden in more densely populated areas while maintaining data reliability. When all sampling rates were taken into account across the Nation, approximately 1 out of every 6 housing units was included in the Census 2000 sample.

CONFIDENTIALITY OF THE DATA

The Census Bureau has modified or suppressed some data in this data release to protect confidentiality. Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual can be identified. The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed.

Title 13, United States Code. Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to 5 years in prison and up to $250,000 in fines for wrongful disclosure of confidential census information.

Disclosure limitation. Disclosure limitation is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual who provided information under a pledge of confidentiality. Using disclosure limitation procedures, the Census Bureau modifies or removes the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual, the Census Bureau has taken steps to disguise the original data while making sure the results are still useful. The techniques used by the Census Bureau to protect confidentiality in tabulations vary, depending on the type of data.

Data swapping. Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percentage of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and the same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Because of data swapping, users should not assume that tables with cells having a value of 1 or 2 reveal information about specific individuals. Data swapping procedures were first used in the 1990 census and were also used for Census 2000.

ERRORS IN THE DATA

Statistics in this data product are based on a sample. Therefore, they may differ somewhat from 100-percent figures that would have been obtained if all housing units, people within those housing units, and people living in group quarters had been enumerated using the same questionnaires, instructions, enumerators, and so forth. The sample estimate also would differ from other samples of housing units, people within those housing units, and people living in group quarters. The deviation of a sample estimate from the average of all possible samples is called the *sampling error*. The *standard error* of a sample estimate is a measure of the variation among the estimates from all possible samples. Thus, it measures the precision with which an estimate from a particular sample approximates the average result of all possible samples. The sample estimate and its estimated standard error permit the construction of interval estimates with prescribed confidence that the interval includes the average result of all possible samples. The method of calculating standard errors and confidence intervals for the data in this product appears in the section called “Calculation of Standard Errors.”

In addition to the variability that arises from the sampling procedures, both sample data and 100-percent data are subject to *nonsampling error*. Nonsampling error may be introduced during any of the various complex operations used to collect and process census data. For example, operations such as editing, reviewing, or handling questionnaires may introduce error into the data. A detailed discussion of the sources of nonsampling error is given in the section on “Nonsampling Error” in this Appendix.

Nonsampling error may affect the data in two ways: errors that are introduced randomly will increase the variability of the data and, therefore, should be reflected in the standard error; errors that tend to be consistent in one direction will make both sample and 100-percent data biased in that direction. For example, if respondents consistently tend to underreport their incomes, then the resulting counts of households or families by income category will tend to be understated for the higher income categories and overstated for the lower income categories. Such biases are not reflected in the standard error.

Limitations of the Group Quarters Data

By definition, universes that include the total population include both the household population and the group quarters population. For example, the universe defined as the population 15 years and over includes all people 15 years and over in both households and group quarters.

In previous censuses and in Census 2000, allocation rates for demographic characteristics (such as age, sex, and race) of the group quarters population were similar to those for the total population. However, allocation rates for sample characteristics, such as school enrollment, educational attainment, income, and veteran status for the institutionalized and noninstitutionalized group quarters population have been substantially higher than those for the household population since at least the 1960 census. A review of the Census 2000 allocation rates for sample characteristics indicated that this trend continued.

Although allocation rates for sample characteristics are higher for the group quarters population, it is important to include the group quarters population in the total population universe. In most areas, the group quarters population represents a small proportion of the total population. As a result, the higher allocation rates associated with the group quarters population have minimal impact on the sample characteristics for the area of interest. In areas where the group quarters population represents a larger percentage of the total population, the Census Bureau cautions data users about the impact the higher allocation rates may have on the sample characteristics.

Calculation of Standard Errors

Totals and percentages. Tables A through C in this Appendix contain the necessary information for calculating the standard errors of sample estimates in this data product. To calculate the standard error, it is necessary to know:

- The unadjusted standard error for the characteristic (given in Table A for estimated totals or Table B for estimated percentages) that would result under a simple random sample design of people, housing units, households, or families.
- The design factor for the particular characteristic estimated (given in Table C) based on the sample design and estimation techniques employed to produce long form data estimates.
- The number of people, housing units, households, or families in the publication area.
- The observed sampling rate.

The design factor is the ratio of the estimated standard error to the standard error of a simple random sample. The design factors reflect the effects of the actual sample design and the complex ratio estimation procedure used for Census 2000. Percent-in-sample values are provided in Summary File 3. The percent of the population in sample is given in P4, Percent of the Population in Sample. Percent-in-sample values for housing units are provided in H4, Percent of Housing Units in Sample by Occupancy Status. Thus, observed sampling rates for housing units are provided separately for occupied and vacant housing units. Data users should use information in H2, Unweighted Sample Housing Units by Occupancy Status, to determine the most prevalent type of housing unit in a specific geography (occupied or vacant), and use its corresponding percent-in-sample value from H4.

Use the steps given below to calculate the standard error of an estimated total or percentage contained in this product. A percentage is defined here as a ratio of a numerator to a denominator where the numerator is a subset of the denominator. For example, the proportion of Black or African-American teachers is the ratio of Black or African-American teachers to all teachers.

1. Obtain the unadjusted standard error from Table A or B (or use the formula given below the table) for the estimated total or percentage, respectively.
2. Obtain the person or housing unit observed sampling rate (percent-in-sample) for the geographic area to which the estimate applies. Use the person observed sampling rate for population characteristics and the housing unit observed sampling rate for housing characteristics.
3. Use Table C to obtain the appropriate design factor, based on the characteristic (Employment status, School enrollment, etc.) and the range containing the percent-in-sample value defined in step 2. Multiply the unadjusted standard error by this design factor.

The unadjusted standard errors of zero estimates or of very small estimated totals or percentages will approach zero. This is also the case for very large percentages or estimated totals that are close to the size of the publication areas to which they correspond. Nevertheless, these estimated totals and percentages are still subject to sampling and nonsampling variability, and an estimated standard error of zero (or a very small standard error) is not appropriate. For estimated

percentages that are less than 2 or greater than 98, use the unadjusted standard errors in Table B that appear in the "2 or 98" row. For an estimated total that is less than 50 or within 50 of the total size of the publication area, use an unadjusted standard error of 16.

Examples using Tables A and B are given in the section titled "Using Tables to Compute Standard Errors and Confidence Intervals."

Sums and differences. The standard errors estimated from Tables A and B are not directly applicable to sums of and differences between two sample estimates. To estimate the standard error of a sum or difference, the tables are to be used somewhat differently in the following three situations:

1. For the sum of or difference between a sample estimate and a 100-percent value, use the standard error of the sample estimate. The complete count value is not subject to sampling error.

2. For the sum of or difference between two sample estimates, the appropriate standard error is approximately the square root of the sum of the two individual standard errors squared; that is, for standard errors

   $SE(\hat{X})$ and $SE(\hat{Y})$ of estimates $\hat{X}$ and $\hat{Y}$, respectively:

   $$SE(\hat{X} + \hat{Y}) = SE(\hat{X} - \hat{Y}) = \sqrt{[SE(\hat{X})]^2 + [SE(\hat{Y})]^2}$$

   This method, however, will underestimate (overestimate) the standard error if the two items in a sum are highly positively (negatively) correlated or if the two items in a difference are highly negatively (positively) correlated. This method may also be used for the difference between (or sum of) sample estimates from two censuses or from a census sample and another survey. The standard error for estimates not based on the Census 2000 sample must be obtained from an appropriate source outside of this Appendix.

3. For the differences between two estimates, one of which is a subclass of the other, use the tables directly where the calculated difference is the estimate of interest. For example, to determine the estimate of non-Black or African-American teachers, subtract the estimate of Black or African-American teachers from the estimate of total teachers. To determine the standard error of the estimate of non-Black or African-American teachers, apply the above formula directly.

Ratios. Frequently, the statistic of interest is the ratio of two variables, where the numerator is not a subset of the denominator. An example is the ratio of students to teachers in public elementary schools. (Note that this method cannot be used to compute a standard error for a sample mean.) The standard error of the ratio between two sample estimates is estimated as follows:

1. If the ratio is a proportion, then follow the procedure outlined for "Totals and percentages."

2. If the ratio is not a proportion, then approximate the standard error using the formula below.

   $$SE\left(\frac{\hat{X}}{\hat{Y}}\right) = \left(\frac{\hat{X}}{\hat{Y}}\right)\sqrt{\frac{[SE(\hat{X})]^2}{\hat{X}^2} + \frac{[SE(\hat{Y})]^2}{\hat{Y}^2}}$$

Medians. The sampling variability of an estimated median depends on the form of the distribution and the size of its base. The reliability of an estimated median is approximated by constructing a confidence interval. Estimate the 68 percent confidence limits of a median based on sample data using the following procedure.

1. Obtain the appropriate (person or housing unit) observed sampling rate for the specific geographic area. Use this rate to locate the design factor for the characteristic of interest in Table C.