

**Census 2000 Dress Rehearsal
100-Percent Summary File Product Documentation
Data for Sacramento, California and Menominee County, Wisconsin
Without Correction for ICM**

The Census 2000 Dress Rehearsal was conducted in 1998 in three sites—Sacramento, California; 11 counties in South Carolina, including the city of Columbia and the town of Irmo; and Menominee County, Wisconsin, including the Menominee American Indian Reservation. The Census Bureau used statistical sampling and estimation techniques in the Sacramento and Menominee sites. In the South Carolina site, only traditional census methods were used. The Bureau released data products for the Sacramento and Menominee sites including the use of statistical methods.

In the Fall of 1997, the House of Representatives passed the 1998 appropriations bill (Public Law 105-119) concerning the issues of sampling. According to this bill, in both Census 2000 and any dress rehearsal or other simulation made in preparation for the Census 2000, the number of persons enumerated without using statistical methods must be publicly available for all levels of census geography. To comply with the Public Law 105-119, the Census Bureau released the same data products for the Sacramento and Menominee sites without correction for integrated coverage measurement (ICM). ICM is a quality check survey to correct for undercoverage. (See the section “Accuracy of the Data” in the *Census 2000 Dress Rehearsal 100-Percent Summary File Product Documentation* for more details on the statistical sampling and estimation techniques associated with ICM. For subject matter definitions, such as race, geography, etc., see the sections “Definitions of Subject Characteristics” and “Geographic Definitions” in the same document). These products are available through the Internet using the Bureau’s American FactFinder (AFF) system.

Accuracy of the Data

SACRAMENTO

INTRODUCTION

The Census 2000 Dress Rehearsal was the last step of the Census 2000 testing cycle. It was conducted in 1998 in the city of Sacramento, California; 11 counties in South Carolina, including the city of Columbia and the town of Irmo; and Menominee County, Wisconsin, including the Menominee American Indian Reservation. These sites were selected because of their demographic and geographic characteristics to reflect some of the expected Census 2000 conditions and environments. The census-taking methodology employed in each site had a different mix of operational and statistical procedures. In the Sacramento and Menominee sites, statistical sampling and estimation techniques were used, aiming to improve the accuracy of the population count. Census results from these two sites are subject to sampling and nonsampling errors. The sampling and estimation techniques and their associated sampling errors are described below. The South Carolina site used traditional census methods only; census results from this site are not subject to sampling errors, but are subject to nonsampling errors. Comparisons of the results between the sites should *not* be made because a different method was used to arrive at the final results for each site.

MASTER ADDRESS FILE DEVELOPMENT

In order for the Census 2000 Dress Rehearsal to be as accurate, complete, and cost effective as possible, the address list, which serves as the basis for control for the census, must be as accurate and complete as possible. If an address is not on the list, then its residents are less likely to be counted. The Master Address File (MAF) building process for the Census 2000 Dress Rehearsal in Sacramento involved a series of operations that built on each other and ultimately resulted in the address list used to conduct the census.

The initial MAF was built using the 1990 Census's Address Control File (ACF), the United States Postal Service's Delivery Sequence File (DSF), and the Topologically Integrated Geographic Encoding and Referencing (TIGER®) database. The ACF and DSF were merged together to create an initial list of addresses that were then matched to the TIGER® database to geocode each unique record to census blocks and higher level geographic areas. A series of updating operations was then conducted to improve the coverage of housing units and the geocoding accuracy of the housing units on the MAF.

A Targeted Multi-Unit Check was conducted, where the counts of housing units at multi-unit addresses (apartment buildings, rooming houses, etc.) were compared between the ACF and the DSF. Where these counts differed, enumerators visited these addresses to ensure that the census address list had the correct number of units and the correct unit designations.

Also, a Targeted Canvass was conducted. Local officials were asked to identify and prioritize blocks that they expected to contain hidden housing units. These hidden units were units that the post office may not be aware of because they may use single mail drop points, they may be recent conversions from single unit addresses (like a basement or garage apartment), or they may be illegal, purposely hidden units. During canvassing, using the census address list, field staff looked for missing or hidden units in the particular blocks identified by the local officials or in a subset of these blocks, depending on how many were identified.

In the Program for Address List Supplementation (PALS), a list of addresses, provided by local Sacramento officials, was matched to the census address list. Units found on the local list but missing from the census list were provisionally added to the census list.

During the Local Update of Census Addresses (LUCA) operation, local and tribal governments were given the opportunity to review the census address list for accuracy and completeness before the delivery of questionnaires. They had the opportunity to provide information about additions, deletes, and corrections to the list of addresses in their jurisdictions, and to correct geocoding errors. Field verification identified which of the PALS and LUCA addresses were retained.

In the Postal Check operation, United States Postal Service employees verified the accuracy of the MAF by comparing MAF listed addresses to the addresses on their carriers' delivery routes. The operation was limited to 22 zip codes that were entirely within the site and that consisted entirely of mailout/mailback areas. The primary purpose of this operation was to capture recent new construction in time for the mailout of census questionnaires.

Finally, the Be Counted program and the Telephone Questionnaire Assistance (TQA) program yielded some additional housing units that were not previously listed on the MAF. These programs offered alternative options for people to be included in the census if they did not think they were otherwise enumerated. The Be Counted program gave residents access to questionnaires in their local community. People also contacted the Census Bureau through TQA and requested that a form be mailed to them.

NONRESPONSE FOLLOW-UP & UNDELIVERABLE-AS-ADDRESSED VACANT SAMPLE DESIGN

Two types of sampling were conducted in the initial phase of the Census 2000 Dress Rehearsal: sampling for nonresponse follow-up (NRFU) and sampling of undeliverable-as-addressed (UAA) vacant housing units. These sampling operations were implemented only in Sacramento, CA. There was complete follow-up of nonrespondents and UAA vacants in Menominee, WI, since it is an American Indian Reservation, and in Columbia, SC, since that site excluded sampling operations. All American Indian Reservations will have 100 percent follow-up of nonrespondents and UAA vacants because the tribes requested it as part of their government-to-government relationship with the Federal government.

Instead of visiting all addresses that did not voluntarily return a census form, as was done in previous decennial censuses, enumerators in Sacramento followed up a sample of nonrespondents. The sample was designed so that each census tract reached a final completion rate of at least 90 percent. For example, if the initial completion rate in a census tract, defined as

$$\frac{\text{\# of Respondents} + \text{\# of UAA vacants}}{\text{Total \# of Addresses}} \times 100 \text{ percent} ,$$

was 60 percent, then a 3-in-4 systematic sample of nonrespondents was selected to reach the 90 percent completion target. If a census tract had at least an 85 percent initial completion rate, then the NRFU sampling rate was 1-in-3.

The sample design included several special cases that should be noted. First, all nonrespondent addresses in blocks selected for the Integrated Coverage Measurement (ICM) sample were included in the NRFU sample. Second, all nonrespondent addresses that were added to the initial phase address list too late to be mailed a census form were also included in the NRFU sample. Finally, the minimum NRFU sample size in each census tract was two units, unless only one nonrespondent address existed, in which case that address was in the NRFU sample with certainty.

The UAA vacant sample design was similar. UAA vacant units were addresses identified as vacant by the United States Postal Service. The UAA vacant sample could be implemented as a check on the quality of the Postal Service's identification of these units. In addition, the UAA vacant sample was designed to obtain information on basic characteristics for this component of the housing unit inventory. The information obtained from the sampled units was used to estimate the characteristics of the UAA vacant units not in the sample. In each census tract, a 3-in-10 systematic sample of UAA vacants was selected for follow-up. As in the NRFU sample design, all UAA vacants in ICM blocks were included in the sample. The minimum UAA vacant sample size was two, unless only one UAA vacant existed in the census tract, in which case that unit was in the sample with certainty.

NONRESPONSE FOLLOW-UP & UNDELIVERABLE-AS-ADDRESSED VACANT ESTIMATION METHODS

Since not all nonrespondents and UAA vacants were included in field follow-up, an estimation method is required to account for the population and characteristics of the persons and housing units that exist at addresses not selected for either of the samples. In the Census 2000 Dress Rehearsal, this estimation method was the nearest-neighbor hot deck. In general, this method imputes data for each nonsampled address from the nearest sampled address, although several constraints are placed on this imputation:

- The donor sampled address must be in the same census tract and in the same sampling universe as the nonsampled address, meaning a NRFU sampled unit cannot donate to a nonsampled UAA vacant, and vice versa. This preserves the integrity of census tract estimates and the independence of NRFU and UAA vacant census estimates.
- A sampled unit cannot count in the final estimates more times than its weight, where the weight is the ratio of the sampling universe size to the sample size within a tract. This avoids the introduction of bias into the estimates by preventing a donor from being used too many times.
- When possible, donations occur within the same multi-unit structure, such as an apartment building. Much previous research has demonstrated that units in the same apartment building tend to be similar, therefore we can improve the accuracy of our estimates with this constraint.
- Nonrespondents and UAA vacants in ICM blocks cannot be donors because they are in the sample with certainty.

After estimation, each residential address on the initial phase address list has complete housing unit and person data. Therefore, total population and subpopulation estimates for any geographic area are simply sums over the housing units in that area.

SERVICE-BASED ENUMERATION

In the Sacramento site, Service-Based Enumeration (SBE) allowed individuals to be enumerated at shelters, soup kitchens, regularly scheduled mobile food vans, and targeted non-sheltered outdoor locations. There were no regularly scheduled food vans in the Sacramento site. An estimate of the number of individuals using these services was computed using multiplicity estimation. Multiplicity estimation is used when each individual can be linked to one or more enumeration units. In the census context, the enumeration unit is the SBE day. Service facility clients or individuals can be linked to one or more enumeration units or days by using information obtained through the “usage” question. Each individual is assigned a survey weight based on “usage frequency.” The survey weights are used to produce an estimate of this component of the total population. The error due to estimation is *not* factored into the standard error values given below. This component of the enumeration should *not* be interpreted as an estimate of the homeless population.

CONFIDENTIALITY OF THE DATA

To maintain the confidentiality required by law (Title 13, United States Code), the Census Bureau assures that the published data do not disclose information about specific individuals, households, or housing units. For the Census 2000 Dress Rehearsal, the primary means of assuring confidentiality consisted of exchanging the data for similar households. As a result, a small amount of uncertainty was introduced into the estimates of census characteristics. The process was controlled so that the basic structure of the data, the redistricting counts required by law, was preserved.

The data exchange was implemented by selecting a small sample of households from the internal files, and swapping some or all of their data with that of similar households not in the same block. Households in small blocks and households which were unique in their block with respect to the characteristics required for the redistricting counts were sampled at higher rates to provide greater protection against disclosure. The data exchange process was implemented in such a way that the quality and usefulness of the data were preserved.

CALCULATION OF STANDARD ERRORS

Type of Error

Variability arises in all samples, such as the NRFU/UAA sample implemented in the Census 2000 Dress Rehearsal. Estimates would differ if different persons and housing units had been selected for the NRFU/UAA sample. The standard error and the variance (the square of the standard error) are measures of the variation among the estimates from all possible samples and thus are measures of the precision with which an estimate from a particular sample approximates the average result among all possible samples.

In addition to the variability which arises from the sampling procedures, the estimates are subject to nonsampling error (human- and computer-related errors), which may be introduced during each of the many complex processes used to collect, process, and tabulate the data.

Nonsampling error may affect the data in two ways. Errors that are introduced randomly increase the variability of the data and should therefore be reflected in the standard error. Errors that tend to be consistent in one direction will make the data biased in that direction. For example, if respondents consistently tend to underreport their age, their age distribution will be skewed towards the lower age categories. Then the resulting estimate of persons by age category will be below the actual figures. Such biases are not reflected in the standard error.

Calculation of the Error

The calculation of the total variance was intended to capture the variance due to the NRFU/UAA sampling procedures. The NRFU/UAA variance was a measure of the variation created by the nearest neighbor hot-deck imputation used to estimate the nonsampled nonrespondents in the NRFU/UAA procedure. Replication methods were used to calculate the total variance.

To calculate the NRFU/UAA component of the total variance, 300 replicates of the data for a specified geographic area were created. In each of the 300 replicates, a prespecified subset of the NRFU-estimated households were replaced with the household information for the "second nearest neighbor", the household that would have been used according to the nearest neighbor imputation rules if the actual nearest neighbor had not been in the sample. The individuals in the households selected for each replicate were given a weight of one, and those not selected for the replicate were given a weight of zero. The weights were multiplied by each individual's coverage factor, and the new weights were then summed to give 300 estimates of the geographic area's population. The NRFU/UAA component of the variance was then computed from the 300 estimates.

For all direct estimates of the variance, the total variance was the sum of these two components. The standard error is just the square root of the variance.

At the site level, the standard error of total population is reported below, and is the direct estimate of the standard error calculated using the above process.

<u>Site</u>	<u>Total Population</u>	<u>Standard Error</u>
Sacramento, CA	377741	321

Because of the very large number of estimates at lower levels of geography, it was determined that it would not be feasible to provide tables listing the standard error of each published estimate. Instead, it was decided to publish parameters which will allow the user to calculate the standard error for any estimates smaller than the site total population.

The parameters were modeled for each of the redistricting (Public Law 94-171) data items, which are categorized by total population, race, age, and Hispanic or Latino, at levels of geography lower than the site. The parameters were computed using regression models to estimate the relationship between the estimated relative variance and the population estimate. The estimated relative variance is the variance of the estimate (the value of the direct variance, calculated from the two-step process above) divided by the estimate squared. The estimate of interest can be substituted into the generalized variance function equation using the computed parameters to calculate an estimate of the standard error.

To calculate a standard error of an estimate or an estimated proportion, the first step is to select the appropriate pair of **a** and **b** parameters from Table 1, based on an age/race/Hispanic or Latino combination that is most appropriate for the estimate. If more than one pair of **a** and **b** parameters is applicable to the estimate of interest, it is recommended that the user calculate the standard errors using each pair of **a** and **b** parameters and use the largest resulting standard error.

The standard error of an estimate, \hat{x} , is computed using

$$SE(\hat{x}) = \sqrt{\frac{a\hat{x}^2 + b\hat{x}}{1000}} \quad (1)$$

where \hat{x} is the estimated number of persons, and **a** and **b** are the estimated parameters taken from Table 1.

The standard error of the estimated proportion (*not* percentage) of persons, \hat{p} , is computed using

$$SE(\hat{p}) = \sqrt{\frac{1}{1000} \left(\frac{b}{\hat{y}} \right) (\hat{p}(1 - \hat{p}))} \quad (2)$$

where \hat{p} is \hat{x}/\hat{y} , \hat{y} is the base of the estimated proportion \hat{p} , and **b** is the estimated regression parameter taken from Table 1.

For any estimate which is the sum or difference of two or more given estimates, the standard error is the square root of the sums of the squared standard errors for the given estimates:

$$SE(\hat{x}_1 \pm \hat{x}_2 \pm \dots) = \sqrt{SE(\hat{x}_1)^2 + SE(\hat{x}_2)^2 + \dots} \quad (3)$$

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.

If the estimate of interest is a ratio of two values (\hat{x} and \hat{y}), but not a proportion, then the following formula should be used to approximate the standard errors:

$$SE(\hat{x} / \hat{y}) = \frac{\hat{x}}{\hat{y}} \sqrt{\frac{SE(\hat{x})^2}{\hat{x}^2} + \frac{SE(\hat{y})^2}{\hat{y}^2}} \quad (4)$$

This method will, however, overestimate (underestimate) the standard error if the two items in the ratio are highly positively (negatively) correlated.

For the standard error of the median of a characteristic, it is necessary to examine the distribution from which the median is derived, as the size of the base and the distribution itself affect the standard error. An approximate method is given here. As the first step, compute one-half of the number on which the median is based (refer to this result as $N/2$). Treat $N/2$ as if it were an ordinary estimate and obtain its standard error as instructed above. Compute the desired confidence interval about $N/2$. Starting with the lowest value of the characteristic, cumulate the frequencies in each category of the characteristic until the sum equals or first exceeds the lower limit of the confidence interval about $N/2$. By linear interpolation, obtain a value of the characteristic corresponding to this sum. This is the lower limit of the confidence interval of the median. In a similar manner, continue cumulating frequencies until the sum equals or exceeds the count in excess of the upper limit of the interval about $N/2$. Interpolate as before to obtain the upper limit of the confidence interval for the estimated median.

When interpolation is required in the upper open-ended interval of a distribution to obtain a confidence bound, use 1.5 times the lower limit of the open-ended confidence interval as the upper limit of the open-ended interval.

Confidence Intervals

A sample estimate and its estimated standard error may be used to construct confidence intervals about the estimate. These intervals are ranges that contain the average value of the estimated characteristic that results over all possible samples, with a known probability. For example, if all possible samples that could result under the Census 2000 Dress Rehearsal sample design were independently selected and surveyed under the same conditions, and if the estimate and its estimated standard error were calculated for each of these samples, then:

1. Approximately 68 percent of the intervals from one estimated standard error below the estimate to one estimated standard error above the estimate would contain the average result from all possible samples.
2. Approximately 90 percent of the intervals from 1.645 times the estimated standard error below the estimate to 1.645 times the estimated standard error above the estimate would contain the average result from all possible samples.
3. Approximately 95 percent of the intervals from 1.96 times the estimated standard error below the estimate to 1.96 times the estimated standard error above the estimate would contain the average result from all possible samples.

The intervals are referred to as 68 percent, 90 percent, and 95 percent confidence intervals, respectively.

The average value of the estimated characteristic that could be derived from all possible samples is or is not contained in any particular computed interval. Thus, we cannot make the statement that the actual value has a certain probability of falling between the limits of the calculated confidence interval. Rather, one can say with a specified probability or confidence that the calculated confidence interval includes the average estimate from all possible samples.

To calculate the bounds of a 90 percent confidence interval, use:

$$\begin{aligned} \text{Lower Bound of 90 percent CI} &= \text{estimate} - 1.645 * \text{SE}(\text{estimate}) \\ \text{Upper Bound of 90 percent CI} &= \text{estimate} + 1.645 * \text{SE}(\text{estimate}) \end{aligned} \quad (5)$$

To calculate the bounds of a 68 percent or a 95 percent confidence interval, substitute 1 or 1.96 (respectively) for 1.645 in (5).

Confidence intervals also may be constructed for the ratio, sum of, or difference between two sample figures. This is done by first computing the ratio, sum, or difference, then obtaining the standard error of the ratio, sum, or difference (using the formulas given earlier), and finally forming a confidence interval for this estimated ratio, sum, or difference as above. One can then say with specified confidence

that this interval includes the ratio, sum, or difference that would have been obtained by averaging the results from all possible samples.

These estimated standard errors do not include all portions of the variability due to nonsampling error that may be present in the data. The standard errors reflect the effect of simple response variance, but not the effect of correlated errors introduced by enumerators, coders, or other field or processing personnel. Thus, the standard errors calculated represent a lower bound of the total error. As a result, confidence intervals formed using these estimated standard errors may not meet the stated levels of confidence (for example, 68, 90, or 95 percent). Thus, some care must be exercised in the interpretation of the data in this data product based on the estimated standard errors.

A standard sampling theory text should be helpful if the user needs more information about confidence intervals and nonsampling errors.

Examples of Standard Error Computation

Suppose, for example, that a user is interested in the standard error of the population estimate of not Hispanic or Latino Asians in a census tract. One of the redistricting data items is "All Persons, Not Hispanic or Latino, Asian alone or in combination." Assume the associated **a** and **b** parameters are

$$a = .362 \qquad b = 2001.328$$

Assume the population estimate for this redistricting category is 370; then from equation (1) above,

$$\begin{aligned} SE(\hat{x}) &= \sqrt{\frac{a\hat{x}^2 + b\hat{x}}{1000}} \\ &\doteq \sqrt{\frac{(.362)(370)^2 + (2001.328)(370)}{1000}} \\ &\doteq \sqrt{\frac{49557.8 + 740491.36}{1000}} \\ &\doteq 28.108 \approx 28 \end{aligned}$$

Using equation (5), a 90 percent confidence interval for the total number of not Hispanic or Latino Asians in the tract would be

$$\begin{aligned} \text{lower bound} &= \hat{x} - 1.645 * SE(\hat{x}) \doteq 370 - 1.645 * 28.108 \doteq 324 \\ \text{upper bound} &= \hat{x} + 1.645 * SE(\hat{x}) \doteq 370 + 1.645 * 28.108 \doteq 416 \end{aligned}$$

90 Percent Confidence Interval is (324, 416)

Equation (1) can't be used directly to approximate the standard error of an estimate of not Hispanic or Latino Asians under age 18 because the **a** and **b** parameters are *only* published for "Total Population" and "18 and Over." To get the standard error of that estimate, equation (3) needs to be used. The estimate of not Hispanic or Latino Asians age 18 and over is 310, hence

$$\begin{aligned} \text{Pop Under 18} &= \text{Total Pop} - \text{Pop 18 and Over} \\ &= 370 - 310 \\ &= 60. \end{aligned}$$

From above, the standard error of the total number of not Hispanic or Latino Asians is 28.108. Using equation (1) again and letting the parameters for “Persons 18 and Over, Not Hispanic or Latino, Asian alone or in combination” be

$$a = .449 \quad b = 1665.251 \quad \hat{x} = 310,$$

the standard error of not Hispanic or Latino Asians 18 and over can be calculated to be 23.651. Using these values and equation (3),

$$\begin{aligned} SE(\text{under 18}) &= SE(\text{total} - 18 \text{ and over}) = \sqrt{SE(\text{total})^2 + SE(18 \text{ and over})^2} \\ &\doteq \sqrt{28.108^2 + 23.651^2} \\ &\doteq 36.735 \approx 37 \end{aligned}$$

To calculate the standard error of the proportion of not Hispanic or Latino Asians that are 18 & over, we can use equation (2):

$$\begin{aligned} \hat{x} &= \text{estimate of not Hispanic or Latino Asians 18 \& over} \\ &= 310 \end{aligned}$$

$$\begin{aligned} \hat{y} &= \text{estimate of not Hispanic or Latino Asians, all ages} \\ &= 370 \end{aligned}$$

$$\hat{p} = \frac{\hat{x}}{\hat{y}} = \frac{310}{370} \doteq 0.838$$

$$\begin{aligned} SE(\hat{p}) &= \sqrt{\frac{1}{1000} \left(\frac{b}{\hat{y}} \right) (\hat{p} (1-\hat{p}))} \\ &\doteq \sqrt{\frac{1}{1000} \left(\frac{2001.328}{370} \right) (0.838 (1-0.838))} \\ &\doteq 0.027 \end{aligned}$$

Note that the **b** parameter for “all persons” was used instead of the **b** parameter for “18 & over”. This was done in order to be conservative, since the **b** parameter for “all persons” gives a larger standard error than the **b** parameter for “18 & over”.

Standard Errors for Housing Unit Data

For housing units in Sacramento, the standard error due to NRFU estimation was also computed, using the same methods as above. To estimate standard errors for **all** housing unit data items, use the formulas above with the parameters:

$$a = -0.0010 \quad b = 211.2981$$

For example, to estimate the standard error of an estimate of 500 housing units,

$$\begin{aligned} SE(\hat{x}) &= \sqrt{\frac{a\hat{x}^2 + b\hat{x}}{1000}} \\ &= \sqrt{\frac{(-0.0010)(500)^2 + (211.2981)(500)}{1000}} \\ &= \sqrt{\frac{-250.00 + 105649.05}{1000}} \\ &= 10.266 \approx 10 \end{aligned}$$

NONSAMPLING ERROR

In any large-scale statistical operation, such as the Census 2000 Dress Rehearsal, human- and computer-related errors occur. These errors are commonly referred to as nonsampling errors. Such errors include not enumerating every household or every person in the population, not obtaining all required information from the respondents, obtaining incorrect or inconsistent information, and recording information incorrectly. In addition, errors can occur during the field review of the enumerators' work, during clerical handling of the census questionnaires, or during the electronic processing of the questionnaires.

While it is impossible to completely eliminate nonsampling error from an operation as large and complex as the decennial census, the Census Bureau attempts to control the sources of such error during the collection and processing operations. Described below are the primary sources of nonsampling error and the programs instituted to control this error in the Census 2000 Dress Rehearsal. The success of these programs, however, was contingent upon how well the instructions actually were carried out during the census. As part of the Census 2000 Dress Rehearsal evaluation program, both the effects of these programs and the amount of error remaining after their application will be evaluated.

Types of Nonsampling Error

Respondent and Enumerator Error

The person answering the questionnaire or responding to the questions posed by an enumerator could serve as a source of error, although the questions were phrased as clearly as possible based on precensus tests, and detailed instructions for completing the questionnaire were provided to each household. The enumerator may misinterpret or otherwise incorrectly record information given by a respondent, may fail to collect some of the information for a person or household, or may collect data for households that were not designated as part of the sample. To control these problems, the work of enumerators was monitored carefully. Field staff were prepared for their tasks by using standardized training packages that included hands-on experience in using census materials. A sample of the households interviewed by enumerators for nonresponse were reinterviewed to control for the possibility of fabricated data for households being submitted by enumerators.

Processing Error

The many phases involved in processing the census data represent potential sources for the introduction of nonsampling error. The processing of the census questionnaires includes the field review by the crew leader, check-in, and transmittal of completed questionnaires; the coding of write-in

responses; and the computer processing. The various field, coding and computer operations undergo a number of quality control checks to ensure their accurate application.

Nonresponse

Nonresponse to particular questions on the census questionnaire allows for the introduction of bias into the data, because the characteristics of the nonrespondents have not been observed and may differ from those reported by respondents. As a result, any imputation procedure using respondent data may not completely reflect this difference either at the elemental level (individual person or housing unit) or on the average. Some protection against the introduction of large biases is afforded by minimizing nonresponse. Characteristics for the nonresponses were imputed by using reported data for a person or housing unit with similar characteristics.

Reduction of Nonsampling Error

To reduce various types of nonsampling errors, a number of techniques were implemented during the planning, development of the mailing address list, data collection, and data processing activities. Quality assurance methods were used throughout the data collection and processing phases of the census to improve the quality of the data. A reinterview program was implemented to minimize the errors in the data collection phase for enumerator-filled questionnaires.

Several coverage improvement programs were implemented during the development of the census address list and census enumeration and processing to minimize undercoverage of the population and housing units. These programs were developed based on experience from the 1990 Decennial Census and results from the Census 2000 testing cycle.

- Be Counted questionnaires, unaddressed forms requesting all short form items plus a few additional items, were available in public locations for people who believed they were not otherwise counted.
- Forms in Spanish and other languages, in addition to the standard English forms, were mailed to targeted areas that were identified by local leaders as having a high concentration of non-English speakers.
- A well-publicized toll-free phone number was available to answer questions about the forms, and individuals' responses could be taken over the phone.
- An introductory letter was sent to all mailout/mailback addresses prior to the mailing of the initial form. A reminder postcard was also sent to these address, and two weeks after the initial mailing a replacement form was sent as well.
- Under the Local Update of Census Addresses (LUCA) program, local officials had the opportunity to address specific concerns on the Master Address File before mailings began.

Computer and clerical edits, as well as telephone follow-ups, also contributed to improved coverage.

Resolving Multiple Responses

With multiple census forms mailed to addresses and additional ways for people to initiate their enumeration as described above, it was very likely that some would respond more than once. A special computer process was implemented to control the extent of this type of nonsampling error by resolving situations where more than one form was received from an address. The process consisted of several steps. When at least one of these returns was from a "Be Counted" form or a telephone interview, a search of the other returns received for the address was initiated to see if these people had been included on another return. If not, the search for the "Be Counted" and telephone-interviewed people was expanded to include people enumerated on returns for other addresses. Whenever a match was found, one of the person records was marked to be ignored in subsequent processing.

Addresses that still had more than one viable return were analyzed further. Within each of these addresses, comparisons of the person records on each return were made against the person records on the other returns at the same address. Persons found to have been included on two different returns were marked as such, and one of the person records was ignored in subsequent processing.

EDITING OF UNACCEPTABLE DATA

The objective of the processing operation was to produce a set of data that describes the population as accurately and clearly as possible. In a major change from past practice, the information on Census 2000 Dress Rehearsal questionnaires generally was not edited during field data collection operations for consistency, completeness, and acceptability. The exception was enumerator-filled questionnaires, which were reviewed by census crew leaders and local office clerks for adherence to specified procedures.

Incomplete or inconsistent information on the questionnaires was assigned acceptable values using imputation procedures during the final automated edit of the collected data. Imputations, or computer assignments of acceptable codes in place of unacceptable entries or blanks, are needed most often when an entry for a given item is lacking or when the information reported for a person on that item is inconsistent with other information for that person. As in previous censuses, the general procedure for changing unacceptable entries was to assign an entry for a person that was consistent with entries for persons with similar characteristics. The assignment of acceptable codes in place of blanks or unacceptable entries enhances the usefulness of the data.

Another way in which corrections were made during the computer editing process was through substitution; that is, the assignment of a full set of characteristics for persons in a household. When there was an indication that a household was occupied by a specified number of people, but the questionnaire contained no information for the people within the household or the occupants were not listed on the questionnaire, a previously accepted household of the same size was selected as a substitute, and the full set of characteristics for the substitute was duplicated.

Table 1. Parameters for Calculating the Standard Error, Sacramento, CA Site, ICM Excluded

Public Law (Redistricting) Category	All Persons		Persons 18 & Over	
	a	b	a	b
All Persons	0.0001	229.0785	0.0004	122.5173
Hispanic or Latino	-0.0009	619.4874	-0.0003	392.2743
Not Hispanic or Latino	0.0001	246.5105	0.0003	144.5247
White alone	0.0001	237.1904	0.0003	157.4010
White alone or in combination with one or more other races	0.0001	242.1685	0.0003	160.9474
Not White alone or in combination with one or more other races	-0.0005	466.5452	0.0000	273.3888
Not Hispanic or Latino, White alone	0.0001	224.3829	0.0003	162.4045
Not Hispanic or Latino, White alone or in combination with one or more other races	0.0001	230.0142	0.0003	166.3705
Not Hispanic or Latino, not White alone or in combination with one or more other races	-0.0006	482.8838	0.0000	289.6487
Black or African American alone	-0.0008	590.9736	-0.0003	386.6432
Black or African American alone or in combination with one or more other races	-0.0008	577.5107	-0.0003	378.8815
Not Black or African American alone or in combination with one or more other races	0.0000	255.7094	0.0003	156.0407
Not Hispanic or Latino, Black or African American alone	-0.0009	594.7411	-0.0003	390.5985
Not Hispanic or Latino, Black or African American alone or in combination with one or more other races	-0.0008	574.4268	-0.0003	385.4716
Not Hispanic or Latino, not Black or African American alone or in combination with one or more other races	0.0001	250.0809	0.0003	168.3522
American Indian and Alaska Native alone	-0.0020	1021.5968	-0.0008	584.6321
American Indian and Alaska Native alone or in combination with one or more other races	-0.0016	860.4399	-0.0008	561.4789
Not American Indian and Alaska Native alone or in combination with one or more other races	0.0001	227.9121	0.0004	127.3287
Not Hispanic or Latino, American Indian and Alaska Native alone	-0.0022	1112.6835	-0.0009	627.5527
Not Hispanic or Latino, American Indian and Alaska Native alone or in combination with one or more other races	-0.0018	960.2204	-0.0009	597.1450
Not Hispanic or Latino, not American Indian and Alaska Native alone or in combination with one or more other races	0.0001	239.9522	0.0003	146.0640

Table 1. Parameters for Calculating the Standard Error, Sacramento, CA Site, ICM Excluded-Con.

Public Law (Redistricting) Category	All Persons		Persons 18 & Over	
	a	b	a	b
Asian alone	-0.0008	585.4804	-0.0004	418.0724
Asian alone or in combination with one or more other races	-0.0007	542.7382	-0.0003	396.8917
Not Asian alone or in combination with one or more other races	0.0000	268.7069	0.0003	143.9281
Not Hispanic or Latino, Asian alone	-0.0008	579.5957	-0.0004	414.8693
Not Hispanic or Latino, Asian alone or in combination with one or more other races	-0.0007	546.2896	-0.0003	391.0909
Not Hispanic or Latino, not Asian alone or in combination with one or more other races	0.0000	271.4765	0.0003	163.4782
Native Hawaiian and Other Pacific Islander alone	-0.0010	642.3602	-0.0006	506.0587
Native Hawaiian and Other Pacific Islander alone or in combination with one or more other races	-0.0015	831.8191	-0.0008	592.1613
Not Native Hawaiian and Other Pacific Islander alone or in combination with one or more other races	0.0001	228.2456	0.0004	124.0130
Not Hispanic or Latino, Native Hawaiian and Other Pacific Islander alone	-0.0008	585.2690	-0.0007	523.1406
Not Hispanic or Latino, Native Hawaiian and Other Pacific Islander alone or in combination with one or more other races	-0.0016	886.8287	-0.0009	611.0746
Not Hispanic or Latino, not Native Hawaiian and Other Pacific Islander alone or in combination with one or more other races	0.0001	243.7708	0.0003	146.0095
Some other race alone	-0.0015	822.2400	-0.0007	532.9078
Some other race alone or in combination with one or more other races	-0.0014	805.9626	-0.0006	512.0645
Not Some other race alone or in combination with one or more other races	0.0001	241.8350	0.0004	137.4346
Not Hispanic or Latino, Some other race alone	-0.0018	958.8340	-0.0013	772.3829
Not Hispanic or Latino, Some other race alone or in combination with one or more other races	-0.0014	804.8158	-0.0010	644.4247
Not Hispanic or Latino, not Some other race alone or in combination with one or more other races	0.0001	245.0777	0.0003	146.2386
One race	0.0001	229.4379	0.0004	126.5758
Two or more races	-0.0006	507.1160	-0.0004	422.0020
Not Hispanic or Latino, One race	0.0001	240.7561	0.0003	146.6970
Not Hispanic or Latino, Two or more races	-0.0006	513.8186	-0.0005	448.7288

Table 1. Parameters for Calculating the Standard Error, Sacramento, CA Site, ICM Excluded-Con.

NOTES

1. The standard error of an estimate, \hat{x} , is computed using

$$SE(\hat{x}) = \sqrt{\frac{a\hat{x}^2 + b\hat{x}}{1000}}$$

where \hat{x} is the estimated number of persons, and **a** and **b** are the estimated parameters from Table 1.

The standard error of the estimated proportion (*not* percentage) of persons, \hat{p} , is computed using

$$SE(\hat{p}) = \sqrt{\frac{1}{1000} \left(\frac{b}{\hat{y}} \right) (\hat{p} (1 - \hat{p}))}$$

where \hat{p} is \hat{x}/\hat{y} , \hat{y} is the base of the estimated proportion \hat{p} , and **b** is the estimated regression parameter taken from Table 1.

2. By the nature of the formulas used to compute them, the estimated standard errors of totals and proportions approach zero for small estimated totals, and for proportions near zero or near one. However, because of the nature of sampling, even estimated totals or proportions of zero have some sampling error, so an estimated standard error of zero is not appropriate. Therefore, use caution because the estimated standard errors calculated using these formulas may be inaccurate for very small estimated totals and for estimated proportions near zero or one.

MENOMINEE

INTRODUCTION

The Census 2000 Dress Rehearsal was the last step of the Census 2000 testing cycle. It was conducted in 1998 in Menominee County, Wisconsin, including the Menominee American Indian Reservation; the city of Sacramento, California; and 11 counties in South Carolina, including the city of Columbia and the town of Irmo. These sites were selected because of their demographic and geographic characteristics to reflect some of the expected Census 2000 conditions and environments. The census-taking methodology employed in each site had a different mix of operational and statistical procedures. In the Sacramento and Menominee sites, statistical sampling and estimation techniques were used, aiming to improve the accuracy of the population count. Census results from these two sites are subject to sampling and nonsampling errors. The sampling and estimation techniques and their associated sampling errors are described below. The South Carolina site used traditional census methods only; census results from this site are not subject to sampling errors, but are subject to nonsampling errors. Comparisons of the results between the sites should *not* be made because a different method was used to arrive at the final results for each site.

MASTER ADDRESS FILE DEVELOPMENT

In order for the Census 2000 Dress Rehearsal to be as accurate, complete, and cost effective as possible, the address list, which serves as the basis for control for the census, must be as accurate and complete as possible. If an address is not on the list, then its residents are less likely to be counted. The Master Address File (MAF) building process for the Census 2000 Dress Rehearsal in Menominee involved a series of operations that built on each other and ultimately resulted in the address list used to conduct the census.

The initial MAF was created by the address listing operation where field staff went door-to-door to identify the mailing address and physical location of housing units. Following the initial creation of the list, the Local Update of Census Addresses (LUCA) operation was conducted where local and tribal governments were given the opportunity to review the census address list for accuracy and completeness before the delivery of questionnaires. They had the opportunity to provide information about additions, deletes, and corrections to the list of addresses in their jurisdictions, and to correct geocoding errors. Field verification identified which of the LUCA addresses were retained. Field staff then conducted the update/leave operation just prior to census day. In this operation, enumerators visited their assigned areas and canvassed each block, delivering questionnaires to every housing unit they could find. They matched what was found on the ground to the list of addresses compiled up to that point. They updated the register by adding new addresses, deleting addresses they could not locate, and correcting addresses, if necessary.

The Be Counted program and the Telephone Questionnaire Assistance (TQA) program yielded some additional housing units that were not previously listed on the MAF. These programs offered alternative options for people to be included in the census if they did not think they were otherwise enumerated. The Be Counted program gave residents access to questionnaires in their local community. People also contacted the Census Bureau through TQA and requested that a form be mailed to them.

SERVICE-BASED ENUMERATION

In the Menominee site, Service-Based Enumeration (SBE) allowed individuals to be enumerated at shelters, soup kitchens, regularly scheduled mobile food vans, and targeted non-sheltered outdoor locations. Targeted non-sheltered outdoor locations were the only Service-Based Enumeration locations done in the Menominee site. No estimation was done on the count of individuals enumerated in this program. This component of the enumeration should *not* be interpreted as an estimate of the homeless population.

EDITING OF UNACCEPTABLE DATA

The objective of the processing operation was to produce a set of data that describes the population as accurately and clearly as possible. In a major change from past practice, the information on Census 2000 Dress Rehearsal questionnaires generally was not edited during field data collection operations for consistency, completeness, and acceptability. The exception was enumerator-filled questionnaires, which were reviewed by census crew leaders and local office clerks for adherence to specified procedures.

Incomplete or inconsistent information on the questionnaires was assigned acceptable values using imputation procedures during the final automated edit of the collected data. Imputations, or computer assignments of acceptable codes in place of unacceptable entries or blanks, are needed most often when an entry for a given item is lacking or when the information reported for a person on that item is inconsistent with other information for that person. As in previous censuses, the general procedure for changing unacceptable entries was to assign an entry for a person that was consistent with entries for persons with similar characteristics. The assignment of acceptable codes in place of blanks or unacceptable entries enhances the usefulness of the data.

Another way in which corrections were made during the computer editing process was through substitution; that is, the assignment of a full set of characteristics for persons in a household. When there was an indication that a household was occupied by a specified number of people, but the questionnaire contained no information for the people within the household or the occupants were not listed on the questionnaire, a previously accepted household of the same size was selected as a substitute, and the full set of characteristics for the substitute was duplicated.

CONFIDENTIALITY OF THE DATA

To maintain the confidentiality required by law (Title 13, United States Code), the Census Bureau assures that the published data do not disclose information about specific individuals, households, or housing units. For the Census 2000 Dress Rehearsal, the primary means of assuring confidentiality consisted of exchanging the data for similar households. As a result, a small amount of uncertainty was introduced into the estimates of census characteristics. The process was controlled so that the basic structure of the data, the redistricting counts required by law, was preserved.

The data exchange was implemented by selecting a small sample of households from the internal files, and swapping some or all of their data with that of similar households not in the same block. Households in small blocks and households which were unique in their block with respect to the characteristics required for the redistricting counts were sampled at higher rates to provide greater protection against disclosure. The data exchange process was implemented in such a way that the quality and usefulness of the data were preserved.

NONSAMPLING ERROR

In any large-scale statistical operation, such as the Census 2000 Dress Rehearsal, human- and computer-related errors occur. These errors are commonly referred to as nonsampling errors. Such errors include not enumerating every household or every person in the population, not obtaining all required information from the respondents, obtaining incorrect or inconsistent information, and recording information incorrectly. In addition, errors can occur during the field review of the enumerators' work, during clerical handling of the census questionnaires, or during the electronic processing of the questionnaires.

While it is impossible to completely eliminate nonsampling error from an operation as large and complex as the decennial census, the Census Bureau attempts to control the sources of such error during the collection and processing operations. Described below are the primary sources of nonsampling error and the programs instituted to control this error in the Census 2000 Dress Rehearsal. The success of these programs, however, was contingent upon how well the instructions actually were

carried out during the census. As part of the Census 2000 Dress Rehearsal evaluation program, both the effects of these programs and the amount of error remaining after their application will be evaluated.

Types of Nonsampling Error

Respondent and Enumerator Error

The person answering the questionnaire or responding to the questions posed by an enumerator could serve as a source of error, although the questions were phrased as clearly as possible based on precensus tests, and detailed instructions for completing the questionnaire were provided to each household. The enumerator may misinterpret or otherwise incorrectly record information given by a respondent, may fail to collect some of the information for a person or household, or may collect data for households that were not designated as part of the sample. To control these problems, the work of enumerators was monitored carefully. Field staff were prepared for their tasks by using standardized training packages that included hands-on experience in using census materials. A sample of the households interviewed by enumerators for nonresponse were reinterviewed to control for the possibility of fabricated data for households being submitted by enumerators.

Processing Error

The many phases involved in processing the census data represent potential sources for the introduction of nonsampling error. The processing of the census questionnaires includes the field review by the crew leader, check-in, and transmittal of completed questionnaires; the coding of write-in responses; and the computer processing. The various field, coding and computer operations undergo a number of quality control checks to ensure their accurate application.

Nonresponse

Nonresponse to particular questions on the census questionnaire allows for the introduction of bias into the data, because the characteristics of the nonrespondents have not been observed and may differ from those reported by respondents. As a result, any imputation procedure using respondent data may not completely reflect this difference either at the elemental level (individual person or housing unit) or on the average. Some protection against the introduction of large biases is afforded by minimizing nonresponse. Characteristics for the nonresponses were imputed by using reported data for a person or housing unit with similar characteristics.

Reduction of Nonsampling Error

To reduce various types of nonsampling errors, a number of techniques were implemented during the planning, development of the mailing address list, data collection, and data processing activities. Quality assurance methods were used throughout the data collection and processing phases of the census to improve the quality of the data. A reinterview program was implemented to minimize the errors in the data collection phase for enumerator-filled questionnaires.

Several coverage improvement programs were implemented during the development of the census address list and census enumeration and processing to minimize undercoverage of the population and housing units. These programs were developed based on experience from the 1990 Decennial Census and results from the Census 2000 testing cycle.

- Be Counted questionnaires, unaddressed forms requesting all short form items plus a few additional items, were available in public locations for people who believed they were not otherwise counted.

- Forms in Spanish and other languages, in addition to the standard English forms, were mailed to targeted areas that were identified by local leaders as having a high concentration of non-English speakers.
- A well-publicized toll-free phone number was available to answer questions about the forms, and individuals' responses could be taken over the phone.
- An introductory letter was sent to all mailout/mailback addresses prior to the mailing of the initial form. A reminder postcard was also sent to these address, and two weeks after the initial mailing a replacement form was sent as well.
- Under the Local Update of Census Addresses (LUCA) program, local officials had the opportunity to address specific concerns on the Master Address File before mailings began.

Computer and clerical edits, as well as telephone follow-ups, also contributed to improved coverage.

Resolving Multiple Responses

With multiple census forms mailed to addresses and additional ways for people to initiate their enumeration as described above, it was very likely that some would respond more than once. A special computer process was implemented to control the extent of this type of nonsampling error by resolving situations where more than one form was received from an address. The process consisted of several steps. When at least one of these returns was from a "Be Counted" form or a telephone interview, a search of the other returns received for the address was initiated to see if these people had been included on another return. If not, the search for the "Be Counted" and telephone-interviewed people was expanded to include people enumerated on returns for other addresses. Whenever a match was found, one of the person records was marked to be ignored in subsequent processing.

Addresses that still had more than one viable return were analyzed further. Within each of these addresses, comparisons of the person records on each return were made against the person records on the other returns at the same address. Persons found to have been included on two different returns were marked as such, and one of the person records was ignored in subsequent processing.