LONG FORM DESIGN FOR THE U.S. CENSUS 2000 DRESS REHEARSAL AND PLANS FOR CENSUS 2000

Philip M. Gbur, Steven P. Hefter, and Lisa D. Fairchild, U.S. Bureau of the Census
Philip M. Gbur, 7975 Central Park Circle, Alexandria, VA 22309

Key Words: Systematic Sample, Weighting, Raking, Variance Estimation, Successive Differences, Iterative Proportional Fitting

I. Introduction and Background

The U.S. Census Bureau is conducting the Census 2000 Dress Rehearsal (DR) in 1998 in Sacramento, CA; Menominee, WI; and Columbia, SC and surrounding counties. We will use traditional enumeration methods in Columbia, SC with a Post Enumeration Survey (PES) to evaluate coverage. The Census 2000 sampling and estimation plan will be used in Sacramento, CA; that is, sampling for nonresponse followup, vacant undeliverable as addressed followup and integrated coverage measurement (ICM). A modified Census 2000 sampling and estimation plan will be used in Menonimee, WI; we will use sampling for ICM only. The Census 2000 plan provides a one-number census designed to reduce cost and the undercount, especially the differential undercount among racial, ethnic and socioeconomic groups that has been documented in every census since 1940.

A systematic sample of addresses in the dress rehearsal sites will receive a long form questionnaire which collects detailed socioeconomic and demographic characteristics. After the data is collected from these questionnaires it will be weighted using the iterative proportional fitting methodology, also known as raking. Variances will be estimated for a subset of resulting long form estimates using a successive difference replication methodology and generalized for use with all estimates. The following sections present a description of the sample design and the current plans for weighting and variance estimation of the long form questionnaire data for the Census 2000 Dress Rehearsal. We will also describe the components which were changed from 1990 and those which will be examined, and therefore may be revised, for Census 2000. In general, the DR design and the plans for Census 2000 are similar to 1990, but revisions have been introduced to improve selected aspects of the 1990 process and to allow flexibility in supporting a census with or without sampling.

II. Sample Design

A. Overview

The addresses that were to receive the long form questionnaire were chosen by taking a systematic, variable rate sample of addresses from the Dress Rehearsal Decennial Master Address File. The ultimate goal is to sample roughly 17 percent of all addresses nationwide. This is achieved through appropriate application of the selected sampling rates to each governmental unit - GU (such as a city, county, or school district) - or census tract. The rates used are 1-in-2, 1-in-4, 1-in-6 and 1-in-8, and are applied based on a GU's or tract's predetermined measure of size.

B. Dress Rehearsal Design

Application of the long form sampling rates for the dress rehearsal was based on the 1990 census tract delineation, as updated census tracts were not yet available.

The sampling strata cutoffs were chosen based on an analysis of the range of coefficients of variation (CVs) obtained from simulation research. The sampling rates were applied at the block level. For blocks that fell into more than one sampling stratum, we applied the higher sampling rate.

The sampling strata and their cutoff points were:

- 1-in-2 for governmental units < 800 housing units;
- 1-in-4 for governmental units between 800 and 1200 housing units; and if not 1-in-2 or 1-in-4; then
- 1-in-6 for census tracts < 2000 housing units; and
- 1-in-8 for census tracts $\geq$ 2000 housing units.

The cutoff points were selected to result in an expected coefficient of variation (CV) of about ten percent on an estimate of ten percent. To simplify calculations, the double sampling formula of Hansen, Hurwitz, and Madow [3] was used to estimate the CVs using the sampling for the long form questionnaire and for nonresponse followup as the two sampling levels. Thus, the CVs are expected to be lower for the DR sites where there is no sampling for nonresponse.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

To account for a projected 10 percent sample loss, we adjusted the sampling rates accordingly. Sample loss occurs when long forms are returned with only short form data. The adjustment changed the sampling rates to 1-in-1.8, 1-in-3.6, 1-in-5.4 and 1-in-7.2 respectively.

The American Community Survey (ACS) is a rolling monthly survey designed to capture data similar in content to long form data. The goal is to have a representative sample for the entire country every five years as a replacement for the long form in future decennial censuses. We modified the sampling rates in the dress rehearsal areas of South Carolina that overlap with the ACS planned for 1998. To reduce respondent burden, we decided to exclude all ACS first stage sample addresses from the sampling frame. This is approximately 17.5% of the addresses in Kershaw and Richland counties. Since we reduced the universe size, and we designed the long form sample to keep the probabilities of selection equal within strata, the sampling rates for these two counties only were adjusted. The resulting rates are: 1-in-1.485, 1-in-2.970, 1-in-4.455, and 1-in-5.940, respectively.

The following rates were used for certain data collections and special populations:

a. Update/leave areas were sampled according to the sampling rate of the blocks in the assignment area (AA). When an AA included more than one sampling stratum, the higher of the rates was used for the entire AA.

b. Group Quarters were sampled at a 1-in-6 rate.

c. Service Sites (such as shelters and soup kitchens) were sampled at a 1-in-6 rate.

d. The Telephone Questionnaire Assistance (TQA) operation took incoming calls for requests for mailing questionnaires and for interviews. Individuals who telephoned to request a questionnaire or to provide an interview received either their designated form type or were subject to a 1-in-6 sampling rate, depending upon whether they had their census identification number.

e. Addresses added to the mailout universe after the initial sampling were sampled according to the sampling rate of the stratum that the addresses' block was in.

The results of the sampling for the Census 2000 DR are given by site in the table below.

Summary of Dress Rehearsal
Designated Long Form Sample

Site	Universe	Addresses in Sample	Percent in Sample
California	173,736	28,154	16.21
South Carolina	290,289	47,483	16.36
Wisconsin	2,060	292	14.17
Total	**466,085**	**75,929**	**16.30**

C. Changes from 1990

There are three major differences in the long form sample design between the 1990 Census and Census 2000. First, the sampling rate cutoffs for the long form will be based solely on address/housing unit counts, not on a mix of population and housing unit counts as in 1990. Ideally, the cutoffs would be based on population counts but reasonable counts are not available for all areas at the level of geography at which we sample. Therefore, housing unit counts are being used for all areas to maintain a consistency for all geographic areas.

Four sampling rates will be used in 2000. In 1990 three rates were used, 1-in-2, 1-in-6 and 1-in-8. A 1-in-4 sampling rate was added for Census 2000 DR and will be used in Census 2000. This rate is being added to achieve more reliable estimates for GUs that would have been sampled at 1-in-6 using the 1990 rates, and to reduce respondent burden in the medium sized GUs that would have been sampled at 1-in-2.

For sampling, we will treat school districts as governmental units in 2000. In 1990, school districts were not considered in the sampling design. Since school districts may receive funding as separate entities, this is expected to produce more reliable estimates for these areas.

D. Plans for 2000

The basic long form sample for Census 2000 will be an approximate 17 percent systematic, variable rate sample of addresses from the Decennial Master Address File. The four sampling rates used in the DR are expected to be used in Census 2000, although, we will research whether to use expected HU or population counts to determine the sampling rates. One of the major reasons for using HUs in the DR is that it allowed us to use one measure for all sampling levels (governmental units and tracts) since population projections are available only for some governmental units. However, there is some concern that the use of housing units may result in inappropriately low sampling rates for areas with a relatively high HU vacancy rate - such as resort or vacation areas with many seasonal vacants. In addition, regardless of whether HUs or population counts, or a combination of the two, are used, we will need some research to determine whether the cutoffs for the sampling strata used for the DR should be modified.

The sampling rate of 1-in-6 was used for all group quarters (GQs) enumeration for the Census 2000 Dress Rehearsal to simplify implementation. We will explore the feasibility of using variable rate sampling for GQs in 2000 based on the sampling rate of the area in which the GQ is located. Individuals who telephone to provide an interview via the TQA operation, and do not have their census identification number, will be interviewed with a short form questionnaire.

III. Weighting

A. Overview

As in every census since 1940, when we introduced content sampling, the iterative proportional fitting methodology will be used in the Census 2000 Dress Rehearsal to estimate the characteristics of the entire country based on the long form sample. We carry out the iterative proportional fitting methodology, also known as raking, within weighting areas.

B. Dress Rehearsal Design

Weighting areas, the geographic level at which we conduct the weighting, are formed within counties. They are generally in close agreement with census tabulation areas. Tabulation areas are required to have a minimum of 400 sampled persons to form a weighting area. If necessary, small counties with fewer than the prescribed number of cases will be allowed to stand alone as weighting areas.

To ensure that we have a basic minimal sample within the weighting areas, augmentation of the long form sample, using a set of predetermined rules, may occur. This is done to attain a minimum observed sampling rate within each area, reducing the associated variance. Long form data will be imputed from short forms for sample augmentation. Augmentation of sample counts will use the smallest number of addresses needed to reach the desired minimum observed sampling rate within each weighting area. After augmentation, weighting proceeds separately for people, occupied housing units, and vacant housing units.

For each sample unit we set an initial weight equal to the inverse of the observed sampling rate. We then carry out the iterative proportional fitting methodology, also known as raking. Raking is performed in several stages.

For person weighting, for each weighting area, we will form a four-dimensional matrix using household type (such as family with own children and family without own children), sampling rate, whether the person is a householder, race, Hispanic origin, age, and sex. For occupied HUs, we will use three dimensions: household type by size; race and Hispanic origin of the householder by tenure; and sampling rate. The weighted record counts within a cross-classification are summed to produce the interior cell counts. These weighted cell counts are called initially inflated counts. At this point we sum the interior cells to obtain the initially inflated sample marginal totals for each category classification within the matrix.

Before raking, we test the matrices against predefined collapsing criteria. If the uninflated sample category classification totals are not "large" enough, or the ratio of the 100 percent data category classification total to the initially inflated sample category classification total fails a collapsing test, then we will combine classifications with other classifications within the same category. The plan for the dress rehearsal will be nearly identical to that of 1990, with a slight variation to account for the added 1-in-4 sampling rate and the race collapsing procedure.

Raking is an iterative proportional adjustment of the cross-classified cell counts. The interior cell counts within a classification are multiplied by the ratio of the control, post-nonresponse followup, total (for that classification) to the initially inflated sample total (for that classification). An iteration of the raking consists of one stage of adjustment for each dimension. Each stage adjusts all interior cell counts by the appropriate cell ratio. In the dress rehearsal, the raking will continue until the weighted sample marginal is within 0.1 percent of the post-nonresponse followup marginal or upon reaching a total of five iterations, whichever is reached first.

As part of the dual track DR design, coverage factors will be produced for the Sacramento and Menominee sites and used to produce the one number census results. We will apply the coverage factors (which will be calculated for poststrata based on age, sex, race, and Hispanic origin) to the long form weights resulting from the raking procedure to reduce coverage error. Similarly, coverage factors will be produced for housing units for use in long form weighting.

C. Changes from 1990

Block code assignment methodology has changed from 1990. Thus, while in 1990 we started with collection blocks to form initial weighting areas and then switched to tabulation blocks, for 2000 we will only use tabulation block definitions.

Due to changes in the information collected on the short form census questionnaire from 1990 to 2000, some changes were required in the characteristics used in defining the matrices used for raking. In addition, we made some revisions in the collapsing criteria for the raking matrices and for deciding when to stop the raking process. In 1990, the raking was stopped after two iterations. For the Census 2000 DR we will use a stopping criteria of weighted estimates being within a tolerance or five iterations. It is expected that these revisions may result in more consistent estimates between long form estimates and the census counts.

Coverage factors were not used in the long form weighting for 1990. The use of coverage factors for housing units may result in greater discrepancies between long form estimates of housing units and the census counts.

D. Plans for 2000

We will review the DR results to evaluate the effects, if any, of the changes in the stopping criteria for the raking. Based on this evaluation, we may revise the stopping criteria, if necessary.

Contingent upon resource availability (time and staff), alternatives will be examined prior to making a final decision for the weighting methodology for Census 2000. The primary alternatives are a generalized least squares estimator (similar to the methodology in use by Statistics Canada) and a quadratic programming methodology developed by staff at the Census Bureau. We will evaluate the alternatives in terms of their operational feasibility, and their potential effect on long form estimates and variance estimates.

Whether we use the coverage factors for Census 2000 will be dependent upon whether the Bureau ultimately implements a sampling census methodology. If the coverage factors are used in production of the official census counts, then they will be used in long form weighting.

IV. Variance Estimation

A. Overview

The long form sample can be the basis of a myriad of estimates calculated at many geographic levels. The Census Bureau has a commitment to provide estimates of sampling error for all estimates and to minimize burden on data users by not overwhelming them with volumes of error estimates. Thus, for the Census 2000 Dress Rehearsal, we will use a successive difference replication (SDR) methodology to calculate direct variances for a subset of estimates. We will then generalize these variances to produce design factors which may be used by data users for calculating sampling error estimates for long form estimates. Extensive research was done prior to the 1980 Census on alternative variance estimators [1] and we selected the SDR methodology based on these results and experience with the SDR on other projects within the Census Bureau.

B. Dress Rehearsal Design

1. Direct Variances

For the Census 2000 Dress Rehearsal, we will use the SDR methodology to calculate the direct variances. (See [2] for details on the SDR methodology.) For simplicity of implementation we want to use one variance estimation methodology for both dress rehearsal tracks (sampling and no sampling). The successive difference methodology can be specified for both tracks relatively easily and it holds promise for more accurately reflecting the variance of the estimates.

For long form variance estimation, 100 replicates will be formed. Sample units are assigned overlapping pairs of row numbers from a Hadamard matrix of order 100. (See [5] for a description of Hadamard matrices.) These row assignments are used in the calculation of replicate factors, as follows:

$$f_{ir} = 1 + (2)^{-\frac{3}{2}} a_{i+1,r} - (2)^{-\frac{3}{2}} a_{i+2,r}$$

Where:

f_{ir} is the replicate factor for the i-th sample unit and the r-th replicate;
$i = 1, \dots n$; $r = 1, \dots , 100$; and

$a_{i+1,r}$, $a_{i+2,r}$ is the Hadamard matrix value (+1 or -1) which corresponds to the i+1-th or i+2-th row and r-th column for the i-th sample unit.

Replicate factors are multiplied by the initial weights to produce replicate initial weights. These weights are raked and coverage factors are applied where applicable. For Sacramento, the replicate weighting process incorporates the variance of the control totals resulting from sampling for nonresponse by making draws from the distributions for those totals. For Sacramento and Menominee, the variance of the coverage factors applied to the weights will be incorporated by making draws from their distributions. Once replicate final weights are produced, then the SDR method estimates the variance of the estimator through the formula below:

$$Var(\hat{X}) = \frac{4}{100}\Sigma_{r=1}^{100}(\hat{X}_r - \hat{X})^2$$

Where:

$\hat{X}_r$ is the weighted total of the r-th replicate; $r = 1$, ..., 100; and

$\hat{X}$ is the weighted total of the original sample.

The SDR methodology has several expected advantages which caused us to select it for use in the DR. Primarily, it better reflects the systematic nature of the sampling. In addition, it has been researched extensively and is currently being used for the ACS and the Current Population Survey. However, as with the implementation of any new approach, there may be risks. The SDR has not been researched with respect to the specific sample design of the long form and data users may not be familiar with it. In addition, we have not investigated what changes could be expected relative to the 1990 methodology.

2. Generalized Variances

The generalized variance methodology is the same as that used for the 1990 census. It begins with the calculation of design factors. Design factors are the ratio of the standard error, S_{SDR}, from the direct variance estimate for the complex design over the standard error estimate, S_{SRS}, assuming a 1-in-6 simple random sample. The design factor, DF, at the weighting area level is calculated as:

$$DF = S_{SDR} / S_{SRS} .$$

A DF will be calculated for selected data items within each weighting area.

Due to space limitations, the design factors will be made available across four percent-in-sample categories or intervals. The percent-in-sample is defined at the weighting area level to be the percent observed unweighted sample count of persons out of the 100% count of persons, which is equal to the final weighting area observed sampling rate multiplied by 100.

Data items are arranged into groups and subgroups based on characteristic. Breaks for the percent-in-sample categories will be determined by graphing average group design factors versus percent-in-sample values for the three DR sites. The average group design factor is the simple average of the design factors of all the data items within a group at the weighting area level.

For each of the three DR sites, generalized design factors for each group and subgroup will be calculated over each of the percent-in-sample intervals. To compare the metropolitan statistical areas (MSAs) with the nonMSAs, these levels of geography will also be taken into account. Only the SC site will have both MSA and nonMSAs. Thus, for the SC site only, there will be four percent-in-sample intervals and three geographic levels (site, MSA, nonMSA) composing a total of twelve Geographic and Percent-in-Sample Classes (GPSCs). The generalized design factor for each group within a given GPSC is a weighted average design factor.

Data item groups will be examined for homogeneity of variance. Data item design factors which are determined to be outliers may be excluded from the final results.

C. Changes from 1990

The random groups methodology was implemented for the 1990 Census long form variances and, as described above, we will use the SDR methodology for the Census 2000 DR direct variances. The design factor methodology was used in 1990 for the generalized variances.

In preparing for the two-track Dress Rehearsal, variance estimation options must be chosen which can take into account sampling for nonresponse. For the 1990 Census, sampling for nonresponse was also considered although we did not implement it. Research was done to investigate whether or not it would be reasonable to treat the estimated short form marginal totals as if they were counts rather than estimates--that is, ignore the variation in these estimated marginals. The results of the research showed that the variance of these estimated marginals is potentially too large to ignore. See [4] for further details.

D. Plans for 2000

Three options are being considered at this time. They are: (1) a random groups (RG) variance estimator, as used in 1990; (2) a Jackknife (JK) variance estimator; and (3) the SDR approach. The SDR estimator would be carried out in a similar manner as described above. We describe the JK and RG methods below.

The JK estimator is based on the sum of the squared differences of pseudo-subsample estimates from the average of these pseudo-subsample estimates. Initially, g subsamples are systematically selected from the full sample. The i-th pseudo-subsample is composed of the g-1 subsamples left when the i-th subsample is left out. Thus, g pseudo-subsamples are created.

The procedure for the RG estimator starts with systematically subdividing the weighting area samples into g subsamples. For the 1990 Census, g was set to 25. The calculation of the estimated variance for a particular estimate may then proceed through this formula:

$$Var(\hat{X}) = (1-f_o)\,\frac{25}{24}\sum_{i=1}^{25}\left(\hat{X}_i - \frac{\hat{X}}{25}\right)^2$$

Where:

$\hat{X}_i$ is the weighted total of the characteristic in a weighting area based on the records assigned to the i-th subsample;

$\hat{X}$ is the sum of the 25 values of $\hat{X}_i$ -- that is, $\hat{X} = \sum_{i=1}^{25} \hat{X}_i$; and

f_o is the observed sampling fraction in the weighting area - in terms of persons or housing units.

Contingent upon resource availability (time and staff) these alternatives will be examined prior to making a final decision for Census 2000. We will evaluate the alternatives in terms of their operational feasibility, and their potential effect on long form variance estimates.

REFERENCES

[1] Fan, Milton C., et. al., "1980 Census Variance Estimation Procedure," Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 176-181, 1981.

[2] Fay, Robert E. and George F. Train, "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Government Statistics Section of the American Statistical Association, pp. 154-159, 1995.

[3] Hansen, Morris H., William N. Hurwitz, and William G. Madow (1953), Sample Survey Methods and Theory Volume 1 Methods and Applications, John Wiley & Sons, Inc., New York, New York, p. 469.

[4] U.S. Census Bureau, "Variance Estimation in the Event of Sampling for Nonresponse," internal memorandum for Thompson from Griffin, STSD 1990 Decennial Census Memorandum Series #Z-35, May 29, 1987.

[5] Wolter, Kirk M. (1985), Introduction to Variance Estimation, New York: Springer-Verlag.