SAMPLE DESIGN FOR THE 1990 DECENNIAL CENSUS

Alfredo Navarro and Richard A. Griffin
U.S. Bureau of the Census

1. INTRODUCTION

Over the past fifty years, the Bureau of the Census has entirely transformed the decennial census from a 100 percent data collection activity into an operation which collects the bulk of census data on a sample basis. The agency resorted to sampling in order to satisfy the increasing needs for more accurate detailed information about the population and housing units. The basic sampling procedure has been much the same in each census since 1960. Some basic information, referred to as 100 percent data, has been collected for each housing unit and person. A sample of housing units has been selected to receive a questionnaire that, in addition to the basic items, has much more detailed questions on topics such as income, occupation and place of work. In 1960, every fourth household was selected to receive a "long form" or sample questionnaire. Two sample questionnaires were used in 1960 so that sample data were collected at three different rates - 25, 20, and 5 percent - depending on whether data were collected on one or both of the sample questionnaires. In 1970, two sample questionnaires were also used, but the overall sampling rate was lowered to 1-in-5. Prior to 1970, addresses were listed and enumerated by census enumerators through the use of maps and address listing registers. About 60 percent of the population in 1970 was enumerated using the mailout/mailback enumeration technique. In 1980, the overall sampling rate remained about 1-in-5, but small governmental units (less than 2,500 population) were sampled at 1-in-2 with the remainder of the United States sampled at 1-in-6. The mailout/mailback procedure was utilized in the 1980 census to enumerate about 95 percent of the population. The 1990 census will employ mail census techniques to enumerate about 95 percent of the population.

The goals in developing the sample design are examined in this paper as well as the process of how we arrived at them. The issue of the national sample size is discussed. Two sampling plans are compared and the advantages and disadvantages of each plan are discussed. Finally, the 1990 census sample design is presented.

2. GOALS FOR SAMPLE DESIGN

The data user involvement was instrumental in the development of the 1990 Census sampling plan. The Census Bureau received a great deal of input through our advisory committees. A highly dynamic communication process was developed between the Census Bureau's staff and members of the user community. The agency received a significant response to virtually all areas and issues considered for the final design. As a result of this communication, the following general goals were followed in developing the 1990 census sample design.

a. To provide reliable data for plkanning purposes especially for small governmental units. The majority of these small governmental units do not have resources to conduct their own surveys and rely on census data for the planning of future development.
b. To provide reliable data for other types of geographic areas smaller than census tracts. Sample data are produced for several types of geographic areas. Sample data are produced for block groups and user-defined areas such as neighborhoods, traffic analyses zones and school districts. Data for these small geographic areas are important for city and county planners in determining for example, housing, transportation, school and day-care center needs.
c. To provide reliable data for small population groups such as the disabled population, elderly, recent immigrants and racial and ethnic minorities. Members of these groups often live under disadvantaged conditions and depend heavily on government programs for assistance. Therefore, it is very important that their characteristics be reliably estimated by the Census. The goal is to provide reliable data for these small population groups in large geographic areas as well as in small areas.
d. To provide reliability which is close to the levels of reliability provided by the 1980 census.
e. To provide data with levels of reliability close to 1980 for data aggregated across tracts. Often, users group tracts or subtract areas forming higher geographic units for analysis purposes.
f. To maintain or decrease (if possible) the level of respondent burden from the 1980 census considering the other objectives cited above.

3. GUIDELINES FROM THE OFFICE OF MANAGEMENT AND BUDGET (OMB)

During our development of the sampling plan, the OMB suggested the Census Bureau consider the following general guidelines for developing a sampling plan that would reduce respondent burden and possibly improve or maintain data quality for 1990 as compared with 1980.

a. Consider reducing the sample size from the 1980 level as a means of improving overall data quality by reducing nonsampling errors as a result of implementing better controls on a smaller sample.

b. The use of variable rate sampling across geographic areas as a way to improve efficiency.

4. NONSAMPLING ERROR ISSUES

In considering the OMB guidelines and suggestions, the Census Bureau investigated the relationship between sample size and total error. We simulated the effects of sampling and nonsampling error using a simple model incorporating the errors listed below.

4.1. Sources of Errors in the Census Sample Data

There are three basic sources of errors associated with the data collected from the decennial census sample.

a. Sampling Error - Error that arises from the random selection of persons and households to be included in the 1990 census sample.

b. Response Error - Error that arises from respondents completing the sample questionnaire. Respondents may remember certain events incorrectly, falsify some information, lack knowledge of the information requested, or misunderstand what is wanted.

c. Interviewer Error - Error that arises from the interviewers in the data collection process. This error is primarily associated with questionnaires that are not returned by mail. This source of error is sometimes referred to as the correlated component because it results from the tendency of errors to be similar within an interviewer's assignment.

There are, of course, many other sources of error that can occur in the census data, because virtually every activity in the entire census process can be an error source. These three components of the error, however, contribute a significant amount to the total error, and the sampling and response error components are directly affected by reducing or increasing the sample size.

4.2 Expression for the Total Error of a Proportion

In this section, a simple model of the total error (as measured by the Mean Square Error (MSE)) is described. The model includes sampling and nonsampling error variance components and a term reflecting systematic error or bias.

In general, the total error or MSE of an estimated proportion from the census sample can be expressed in an additive model containing terms for the error sources described above. Such models were developed by, for example, Hansen, Hurwitz, and Bershad [2] and extended by Biemer [3].

Using this model, the total MSE of an estimated proportion (p), obtained from the census sample design and collection methodology can be approximated as follows:

$$\mathrm{MSE}\ (\hat{p}) = [\{\ (1-f)SV + SRV\ \}/fN + \{\ (1-R)CC_{NR}\ \}/k] + B^2$$

Where;

$\hat{p}$ = estimated proportion (assumed .10)

R = long form mail return rate (assumed .70)

f = sampling fraction (1/6, 1/10 or 1/20)

N = total units

SV = the sampling variance component

SRV = the simple response variance component

CC_{NR} = correlated component among the nonmail returns in an interviewer's follow-up assignment

k = the number of interviewers

B = the bias of the estimated proportion

4.3 Evaluation of the Model

The expression for the total mean square error given in the previous section, shows that it is a function of the sampling rate (f), the sampling variability (SV), two nonsampling variance components (SRV and CC_{NR}), the bias (B^2) and the size of the area (N). Values were assigned to each component to produce a value for the total MSE. The values were assigned as follows:

a. To produce estimates of the two nonsampling error variance components of the total mean square error, data from previous census evaluation studies were used. These include the 1970 and 1980 Content Reinterview Studies [4], [5], and the 1970 Enumerator Variance Study [6]. These data were developed for low, medium and high levels of census nonsampling error. It should be noted that the 1980 data include the results of the imputation for missing data.

b. The sampling error component was developed based on a 10 percent characteristic.

c. The size of the area for which estimates of total error were produced was assumed to be 4,154 persons. This is the average size of tracts that contain between 1,000 and 2,500 housing units. Over 60 million housing units will fall in these tracts in 1990.

d. The bias component was estimated from the 1970 and 1980 Content Reinterview Studies [4] and [5]. These studies showed that for many characteristics the bias was negligible, that for others it ranged to about 10 percent of the characteristics, and that for several it exceeded the 10 percent level. In the development that follows, the bias component is assumed to be either negligible or to be 10 percent of the characteristic (i.e., the absolute bias is one percentage point).

e. Three sampling rates are compared in the following analysis: 1-in-6, 1-in-10, and 1-in-20.

4.4 Results

Figures A and B, each combine the above data to display the components of the MSE for the three sampling rates. Figure A gives the MSE for the low level of nonsampling variance and for characteristics that have no bias. Figure B gives the MSE for the same level of nonsampling variance, but for characteristics that have a 10 percent bias component.

Several important relationships between the sampling rate and the components of sampling and nonsampling error are shown in these two figures. The sampling error and the simple response variance components of the MSE (SV and SRV) are significantly affected by increasing or decreasing the sampling rate. The correlated component due to interviewers is not appreciably affected by the sampling rate because it is a function of the number of interviewers required for an area and is relatively constant. This is the case, since in the context of a census, changes in the sampling rate only affect the distributions of long and short forms in a followup workload and not the size of the workload.

The bias component may be reduced by extensive precensus questionnaire and procedure testing, but is not affected by changes in sample size. Thus, the bias component is constant for all sampling rates.

Figures A and B show that reducing the sampling rate will significantly increase the total mean square error. Furthermore, the only component of this error that could realistically be reduced by the application of increased resources (e.g., training, more qualified personnel, better quality control, etc.) is the correlated component of enumerator variance (CC_{NR}). Figures A and B indicate that even if the correlated component was eliminated from the 1-in-10 and 1-in-20 designs, the total error would still be higher than for the 1-in-6 design.

Figures A and B show that errors that are a function of sample size are the dominant component of the total 1990 census error. An increase in the sampling rate clearly will bring about a reduction in the total error for census sample characteristics. The same situation is observed under the scenario of high levels of nonsampling error.

As a result of this analysis, the Census Bureau and the OMB agreed not to reduce the sample size from the 1980 level. Thus, we will be able to provide reliable data to meet the purposes of the 1990 Census.

5. COMPARISON OF SAMPLING PLANS

In 1980, governmental units with a population of fewer than 2,500 persons were sampled at 1-in-2. In order to have close to the same level of reliability in 1990, it was decided that any sampling plan considered would have at least a 1-in-3 sample rate for these small governmental units.

A sampling plan was developed by the Census Bureau basically using census tracts, block numbering areas (BNAs) and small governmental units including small minor civil divisions (MCDs) in selective states as design areas. This plan is referred to as the 10 Percent Equal CV Plan. Housing units are stratified into 4 tract/BNA strata and 2 small governmental unit strata. Each strata is defined based upon sizes of the design areas. The plan provides equal coefficients of variation (CV) for an average size of design area within each strata. In other words, all design areas within a given strata are sampled at a uniform sampling rate, therefore, design areas of different size within the given strata are sampled proportionately. A sampling rate was defined for each strata so that the estimate for a 10 percent population characteristic in an average size design area would have a standard error of 1 percentage point (or a 10 percent CV). Essentially, the plan called for sampling small governmental units at either 2-in-3 (less than 1,000 population) or 1-in-3 (less than 2,500 population). Tracts and BNAs were statified by size, 4 strata were defined, as follows.

Strata (HUs)	Sampling Rate
1- 999	1-in-3
1000-2499	1-in-6
2500-3499	1-in-8
3500-over	1-in-12

A uniform sampling rate design was suggested by the American Statistical Association Advisory Committee as such a design does not sacrifice precision of results for one group for the benefit of another [7]. Considering the sample size constraint and the decision to have a large sampling rate for small governmental units, the Census Bureau developed a plan for consideration, which we shall refer to as the Two Rate Plan, calling for a 1-in-2 sampling rate for small places and a 1-in-7 sampling rate elsewhere.

5.1. 10 Percent Equal CV Plan Vs. the Two Rate Plan

These two plans were simulated at the national level using 1990 housing unit projections based on 1980 data. These two plans were also simulated in California and Colorado using 1980 data by race and Spanish origin for several geographic levels.

a. National Simulation - The 10 Percent Equal CV Plan requires a total sample of approximately 19.6 million HUs. The Two Rate Plan requires a sample of only 17.75 million, however, the plan generates larger CVs in the smaller design areas (less than 1,000 HUs). For small subpopulation groups in large urban tracts, the 10 Percent Equal CV Plan produces less precise data than the two rate scheme. Table 1 and 2 show the distribution of national sample size for the 10 Percent Equal CV Plan and the Two Rate Plan by type of area[8].

b. Simulation for California and Colorado Using 1980 Data - The two designs were simulated for data by race and Hispanic origin. We found that the overall level of the precision of data that results from either of the designs is similar. The CVs for a 10 percent population characteristic by race and Hispanic Origin were calculated at the county, tract and place geographic level. As expected, the 10 Percent Equal CV Plan performs better in the less than 1,000 HU and 1,001-2,500 HU size categories while the Two Rate Plan out performs the 10 percent equal CV plan in the 2,501-3,500 HU and more than 3,500 HU size categories.

5.2. Summary of Results

The 10 Percent Equal CV Plan meets the needs of providing adequate data for many users. Precise data is provided for all design areas. However, data users indicated they did not need the extra precision over 1980 for smaller design areas. They felt it would be better to put more emphasis on maintaining close to the 1980 levels of precision and on precision for sub-population groups and sub-areas of tracts and BNAS. Thus, the use of the small sampling fractions (1-in-10 and 1-in-12) in larger tracts and BNAS was discarded. The 10 percent equal CV plan was ruled out for implementation in 1990.

The Two Rate Plan would provide data with slightly less precision than in 1980 for a majority of the design areas due to a 1-in-7 sampling rate other than for small governmental units as contrasted to a 1-in-6 rate in 1980. We would like to have as many design areas as possible sampled at the same rate as 1980. Using a 1-in-6 sampling rate for all areas outside small governmental units would result in to large a sample. We developed a plan to sample as many design areas as possible at the same rate as 1980 and satisfy the sample size constraint by slightly undersampling (1-in-8) very large tracts and BNAs while sampling other tracts and BNAs at 1-in-6 and small governmental units at 1-in-2. The slight undersampling of large tracts and BNAs will still provide reliable data for sub-groups and subareas of these tracts and BNAS.

6. 1990 SAMPLE DESIGN

The 1990 census sample is designed to provide sufficient precision for small areas and small subpopulation groups in the larger design areas. In addition to this, the plan maintains the 1980 levels of reliability for sample estimates for a large majority of the design areas.

6.1. Description of the Sampling Plan

The design is described as follows:

- Approximately 17.7 million households will be sampled.
- For mailing areas, the sampling scheme is as follows:
- Governmental units with a population of fewer than 2,500 persons will be sampled at the rate of 1-in-2. Governmental units include incorporated places, counties and functioning MCDs which provide a wide array of governmental functions. States that include these MCDs are Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Wisconsin.
- Tracts and block numbering areas (BNAs) with a HU count below 2,000 will be sampled at 1-in-6 for those portions not in governmental units with a population less than 2,500. The size level for which tracts and BNAs will be sampled at 1-in-6 was determined by examination of precensus housing unit counts.
- Tracts and BNAs with a count of housing units over 2,000 will be sampled at 1-in-8 for those portions not in governmental units with a population less than 2,500.
- For list/enumerate areas, governmental units with a population fewer than 2,500 persons will be sampled at a rate of 1-in-2, and all other areas will be sampled at a rate of 1-in-6. A 1-in-6 rate will be used for all of Puerto Rico. List/Enumerate is the census methodology used in sparsely populated areas where an enumerator creates a address list while collecting a completed questionnaire from each household.
- Tribal Jurisdiction Statistical Areas, American Indian reservations and Alaska Native villages will be sampled like all other governmental units with the sampling rate varying

according to the size of the Indian and Alaska Native populations, as measured in the 1980 Census. Trust lands will be sampled according to the guidelines set for their associated Indian reservations.

- Census Designated Places (CDP) in Hawaii will be sampled like all other governmental units.
  Census Designated Places are densely populated centers without legally defined corporate limits or corporate powers or functions. Ideally, a CDP should have an overall population density of at least 1,000 persons per square mile. There are no incorporated places in Hawaii. Incorporation of places is prohibited by law in Hawaii. In the 1980 census, CDPs in Hawaii were treated like any other governmental unit. In keeping with this tradition, it was decided to sample CDPs in Hawaii based upon the population.
- All persons in group quarters will be sampled at a rate of 1-in-6.

6.2. The plan, as described above, has the following desirable features:

a. Our current best estimate of the distribution of housing units by design area size, shown on Table 5, indicates the following:
   1. About 60 percent of the 1990 estimated number of housing units will be sampled at the same rate as in 1980.
   2. About 75 percent of all design areas will be sampled at the same rate as in 1980.
   This is in accordance with our objective of maintaining the 1980 levels of reliability as much as possible.

b. Tracts and BNAs will be treated in the same fashion.

c. The loss in precision for data aggregated above the tract/BNA level is very little as compared to the 1980 sampling plan.

d. Data for minority and small population groups for large design areas will not be adversely affected.

7. CONCLUSION

The variable rate 10 Percent Equal CV Plan originally proposed by the Census Bureau would have provided, on average, about the same reliability for all design areas used in developing the plan. However, data for small subpopulation groups in larger areas would be less precise than in 1980. We developed an alternative to this plan that preserves the use of variable sampling rates and at the same time satisfies the objective of providing data with reliability which is close to the levels of precision provided by the 1980 census sampling scheme. That is the case particularly for data for small design areas and for subpopulations of larger design areas.

The 1990 Census Sample design, as proposed will result in a total sample of about 17.7 million housing units.

For a tract of more than 2000 HUS, the CV will increase by about 18 percent as compared to 1980. Areas containing about 60 percent of all housing units will be sampled at the same rate as in 1980. In addition, 75 percent of all design areas will be sampled at the same rate as in 1980.

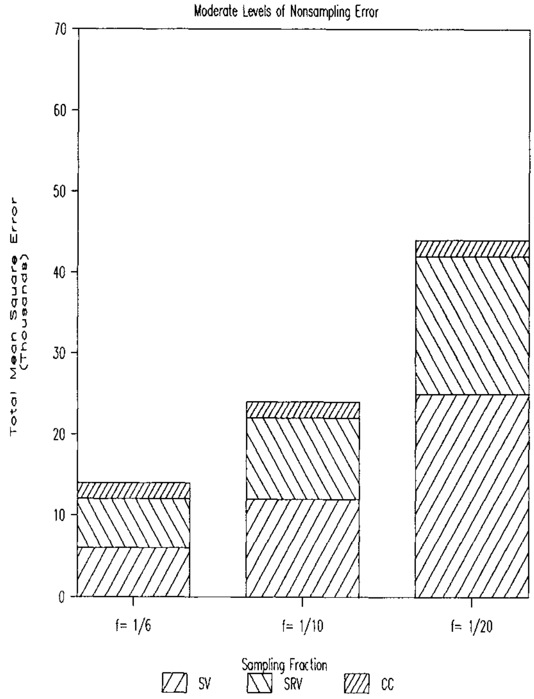
Acknowledgements

Significant contributions to this research were made by Charles D. Jones, John H. Thompson and Henry F. Woltman from the Bureau of the Census. The authors want to thank Ardi Milstead for her expertise in preparing this manuscript.

References

1. Jones, Charles D., Documents Concerning Content and Sample Design for the 1990 Census of Population and Housing, April 14, 1988.
2. Hansen, M.H., Hurwitz, W.N., and Bershad M., Measurement Errors in Censuses and Surveys, Bulletin of the International Statistical Institute, 38, 2, 359-374, Statistical Institute, 38, 2, 359-374, 1961.
3. U.S. Bureau of the Census, Evaluating Censuses of Population and Housing, Statistical Training Document, ISP-TR-5, 1985.
4. U.S. Bureau of the Census, Census of Population and Housing: 1970, Evaluation and Research Program, Accuracy of Data for Selected Housing Characteristics as Measured by Reinterviews, PHC(E)-10, 1975.
5. U.S. Bureau of the Census, 1980 Census of Population and Housing, Evaluation and Research Reports, Content Reinterview Study: Accuracy of Data for Selected Population and Housing Characteristics as Measured by Reinterview, PHC(80)-E2, 1986.
6. U.S. Bureau of the Census, Census of Population and Housing: 1970, Evaluation and Research Program, Enumerator Variance in the 1970 Census, PHC(E)-13, 1979.
7. Hansen, Morris, Allocation of the 1990 Census Sample, April 21, 1988. Draft Statement for ASA Census Advisory Committee.
8. Swan, Carolyn, Documentation for 10% CV Plans and Proportional Allocation Design, Under Consideration for the 1990 Census, October 6, 1988. Internal Census Bureau memorandum.

A. TOTAL MEAN SQUARE ERROR



B. TOTAL MEAN SQUARE ERROR

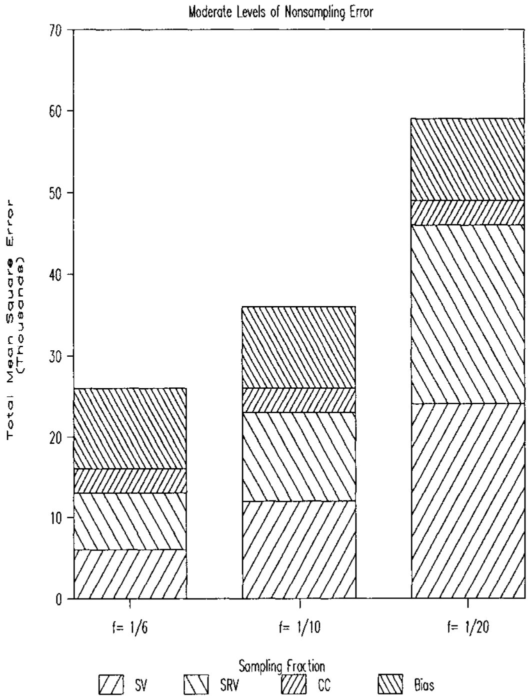


Table 1
10% Equal CV Plan (Million)

Area Type	HUs	Sample HUs	Sample Rate
Untracted (BNA)	16.2	3.2	.196
Tracted	83.2	13.4	.161
Sm. Places (MCDs)	6.6	3.0	.465
Total	106.0	19.6	0.185

Table 2
Two Rate Sampling Plan (Million)

Area Type	HUs	Sample HUs	Sample Rate
Untracted (BNA)	16.2	2.3	0.143
Tracted	83.2	11.9	0.143
Sm. Places (MCDs)	6.6	3.3	0.500
Total	106.0	17.5	0.165

Table 3
Est. 1990 Distribution of HUs by Size and Type of Geographic Area (M)

HUs in Tract/ BNA	No. of Areas	Per-cent	No. of HUs	Per-cent	Aug. Area Size
Sm. GUs	20.0	25.1	8.0	7.5	400
Tract/BNA					
1-999	16.1	20.2	7.8	7.3	484
1000-1499	15.4	19.3	18.8	7.3	1221
1500-1999	13.4	16.7	23.6	22.3	1764
2000-2499	7.4	9.4	17.6	17.2	2366
2500-3499	5.5	6.9	18.9	17.8	3468
3500-over	1.9	2.4	11.3	10.6	6079
Total	79.7	100.0	106.0	100.0	1331