NEW APPLICATIONS OF SAMPLING AND ESTIMATION
IN THE 1995 CENSUS TEST

Mary H. Mulry and Rajendra P. Singh[1]
Bureau of the Census, Washington, DC 20233

Key Words: Coverage measurement, 2000 Census

1. Introduction

The Census Bureau is testing a combination of counting with sampling and estimation for producing census numbers for the size of the population in the 1995 Census Test. The new approach is under consideration for the 2000 Census. The change from using counting alone is motivated by the fact that counting has grown to very difficult and very expensive, particularly for some groups.

Since 1970, the census methodology for counting has been based on people returning a questionnaire with a list of the people in their household by mail. When a household did not return the questionnaire by mail, the Bureau sent interviewers to their home to collect their information in person. This operation is known as nonresponse followup (NRFU). Each and every dwelling not returning a questionnaire was visited. Some had to be visited many times which increased the expense. Even with this effort, some people were not counted. Some reasons for this may be that people are seldom home, or they move often, or they do not have a permanent address. Other people want to avoid any contact with the government.

Another contributor to the increase in the cost of nonresponse followup is the increase in the percentage of the households that required followup. The response rate for the mail questionnaire declined from 75 percent in 1980 to 65 percent in 1990 (Treat, 1993). The difficulties with nonresponse followup have been major factors in the increase in the cost of the census.

The cost of taking the census has grown from 0.8 billion in 1970, to 1.6 billion in 1980, and finally to 2.8 billion in 1990, all in 1993 dollars (Neece and Pentercs, 1993). The Government Accounting Office (1992) has projected that the 2000 Census would cost $4.8 billion if conducted with the traditional counting methodology. In spite of the rising cost, the Census Bureau has measured an error in the census count. For example, the Census Bureau estimates that the overall 1990 Census count was 1.6 percent too low, and the count for Blacks was 4.4 percent too low. Although there has been a steady improvement in census taking since 1940, the percentage error in the count of blacks has remained aproximately 3 to 4 percent greater than for the country as a whole. This type of difference in accuracy is called the differential undercount.

Over the years, the Bureau continually has made improvements in the counting methodology. However, the rising cost and the persistent differential undercount has caused the Congress to direct the Census Bureau to reduce cost and the differential undercount. Combining counting with sampling and statistical estimation is the Census Bureau's solution to these problems.

The plans for the 1995 Census Test call for applications of sampling and estimation at two points in the census process. The first one is conducting followup interviews for only a sample of the nonrespondents to the mail questionnaires. The Census Bureau will not try to contact all the nonrespondents as in previous censuses. The sampling and estimation based on nonresponse followup is expected to lower the cost of the census. The second application of sampling and estimation is a coverage measurement survey at the end of nonresponse followup. The results of the estimation based on this survey will be incorporated into the census numbers. The end product is known as the one-number census. The methodology of integrated coverage measurement (ICM) is expected to reduce the differential undercount.

Currently the Bureau is planning to test this combination of counting and estimation in the 1995 Census Test which is being conducted in three locations: New Jersey; Oakland, California; and six parishes in northwest Louisiana. The results of the 1995 Census Test will provide a basis for the Census Bureau to evaluate the proposed methodology. The decision on the methodology for the 2000 Census will be based on an assessment of the accuracy of the census numbers and the cost of producing them.

This paper describes the methodology under development and the plans for its evaluation. Section 2 describes the plans for the mail questionnaires and the sampling for nonresponse followup. Section 3 discusses the integrated coverage measurement design. The sample design and estimation for

nonresponse followup and integrated coverage measurement is discussed in Section 4. The evaluation plans are described in Section 5 with an overall summary in Section 6.

2. The Mail Questionnaires and Sampling for Nonresponse Followup.

The design for the 1995 Census Test and the 2000 Census is based on a Master Address File (MAF). The MAF will be a list of addresses that will be compiled and then kept continuously updated in most areas. The list initially will be assembled using information from the 1990 Address Control File (ACF) and the U. S. Postal Service (USPS). Subsequent corrections will be based on information from field operations, the USPS, state and local governments, and administrative records.

In the 1995 Census Test and the 2000 Census, the MAF will provide the list for mailing the census questionnaires in most areas. In the rural areas such as the test site in Northwest Louisiana, questionnaires will be delivered by Census Bureau employees. The respondents will fill out the questionnaires and return them by mail. Group quarters are enumerated in a separate operation.

The 1995 Census Test is the first opportunity to test a new mailing procedure in a census environment. This scenario has shown improvements in response rates in national tests conducted the past few years (Dillman, Clark, and Treat, 1994). The mail response rate is still important even with the addition of sampling and estimation. When the number of people who are enumerated initially is high, there is less uncertainty in the census numbers.

The new strategy for the mailing of questionnaires to housing units has four stages. First a pre-notice is sent on February 27 to inform the household that the census questionnaire is coming. Two days later on March 1, the Census Bureau mails the census questionnaire. Census Day is March 4. Five days later on March 6, a reminder notice which is also a thank-you note is sent. When the questionnaire is not received within two weeks after the first questionnaire, a second questionnaire is mailed during the period or March 20 - 27.

The vacant and nonexistent housing units will be identified by a post-master return of the first questionnaire. The status of these housing units will be confirmed by interviewers who visit the dwelling. If the housing unit proves to be occupied as of Census Day, the interviewer will conduct an interview.

If the second questionnaire is not returned within three weeks and the housing unit is not shown to be vacant or nonexistent, the housing unit is considered to be a nonresponse unit. A sample of the nonresponse housing units will be selected for followup. In the urban sites, an attempt to interview the residents will be made by telephone using Computer Assisted Telephone Interviewing (CATI) if the Census Bureau has been able to obtain a telephone number for the housing unit. If there is no phone number or CATI attempts are unsuccessful in obtaining a completed form, then an interviewer will make a personal visit and conduct an interview with a paper questionnaire.

One goal of the 1995 Census Test is to test two basic sampling designs for nonresponse followup. The issue is whether to select a sample of nonresponding housing units without regard to geographic clustering or to select a sample of blocks and include all the nonresponding housing units in the blocks in the sample. Overall the sampling rate for the test will be one-third of the housing units which did not respond with one-sixth allocated to each design. More details of the sample selection are given in Section 4.

Each type of sample design appears to have its advantages and disadvantages. Simulations with the 1990 Census show that the sample based on housing units apparently has less bias and variance for small areas (Isaki, Tsay, and Fuller, 1994). The block sample appears easier to implement in conjunction with integrated coverage measurement because it requires a block sample. However, research to improve the estimation with the block sample is in progress. The difference in field costs for the two designs is unclear so the test is being designed to collect information on cost for such a comparison. Speculation is that the block design would cost less because it is more clustered and would require less expense for travel by interviewers.

3. Integrated Coverage Measurement

Coverage measurement surveys have been used by the Census Bureau in past traditional censuses as possible adjustments to counts and for evaluations. The 1995 Census Test is the first time the results of a coverage measurement survey will be included as part of the census process and the production of the census numbers.

One concern about integrated coverage measurement for 2000 is whether it can be completed by the deadline of December 31, which is when the Census Bureau is required to deliver census numbers for states to the President. Some aspects of the design are aimed at speeding up the process by using improved technology as much as possible. New

technology makes it possible to eliminate a separate keying operation and to automate access and control of data.

The sample for the ICM will be a block sample, and the same sample of blocks will be used for the within housing unit and the whole housing unit parts. Every housing unit within the sample blocks will be in the sample. In the 1995 Census Test, all the nonrespondents in the blocks selected for the ICM sample will be included in the followup. The ICM sample needs to be a block sample to facilitate the identification of missed housing units. There are more details about the sample selection in Section 4.

The data collection is designed to examine two types of estimation, CensusPlus and dual system. Both methods of estimation which are described more in Section 4 use data collected on census questionnaires and in the ICM, but somewhat differently. The ICM operation first requires that an independent listing of the sample blocks will be performed prior to the mailing of the questionnaires. The two lists of housing units will be matched, and the discrepancies sent to the field for reconciliation. If a housing unit is on the MAF but missed by independent listing, its existence will be confirmed. If a housing unit has been included on the MAF in the sample block by mistake, that also will be noted. The final product will be an enhanced address list which included the independent listing and additions of existing units which it missed.

The enhanced housing listing will provide the basis for the ICM interviewing. The interviews will be conducted at the end of nonresponse followup, possibly overlapping about a week, using Computer Assisted Personal Interviewing (CAPI). First, the interviewer will first collect a roster of the people living at the housing unit on Census Day. After the roster is completed, the CAPI system will provide the interviewer the names of the people on the census questionnaire for that address. This method which uses relatively knew technology assures that the roster the ICM collects is independent of the census questionnaire.

During the interview, the interviewer will collect information about the people who do not match. Information for determining the residency according to census rules will be collected for the all people on the census questionnaire and the ICM independent roster. The interviewer will reconcile the differences in the rosters from the census questionnaire and the ICM. The reconciliation and the answers to the questions about residency rules will produce list of the people who should have been enumerated at the housing unit, called a resolved roster. People may added to the resolved roster as part of the reconciliation process even though they did not appear on the census questionnaire or the ICM independent roster. The CAPI machine will record the stage at which people on the resolved roster were added.

The resolved rosters, after assignment of residency status, review, and unduplication in the processing office, will provide the data for the CensusPlus estimator about July 21. More field work and office processing is necessary for the data for DSE. The ICM independent roster constitutes the population sample (P sample). The rosters of the census questionnaires constitutes the enumeration sample (E sample). There will be a computer and clerical matching operation which will use the information from the ICM interview to resolve the enumeration status of both P-sample and E-sample people. There will be a followup of cases that can not be resolved. For example, the followup will include people listed on a census questionnaire who do not appear on a resolved roster and no other information collected during the ICM interview determines whether they were correctly or erroneously enumerated. The data for the DSE is expected to be ready about October 6.

4. Sampling Design and Estimation

The sampling design for the 1995 Census Test accomodates sampling and estimation for nonresponse followup and ICM. The estimation based on the sampling of nonrespondents will be made using the data from the housing unit sample, the block sample, and possibly from a combination of the two samples. Two methods of estimation will be explored for ICM, CensusPlus estimation and dual system estimation.

The comparison of the CensusPlus and DSE focuses on the fact that they have different underlying assumptions. CensusPlus estimation is based on the assumption that the ICM finds the "truth" in the sample blocks, the truth being the resolved rosters from the field reconciliation. The DSE assumes the independent roster collected during the independent part of the ICM interview is another independent list, but not necessarily the "truth." The DSE estimates people on neither list while CensusPlus finds them trough the reconciliation process during the interview. The two estimates of the additional people will be the basis of comparison of the two methodologies.

The sample selection requires four steps. The first step is to combine blocks to form clusters of 30 to 40 housing units. The block clusters will be

used as the sampling units. Blocks with more than 40 housing units will not be combined with other blocks, and neither will they be split or subsampled. Next the block clusters will be stratified and divided into two panels. Half of each stratum will go to each panel. The panels are for the sampling for nonresponse, one for the unit sample and one for the block sample. Then a new stratification for sampling will be defined in each panel. One half of the ICM sample will be selected from each panel. The last step is to select the samples for nonresponse in each panel from the clusters not selected for the ICM. While the specifics of the sample selection for nonresponse are still under development, the overall sampling rate will be one third. The number of block clusters selected for ICM will be 150 in Oakland, 100 in Paterson, and 100 in northwest Louisiana.

Table 1. ICM Sample Sizes

	Sample Clusters	Estimated Total Clusters
Oakland	150	2269
Paterson	100	716
NW Louisiana	100	1052

Using these samples, estimates for entire sites and for poststrata defined by age/sex/race/Hispanic ethnicity/tenure within sites will be made after nonresponse followup and again after the ICM data is collected. The estimation after nonresponse followup, including the compensation for any missing data, is still under development for each sample design and for a combination of the two sample designs. One method under consideration is essentially an imputation for housing units not included in the sample similar to the one investigated by Isaki, Tsay, and Fuller (1994). The estimation for nonresponse will fit the following format. Define

$\hat{C}_{bj} =$ estimate for poststratum j in the bth block from mail returns and nonresponse followup.

The estimate $\hat{C}_{bj}$ has three components,

$$\hat{C}_{bj} = COUNT_{bj} + \hat{NR}_{bj}$$

where

$COUNT_{bj} =$ count for poststratum j in b^{th} block from the mail returns, and housing units included in nonresponse followup,

$\hat{NR}_{bj} =$ estimate for poststratum j in the bth block for nonresponding housing units not included in the followup.

Now let's turn our attention to the estimation for ICM. For the CensusPlus estimator, define

$RE_{bj} =$ resolved count for poststratum j in the bth block.

The ratio estimator of the total population for poststratum j is

$$\hat{P}_j = \hat{B}_j \hat{C}_j.$$

where

$$\hat{C}_j = \sum_b \hat{C}_{bj} = \sum_{b \in S_{ICM}} \hat{C}_{bj} + \sum_{b \notin S_{ICM}} \hat{C}_{bj}$$

$$\hat{B}_j = \frac{\sum_{b \in S_{ICM}} W_b RE_{bj}}{\sum_{b \in S_{ICM}} W_b \hat{C}_{bj}}$$

$W_b =$ ICM sample weight for b^{th} block.

$S_{ICM} =$ set of blocks in the ICM sample.

Another formulation of the estimation was done by Wright(1993) who set it in the framework of the simple proportional regression model. This characterization provides a structure for validating modeling assumptions.

For the Dual System Estimator in poststratum j, DSE_j, define the following:

$\hat{C}_j =$ estimate from mail returns and nonresponse followup

$\hat{I}_j =$ estimate of persons who are unmatchable because of insufficient information

$\hat{EE}_j$ = estimate of number of erroneous enumerations

$\hat{N}_{pj}$ = estimate of population from the P-sample in ICM

$\hat{M}_j$ = estimate of the number of matched persons

Then

$$DSE_j = (\hat{C}_j - \hat{I}_j - \hat{EE}_j)\frac{\hat{N}_{pj}}{\hat{M}_j}$$

The application of DSE in the 1995 Census Test differs from the 1990 Post Enumeration Survey in two important ways. One way is in the definition of the P-sample population. In 1995, the P-sample population is comprised of the people who lived in the sample blocks on Census Day. In 1990, the P-sample population was defined by the people who lived in the sample blocks at the time of the PES interview. The new definition eliminates the large operation in 1990 of matching movers into the sample blocks at their Census Day address. However, the 1995 definition includes people who move out of the sample blocks before the ICM interviews. Data for the out-movers will be collected from proxies and by tracing. The other difference is that the reconciliation of the two lists is performed by interviewers in the field in 1995. In the 1990 PES, the independent roster was collected by one interviewer and matched to the census in the processing office. Cases that could not be resolved were sent back with a different interviewer to collect more information. The 1995 method of reconciliation uses technology that was not readily available in 1990.

5. Evaluations of Sampling and Estimation

The Census Bureau has set the goal of deciding on the basic methodology for the 2000 Census by the end of 1995. Therefore, not only do the final test census numbers have to be produced by December 31, 1995, but the evaluations needed to make the decision about the census methodology also are required. Currently, seventeen evaluation projects have been proposed and are under review. The ones with the highest priority are discussed below.

A major goal of the evaluation of the sampling and estimation for this decision is to verify that the assumptions underlying the data collection and estimation are valid. One way of accomplishing this is to verify that the data flowing into the final estimation is correct. To meet one aspect of this goal, the evaluations need to assess whether the erroneous census enumerations are identified accurately and whether people missed by the mail questionnaires and nonresponse followup are found. Another aspect of the evaluations will be to assess the effectiveness of the estimation schemes under consideration which requires examining the final census numbers. Although the evaluations are still in the planning phase, the initial indications are that the primary tools will be reinterviews, rematches, and analysis of the data collected.

A major criterion for the evaluation is whether the ICM is adding people in the traditionally undercounted groups. This analysis is aimed at assessing whether the ICM is reducing the differential undercount. The final census numbers will be compared with the population estimates after nonresponse followup to determine whether people in the traditionally undercounted groups are being added by the ICM. The sex ratios of the final census numbers compared to those from the mail questionnaires alone, to those from the mail questionnaires and nonresponse followup, and to those from the 1990 Census will shed light on the overall effectiveness of the operations. Also, other dual system estimators such as the split dual system estimator (Bell, 1993) and the logistic regression dual system estimator (Alho, Mulry, Wurdeman, and Kim, 1993) may be calculated and compared to the ratio estimator and the dual system estimator.

Three of the studies which focus on the ICM interviewing are the evaluation interview, the evaluation of the out-movers, and the reconciliation bias study. The goal of the evaluation interview is to examine the accuracy of the ICM resolved roster. The plan is to use cognitive interviewing methods to obtain explanations of differences between the census questionnaire, the ICM independent roster, and the resolved roster.

The evaluation of the out-movers is aimed at the assumption that accurate determination of the Census Day residency of these people made using proxies and tracing. The study will do more extensive tracing of a sample of the out-movers and compare the results to the ICM decisions on the the Census Day residency of the people. The other method will involve identifying out-movers using data from other sources such as a Census Bureau survey or the Postal Service. A comparison with ICM interviews will be made to determine if ICM also

identified these movers.

Another study examines the effect of the field reconciliation of the census questionniare and the ICM roster by the interviewer. There may be a tendency by the interviewer or the respondent to make the resolved roster consistent with the ICM independent roster. One approach measures interviewer by performing the ICM interview and then having another interviewer return and conduct the reconciliation part of the interview following the independent roster. An alternative approach measures differential bias by performing the ICM interviews for a sample of blocks with 100 percent NRFU but not in the ICM sample and having the reconciliation done independently by another interviewer.

A review of the matching and final coding of cases both for CensusPlus and DSE is aimed at evaluating another aspect, the data processing. More highly trained personnel will examine the final coding to determine errors were made in the final office review for both estimation schemes. In addition, a tracing of cases through the entire matching process will show any underlying patterns which indicate problems. The automation of the ICM operations permits the detailed record keeping this analysis requires.

6. Summary

Combining counting and sampling is expected to be much less expensive than attempting to interview everyone and more accurate. However, measuring part of the population using sampling introduces some uncertainty into the numbers. The size of the sample effects the cost and the margin of error. As the size of the sample increases, the cost increases, and the percentage margin of error decreases. If the cost is increased too much, there is no advantage to sampling.

The censuses that only use counting have been subject to differential coverage error which is a systematic error that effects all levels of geography. Differential coverage error is expected to have much less impact with the combination of counting with sampling and statistical estimation than in previous censuses, but there will have some uncertainty surrounding the census numbers. The uncertainty is caused by sampling instead of interviewing the entire population. This error from sampling is not systematic and the effects vary depending on the particular sample that is taken. The results will be accurate if the sample is representative of the entire population, but less accurate otherwise. The uncertainty also increases as the level of geography

decreases. It may be considerable at the block level.

The 1995 Census Test provides an opportunity to test the proposed methodology. The Census Bureau will be able to learn about the implementation and cost of combining counting with sampling and estimation. At the end of 1995 when the results of the Census Test have been reviewed, the Census Bureau will announce the basic methodology for the 2000 Census.

[1] This article reports results of research undertaken by staff members of the Census Bureau. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

References

Alho, Juha, Mulry, Mary H., Wurdeman, Kent, Kim, Jay (1993) Journal of the American Statistical Association, 88, 1130-1136.

Bell, William R. (1993) "Split Dual System Estimation," Statistical Research Division, Bureau of the Census, unpublished manuscript.

Dillman, Don A., Clark, Jon R., Treat, James B. (1994) "Influence of 13 Design Factors on Completion Rates to Decennial Census Questionnaires," Proceedings of the 1994 Annual Research Conference, Bureau of the Census.

General Accounting Office (1992) "Decennial Census: 1990 Results Show Need for Fundamental Reform," Report to Congressional Requesters GAO/GGD-92-94, June 1992.

Isaki, Cary T., Tsay, Julie H., and Fuller, Wayne A. (1994) Proceedings of the 1994 Annual Research Conference, Bureau of the Census.

Neece, Lorraine and Pentercs (1993) "1970, 1980, 1990 Decennial Census Cost Comparison," Year 2000 Research and Development Staff, Bureau of the Census, unpublished manuscript.

Treat, James B. (1993) "1993 National Census Test Appeals and Long-Form Experiment, Appeals or Short-Form Component," Final Report, DSSD 2000 Census Memorandum Series #E-62, Bureau of the Census, October, 1993.

Wright, Tommy (1993) "CensusPlus: A Sampling and Prediction Approach for the 2000 Census of the United States," Bureau of the Census, unpublished manuscript.