Impacts of Sampling for Nonresponse Follow-up and Integrated Coverage Measurement on Census Methodology for a One Number Census

Rita J. Petroni, Michael Ikeda, and Rajendra P. Singh[1]
Bureau of the Census, Washington, DC 20233-0001

KEY WORDS: NRFU Sampling, Coverage Adjustment, Census Estimates

I. What Is Covered?

The 1995 Census Test is researching two fundamental changes to traditional U.S. Census methodology - following-up only a sample of nonresponding households (NRFU sampling) and integrating coverage measurement into estimation. (Thompson, 1994.) The introduction of these two fundamental changes creates the need to re-evaluate and change other census methodologies to ensure statistically valid census estimates.

This paper describes the 1990 Census, the 1995 Census Test, and the implications of adopting the two fundamental changes for other census methodologies. It examines the implications of these fundamental changes for using late census returns, census primary selection and search/match methodology, and census imputation methodology and how the Census Bureau is implementing these methodologies for the 1995 Census Test. It also proposes research and alternatives to these methodologies in an attempt to improve estimates in the 2000 Census.

II. Fundamental Changes and Their Implications

A. The 1990 Census

In the 1990 Census, the Census Bureau collected census data in all blocks, followed-up all nonresponding units, included late census returns through the end of December 1990, conducted primary selection and search/match operations (see Section III.B.), and carried-out editing and imputation. Following these operations, we obtained and published census counts.

Independent of the census, we selected a sample of census blocks, collected data in the selected blocks, and used data from the sample blocks to measure net coverage error.

B. The 1995 Census Test

We are implementing the 1995 Census Test in three sites - Oakland, California; Paterson, New Jersey; and Northwest Louisiana. Due to budget constraints, the full test is conducted only in the Oakland site. Hence the remaining discussion applies only to that site.

The Census Bureau selected a sample of census blocks called Integrated Coverage Measurement (ICM) blocks. We will use independently collected data from the ICM blocks (ICM data), in conjunction with census collected data from all ICM and non-ICM census blocks in the site, to obtain one number census estimates.

For the 1995 test, the Census Bureau collected census data in all test site blocks, followed-up a sample of nonresponding units in non-ICM blocks, followed-up all nonresponding units in ICM blocks, excluded most census returns received after the date when the nonresponse universe was created, conducted primary selection and search/match operations, and carried out editing and imputation.

In the test, we research two alternative nonresponse follow-up sample designs - block sampling and unit sampling. For this purpose, we assigned each census block to one of two experimental panels - block and unit - so that each panel is representative of a site. We then selected ICM blocks from the two panels. This resulted in a partition of census blocks into three mutually exclusive universes:

-ICM blocks (all nonresponding units are followed-up)

-Block sample universe blocks: non-ICM blocks from which a sample of blocks was selected for follow-up of all nonresponding units

-Unit sample universe blocks: non-ICM blocks from which a sample of nonresponding units were selected for follow-up (Navarro, Bates, Scott, and Isaki, 1995.)

We also research two alternative ICM estimation methodologies - Census Plus and Dual System Estimation. (Thompson, 1994.)

For the site, we use all census blocks to obtain initial census estimates for 56 post-strata defined by age, race, hispanic origin, sex, and tenure. Call these estimates C_s where: $C_s = C_s(I,R) + C_s(I,NRFU) + C_s(B,R) + C_s(B,NRFU) + C_s(U,R) + C_s(U,NRFU)$, $C_s(I,R)$, $C_s(B,R)$, and $C_s(U,R)$ are, respectively, census counts for post-stratum s from the response universes of ICM blocks, block sample universe blocks, and unit sample universe blocks and $C_s(I,NRFU)$, $C_s(B,NRFU)$, and $C_s(U,NRFU)$ are, respectively, census estimates for post-stratum s from

the nonresponse universes of ICM blocks, block sample universe blocks, and unit sample universe blocks.

The initial site-level estimate is the sum of the 56 post-stratum-level initial census estimates.

We increase the reliability of the 1995 Census Test estimates in two ways. First, we combine sample from all universes. Second, rather than weighting the sample in each nonresponse universe by the inverse probability of selection within the universe, we will weight ICM and non-ICM sample cases in the unit panel by the inverse of the overall probability of being selected from the unit panel. We will weight ICM and non-ICM sample cases in the block panel in a similar way. (Ikeda, 1995a, Navarro, 1995.)

In ICM blocks, we collected ICM data using procedures designed to obtain more accurate coverage. The Census Bureau will use results from these blocks for coverage measurement and integrate them with the initial post-stratum level census estimates to obtain one number census estimates for the site and each post-stratum. For a post-stratum, the one number census estimate is:
$CO_s = [(I_s)/(CI_s)]*C_s$ where I_s estimates the census population for post-stratum s using ICM collected data and CI_s is the census count from the response and nonresponse universes of ICM blocks for post-stratum s weighted up to estimate the post-stratum for the site. (Schindler, 1995.)

The one number site-level census estimate is the sum of the 56 post-stratum level one number census estimates. We will publish the one number census estimates only. The initial estimates will not be published.

C. What Are The Implications of the Fundamental Changes?

Because of the two fundamental changes, the Census Bureau must move to new approaches for other census operations. It is important that the methodologies the Bureau uses to obtain CI_s and C_s be carried-out so that CI_s and C_s are consistent to obtain statistically valid one number census estimates. We must implement the methodologies so that $E(CI_s) = E(C_s)$. This means the Bureau needs to use late returns, implement primary selection and search/match methodology, and implement census imputation methodology in the same or equivalent manner for ICM blocks and non-ICM blocks.

For the 1995 test, it also means we must carry-out the methodologies in the same or equivalent manner in ICM blocks and in blocks in the unit and block sample universes. For the 2000 census, the Bureau will have decided between unit and block NRFU sampling and can implement the methodologies in ways best suited to the selected sampling method.

III. Impact On Census Methodology For The 1995 Census Test

A. Uses Of Late Returns

For the 1990 Census, we used as much late census data as possible to ensure as complete coverage as possible. The Bureau processed returns received from census day (April 1, 1990) through the end of December 1990. We processed on time and late returns through primary selection and search/match operations when needed and through edit and imputation.

For the 1995 Census Test, we considered a census return to be late if it was received after the date we created the NRFU universe. Census day was March 4, 1995. The NRFU universe was created April 13, 1995. With a few exceptions, we did not use returns received after that time. (Ikeda, 1995a.) Thus, the 1995 Census Test disregards most census returns received after NRFU universe creation.

Ideally, we want to use late returns. However, no data existed to explore whether we could obtain unbiased estimates if we use late returns. We were concerned that 100% NRFU blocks may receive proportionally more late returns than other blocks; and hence, introduce bias and also affect comparison of the unit and block designs. Since we can obtain unbiased estimates and avoid contamination of comparisons between the unit and block designs by rejecting late returns, we decided not to accept late returns in non-ICM blocks.

Since we must use the same or equivalent procedures in ICM and non-ICM blocks to ensure that $E(CI_s) = E(C_s)$, we decided not to accept late returns for any blocks.

The Census Bureau expects to conduct research on how to make more use of late census returns in the 2000 Census.

B. Primary Selection and Search/Match Methodology

The Census Bureau conducted a search/match operation during the 1990 Census to ensure the correct enumeration of certain subpopulations and/or to ensure that all persons living away from their usual residence on census day were enumerated at their usual residence. To accomplish this, the Bureau searched the census questionnaire at a person's reported usual residence to determine if they were counted there. If they were not, we added them to the census at that address. (Wajer, 1993a.)

Additionally, the Census Bureau applied an algorithm to select persons to assign to an address

identification number (ID) when multiple census forms were included on census files for an ID. The algorithm selected one vacant record or one form and any supplemental forms associated with the selected form to represent the ID. (Love, 1990.) From late October through the end of December 1990, the Bureau reviewed data captured questionnaires whose data records were not selected by the algorithm to represent a given household. The Census Bureau sent the selected and non-selected persons in these IDs through a modified search operation to determine if the non-selected persons were counted in the census. The modified operation allowed searching in up to ten neighboring addresses to the ID. If we found that non-selected persons were not counted in the census, we added them to the census. Also, in some cases, if a selected person was counted elsewhere, we removed the duplication of the person from the census. (Wajer, 1993b.)

The October-December 1990 review added or moved 350,448 persons in 161,541 housing units and corrected for duplication at about 9.7% of these housing units. (Wajer, 1993b.) This suggested that future censuses should attempt to refine the algorithm to handle the varying circumstances causing different reported rosters for a given household. As a result, the initial plan for the 1995 Census Test was to search, match, and unduplicate within and between all IDs in census blocks. However, because of sampling for NRFU in non-ICM blocks in the 1995 Census Test, we do not have data to search and match between all census IDs in non-ICM blocks. To ensure that $E(CI_s) = E(C_s)$, the 1995 primary selection and search/match operations search, match, and unduplicate the same way in ICM and non-ICM blocks. Additionally, to keep the search, match, and unduplication process as simple as possible for the 1995 Census Test and to maintain consistency in processing procedures between the unit sample block universe and the block sample universe, the Bureau decided to search, match, and unduplicate persons reported on different census forms within an ID only. No attempt is made to unduplicate persons within IDs which receive only one return.

For the 1995 Census Test, we considered implementing two other alternatives which allow for more potential for unduplication. One was to:

(a) search, match, and unduplicate within IDs in the nonresponse universe;

(b) search, match, and unduplicate within and between census IDs within blocks in the response universe of ICM blocks, unit sample universe blocks, and block sample universe blocks.

The other was to:

(a) search, match, and unduplicate within census IDs in the nonresponse universe of unit sample universe blocks;

(b) search, match, and unduplicate within and between census IDs within blocks in the nonresponse universe of ICM blocks and block sample universe blocks;

(c) search, match, and unduplicate within and between census IDs within blocks in the response universe of ICM blocks, unit sample universe blocks, and block sample universe blocks. (Petroni, 1994.)

The Bureau expects to conduct research to identify the best approach for the 2000 Census.

C. Census Imputation Methodology

For the 1990 Census, the Census Bureau imputed nonrespondents to the census after 100% NRFU, using a hot deck procedure that substitutes the nearest previous responding unit having the same race and household size for the nonresponding unit. The nearest responding unit may or may not have been included in NRFU. (Treat, 1994.)

The 1995 Census Test imputes for nonrespondents in two phases. First, the 1995 Census Test imputes the characteristics of nonrespondents to the NRFU sample from respondents to the NRFU sample from the same block. If there are no respondents to the NRFU sample in the same block, the traditional census procedure is used. (Spencer, 1995.) Second, since we expect sample and nonsample cases in the NRFU universe to be more alike than cases not in the NRFU universe, we impute nonsample NRFU universe cases from sample NRFU universe cases only.

We need to impute nonrespondents to the NRFU sample so that $E(CI_s) = E(C_s)$. If the Bureau were to use the traditional imputation approach in all blocks, then different donor pools may be used in ICM and non-ICM block imputation, introducing biases.

The Census Bureau is considering to conduct research to identify an appropriate approach for the 2000 Census.

IV. Research For The 2000 Census

A. The 2000 Census

For the 2000 Census, the Bureau expects to select a sample of ICM blocks to assist in obtaining a one number census estimate. We also expect to collect census data in all blocks, follow-up a sample of nonresponding units in non-ICM blocks (using either the block or unit design, but not both), follow-up all nonresponding units in ICM blocks, include late census returns, conduct primary selection and search/match operations, and carry-out editing and imputation. Under this scenario, we will obtain initial

post-stratum census estimates for currently unspecified geographic levels, C_s, where:
$C_s = C_s(I,R) + C_s(I,NRFU) + C_s(N,R) + C_s(N,NRFU)$, $C_s(I,R)$ and $C_s(N,R)$ are, respectively, the census counts of a geographic level for post-stratum s from the response universes of ICM and non-ICM blocks and $C_s(I,NRFU)$ and $C_s(N,NRFU)$ are, respectively, the census estimates of a geographic level for post-stratum s from the nonresponse universes of ICM and non-ICM blocks.

The one number census estimate for the geographic level will be the sum of the post-stratum level one number census estimates for the geographic level.

Currently for the 2000 Census, the Bureau plans to again select and follow-up a sample of the nonrespondents. The exact definition of the nonresponse universe is under discussion.

Again, we will collect data independently in ICM blocks. The Census Bureau will use ICM data for coverage measurement and integrate them with the initial census estimates to obtain one number census estimates. The initial estimates will not be published.

In anticipation of this census design, we propose to do research in the next few years to determine appropriate approaches to using late returns and to implementing primary selection and search/match methodology and census imputation.

B. Potential Alternatives And Research For 2000

1. Use of Late Returns

For the 2000 Census, we expect to use late census returns as fully as possible, even for units in the NRFU universe. Under the 2000 Census design, late returns are those received after the NRFU universe is created. The Bureau expects to use 1996 Reengineered Census Test data to research how to use late census returns and at the same time obtain unbiased estimates.

If the Bureau selects the block design for the 2000 Census, we will examine whether the proportion of late returns to the response universe is the same in 100% and 0% NRFU blocks of the block sample universe and in the ICM blocks. If they do differ, we will examine whether the difference would have resulted in a different count and distribution of persons. If the proportions do not differ or if they do and the counts and distributions of persons do not differ, we have evidence supporting the use of all late returns.

If this occurs, we anticipate using late returns for sample and non-sample NRFU IDs which receive them and excluding them from the imputation process. If the proportions, counts, and distributions do differ, we will research modelling to account for late returns in the response universe. One possibility is to obtain census counts with and without inclusion of late returns in 100% NRFU blocks, create a model from the two results, obtain census counts excluding late returns in 0% NRFU blocks, and apply the model to these counts from the 0% NRFU blocks.

In the event that counts and distributions differ by a small amount, we plan to compare results of using all late returns to results from implementing the modelling approach to select the appropriate way to handle late returns.

If the Bureau selects the unit design for the 2000 Census, we will do similar research. However, in this case, we will compare the proportions, counts, and distributions between ICM and non-ICM blocks for response and NRFU universes and, if necessary, research modelling for both the response and NRFU universes.

2. Primary Selection and Search/Match Methodology

For the 2000 Census, our goal is to implement a search, match, and unduplication approach that will allow us to use as much information as possible to resolve who belongs to an ID while maintaining statistically valid estimates. In preparation, we expect to use 1995 Census Test data to simulate the response and NRFU universes of the anticipated 2000 Census NRFU design and will use the simulation to identify an appropriate search, match, and unduplication procedure.

We expect to first research whether the particular search/match alternative does in fact affect census counts and whether the various alternatives are operationally feasible. If the various alternatives are feasible, we will research the impact of the various alternatives on the bias and variance of estimates. Also if they are feasible, then we will expand the research to explore the use of extended search/match procedures and modelling. The research will evaluate bias and variance properties of estimates resulting from the models to evaluate the appropriateness of modelling and to choose a model, if one is feasible.

For the block design, we will:

1) obtain census counts for the site and each post-stratum from the ICM blocks for response and NRFU universes when searching and matching is done only within IDs (Option 1); call these counts, respectively, C1(R) and C1(NRFU); and let C1 = C1(R) + C1(NRFU).

2) develop procedures to search, match and unduplicate within and between IDs in the response universe of a block and to search, match, and unduplicate within and between IDs in the NRFU

universe of a block (Option 2); apply the procedures to the census data obtained for the 1995 ICM blocks; obtain census counts for response and NRFU universes; call these counts, respectively, C2(R) and C2(NRFU); and let C2 = C2(R) + C2(NRFU).

3) develop procedures to search, match, and unduplicate within and between IDs in the block regardless of response or NRFU universe (Option 3); apply the procedures to the census data obtained for the 1995 ICM blocks; obtain census counts for response and NRFU universes; call these counts, respectively, C3(R) and C3(NRFU); and let C3 = C3(R) + C3(NRFU).

4) compare C1(R), C2(R), and C3(R); C1(NRFU), C2(NRFU), and C3(NRFU); and C1, C2, and C3 to examine the impact of extended searching.

Depending on the operational feasibility of the various alternatives, we will explore the use of modelling. The Bureau may research the following modelling approaches.

-Implement options 1 and 3 in blocks where we conducted 100% nonresponse follow-up and use results to obtain a model. Then, implement option 1 in blocks which were not selected for NRFU and apply the model to these blocks to estimate C3(R) and C3. Finally, implement option 3 in blocks which had no nonresponse to obtain C3.

-Implement options 2 and 3 in blocks where we conducted 100% nonresponse follow-up and use results to obtain a model. Then, implement option 2 in blocks which were not selected for NRFU and apply the model to these blocks to estimate C3(R) and C3. Finally implement option 3 in blocks which had no nonresponse to obtain C3.

Depending on the processing stage at which we obtain C1, C2, and C3, the first model may be applied to 1995 Census Test results to obtain revised census estimates which we will compare to original census test estimates.

For the unit design, in ICM blocks we will compare C1(R), C2(R), and C3(R); C1(NRFU) and C3(NRFU); and C1, C2', and C3 to examine extended searching under unit sampling. (C2' = C2(R) + C1(NRFU).) Depending on the operational feasibility of the various alternatives, we will research models similar to those for the block design. In this case, however, the second model will define option 2 to be search, match, and unduplicate within and between IDs in the response universe and to search, match, and unduplicate only within IDs in the NRFU universe. Also, we will develope separate models for the response and NRFU universes.

Ikeda(1995b) proposes the following possible modelling strategies. The methods are applicable to both the block and unit samples. The basic procedures for the response portion are similar in the two sample designs.

-Use the 100% nonresponse follow-up blocks to calculate weighted control totals for various household types and then use the control totals to adjust the weights of households of the various types in blocks not in the nonresponse follow-up sample.

-Calculate ratio adjustments for each NRFU stratum and household type. The numerator of the adjustment is the mean number of respondent addresses in blocks where we conducted 100% nonresponse follow-up for the NRFU stratum and household type. The denominator is the mean number of respondent addresses in blocks where we conducted less than 100% nonresponse follow-up for the NRFU stratum and household type.

-Use the 100% nonresponse follow-up blocks to calculate weighted control totals for various person categories and use these control totals to adjust the weights of persons in the person categories in the remaining blocks.

-Adapt Fuller, Isaki, and Tsay's (1994) method for modelling the nonresponse portion to model the response portion.

3. Census Imputation Methodology

Research of census imputation for nonresponding cases in the NRFU sample will use 1990 census data and take into account the expected NRFU sampling rate for 2000 as well as whether the unit or block design is selected for 2000. We selected 2-in-7 NRFU samples of non-ICM blocks and units in Oakland and 1-in-6 NRFU samples of non-ICM blocks in Paterson and Northwest Louisiana for the 1995 Census Test. (Navarro, Bates, Scott, and Isaki, 1995.) However, for the 2000 census, we may sample at a smaller rate.

Under the unit design, if we sample at a smaller rate in 2000, we should obtain one or fewer NRFU units in a block quite frequently. In that case, when a NRFU sample unit is a nonresponse, we will have no other sample NRFU units in the block to use as a donor. We need to conduct research to determine whether to expand the donor pool to responding NRFU units in other blocks or to use respondents in the same block as the donor pool. One possibility is to include only respondents with a large number of callbacks in the donor pool. The research will consist of first determining whether within a block the characteristics of respondents and nonrespondents are the same. Secondly, if the characteristics differ, we need to determine whether we can obtain "better" donors for the nonrespondents to the NRFU sample by using NRFU sample cases from other blocks. In

this case, we will determine whether nonrespondents from different blocks are the same. We will also determine whether the nonrepondents in a block are more like the respondents in the block or the nonrespondents in other blocks. The outcome of this research will also determine whether the donor pool for nonsample NRFU cases should be from both the self-response universe and NRFU sample cases.

Depending on the outcome of this exploratory research, the Bureau may opt to impute for nonrespondents to the NRFU sample using a model-based imputation approach such as is being developed by Schafer (1995). The Bureau may also opt to use all respondents as the donor pool followed by an adjustment for differences in respondents and nonrespondents. (Causey, 1995.)

Under the block design, we will have fewer instances where there are no responding NRFU sample units to use as donors within a block. However, research similar to that described for the unit design is still relevant and may be conducted.

V. Conclusions

The 1995 Census Test experience demonstrates that the two anticipated fundamental changes to the 2000 census have implications for other census methodologies. To ensure the 2000 Census provides the best quality census estimates, we need to use the decisions reached with the 1995 Census Test to research how best to redesign other census methodologies to obtain statistically valid census estimates.

ACKNOWLEDGEMENTS: The authors are grateful to Charles Alexander, Dennis Schwanz, Susan Love, Susan Ammenhauser, Maureen Lynch, and Ellen Katzoff for providing valuable comments and to Lisa Mundy and Hazel Beaton for preparing this manuscript for publication.

REFERENCES:

CAUSEY, B. (1995) "Numbers of Persons in Nonrespondent Households as Needed for Block Person Totals," Draft Internal Document, Bureau of the Census.

FULLER, W.A., ISAKI, C.T., Tsay, J.H. (1994) "Design and Estimation for Samples of Census Nonresponse," Proceedings of the 1994 Annual Research Conference and CASIC Technologies Interchange, Arlington, Virginia.

IKEDA, M. (1995a) "Computer Specifications for Site Level Post NRFU Estimates for the 1995 Census Test," 1995 Census Test Memorandum Series IS-7--Revision 1, Bureau of the Census.

IKEDA, M. (1995b) "Proposal for Research into Effects of Alternative Search Match Primary Action Algorithms," Draft Internal Memorandum, Bureau of the Census.

LOVE, S.P. (1990) "Description of the Decennial Census Algorithm for the Selection of the Primary and Supplemental Records from the 1990 FOSDIC Data Capture Files," Internal Memorandum from Love through Jackson for Miskura dated November 27, 1990, Bureau of the Census.

NAVARRO, A. (1995) "Specification for Computing Nonresponse Followup Sampling Weights," DSSD 1995 Census Test Memorandum Series #A-22, Bureau of the Census.

NAVARRO, A., BATES, L.M., SCOTT, J.B., ISAKI, C.T. (1995) "Computer Specifications for Sampling the Nonrespondents in the 1995 Census Test (Block and Unit Samples)," DSSD 1995 Census Test Memorandum Series #A-17, Bureau of the Census.

PETRONI, R.J. (1994) "Search Match/Primary Selection Algorithm Recommendation," Internal Memorandum from Singh for Thompson dated September 13, 1994, Bureau of the Census.

SCHAFER, J.L. (1995) "Model-Based Imputation of Census Short-Form Items", presented at the 1995 Annual Research Conference, Bureau of the Census.

SCHINDLER, E. (1995) "Computer Specifications for ICM Estimation for the 1995 Census Test," 1995 Census Test Memorandum Series IS # 8--Revision 1, Bureau of the Census.

SPENCER, G. (1995) "Revised 1995 Census Imputation Specifications," Internal Memorandum from Spencer for Phillipp dated June 19, 1995, Bureau of the Census.

THOMPSON, J.H. (1994) "1995 Test Census Design," presented at the 1994 annual meeting of the American Statistical Association, Toronto, Canada.

TREAT, J.B. (1994) "Summary of the 1990 Census Imputation Procedures for the 100% Population and Housing Items," DSSD 1990 REX Memorandum Series # BB-11, Bureau of the Census.

WAJER, S.C. (1993a) "Results from the 1990 Search/Match Operation: Add Rates and Erroneous Enumeration Rates by Search Form Type," DSSD 1990 REX Memorandum Series # T-24 and # U-7, Bureau of the Census.

WAJER, S.C. (1993b) "Final Results from the 1990 Primary Selection Algorithm Review," DSSD 1990 REX Memorandum Series # LL-8, Bureau of the Census.

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