Validation of Causality for the Dual System Estimation Procedure in the Census 2000 Dress Rehearsal

Sam Hawala, Bureau of the Census
Planning Research and Evaluation Division, Room BH120/2, Washington, D.C. 20233

1. SUMMARY

The Census 2000 Dress Rehearsal used the dual-system estimation method in producing official numbers for Sacramento, Menominee and for the purpose of coverage measurement in South Carolina. The application of the dual system estimation method requires assuming that there are two independent lists of the population. The first list is the original census enumeration, and the second is a list of those covered by the sampling frame for the sample of the Integrated Coverage Measurement survey in Sacramento and Menominee, and the Post Enumeration Survey in South Carolina.

The independence assumption can fail due to causal dependence, or contamination, between the two lists. Contamination occurs when the event of an individual's inclusion or exclusion from one list affects the probability of their inclusion in the other list. For example, some people who did answer the census may not cooperate with the Integrated Coverage Measurement/Post Enumeration Survey, thinking that they had helped enough. Occasionally a survey sample block resident is asked a question before census day to confirm a survey address listing, making them aware of the census before it actually happens. A small fraction of census followup contacts are made after the beginning of the survey interviewing.

In this evaluation we engaged in testing the validity of the assumption of no contamination. We tested whether Integrated Coverage Measurement/Post Enumeration Survey areas differ from areas where no survey was done. To test for differences between survey blocks and non-survey blocks in the census data, survey blocks are matched with non-survey blocks. The matching was done with respect to the size, or pre-census number of housing units in the block. We matched each Integrated Coverage Measurement/Post Enumeration Survey block with several independent non-survey blocks. After the matching was performed, differences in relevant block level variables were estimated and tested. Estimates of the effects of the Integrated Coverage Measurement/Post Enumeration Survey are found by studying differences in the responses to the census for survey and non-survey blocks. Relevant data are extracted from the census files and aggregated to block level records. The resulting data are then tested for possible differences in the responses.

No significant differences were found in housing unit status and respondent reaction indicators. Overall, no differences attributable to the survey were found when comparing blocks in which the survey was conducted and matched blocks not included in the survey area. Our tests found no evidence of contamination of the census data by the Integrated Coverage Measurement/Post Enumeration Survey.

2. BACKGROUND

The independence assumption is an important factor in the Census 2000 Dress Rehearsal estimation procedure. Violation of the independence assumption will introduce bias into the resulting population estimates (Sekar and Deming 1949) and (Wolter 1986).

The estimation procedure is the capture-recapture (dual system) method which relies on two independent phases. The census, or initial enumeration, phase and the Integrated Coverage Measurement (ICM) or Post Enumeration Survey (PES) phase. (See Appendix for an overview of ICM/PES). From this point on, what I say for ICM is also true for PES. One can combine the data from these two phases in a 2×2 table of counts cross-classifying presence or absence in the initial phase enumeration with presence or absence in the ICM phase. One cell is missing - the fourth cell represents the count of those missed in both the initial phase and the ICM phase. The dual-system technique assumes independence which translates into a fixed odds-ratio for this table and estimates the fourth cell from this odds-ratio.

	CENSUS ENUMERATION		
ICM	IN	OUT	TOTAL
IN	N_{11}	N_{12}	N_{1+}
OUT	N_{21}	$N_{22} = ?$	N_{2+}
TOTAL	N_{+1}	N_{+2}	N_{++}

Note: Persons in the ICM sample of blocks counted in the initial phase are referred to as the E- sample, and persons counted in the ICM phase as the P-sample.

There are three components to the independence assumption: Causality, Homogeneity, and Autonomy.

• Causality: The probability of an individual being included in the P-sample is not altered by inclusion in the E-sample, and vice-versa. Failure of this assumption can result in a bias in the estimation of the number of people missed in the census. For example, if inclusion in the P-sample tended to increase an individual's chance of being included in the E-sample, then an understatement of the undercount would result. If

inclusion in the P-sample tended to decrease an individual's chance of being included in the E-sample, then an overstatement of the undercount would result.

• Homogeneity: The probability of inclusion in the E-sample and the probability of inclusion in the P-sample are equal for all individuals. Failure of this assumption is generally thought to result in an understatement of the undercount.

• Autonomy: The inclusion in the E-sample or the P-sample for an individual is not affected by the inclusion of other individuals. There is no clustering effect due to households, blocks, geographic region, etc. Failure of this assumption is generally thought to increase the level of random error associated with the estimated undercount.

The first two components are necessary, but not the third, for the dual system estimation model. The attempt to control biases due to lack of independence is done in several ways. The causality assumption is addressed by conducting the ICM independently from the initial phase after the major initial phase operations are completed. The homogeneity assumption is addressed by preparing dual system estimates for poststrata where the inclusion probabilities are thought to be similar. Finally, the autonomy assumption is known to fail but does not by itself cause a bias. Jackknifing block clusters for standard error estimates takes care of the lack of autonomy. The present evaluation focuses on assessing whether the causality assumption is met. The study proposes to determine if ICM had any effect on the initial phase results.

The initial phase could be affected by respondent reaction to ICM field operations, or through a difference in processing of initial phase data for ICM sample blocks. To insure some measure of independence, all ICM operations should begin at the conclusion of census operations. Since this is not feasible given the complexity and timing requirements of ICM, some ICM operations begin before the census is complete.

For the most part, it seems reasonable to presume that the ICM could not influence initial phase results, because initial phase enumerations are carried out before most respondents are aware of the ICM. However the following actions may happen. An ICM sample block resident is asked a question before Census Day to confirm an ICM address listing. Census followup contacts are made after the beginning of ICM interviewing. The initial phase results may also be affected in Sacramento because there is a 100of sampling for nonresponse followup in non-ICM blocks. These actions alone are sufficient causes to pursue the contamination issue.

The question is then one of assessing whether the ICM procedures have an effect on responses to the initial phase. Those initial phase response variables which can be affected by failure of the independence assumption are investigated using hypothesis testing. Large differences between initial phase data for the ICM and the non-ICM blocks may be an indication of the failure of the independence assumption.

3. METHODOLOGY

We tested whether ICM areas differ from areas where no ICM was done, using specific census variables. To test for differences between ICM and non-ICM blocks in the initial phase data, the 1990 contamination study (Davis 1990) and the study done for the 1995 census test (Griffiths 1996) both selected a sample of ICM blocks and matched each of these blocks with one non-ICM block. For the 1990 study the matching was done with respect to the number of housing units as measured by pre-census counts. For the 1995 study the matching was based on a model which predicted block nonresponse rates, and the blocks were matched on similarity of predicted nonresponse rates. After the matching was performed, differences in relevant 1990 or 1995 block level variables were tested. Non-ICM blocks acted as a control group, and the ICM blocks acted as a treatment group, or group of cases.

For the Census 2000 Dress Rehearsal, we matched each ICM block with several independent controls. The use of more than one non-ICM block, for each ICM block, provide valuable information in addition to increasing power and precision (Biometrics 1969), especially considering that the pool from which to choose non-ICM blocks is large enough to allow for the selection of at least two non-ICM blocks for each ICM block.

Estimates of the effects of ICM are found by studying differences in the responses to the initial phase for ICM and non-ICM blocks. Relevant data are extracted from the initial phase files and aggregated to block level records. The resulting data are then submitted for estimation and testing for no differences in the responses.

3.1. **Matching the ICM Blocks.** The non-ICM blocks are selected with the criteria that paired blocks are within the same cell, i.e. same site and county, share the same type of data collection method (Mailout/mailback or Update/leave) and sampling stratum (SS). Moreover, we apply a modified nearest available pair-matching method (Rubin 1973a) . The method first puts the ICM blocks of a cell in a random order, then assigns the closest match for each ICM block, from the yet unmatched and randomly ordered non-ICM blocks of the cell. We run the procedure twice initially, defining closest match to mean that the blocks have identical pre-census number of housing units. We run the procedure again two more times, defining closest match to mean that the blocks have pre-census number of housing units that are within a fraction of the pooled standard deviation of these numbers for the two groups of ICM and non-ICM blocks in the cell.

The two groups of blocks to be compared will obviously differ in some variables that are not considered in the matching procedure. Notwithstanding, the randomizations

performed in our matching procedure can diminish to a tolerable level the average effects of the remaining disturbing variables (Cochran 1965). Ideally one would match blocks with the same base response values, or response values observed if ICM is not conducted in the block. We do not however observe the base response values for blocks where ICM is conducted.

3.2. **Extraction of Data from Initial Phase Records.** The initial phase data used in the analyses are extracted from the final census data files. The data extracted from these files are comprehensive yet focused on relevant measures of potential ICM impact. Some information on the files, such as street address or name, are irrelevant to the focus of this study and are therefore not considered for the analysis. In order to be included in the preliminary list of variables under consideration for hypothesis testing, a variable must:

- be related to respondent reaction,
- be related to a potential difference in processing

All variables on the Census Edited File which meet the above criteria are extracted for the preliminary list of variables. Edit and substitution flag variables are not part of this list.

3.3. **Computation of Block-Level Variables.** Person and housing unit data from the Census Edited Files are collapsed to the block level. Most of the block level variables are expressed as a proportion of units in a block rather than as a count. The denominators could be, as appropriate, one of either, the total number of housing unit IDs, the total number of 'data defined' persons, or the total number of persons in the block.

3.4. **Organization of the Variables.** The variables used are grouped in either population coverage, housing unit status, or respondent reaction variables.

- Population coverage: This group of variables includes: Proportion Householders Proportion Females
- Housing Unit Status: This group of variables includes: Proportion of Occupied Units Proportion of Vacant Units
- Respondent Reaction: This group of variables includes a measure of the respondent's promptness in returning the census questionnaires, and the proportions of the different forms used. Average Number of Days For Mail Return Proportion of Long Form Returns Proportion of Be Counted Forms Equivalent

3.5. **Estimation and testing for the effect of ICM.** The objective is to estimate the effect of ICM on initial phase response variables. We use the difference between

the proportion in the ICM block and the average proportion in the set of matched non-ICM blocks. We calculate the estimates and perform the tests independently for each cell. A cell is defined by the site, the county, the type of data collection method, and the ICM sampling stratum.

4. LIMITATIONS

This project examines differences in response measures; however, a difference in response between ICM and non-ICM blocks does not necessarily indicate a violation of the causality assumption. If we can assume that response characteristics and probability of inclusion are related, then a finding of no difference in response measures between ICM and non-ICM blocks would imply that ICM does not affect E-sample inclusion probabilities. This would give us reason to have confidence in the causality assumption.

5. RESULTS

Based on large sample t-tests and an overall 10% level of significance no differences were observed between areas where the ICM survey was conducted and areas where the survey was not done.

6. CONCLUSIONS/RECOMMENDATIONS

6.1. **Conclusions.** The main purpose of this study originates from the concern that knowledge of or exposure to ICM/PES operations may have affected some reactions in ICM block residents and members of field and processing staffs which influenced responses to the census. Any systematic impact on census outcomes, or contamination, would introduce an error or bias into coverage measurement estimates. We engaged in comparisons of initial phase responses to suitable variables, measured where ICM/PES was done and where it was not done, to discover if the concern is warranted. A conclusion is that no differences attributable to the ICM or PES have been found.

6.2. **Recommendations.** In Census 2000 the survey comparable to the ICM/PES will be called the Accuracy and Coverage Evaluation Survey (A.C.E.). The recommendations to continue prevention of contamination are:

During listing for the A.C.E. survey, the level of contact should remain as low as possible.

Overlap between the A.C.E. survey and census field operations should be minimized, using procedures and scheduling similar to that of the dress rehearsal.

In general, the census treatment of households included in the A.C.E. survey should be as similar as possible to that of households not included in the survey. The differences between the A.C.E. survey and census processing should be transparent not only to the survey households, but also to those conducting the survey and processing the survey data. Special diligence is important during the time when census and survey activities overlap.

7. REFERENCES

• Cochran, W. G. (1965) "The Planning of Observational Studies of Human Populations", *Journal of the Royal Statistical Society*, Series A 128, 234-255.

• Davis, M. (1990) "Final Report for Post Enumeration Survey Evaluation Project P14, Part I: Independence of the Census and P Sample: Comparison of Blocks." Unpublished report, July 9, 1991, Washington, DC: Bureau of the Census.

• Griffiths, R. (1996) "Results from the 1995 Census Test: The Contamination Study - Project 11." Unpublished report, March 8, 1996, Washington, DC: Bureau of the Census.

• Rubin, B. D. (1973a) "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159-183.

• Sekar, C. C. and Deming, W. E. (1949), "On a Method of Estimating Birth and Death Rates and the Extent of Registration," *Journal of the American Statistical Association,* 44, 101-115.

• Wolter, K. M. (1986), "Some Coverage Error Models for Census Data," *Journal of the American Statistical Association*, 81, 338-353.

8. Appendix: Census 2000 Dress Rehearsal Overview

The Census 2000 Dress Rehearsal is the culmination of the Census 2000 testing program which began shortly after the 1990 Census was completed. It was conducted in Columbia, South Carolina and eleven surrounding counties; Menominee County, Wisconsin; and Sacramento, California. Each dress rehearsal site was selected because of its demographic and geographic characteristics to provide experience with some of the expected Census 2000 environments. Each site is using a different mix of census and statistical procedures. The dress rehearsal will provide information to assess procedures used in the individual sites but not for comparisons between sites.

8.1. **Site Selection.** The Columbia, South Carolina site includes the city of Columbia in its entirety, including a small portion in Lexington County; the entire town of Irmo, which is in Richland and Lexington Counties; and the following 11 contiguous counties

in North Central South Carolina: Chester, Chesterfield, Darlington, Fairfield, Kershaw, Lancaster, Lee, Marlboro, Newberry, Richland, Union.

This site was selected because it contains living situations and socioeconomic characteristics that are not found in a predominately urban environment. The site provides the only opportunity to test procedures for developing the census address list in an area containing a mixture of house number/street name and rural route and box number addresses. Since this site has a relatively high proportion of African Americans, it provides an opportunity to test Census 2000 procedures for reducing the differential undercount for this racial group. The site selected represents the size of typical local census offices planned for Census 2000, which was necessary to provide an understanding of the effectiveness of census operations.

The Menominee County, Wisconsin site includes the Menominee American Indian Reservation. This site was selected because of the Menominee Reservation. A very high proportion of the population living on the Reservation are American Indians. Also, the Census Advisory Committee on the American Indian and Alaska Native Populations recommended the Menominee Reservation. Conducting the dress rehearsal on an American Indian reservation allows for the testing of Census 2000 methodologies for reducing the differential undercount for reservations, which exceeded 12 percent in the 1990 Census.

The Sacramento, California site was selected because it contains great diversity among racial and ethnic groups. Selecting a site with a diverse population provides the opportunity to test Census 2000 methods designed to reduce the differential undercount and produce an accurate census for all components of the population. Sacramento is a primary media market, which allowed a full test of the paid promotion program. Also, this site represents the size of typical urban local census offices planned for Census 2000, which allowed for an understanding of the effectiveness of census operations in this type of office.

The relative sizes of each of the sites are estimated to be:

	Sacramento	South Carolina	Menominee
Housing Units	153,000	252,000	1,700
Population	374,000	667,000	4,600

8.2. **Methodology by Site.** The Dress Rehearsal involved operational testing of the Headquarters, Regional Census Center, Local Census Office, and Data Capture Center procedures and systems in a census-like environment. Several procedures new and enhanced since the 1990 Census, such as user-friendly forms, digital capture of forms, statistical sampling and estimation, and paid advertising, were tested individually prior to the Dress Rehearsal.

	Sacramento	South Carolina	Menominee
Mailout/mailback	X	X	
Update/leave/mailback		X	X
100% Nonresponse followup		X	X
Sample Nonresponse followup	X		
Integraged Coverage Measurement	X		X
Post Enumeration Survey		X	

8.3. **Mailout/Mailback and Update/Leave/Mailback.** Two questionnaire delivery methodologies were used in the Dress Rehearsal. The mailout/mailback methodology involved delivery by the United States Postal Service. There were four components of the mailout/mailback delivery: an advance letter, an initial questionnaire, a reminder card, and a "blanket" replacement questionnaire (mailed to all addresses). The update/leave/mailback methodology involved Census Bureau enumerators delivering the questionnaires at the same time they updated maps and the list of addresses. In addition to the delivery of questionnaires by Census enumerators, the U.S. Postal Service delivered an advance letter and a reminder card to all "Postal Patrons" within the update/leave area. Under both delivery methodologies, respondents were asked to mail back their questionnaires in provided envelopes.

Short and long form questionnaires were included in both delivery methodologies. Every household received either a short or a long form. The observed rates for the dress rehearsal sites varied since the sampling rate varied by size of place. Using the same sampling plan for the entire nation for Census 2000, we would expect about 17 percent of all households to receive a long form.

8.4. **Nonresponse Followup - 100%.** In Columbia, South Carolina and Menominee, Wisconsin a 100 percent followup for households not returning their Census questionnaires was used. This procedure sent enumerators to all addresses and/or locations that received a mailed questionnaire, or had a questionnaire delivered in person, and did not return it by a specified date. This followup operation continued until a response from the household was received, or proxy information was obtained.

8.5. **Sample Nonresponse Followup.** In Sacramento, California a sample nonresponse followup was used. Instead of following up all addresses that did not return a census form, enumerators followed up a sample of nonrespondents. The sample was designed so that each census tract reached a final completion rate of at least 90 percent.

For example, if the initial completion rate in a census tract was 60 percent, then a 3-in-4 systematic sample of nonrespondents was selected to reach the 90 percent

completion target. If a census tract had at least an 85 percent initial completion rate, then the sampling rate was 1-in-3.

Followup continued for these sample-selected households until a response is received, or proxy information was obtained. Through statistical estimation techniques, responses for all of the other nonresponse households were derived from the sample responses. Additionally, all nonrespondent addresses that were added to the address list too late to be mailed a census form were also included in the nonresponse followup sample.

8.6. **Integrated Coverage Measurement (ICM).** This operation is independent of the Dress Rehearsal operation and was used in Sacramento, California and Menominee, Wisconsin. There are three phases to this operation: Housing Unit phase, ICM Interview phase, and Person Matching and Reconciliation phase. The Housing Unit phase compiled a list of housing units (within selected sample blocks) confirmed to be in existence on census day. This list was created independently of the Dress Rehearsal, using a workforce that was independent of the one used for Dress Rehearsal. Enumerators used this list to conduct the second phase. During the Interview phase, enumerators collect information about current residents and those who moved out of the sample block between census day and the time of the interview. This interview is done by personal visit or telephone. The Person Matching and Reconciliation phase involves matching persons enumerated in the ICM process with persons enumerated during the Dress Rehearsal operations. Selected cases warranted additional interviews to reconcile differences in recorded data.

After completing all matching, estimates were developed of the number of people missed or duplicated during the Dress Rehearsal operations by means of statistical methods. These ICM results were incorporated into the final Dress Rehearsal enumeration to produce "one number" estimates of the population.

8.7. **Post Enumeration Survey (PES).** This operation was independent of Dress Rehearsal operations and was used in Columbia, South Carolina. Like the ICM, it was used to develop an estimate of the undercount from Dress Rehearsal operations.

The field procedures used for PES were very similar to those used during the ICM. The difference between the two is that the PES estimates were not used to produce the final Dress Rehearsal results.