ERROR PROFILE FOR THE CENSUS 2000 DRESS REHEARSAL

Susanne L. Bean, Katie M. Bench, Mary C. Davis, Joan M. Hill, Elizabeth A. Krejsa, David A. Raglin,
U.S. Census Bureau

Joan M. Hill, U.S. Census Bureau, Room 1002-F FOB 2, Washington, DC 20233

Keywords: Integrated Coverage Measurement/Post Enumeration Survey, Matching Error Study, Evaluation Followup Interview, Data Collection Mode Study, measurement error.[1]

1. INTRODUCTION

The error profile examines specific sources of error corresponding to the Census 2000 Dress Rehearsal Integrated Coverage Measurement/Post Enumeration Survey (ICM/PES) that are feasible to measure given the design of the ICM/PES. A sample of ICM/PES block clusters in each site was selected (187 total block clusters across three sites) to assess the magnitude of nonsampling error. This is known as the evaluation cluster sample. The errors with regard to the 'one-number census' in Sacramento, CA and Menominee, WI may occur in the initial dress rehearsal enumeration operation (i.e., initial phase), the ICM enumeration (i.e., final phase), or both. Similarly, the errors measured within the South Carolina site may be found in both the census enumeration and the PES activities. In all three sites, the objective of the error profile is to measure error in the ICM/PES process.

The individual sources of error that are isolated and examined separately in this report are data collection (in both the E-sample and the P-sample[2]) and instrument error, certain errors in the processing of data (the focus here is errors from the ICM/PES clerical matching operation), and the effects of alternative data collection modes. These survey measurement and processing errors are evaluated using the following three tools: Matching Error Study, Evaluation Followup Interview, and the Data Collection Mode Study.

Although production and evaluation operational problems made it impossible to conduct any of these studies as originally intended, the error profile evaluation yielded some interesting results.

2. MATCHING ERROR STUDY

One source of processing error in the ICM/PES is clerical person matching error. People collected in the ICM/PES in a cluster are matched to people found by the initial phase in the same cluster. The first step in this process is a computer match, where obvious matches are made and possible matches are identified. The possible matches and remaining nonmatches are then matched clerically to find the less obvious matches, first by lower-level matchers, and then by matching experts.

The Matching Error Study (MES) attempts to measure the error in the clerical matching process by having expert matchers rematch persons within each block cluster in the evaluation cluster sample. The results from the rematching operation are compared to the production results to find differences in match status.

The discrepancy rates between the production and ICM/PES matching operations were less than one percent in each of the three sites: Sacramento, South Carolina, and Menominee. Presumably they would have been higher if the matching experts had not performed a 100 percent quality assurance during the production matching operation.[3] However, the relatively small matching error does suggest that the matching expert coding is highly reliable.

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

[2] An E-sample housing unit is a housing unit which is counted in the initial phase at the time the person matching begins and is in an area included in the ICM/PES sample (Childers, 1998). A P-sample housing unit is one that is listed in the ICM/PES listing book and is confirmed to exist in a block cluster in an ICM/PES sample area.

[3] The original design of the MES was to use matching experts to do a sample quality assurance review of the work done by the clerks and technicians. However, due to last minute production changes which resulted in the matching experts' review of every cluster, the focus of the study shifted to an assessment of matching reliability.

According to the Census 2000 design, after matchers have passed an initial 100 percent quality assurance, matching experts will perform quality assurance on only a sample of cases during production matching. Therefore, the Census 2000 study of matching error is expected to measure the actual magnitude of matching error in the Accuracy and Coverage Evaluation and its subsequent effect on the Census 2000 Dual System Estimation.

3. EVALUATION FOLLOWUP INTERVIEW

The Evaluation Followup Interview (EFU) measures aspects of two types of survey error. The first type is measurement error, the error introduced into the survey process by the interviewer, respondent, and instrument. That error is measured by the Evaluation Person Followup Reinterview, a replication of the ICM/PES Person Followup Interview in a subset of the clusters in the evaluation cluster sample. The production ICM/PES Person Followup Interview is conducted when people from the initial phase and the ICM/PES do not match after the initial clerical person matching operation. The interview collects information to ensure that all correct matches are made and correct residence status are set.

This evaluation attempts to obtain the true match and residence status by giving the clerical matchers a second set of data from the Evaluation Person Followup Reinterview, along with the production Person Followup Interview data, to use when determining the final match and residence status of each person. The comparison of these results with the production data provides an estimate of measurement error in the ICM/PES data.

The second type of error the EFU attempts to measure is production error due to the decision to not conduct a Person Followup Interview for certain people who did not match between the initial phase and the ICM/PES. This study is called the Person Followup Criteria Evaluation. Research in previous Census tests suggested not including these people in the Person Followup Interview, but to code them as residents in the cluster because it was doubtful that useful information would be gleaned from the Person Followup Interview.

The underlying assumption of the Evaluation Followup analysis is that the EFU process results in residence status and match codes that are closer to truth than results of the ICM/PES process. There were several operational components aimed at meeting this assumption: 1) The field operation was conducted using experienced interviewers, 2) An automated system was used to print person workload information and form questions simultaneously for the PFU Criteria Evaluation, thus eliminating error associated with clerical transcription/preparation of the EFU forms, 3) matching experts were used in lieu of matching clerks to assign match codes and residence status, and 4) information from all previous contacts was used to resolve evaluation cases.

Despite the best efforts of the evaluation staff to design the EFU to produce accurate error estimates, limitations should be considered when interpreting results. Because of processing demands, the EFU interview was not in the field until approximately eight months after Census day. The ability to establish "truth" for the portion of the population that we are most concerned with eight months after Census day is highly questionable due to recall error. Time lag limitations also apply to the ICM/PES production results, since the ICM/PES Person Followup interview directly precedes the Evaluation Followup Interview. As a result, error attributable to the ICM/PES may be understated. An additional limitation of the EFU Interview is the data collection mode, that is, paper and pencil interview as opposed to a Computer Assisted Personal Interview (CAPI). The paper form limits the ability to ask complex question skip patterns, which could be implemented with a CAPI instrument. The decision to use a paper instrument was driven by unavoidable timing and resource restrictions. In addition to these limitations, the EFU could not be conducted as originally intended and certain analysis could not be performed due to operational problems with the evaluation.

3.1 Evaluation Person Followup Reinterview

If we assume the EFU data are closer to the truth than the production data, residence status changes from ICM/PES to EFU indicate that ICM/PES was in error. To evaluate the magnitude of measurement error introduced by the survey process, we produced crosstabulations of the residence status of each person as determined after the EFU versus the final ICM/PES residence status. When there was a conflict, information collected in the EFU was used to help determine the source of such conflict and the resolution. Based on these crosstabulations, gross difference rates are estimated.

The gross difference rate (GDR) (Forsman and Schreiner, 1991) is the proportion of people whose ICM/PES and EFU assigned codes differed. The GDR is calculated by dividing the total number of conflicts by the total number of cases. Interpretation of the GDR is subjective.

Regarding match codes, the GDR was 9.7 percent (SE=4.6 percent) in Sacramento and 7.5 percent (SE=3.1 percent) in South Carolina. For the residence status, the GDR was 15.2 percent (SE=5.9 percent) for Sacramento and 9.0 percent (SE=2.8 percent) in South Carolina. For Menominee, the GDR was 11.3 percent (no standard error estimate was produced) for both match and residence status.

These results are in the range of moderate concern, but given the reduced sample of clusters for the Evaluation Person Followup Reinterview due to evaluation operational problems (49 of the 187original evaluation block clusters), no specific conclusions can be made from these results. The Census 2000 Evaluation Followup Interview design will take these results into consideration.

3.2 Person Followup Criteria Evaluation

The Person Followup (PFU) Criteria Evaluation was conducted using the EFU form which was a modified Person Followup Interview form. It collected information about all people in the evaluation sample clusters who did not match initial phase people but were excluded from the Person Followup Interview. These results were compared to the production results with regard to changes to match and residence status codes.

In addition, the PFU Criteria Evaluation results were used to recompute the Dual System Estimates (DSEs) which are compared to the production DSEs (based on ICM/PES production data from within the evaluation cluster sample). The purpose was to quantitatively evaluate the effect of the criteria decision on the population estimates and/or coverage factors. The difference between the production DSEs and the recomputed DSEs represent the change in population estimates based on the inclusion of the additional nonmatch cases in followup, assuming the PFU Criteria Evaluation represents truth.

Tables 1 and 2 show the DSEs calculated from the production data within the sample clusters, the DSEs calculated from the PFU Criteria Evaluation results, the differences, and whether or not those differences are significantly different than zero. Data are shown for the site total, tenure, and race/ethnicity. Estimates for age/sex groups were also computed but are not shown in the tables. The DSE population estimates have been corrected using iterative proportional fitting (i.e., raking), and are based on the PES-C estimation method used in the dress rehearsal (Childers, 1998).

No Evaluation PFU Reinterview results were included in these tables. Only PFU Criteria Evaluation changes are included because the PFU Criteria Evaluation sample covered all 187 clusters, whereas the Evaluation PFU Reinterview sample covered a smaller number of clusters. In addition, the main objective is to determine the effect on the estimates by not sending the PFU Criteria Evaluation cases to PFU.

The DSEs shown in the tables, both production and PFU Criteria Evaluation-based, contain data from the 187-cluster evaluation sample and are not equivalent to the official population estimates. These estimates have been weighted to represent the whole sites. Note these DSEs only include people eligible for ICM/PES; group quarters and service-based enumeration areas are not included. Calculations for Menominee are not given because the number of sample clusters in Menominee is small.

Significance was determined at $\alpha = .10$, which is the Census Bureau standard, using the Dunn method of controlling for multiple comparisons. With the Dunn method, the alpha level was divided by the number of comparisons to be made: one for the total, two for tenure, seven for race/ethnicity, and six for age/sex (Toothaker, 1993), to come up with the significance level used in the tests.

As shown in Tables 1 and 2, there were no significant differences in the DSEs calculated using production results versus the PFU Criteria Evaluation results for the 187 Evaluation clusters weighted up to the Census 2000 Dress Rehearsal sites. The results for the age/sex groups were similarly non-significant.

The computation of the DSEs after the PFU Criteria Evaluation interview includes the results of following up on specific cases not followed up in PFU. Since there were no statistically significant differences in the DSEs, there is no reason to believe that sending the PFU Criteria Evaluation cases to PFU affects the DSEs (assuming PES-C estimation methodology). Hence, there is no evidence that the Census Bureau should not use the same criteria to followup people in Census 2000.

Table 1: Comparison of DSEs for Poststrata Marginal Variables, Sacramento

Subgroup	DSEs for Eval Clusters Using Production Results	DSEs for Eval Clusters Using Evaluation Results*	Difference	p-value	Signif
Site Total	403,105 (12,119)	401,483 (12,506)	-1,623 (1,978)	0.41	No
Owner	202,434 (7,968)	202,354 (7,921)	-80 (586)	0.89	No
Renter	200,671 (7,189)	199,129 (7,619)	-1,542 (1,470)	0.29	No
NH White/Other	178,712 (9,195)	177,935 (9,039)	-777 (706)	0.27	No
NH Black	64,025 (3,425)	64,686 (3,649)	661 (395)	0.09	No
NH Amer Ind/Alas Nat	13,156 (541)	13,148 (543)	-8 (57)	0.89	No
NH Native Haw/Pac Isl	2,859 (109)	2,859 (110)	0 (12)	0.97	No
NH Asian	62,291 (1,949)	60,734 (2,780)	-1,557 (1,359)	0.25	No
Hispanic	82,063 (2,904)	82,120 (2,913)	57 (342)	0.87	No

Note: Standard errors are in parenthesis
* Only the PFU Criteria Evaluation results were used in the calculation of the DSEs.

Table 2: Comparison of DSEs for Poststrata Marginal Variables, South Carolina

Subgroup	DSEs for Eval Clusters Using Production Results	DSEs for Eval Clusters Using Evaluation Results*	Difference	p-value	Signif
Site Total	756,533 (46,153)	754,258 (46,104)	-2,275 (2,256)	0.31	No
Owner	554,834 (34,915)	553,362 (36,230)	-1,472 (1,407)	0.30	No
Renter	201,699 (19,039)	200,896 (18,370)	-803 (1,550)	0.60	No
NH White/Other	407,923 (16,047)	405,312 (16,128)	-2,611 (1,968)	0.18	No
NH Black	321,225 (32,824)	321,542 (32,879)	317 (667)	0.64	No
NH Amer Ind/Alas Nat	4,447 (450)	4,452 (451)	4 (10)	0.66	No
NH Native Haw/Pac Isl	435 (44)	436 (44)	0 (1)	0.73	No
NH Asian	7,609 (762)	7,614 (760)	5 (23)	0.83	No
Hispanic	14,893 (1,508)	14,903 (1,506)	10 (41)	0.81	No

Note: Standard errors are in parenthesis
* Only the PFU Criteria Evaluation results were used in the calculation of the DSEs.

4. DATA COLLECTION MODE STUDY

This study attempts to measure error due to collecting ICM/PES Person Interview data over the telephone from the interviewer's home using the computer-assisted instrument as opposed to collecting the data using the same instrument during a personal visit.

The ICM/PES Person Interview was CAPI. It was designed to be conducted in person by the interviewer after the completion of the initial phase Nonresponse Followup to avoid contaminating the initial phase data in ICM/PES clusters. However, to alleviate tight schedule demands it was decided to collect data for selected cases by telephone using the (CAPI) instrument before the Nonresponse Followup was finished and ICM/PES personal visits began. The selected cases included those people who responded to the initial phase by mail early in the process and provided a phone number.

The study was conducted by not allowing data to be collected by telephone for half of the eligible cases in the evaluation sample clusters, while attempting to collect the data by telephone for the other half. The phone and personal visit cases were paired as the sample was selected, and the percentage of matches to initial phase people and item nonresponse rates were compared to attempt to measure if there were significant differences by the mode of data collection in our population of interest.

Because of production problems, the sample size for this evaluation is too small to make any strong conclusions, but we found no evidence that the mode of data collection affected the person match rates or the item nonresponse rates.

5. NET ERROR

The original error profile study design included plans for examining the net effect of a subset of nonsampling error sources. The methodology involved estimating a net nonsampling error and combining it with the sampling, or random, error which occurs because only a sample of blocks (and households) is observed in the ICM/PES (Spencer, et. al., 1998). The subset of nonsampling errors that would have been incorporated into the net effect are as follows: (i) errors in the collection of data, (ii) matching errors, and (iii) random nonsampling error related to estimation which includes the effects of heterogeneity bias, synthetic estimation error, and ratio-estimator bias.

The planned methodology for computing the net error estimates involved using the results of the EFU in conjunction with the MES to determine more accurate residence status and match codes for everyone who was interviewed in the initial and final phases in evaluation sample blocks on Census Day. This process would be used to obtain a lower bound for net error[4], which would include the subset of the components of nonsampling error delineated above.

Due to the changes in the ICM/PES matching QA program during production (i.e., 100 percent QA by matching experts) and a predetermined coding specification for a portion of the production workload (i.e., the classification of whole household ICM/PES nonmatching persons from non-proxy interviews as residents), the estimate of matching error based on the MES and the estimate of data collection/instrument error based on EFU are thought to underestimate the actual error. The magnitude of the separate component errors, which feed into the net error estimate, are substantially smaller than what is expected in a national decennial census. Based on the proposed net error estimation methodology, the lower bound for the net nonsampling error would similarly underestimate the true value. Thus, the net error estimates are not included in this paper but will be provided at a later date.

For Census 2000, the net error lower bound, as well as the component error estimates, are not expected to underestimate the true value as much as in the Census 2000 Dress Rehearsal. Regarding the production matching operation, the QA program is not designed to include a 100 percent block cluster rematch by the matching experts. In addition, the 2000 Census Evaluation Followup Interview is currently being redesigned to ensure an accurate and reliable estimate of data collection error.

6. RELATED RESEARCH

For the 1988 dress rehearsal census of St. Louis and east-central Missouri, Mulry and Spencer offer estimates of the errors of the census, DSE, and undercount estimates. Their 'total error' methodology includes decompositions of error based on the PES into components and summarizing the combined effect of

[4] This "lower bound" is not a mathematical lower bound associated with a confidence interval. Instead, this phrase is used to indicate that the net error estimate is a subset of the total net error.

the component errors in a total error estimate (Mulry and Spencer, 1991).

The total error methodology is similar to the original design of Census 2000 Dress Rehearsal Error Profile in the general format in which the error related to the respective post-coverage surveys (i.e., the 1988 PES and 1998 ICM/PES) is described. The Error Profile estimates the magnitude of a subset of individual sources of error and originally intended to incorporate these isolated errors into a net error estimate which serves as a lower bound for total error. The total error model also examined errors individually and attempted to create a combined estimate.

However, the 1988 Dress Rehearsal total error model and the Error Profile differ in at least one critical way, the scope of the study. The total error approach to estimating error in the DSE is to try to identify all the sources of error, estimate their magnitudes, and study their propagation through the estimation process (Mulry and Spencer, 1991). The seven individual components of error included in the total error strategy are model error, P-sample matching error, error in the P-sample reported census day address, P-sample fabrications, error in the measurement of erroneous enumerations, imputation error, and sampling error. The total error model also considers mixed error, that is, error that arises from a mixture of a kind of measurement error known as balancing error and failure of assumptions, but concludes that mixed error is negligible for the 1988 PES.

This comprehensive and ambitious philosophy differs from the Error Profile in that only a subset of nonsampling error is isolated and estimated in this report. The net error portion of the Error Profile was originally designed to incorporate errors in the collection of data (somewhat comparable to the total error component of errors in the measurement of erroneous enumerations), matching error (comparable to the total error P-sample matching error component), and random nonsampling error related to estimation which includes the effects of heterogeneity bias, synthetic estimation error, and ratio-estimator bias. The Error Profile included the study of data collection mode effects, which is not applicable to the 1998 PES. Other error sources (i.e., fabrication, imputation and sampling error) are not included in the error profile but are examined in separate evaluation reports.

Given this important difference (as well as several others not mentioned here) between the total error model and the Error Profile, 1988 and 1998 dress rehearsal error results should not be compared. The Error Profile was never intended to be a comprehensive, exhaustive delineation of all nonsampling and sampling errors in the ICM/PES, but rather a snap-shot of the major nonsampling errors associated with the final population estimates.

7. REFERENCES

Childers, Danny R. (1998) The Design of the Census 2000 Dress Rehearsal Integrated Coverage Measurement (draft), Bureau of the Census internal memorandum dated May 27, 1998.

Forsman, G. and Schreiner, I. (1991) The Design and Analysis of Reinterview: An Overview. Measurement Errors in Surveys. Ed. Paul P. Beimer et al. New York: John Wiley & Sons, Inc. 279-302.

Mulry, M. H. and Spencer, B. D. (1991) Total Error in PES Estimates of Population. *Journal of the American Statistical Association* **86,** 839-863.

Spencer, B. D., Hill, J. M., Haines, D. E. (1998) Accuracy of Block-Level Estimates of Population (preliminary draft), memorandum dated May 31, 1998.

Toothaker, L. (1993) Multiple Comparison Procedures. Ed. Michael S. Lewis-Beck, Newbury Park: Sage Publications.