ERROR PROFILE FOR PES-C AS IMPLEMENTED IN THE 2000 A.C.E.

Mary H. Mulry and Rita J. Petroni[1]
Statistical Research and Planning, Research, & Evaluation Divisions, Census Bureau, Washington, DC 20233

KEY WORDS: Census, Undercount, Evaluation

1. Introduction

The 2000 Accuracy and Coverage Evaluation Survey (A.C.E.) provided estimates of coverage error for Census 2000. The A.C.E. estimation used the PES-C version of dual system estimator (DSE) with the data collected by the A.C.E. The estimates of coverage error are subject to nonsampling error as well as sampling error. Although the error components for the DSE were studied after the 1990 Census, the introduction of PES-C necessitated identifying new error components for the P sample.

The size of the population from Census 2000 was 281,421,906. The A.C.E. undercount rate estimate published in March 2001 was 1.18 percent. However, the Census Bureau came to believe that the A.C.E. was flawed and in October 2001 issued 0.06 percent as the 'revised early approximation' of the undercount rate (Thompson, Waite, and Fay 2001).

This paper focuses on the March 2001 A.C.E. estimate and identifies the various possible errors in the use of PES-C and classifies them by type and source. The discussion also includes methods of measuring the magnitude of these errors. Previous work on error components for PES-C does not focus on categorizing them by source but rather on the formulation of a total error model (Spencer 2000 and Petroni 2001).

2. Overview of A.C.E. Estimates

The A.C.E. is really two sample surveys, a sample of census enumerations called the E-sample and a sample of the population called the P-sample. The P-sample interviews were conducted between May and September 2000 with a followup in November. The samples overlapped on 11,303 block clusters with 311,029 housing units in the E-sample and 300,913 housing units in the P-sample.

Census 2000 is the first census for which the Census Bureau used the PES-C version of the DSE for estimating census coverage. The 1980 and 1990 implementations of the DSE methodology used the PES-B version. The difference in the versions of the DSE is in the definition of the P-sample and in particular, the treatment of movers . For PES-B, the members of the P sample are those people who live in the sample blocks at the time of the P-sample interview while PES-A defines the P-sample as those people who live in the sample blocks on Census Day. PES-B includes the inmovers in the P-sample and searches for their Census enumeration at the address where they lived on Census Day, which leads to a matching operation extended across the whole country. PES-A includes the outmovers, which confines the matching operation to the area surrounding the sample blocks, but means the information concerning the outmovers is collected by proxy interviews that are usually less reliable than self-responses. PES-C attempts to combine the best of PES-A and PES-B by using the inmovers to estimate the number of movers and the match rate for the movers using the outmovers.

PES-C actually has more nonsampling error components than PES-B (Mulry and Spencer 1993). The reason is there are two types of movers as well as nonmovers. These categories contribute to the DSE in different ways. The nonmovers and the outmovers were matched to the census while the inmovers were not. The nonmovers, outmovers, and inmovers also may not have been listed on the household roster correctly depending on whether they actually resided at the address. Errors may occur in mover status as well as residency status and match status. When they occur, errors may be in mover status, residency status, or match status, giving rise to many error components.

For the estimator, we first define the following notation for each poststratum, h.

C_h = census "count"
II_h = number of persons imputed into the original enumeration
$\hat{I}_{E,h}$ = estimated number of enumerations with insufficient information for matching
$\hat{E}_{E,h}$ = estimated number of erroneous enumerations

[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.

$\hat{N}_{E,h}$ = estimated population size from the E-sample

$\hat{CE}_h$ = estimated population size who could possibly be matched

$\hat{CE}_h = \hat{N}_{E,h} - \hat{I}_{E,h} - \hat{E}_{E,h}$

$\hat{N}_{P,h}$ = estimated size of the P sample population

$\hat{M}_h$ = estimated number of the P sample population enumerated in the census

The dual system estimator for the population size N_h in poststratum h is defined by

$$\hat{N}_h = (C_h - II_h)(\hat{CE}_h/\hat{N}_{E,h})(\hat{N}_{P,h}/\hat{M}_h).$$

The 2000 A.C.E. used the PES-C formulation of the dual system estimator which uses the number of inmovers to estimate the number of outmovers, but uses the match rate for the outmovers to obtain the estimate of the number of outmovers that match the census. The poststratum index h is suppressed in the following formula.

$$(\hat{N}_P/\hat{M}) = (\hat{N}_n + \hat{N}_i)/(\hat{M}_n + (\hat{M}_o/\hat{N}_o)\hat{N}_i).$$

where

$\hat{N}_n$ = estimated number of nonmovers

$\hat{N}_o$ = estimated number of outmovers

$\hat{N}_i$ = estimated number of inmovers

$\hat{M}_n$ = estimated number nonmovers enumerated in the census

$\hat{M}_o$ = estimated number outmovers enumerated in the census

When a poststratum had less than 10 outmovers, the PES-A version of the dual system estimator that does not use inmovers was employed as follows.

$$(\hat{N}_P/\hat{M}) = (\hat{N}_n + \hat{N}_o)/(\hat{M}_n + \hat{M}_o).$$

The census coverage factor for poststratum h is defined as $\hat{A}_h = \hat{N}_h/C_h$. The census estimate for area j is $N_{unadj,j} = \sum_h C_{h,j}$ and the A.C.E. estimate is $\hat{N}_{adj,j} = \sum_h \hat{A}_h C_{h,j}$. The estimate of undercount in the population size of area j is $N_{adj,j} - N_{unadj,j}$ and the estimate of the corresponding undercount rate is

$$(N_{adj,j} - N_{unadj,j})/N_{adj,j}.$$

3. Sources of Error in A.C.E. Estimates

The A.C.E. estimates are subject to a variety of possible sources of error: sampling error, data collection and survey operations error, missing data, error from exclusion of late census data and data with insufficient information for matching, contamination error, correlation bias, synthetic estimation bias, inconsistent poststratification, and balancing error.

3.1. P-Sample Matching Error and E-Sample Processing Error

The term "P-sample matching" has been used to describe the search of the census records for enumerations for P-sample respondents. The P-sample respondents are designated as matching an enumeration in the census or as not enumerated in the census. The counterpart for the E sample is called "E-sample processing" where census enumerations are designated as correctly enumerated or erroneously enumerated. When the status of a P-sample or E-sample case can not be determined, it is designated as unresolved.

"P-sample matching error" refers to the net effect of errors during the processing that affect the determination of whether a P-sample person matches a census enumeration. Likewise, the net effect of errors in assigning enumeration status to E-sample enumerations during the office processing is called "E-sample processing error".

Errors may occur in either direction. P-sample people not in the census may be designated as matching a census enumeration, called a "false match," or people may be designated as not enumerated in the census although they are, called a "false nonmatch". E-sample enumerations may be falsely assigned a correct enumeration status, called a "false correct enumeration," or enumerations may be incorrectly designated as an erroneous enumeration, a "false erroneous enumeration."

Matching error also encompasses errors in the size of the P-sample population that may happen during the processing of the P-sample. These errors also may occur in either direction. A person included as a member of a household may really reside at another location or not be in the population of interest. For example, the census residency rules consider family members away at college to reside at their college address. A family member in a nursing center is considered to be in the group quarters population, which is not part of the population of interest. Vice versa, a person with two homes, may be designated as living at the other home, but really live at the one in the sample.

In the application of PES-C, respondents have the potential of many more statuses than was possible in the processing of the P-sample in 1990. The reason is that a P-sample respondent may be a nonmover, an outmover, an inmover, or an out-of-scope person. The nonmovers and outmovers have another characteristic, which is resident or nonresident. A person who is living at the sample address on Census Day is called a resident.

Errors in mover status may go in all directions. A person designated as a nonmover may be an inmover or an outmover. All combinations of errors may happen and affect the DSE in different ways.

P-sample matching error effects both the estimates of nonmovers and inmovers in the estimate of the size of the P-sample population. In addition, matching error effects the estimates of the number of nonmover matches, the number of outmovers and outmover matches, and the number of inmovers in the estimate of the number of matches. E-sample processing error effects the estimate of the number of erroneous enumerations. The estimates of the net error due to error in P-sample matching and E-sample processing are based on the Matching Error Study (MES). The MES used a subsample of the A.C.E. sample called the Evaluation Sample that contained 2,259 block clusters. The estimated bias in the DSE from the P-sample matching error and E-sample processing error combined was -483,938 (Bean 2001).

3.2. P-Sample and E-Sample Data Collection Error

Errors may occur during the data collection. While an interview is in progress, the respondent may make an error in answering a question, or the interviewer may make an error in asking a question or recording the answer. Errors also occur when an interviewer goes to the wrong address. Regardless of whether the error is caused by the respondent, the interviewer, or a combination of the two, such errors may cause the matching operation to assign mover status, residency status, or match status incorrectly to a person on the household roster. The A.C.E. interviewer collects both a Census Day roster and an interview day roster. A person who resides at the household on both days is classified as a nonmover. A person who lived there only on Census Day is an outmover while a person who lived there only on interview day is an inmover. Persons classified as outmovers and nonmovers may or may not have been a resident at the address on Census Day. Errors in the mover status, residency status, or other errors may cause the matching operation to fail to determine that a person was enumerated and to classify the person as a nonmatch incorrectly.

Sometimes people listed on household rosters do not exist. For example, an interviewer who is having trouble contacting the residents of a housing unit may copy the name from a mail box. This type of error is called "P-sample fabrication". Usually fabricated households cause an underestimate of the match rate because they are smaller than the average household size and do not match.

A special type of E-sample data collection error is the failure to identify duplicate enumerations. The processing includes a search for duplicate enumerations within the block cluster and the surrounding blocks. Duplicate enumerations outside the block cluster and surrounding blocks are more difficult to find. Identifying these duplications requires the respondent to provide information concerning another address where a household member may also be enumerated. Errors may occur when the respondent does not understand the residency rules or is unaware that a household member may be enumerated at another address. The situations prone to causing duplicate enumerations include college students enumerated at their family home and their college address, children in joint custody agreements enumerated at both parents' addresses, and people with two residences (Adams & Kresja 2001).

Another type of field error occurs during the listing of the housing units for the census or for the P-sample. The housing units listed as being in the sample block may be in another block or vice versa. These types of errors are called "geocoding error". To account for minor geocoding errors in 2000, the search for matches occurred within all block-clusters and also in surrounding blocks for a sample of the cases with geocoding errors recorded in the E- sample– a design called "Targeted Extended Search".

P-sample fabrication and data collection error affect both the estimates of nonmovers and inmovers in the estimate of the size of the P-sample population. In addition, fabrication and data collection error effect the estimates of the number of nonmover matches, the number of outmovers and outmover matches, and the number of inmovers in the estimate of the number of matches. E-sample data collection error effects the estimate of the number of erroneous enumerations.

The Evaluation Followup (EFU) was conducted on a subsample of the E- and P-samples in the Evaluation Sample (Raglin and Krejsa 2001a). The best coding of the EFU (Adams and Krejsa 2001) found the A.C.E. had coded 1.4 million too many correct enumerations in the E-sample while the Computer Census Duplication Study estimated 2.9 million too many correct enumerations (Fay 2002). The estimated bias in the DSE from these errors in the E-sample is 1.6 million from the EFU and 3.2 million from the Computer Census Duplication Study. The EFU results for the P sample showed A.C.E. underestimated the resident nonmovers by 0.7 million and overestimated the resident outmovers and inmovers by 1.1 million and 0.5 million, respectively. These results are surprising and believed to be related to a design flaw in the EFU questionnaire. The estimated bias in the DSE from the error in mover status alone is -427,026 (Raglin and Krejsa 2001b).

3.3. Missing Data

A. C. E. data may be missing for a variety of reasons – some A.C.E. interviews fail to take place, some households provide incomplete data on questionnaire items, and in some cases the information for classification as a match or nonmatch is ambiguous. The methods used to compensate for missing data effectively assume that the match status for the case with missing data is equal on average to the status for cases that are similar except that they have complete data. Missing data on characteristics are imputed from otherwise similar cases with complete data. Nonresponse

weighting adjustments are used to account for sampled but non-interviewed households. The P-sample matching and E-sample processing operation assigns "Unresolved" enumeration status to a case when the available data is inadequate to determine whether the person is enumerated in the census, and a probability of being correctly enumerated is imputed for such cases.

Also, error in the resolved cases causes error in the imputations because the resolved cases are used to form the imputations. Even if the imputation model were perfect, the imputations will have error if the data used to fit the model has error. This type of error is called "imputation error due to data error."

The variance component due to imputation for missing data has three components.

V_M = variance due to imputation
$= V_{RA} + V_B + V_I$
V_{RA} = variance due to the imputation model selection
V_B = variance due to the model parameter estimation
V_I = within- person imputation variance.

The imputation variance components due to parameter estimation and within person estimation are included in the sampling error estimates, leaving the variance due to model selection to be estimated separately. The missing data variance-covariance matrix is added to the sampling variance-covariance matrix to obtain a variance-covariance matrix for the coverage correction factors that contained the random error due to sampling and imputation for missing data.

To estimate the variance component, we use the results of the Analysis of Reasonable Alternative Imputation Models**.** The study includes alternative models for the imputation of enumeration status, residency status, and the P-sample noninterview adjustment.

The standard error of the A.C.E. estimate due to imputation model selection when combining ignorable and nonignorable alternative missing data models (531,751) is higher than the sampling error (378,222). Using only ignorable models, the standard deviation (384,115) is approximately the same magnitude as the sampling error. (Keathley, Kearney, and Bell 2001)

3.4. Sampling Error

Sampling error gives rise to random error, quantified by sampling variance, and to a systematic error known as ratio-estimator bias. The sampling variance is present in any estimate based on a sample instead of the whole population. Ratio-estimator bias arises because even if X and Y are unbiased estimators, X/Y typically is biased.

Random sampling error is reflected in the estimated variance-covariance matrix of the coverage correction factors. The covariance matrix is estimated by the Census Bureau's sampling-error software applied to the A.C.E. data. The software also can be used to produce estimates of ratio-estimator bias. The A.C.E. sampling error for the national estimate is 378,222 (Keathley, Kearney, and Bell 2001). No estimate of the ratio estimator bias is available, but it is expected to be small.

3.5. Correlation Bias

If there is variability of the enumeration probabilities for persons in the same poststratum or if there is a dependence between enumeration in the census and in the A.C.E. – e.g., people less likely to be enumerated in the census may also be less likely to be found in the A.C.E. – then correlation bias may arise. Correlation bias is most likely a source of downward bias in the DSE. Evidence of correlation bias in national estimates comes from sex ratios (males to females) for A.C.E. estimates that are low relative to ratios derived from demographic analysis of data on births, deaths, and migration. Robinson (2001) describes the construction of the demographic analysis estimates.

The information from demographic analysis is insufficient to estimate correlation bias at the poststratum level, however, and alternative parametric models have been used to allocate correlation bias estimates for national age-race-sex groups down to poststrata. Estimates of correlation bias at the national level provided by demographic analysis information also account for possible error from groups whose probabilities of enumeration are so low that the DSE will fail to account for them. The estimates of correlation bias based on sex ratios are affected by error in the demographic-analysis sex ratios and by possible other biases in the sex ratios in the DSE. The assumptions and model underlying the measurement of correlation bias are discussed in detail in a paper by Bell (2001).

Correlation bias in A.C.E. caused a significant underestimate for all adult Black males and Nonblack males 30+. For Nonblack males 18-29, the data were inconsistent, but indications were that the bias reasonably can be assumed to be zero. For Blacks, the error rates were 6.91% for 18-29, 8.26% for 30-49, and 4.95% for 50+. For Nonblacks, the error rates were 0.85% for 30-49 and 0.79% for 50+. The estimated bias in the DSE due to correlation bias is -1.26 million (Bell, 2001)

3.6. Excluded-data Error from Census Reinstates

The DSE treats late census data as non-enumerations. Thus, duplicate enumerations among the late data do not contribute to census data but valid enumerations among the late data are treated as census misses and are estimated by the DSE. If the late census data were excluded from the entire A.C.E. processing and estimation, no new source of error would be present. The A.C.E. estimates do partially incorporate late census data, by including them in the census in synthetic estimation for an area but excluding them from the computation of the DSE. This use of late data affects the estimates for areas with disproportionately many or few

late adds, with an effect that is similar to synthetic estimation error. In addition, the exclusion of late census data from the E-sample could bias the estimates at the poststratum level. There are two conditions that have to be met for the exclusion of the late adds from the processing of the A.C.E. not to bias the dual system estimates at the poststratum level:

- The P-sample covers the correct enumerations among the late adds at the same rate as other correct enumerations.
- The late adds occur in the E-sample at the same rate as they occur in the census (excluding the imputations)

Exclusion of reinstated cases caused A.C.E. to overestimate the coverage rate by only 0.034% to 0.082%. Thus, the estimated bias in the DSE due to excluding the reinstated cases is between 96,792 and 233,441 (Raglin 2001).

3.7. Contamination Error

Contamination occurs when the A.C.E. selection of a given block cluster alters the implementation of the census there and affects enumeration results, e.g, by increasing or decreasing erroneous enumerations or by increasing or decreasing coverage rates. Contact with residents of the sample blocks during the listing for the P-sample may cause them to not respond to the census because they think that the listing contact is a response to the census. The Census Bureau compares census response characteristics for sample blocks with those not in the sample to assess whether contamination error may be present. The analysis indicated that contamination error was negligible. (Bench 2001)

8. Synthetic Estimation Bias

The A.C.E. estimation methodology for small areas relies on a method called synthetic estimation to provide the same coverage correction factor for all enumerations in a given poststratum, regardless of whether the enumerations are from the same geographic area. Synthetic estimation bias arises when the census from different areas but in the same poststratum should have different coverage correction factors.

Synthetic estimation may cause a bias in the estimates from the A.C.E. for an area. Error from synthetic estimation does not affect the dual system estimate for a poststratum, only areas within a poststratum. A study with artificial populations using 1990 data showed decisions on whether census or adjusted state counts had less error using squared error loss functions were not affected by synthetic bias. (Griffin and Malec 2001)

3.9 Inconsistent Poststratification

The computation of the correct enumeration rate requires census enumerations to be assigned to poststrata, and the computation of the match rate requires P-sample enumerations to be assigned to poststrata. When the assignments are not made consistently for the two samples, error arises in the match rate. Inconsistent assignments to poststrata may be caused by mis-reporting of characteristics used in poststratification.

Cases prone to inconsistent classification are those where there is a different respondent for the household in the census and the A.C.E. For example, a household member's age or race may be reported differently in a self-response than when another household members responds for the person. However, individuals may report their age or race differently depending on the circumstances at the time of the response. Such inconsistencies also may be due to computer processing errors.

The matches in the A.C.E. sample provide a source of data for estimating the error due to inconsistent poststratification. An estimate of the error for a poststratum may be formed by assuming the inconsistency rate observed in the matches also holds for those not matched. The error is expected to be small, but no estimate is available at this time.

3.10 Error from Estimating Outmovers with Inmovers

This error is unique to the PES-C model used in the ACE. For the PES-C model, the members of the P-sample are the residents of the housing units on Census Day. There is some difficulty in identifying all the residents of all the housing units on Census Day because some move prior to the A.C.E. interview. The A.C.E. interview relies on the respondents to identify those who have moved out, the outmovers. Since the outmovers are identified by proxies, many of the outmovers are not recorded. Therefore, the estimate of outmovers is too low. PES-C uses the number of inmovers to estimate the number of outmovers. The inmovers are those who did not live in the sample blocks on Census Day, but moved in prior to the A.C.E. interview. Theoretically the number of inmovers in the whole country should equal the number of outmovers. However, the number of inmovers may not equal the number of outmovers in a poststratum because of circumstances such as economic conditions causing more people to move out of an area than to move into an area. The error due to using the inmovers to estimate the outmovers affects the estimates of the size of the P-sample population and the number of matches. The error is expected to be small, but no estimate is available at this time.

3.11 Balancing Error

Balancing error must be addressed in the design of the search areas used to search for E-sample correct enumerations and P-sample matches. Limiting the search for correct enumerations and matches is necessary because the matching operation cannot search the entire census. By limiting the search area, a small percentage of correct enumerations will not be found and a small percentage of

matches will not be found. This causes an underestimate of the correct enumerations and an underestimate of the matches. However, the estimate of the net error is not biased as long as the percentage error in the correct enumerations equals the percentage error in the matches. The A.C.E. design avoids balancing error by choosing the same block clusters for the E-sample and the P-sample and drawing the search areas consistently.

There is not a separate measurement of balancing error. Any balancing error that may arise during the implementation of the A.C.E. will be included in the measurement of data collection error.

References

Adams, Tamara, and Krejsa, Elizabeth A. (2001) "ESCAP II: Results of the Person Followup and the Evaluation Followup Forms Review." Executive Steering Committee For A.C.E. Policy II, Report No. 24. dated October 12, 2001. Census Bureau.

Bean, Susanne L. (2001) "ESCAP II: Accuracy and Coverage Evaluation Matching Error." Executive Steering Committee For A.C.E. Policy II, Report No. 7. dated October 12, 2001. Census Bureau.

Bell, William (2001) "ESCAP II: Estimation of Correlation Bias in 2000 A.C.E. Using Revised Demographic Analysis Results" Executive Steering Committee For A.C.E. Policy II, Report No. 10. dated October 13, 2001. Census Bureau.

Bench, Katie (2001) "ESCAP II: Conditioning of Census 2000 Data Collected in Accuracy and Coverage Evaluation Block Clusters." Executive Steering Committee For A.C.E. Policy II, Report No. 14. dated October 19, 2001. Census Bureau.

Fay, R. (2002) "ESCAP II: Evidence of Additional Erroneous Enumerations from the Person Du0plication Study." Executive Steering Committee For A.C.E. Policy II, Report No. 7. dated March 27, 2002. Census Bureau.

Griffin, Rick and Malec, Donald (2001) "ESCAP II: Sensitivity Analysis for the Assessment of the Synthetic Assumption." Executive Steering Committee For A.C.E. Policy II, Report No. 23. dated October 12, 2001. Census Bureau.

Keathley, Don, Kearney, Anne, and Bell, William (2001) "ESCAP II: Analysis of Missing Data Alternatives for Accuracy and Coverage Evaluation." Executive Steering Committee For A.C.E. Policy II, Report No. 12. dated October 11, 2001. Census Bureau.

Mulry, Mary H. and Spencer, Bruce D. (1993) "Accuracy of the 1990 Census and Undercount Adjustments." *Journal of the American Statistical Association*, 88, 1080-1091.

Petroni, Rita J. (2001) " Measuring Quality in the U. S. Census 2000 Dual System Estimator Using a Total Error Model." *Proceedings for The International Conference on Quality in Official Statistics*, Statistics Sweden, Stockholm, May 14-15, 2001.

Raglin, David A.(2001) "ESCAP II: Effect of Excluding Reinstated Census People from the A.C.E. Person Process." Executive Steering Committee For A.C.E. Policy II, Report No. 13. dated October 9, 2001.

Raglin, David A. and Krejsa, Elizabeth A. (2001a) "ESCAP II: Evaluation Results for Changes in A.C.E. Enumeration Status." Executive Steering Committee For A.C.E. Policy II, Report No. 16. dated October 15, 2001. Census Bureau.

Raglin, David A. and Krejsa, Elizabeth A. (2001b) "ESCAP II: Evaluation Results for Changes in A.C.E. Enumeration Status." Executive Steering Committee For A.C.E. Policy II, Report No. 16. dated October 15, 2001. Census Bureau.

Robinson, J. Gregory (2001) "ESCAP II: Demographic Analysis Results." Executive Steering Committee For A.C.E. Policy II, Report No. 1. dated October 13, 2001. Census Bureau.

Spencer, Bruce D. (2002) "Total Error Model for Census 2000: How Component of Error Can Be Estimated from the Bureau's Planned Evaluation Studies." Paper prepared by Abt Associates Inc. and Spencer Statistics, Inc. under Task Number 46-YABC-7-00001, Activity 12.18, Contract Number 50-YABC-7-66020. March 20, 2002.

Thompson, J., Waite, P. and Fay, R. (2001) "ESCAP II: Basis of 'Revised Early Approximation' of Undercounts Released Oct. 17, 2001." Executive Steering Committee For A.C.E. Policy II, Report No. 9a. Dated October 26, 2001. Census Bureau.