UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

October 30, 2008

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-E-20

MEMORANDUM FOR Donna L. Kostanich, Co-Chair
Census Coverage Measurement Operational Integration Team

Magdalena Ramos, Co-Chair
Census Coverage Measurement Operational Integration Team

From: Richard A. Griffin *(signed)*
Chair, Census Coverage Measurement Estimation Subgroup

Prepared by: Richard A. Griffin and Tom Mule
Decennial Statistical Studies Division

Subject: Spurious Events in Dual System Estimation

The attached document on "Spurious Events in Dual System Estimation" is being provided as background material for the Census Coverage Measurement Estimation Workshop to be held in January 2009.

Attachment

cc:
DSSD CCM Contacts List

Spurious Events in Dual System Estimation

I. Introduction

The U.S. Census Bureau uses dual system estimation to estimate the population of people in the United States. Dual system estimation produces valid population estimates as long as certain assumptions based on the chosen model hold. The application of dual system estimation for the Accuracy and Coverage Evaluation (A.C.E.) of Census 2000 showed the bias implications when one of those assumptions concerning spurious reports is violated.

Section II provides some documentation on dual system estimation and the spurious events assumption. This section lists the measures taken in the 2000 A.C.E. as part of the production of the March 2001 estimates to address this assumption and minimize the bias from its violation. The issue of how to minimize the bias based on the inclusion of spurious reports in either system has been addressed in several published papers on capture-recapture methods. This section also documents from these papers how the relative bias in the estimate can be assessed based on the spurious reporting in either or both systems.

Section III documents how the evaluations of the A.C.E. in 2001 and 2002 identified a bias in the March 2001 estimates due to the reporting of spurious events and how revisions were made to the estimates in October 2001 and March 2003 to compensate for this large identified bias.

Section IV examines these adjustments to the dual system estimates in A.C.E. Revision II to compensate for spurious reports. These adjustments were based on using the results of linking the sample cases to enumerations outside of the search area and may have introduced other biases. These biases may have been caused by a) possible introduction of operational dependence, and b) mis-specification of the adjustment corrections. For the A.C.E. Revision II, the possible introduction of biases from these usages was acceptable as we addressed the large bias caused by spurious reports. By identifying the possible additional biases based on the A.C.E. Revision II usage, we can assess what biases may be present based on the estimation plans for 2010 Census Coverage Measurement (CCM). We can also assess based on various possible outcomes of these different biases whether a correction in some instances may do more harm than good.

Section V examines the CCM plans for the evaluation of net error in the 2010 census. It notes the several enhancements that the CCM is either currently doing for the interviewing or considering for the estimation operations. The interviewing and followup are being changed to use questions to better identify these spurious reports. These include questions to identify possible other addresses where the person may have been enumerated, addresses where a person may have moved from recently and situations where someone cycles between two or more addresses. The CCM estimation is proposing to treat certain cases with links outside the search area as unresolved. The currently planned logistic regression model for missing data will be expanded to account for these unresolved cases. We are seeking input on ideas on improving our estimation for spurious events in general and specifically for dealing with links outside the search area.

II. Dual System Estimation and Spurious Events

Wolter (1983, 1986) presents several alternative models for the assessment of coverage error in censuses of human populations. The Census Bureau implements the Petersen Model (M_t) within post-strata. As Wolter notes, it has been called the Schnabel estimator, Lincoln index, the Chandrasekar-Deming method and the dual-system estimator. Wolter refers to applying this model within post-strata as the M_{th} model. In 2010 Census Coverage Measurement, the Bureau will apply a version of the M_t model that uses logistic regression instead of post-stratification to attempt to minimize the heterogeneity and serve as the synthetic vehicle for subdomain estimates. The Habermann et al (1997) analysis using logistic regression showed the initial viability of this estimation approach. For more details on the CCM estimation, see Griffin (2005) and Mule (2008).

Regardless of the alternate models examined in Wolter, there are several assumptions that apply to each for valid estimates to be produced. One of the assumptions whose violation in the A.C.E produced a large bias was the violation of the "Spurious Events Assumption." El-Khorazaty et al (1977) document this bias as the "Effects of Over-coverage Bias" (Section 2.2). Seltzer and Adlakha (1974) document it as the bias due to spurious reports.

Wolter (1986) states the Spurious Events Assumption as:

> Both List A and List B are void of spurious events or have been eliminated prior to estimation. This means that all errors are avoided in recording both the census and the survey results. In practice, important spurious events that do occur include (a) duplicates on the census list, (b) reports of nonexistent cases in either the census or the sample survey, and out-of-scope cases, such as an individual born after the reference period, that are recorded erroneously in the census.

To minimize this bias, there are several suggestions. Chandrasekar and Deming (1949), Coale (1961) and Seltzer and Adlakha (1974) suggested a followup for non-matched cases to find and remove events wrongly recorded. Seltzer and Adklakha (1974) recommend the use of concepts "that are unambiguous and meaningful". Marks (1971) suggested that if both sources are samples that "larger sampling units (with well-defined boundaries) should be selected for one of the data collection sources and one-way matching applied over the entire area."

The Bureau had implemented several measures to minimize this bias in the 1990 Post Enumeration Survey and the 2000 A.C.E. Hogan (2003) documents the theory and design of the Accuracy and Coverage Evaluation. The sampling unit is one or more contiguous census collection blocks called a block cluster. The block clusters were designed to average 30 housing units and collapse over invisible boundaries as much as possible. Application of the dual system model for the 1990 Census is documented in Hogan (1992, 1993).

The A.C.E. utilized Computer Assisted Personal Interviewing (CAPI) during the A.C.E. Person Interviewing. This made falsification difficult by "time stamping" the interview and recording

every key stroke. We instituted a quality assurance process to minimize other sloppy or dishonest A.C.E. interviewing. Some cases were sent to the field for a followup interview to gather further information after the initial matching was complete. Followup is only useful if it provides more accurate or consistent responses. To provide better responses, followup uses better resources, for example: a) better respondents (household vs proxy), b) a better trained, supervised or quality-controlled interviewer or c) better questions or interview procedures.

The Census Bureau attempted to further fulfill the requirement of the spurious events assumption in 2000 by first defining and estimating the set of individuals who were "correctly" enumerated in the census. Hogan (2003) documents the four dimensions below that are used to measure this context of correct enumeration:

1. Appropriateness
2. Uniqueness
3. Completeness
4. Geographic correctness

"Appropriateness" means that the person should be included in the census. People who die before or who were born after the census reference date (April 1 in the U.S.) are not part of the population (universe) to be measured. Similarly, records that refer to fictitious "people," tourists or animals are out-of-scope.

"Uniqueness" refers to the fact that we wish to measure the number of people included in the census, not the number of census records. If more than one record refers to a single person, the count of records must be reduced for purposes of the dual system estimation.

"Completeness" means that the census record must be sufficient to identify a single person. If it lacks sufficient identifying information, we cannot determine for all of those records whether the person was appropriately and uniquely included in the census, nor can we determine whether he or she was also included in the survey.

"Geographic correctness" means that people are included in the census where they should be included. Enumerations whose correct location is outside the defined search area that are counted in the census are not considered correctly included in the census for dual system estimation.[1] This area must be searched during the matching process as well as searched for duplication. As the size of the small area increases, the complexity increases and the chance of

---

[1] For component error estimation an attempt is made to determine the enumeration status of these geocoding errors based on the geographic area of interest. For example for national estimates an enumeration is correct for component errors if the correct location is anywhere in the Unites States.

false matches grows. Normally a small area is defined such as a block cluster and the surrounding ring of collection blocks. A block cluster is a single collection block or a group of contiguous collection blocks.

In order to estimate the proportion of people who were correctly enumerated based on the P-sample (persons enumerated during the CCM person interview) results, we apply the same four dimensions to the P-sample cases as we did to a census case to be a correct enumeration. This application to the P sample allows us to identify and remove out-of-scope records from the estimation.

Seltzer and Adlakha (1974) show that spurious events in either or both systems will bias the dual system estimates of the population size in an upward direction. They show that assuming statistical independence of the two systems and that no matching errors occur, the relative size of this bias is equal to:

$$Relative bias = \frac{u}{v_1 v_2} - 1$$

where u = proportion of total matches based on valid reports (no spurious events included)
v1 = proportion of all events of system 1 (Correct Enumerations based on Census and E sample) that are valid reports
v2 = proportion of all events of system 2 (P sample) that are valid reports.

III. 2000 A.C.E. Results

The original A.C.E. estimates released in March 2001 (see Davis (2001)) showed an estimated undercount of 3.3 million people. This contrasted with a demographic analysis estimate of a 1.8 million overcount[2]. The U.S. Census Bureau's Executive Steering Committee for A.C.E. Policy (ESCAP) examined the differences between the two results. The first round of examination (ESCAP I) led to the Bureau's recommendation and decision that the unadjusted census data be released as the Bureau's official redistricting data (see Barron (2001)).

Between March and October 2001, the ESCAP examined the issue of whether adjusted data should be used for non-redistricting purposes (ESCAP II). The ESCAP concluded that since the A.C.E. had overstated the census undercount by at least 3 million people that unadjusted census data should also be used for non-redistricting purposes. Fay (2002) estimated that the A.C.E. had overestimated the number of correct enumerations in the census by 2.9 million. His conclusion estimates an enumeration is correct for component errors if the correct location is anywhere in the United States. was based on examining results from the Measurement Error Reinterview study (Adams and Krejsa 2001) combined with duplicate link information from the Person Duplication Study (Mule 2001). Feldpausch (2001) had shown in her ESCAP II analysis that the correct enumeration rate for E-sample cases with duplicate links outside the search area was

[2]Base Demographic Analysis estimate in ESCAP I, B-4

considerably higher than expected. Fay computed a one-cell DSE with the correct enumeration rate changed from 95.28% to 94.17% based on 2.9 million additional erroneous enumerations. He estimated a preliminary new DSE of 273.0 million which was a reduction of approximately 3.2 million. Based on the change of 3.2 million, the original DSE had a relative bias of 1.16%. Fay's analysis was one way of showing how the spurious events of erroneous census cases being treated as correct enumerations produced an upward bias in the results.

Using formula 1 from Seltzer and Adlakha, we can generate the same relative bias result as well. At this point, we had only examined spurious inclusions in System 1, the Census E sample, so we were assuming that there are no spurious matches (u) or invalid reports (v_2) in the System 2 (P sample).

We can estimate v_1 and then the relative bias by:

$$v_1 = \frac{0.9417}{0.9528} = 0.9884$$

$$relativebias = \frac{1}{0.9884 \times 1} - 1 = 0.0118 \; or \; 1.18\%$$

Since preliminary revised estimates released in October 2001 (see Thompson (2001)) still showed a differential undercount for renter and minority populations, the Bureau implemented the A.C.E. Revision II in 2002 to produce a final set of estimates. Fay's and other analyses focused on spurious events for the Census E sample (sample of census enumerations in areas corresponding to those enumerated for the P-sample) system, but there was also evidence that similar problems may have affected the P sample as well. Raglin and Krejsa (2001) showed that there were measurement errors in determining residency and mover status. The A.C.E. Revision II also included the analysis and adjustment for the possible inclusion of spurious events in the P sample.

The A.C.E. Revision II accounted for these spurious inclusions in both systems by a) correcting measurement error by recoding a subsample of the E and P samples using both the original A.C.E. data (initial interview and Person Followup results) and the Evaluation Followup information[3] and b) adjusting the usual dual system estimate formula for E and P- sample cases that had a link to a census enumeration outside of the search area[4]. There is a history of adjusting estimates for an initial result based on the results of more detailed examination of a subsample (Cochran (1977)). The second part of adjusting the usual dual system estimate formula was

---

[3]The Evaluation Followup Interview was conducted on a 10 percent subsample of the A.C.E. in January 2001 to provide the data for the Matching Error Study evaluation.

[4]U.S. Census Bureau (2004) fully documents the work done during 2002 for both of these operations and how they were incorporated into the A.C.E. Revision II estimates.

analyzed and debated especially for the P-sample cases which were not examined during the ESCAP II work.

Bell (2003) examined the issue of how to estimate the number of correct enumerations in the Census when you have links to other enumerations outside the search area. Since the error in the estimate depends on the true value of the cases which is unknown, his work provides a theoretical examination of groupings of cases based on the their coded status, true status and whether there is a duplicate link of the case to another enumeration outside the search area. This work assumed that all of the duplicate links could be identified.

Based on this and other assumptions, he examined four alternative estimation options for the cases where the duplicate link is to another housing unit. He concluded that *no estimation approach would produce an unbiased estimate of the correct enumerations*. Each estimation approach would have some bias and the selection of a particular approach should be based on minimizing the bias. Two cases stood out for minimal bias in this case. Option 2 used by Fay in his ESCAP II analysis treated all linked cases as ½ correct regardless of whether they were correct or erroneous. Option 4 (the method used in A.C.E. Revision II) treated the coded erroneous cases as erroneous and determined the factor (z_t probability in A.C.E. Revision II estimation) that needed to be applied to the correct enumeration (CE) cases so the same aggregate CE estimate as Option 2 was produced. Bell also examined Fay's approach of making all E-sample cases with a duplicate link to a group quarters where they could not claim a usual home elsewhere (UHE) as erroneous enumerations. Since this analysis assumed that all duplicate cases were identified, methods were developed to assign a probability of a case being a duplicate based on the Further Study of Person Duplication results (Feldpausch (2001)). Bell's theoretical analysis of E-sample cases having duplicate links was the basis of how the estimate of correct enumerations was adjusted for spurious inclusions for the A.C.E. Revision II.

The A.C.E. Revision II implemented the adjustment based on classifying sample cases with these links based on different "linked situations." Details of this adjustment are provided in the Appendix.

The final A.C.E. Revision II results determined the following errors because of spurious inclusions:

- the correct enumerations were overestimated by 4.7 million. The proportion of correct enumerations in the original A.C.E. based on valid reports (v1) was 98.15 percent.
- Matched P-sample cases were overestimated by 1.2 million. The proportion of match
- P-sample cases in the original A.C.E. based on valid reports (u) was 99.5 percent.
- The P-sample total cases were overestimated by 2.4 million. The proportion of P-sample total cases in the original A.C.E. based on valid reports (v2) was 99.09 percent.

Using Formula 1, the final relative bias because of spurious events in the A.C.E. estimation based on the A.C.E. Revision II is:

$$relativebias = \frac{9950}{0.9815 \times 0.9909} - 1 = 0.0231 \; or \; 2.31\%$$

Based on the A.C.E. Revision II, this estimated that the relative bias was approximately twice as much as previously believed based on the ESCAP II analysis. Most of the relative bias estimate from A.C.E. Revision II can be traced to the overestimate of correct enumerations. In examining Seltzer and Adlakha's relative bias formula, this makes sense since spurious events of matches in the P sample will impact both the numerator and denominator of the relative bias formula. The change in the dual system estimate based on spurious inclusions in the second system is directly related to the spurious events of P-sample nonmatches. Mule (2003) documented the same resulting change in net error estimates based on the A.C.E. Revision II work. Table 1 in his document shows the adjustments for spurious events by coding and person duplication corrections resulted in a reduction in the dual system estimate of 6.3 million. A reduction of 6.3 million corresponds to a relative bias of 2.3 percent.

IV. Other Biases Possibly Introduced by A.C.E. Revision II Corrections

The A.C.E. Revision II implemented adjustments to correct for the large bias due to the failure of the spurious report assumption. In this section, we will identify if any of these adjustments based on the use of links to census enumerations may have violated other assumptions of the dual system estimate model. These violations may have resulted in other biases in the population estimate. In 2000, the bias due to spurious events was so large that the introduction of possible additional biases by those corrections was an acceptable result.

However in looking forward to the 2010 estimation, there is a question about whether we need to have a spurious report adjustment like we did in A.C.E. Revision II. This 2010 adjustment would be based on sample cases which have a link to a census enumeration outside of the search area. A.C.E. showed that an adjustment seems warranted when a large number of spurious events are identified. By identifying the other potential biases that could be introduced by these adjustments, we can try to assess the impact of these additional other biases for different magnitudes and assumptions of spurious reporting.

The A.C.E. Revision II adjustments could have produced the following possible violations of the DSE assumptions:

a) Operational Dependence

Hogan (2003) stresses the DSE model will work if there is operational independence. One of the usual focuses of operational independence is taking great care to prevent the same field staff from working the same areas for both the census and the coverage survey. There are other focuses as well that Hogan addresses and one of them is followup. When discussing which cases to send to followup, he notes that the decision of which cases to send can compromise operational independence since followup may selectively change the defined "correct location" for certain cases. He advises choosing cases by balancing the need for accurate and consistent information with the need for independence by following up cases only when better information was likely.

The spurious report adjustment as implemented in A.C.E. Revision II can be thought of as a selective identification of a followup situation. Since these identified cases are based on the result of finding another enumeration in the census, this selective identification has the potential of introducing operational dependence. This could impact both the correct enumeration status of the E sample and the P-sample inclusion and match status.

b) Nonresponse Assumption (Missing Data for Unresolved Statuses)

The sample cases in A.C.E. Revision II that had links outside of the search area can be thought of as followup cases where the enumeration status in the E sample and Census Day status in the P sample had not been resolved. Since we had exhausted the interview data available, these cases are essentially unresolved as a group. The duplicate link adjustments for A.C.E. Revision II was a way of implementing a missing data adjustment for this group of cases.

Hogan (2003) cautions that, if the followup operation results in a noninterview, further biases can be introduced depending on the missing data models applied to the cases. If we applied the wrong "missing data" process to these cases in the A.C.E. Revision II adjustments, we may introduce a resulting bias in the population estimates. This could impact both the correct enumeration status of the E sample and the Census Day residence of the P sample.

V. 2010 Planning

Based on the experiences of the A.C.E., the Bureau has been planning and preparing for the 2010 Census Coverage Measurement (CCM) program. For the A.C.E., this additional work showed that the extent of the bias due to the inclusion of spurious events could not be corrected by recoding utilizing only the A.C.E. and Evaluation Followup Information. To empirically reduce the bias required adjusting the E and P-sample cases that linked to an enumeration outside of their net error search area. In order to help determine if an adjustment for spurious events is needed for the 2010 dual system estimation, we will review some of the enhancements to the CCM program.

First, the CCM Person Interview and Followup has been enhanced to utilize the strengths and lessons learned from the 2000 A.C.E. and the Evaluation Followup interviews. There are more probing questions so possible other addresses where the person in question may have been enumerated will be identified during the interview. Also included are probing questions about the possibility of joint custody, child in group quarters, cycling between addresses (work or vacation) and others have been added. All of this will be used to attempt to resolve the enumeration status of the sample cases with duplicate links outside the search area. Thus, the CCM has designed an interview to obtain as accurate results as possible and minimize the inclusion of spurious events.

For A.C.E. Revision II, the national computer matching information was utilized after all coding and missing data operations were already done. Another change for 2010 is that we will be implementing a national computer matching search prior to the clerical matching. Research on nationwide matching of Census 2000 data is provided in Ikeda (2007) and Ikeda (2008). Results of this research have been used in developing methodology for 2010. This search will utilize mover or alternate addresses provided by the Person Interview respondent to help determine if two enumerations are the same person. This matching will allow the matching operation to possibly account for this information during the matching and followup. This duplicate information can be used in possibly selecting cases for followup where the Person Interview may have been done by a proxy interview. This process will also result in more duplicate link cases being resolved and it is likely that there will be very few resolved duplicates that do not actually represent the same person. Details of matching are given in Moldoff (2008).

Based on these two changes, the CCM estimation is assuming there are no spurious reporting errors in either system if a) the sample case links to an enumeration at another address outside the search area and b) that address was reported during the Person or Followup Interview. One of the issues with the use of duplicate links in the A.C.E. Revision II was that we were not able to ask questions about which of the two addresses associated with a case was correct. For 2010 CCM, if the other address has been provided by the respondent, we will be able to ask the series of questions to determine where the person should have been counted. Since we will be able to do that, we are assuming that these cases have no errors of spurious reporting.

Third, we are implementing a version of PES Procedure B. PES B determines the P sample based on the residents of the sample housing unit at the time of the Person Interview. Thus, nonmovers and inmovers are in the P sample and matched to their Census Day address. For nonmovers the Census Day address is in the sample block cluster. For inmovers, the Census Day address is in another block cluster. Since inmovers are residing in the sample housing unit at the time of the interview, the need for proxy respondents is diminished. P-sample interviews for inmovers with a link to a census enumeration in their Census Day block cluster are not spurious events. This is different from A.C.E. that utilized PES C. The A.C.E. collected information on the residents of the sample address as of Census Day. Thus, nonmovers and outmovers were in the P sample and matched in the sample block cluster search area to obtain the match rate for the mover population. The Census Day address for outmovers was in the sample block cluster. Proxy interviews were more likely since the outmovers were no longer residing at the sample address. The number of movers was estimated by determining the residents of the housing unit on interview day who had moved in since April 1st.

Since the A.C.E. was matching all cases (nonmovers and outmovers) in a small geographic block cluster area, there was very little matching error of P-sample cases to the census enumeration in the block cluster. As stated earlier, the main error in A.C.E. was the spurious inclusion of cases who should not have been determined to be correct in the census or a Census Day resident in the P sample. Since we are now doing a version of PES-B, we will be matching the inmovers to their Census Day address. While this has the potential to reduce the inclusion of spurious events (especially in the P sample for movers), it does raise the possibility that we may have more

matching error for movers than in the A.C.E. However, we feel that the reduction of the need for proxy interviews with PES B more than compensates for any increased matching error.

The CCM estimation is planning on handling unresolved cases with links outside the search area as a missing data problem for the 2010 estimation. These cases would have links to another enumeration but the address of the other possible enumeration was not provided during the Person or Followup Interview. We can think of two groups of these unresolved links. The first is where the records linked together have been confirmed by the clerical matching staff. The second is where the two records linked together could not be resolved as representing the same person. Resolved links can also be classified in one of these groups. This information can be used as an independent variable for missing data imputation.

Sample cases with these linked situations will be treated as having unresolved status. For the E-sample cases, these will have an unresolved enumeration status. For the P-sample cases, they will have an unresolved residence status (should they be included in the P sample as residents at the sample address). The current plan is for a logistic regression model to be developed to include covariates to impute a probability of being a correct enumeration in the E sample or a resident in the P sample. As was done for the1990 Post-Enumeration Survey and the 2000 A.C.E., the 2010 missing data approach will use covariates based on the before followup group. One thing this does is distinguish cases by whether they matched or not to a correct enumeration in the census.

The A.C.E. Revision II adjustment required the assignment of a probability of whether the cases linked together were really the same (p_t). This determination was based on a lot on work in developing a model to apply to the results of exact matching of first name, last name and month and day of birth to determine a probability that the linked cases were the same person (see Fay (2002)). To identify all duplicates, we will attempt to identify single link situations where the names or birth dates may not agree exactly. This is possible because of utilizing other reported addresses that were not available for the A.C.E. matching. The 2010 CCM may have unresolved links that would require new models and approaches than were used for A.C.E. Revision II. The proposed 2010 missing data approach does not require assigning the explicit probabilities of cases being the same person that were required in A.C.E. Revision II. This is another reason why we feel the proposed 2010 approach is best.

References

Adams, T. and Kresja, E. (2001), "ESCAP II: Results of the Person Followup and Evaluation Followup Forms Review", ESCAP II report 24.

Bell, W. (2003), "Alternative Options for Tabulating Estimates of Census Correct Enumerations Allowing for Duplicate Links", DSSD A.C.E. Revision II Memorandum Series, PP-3.

Barron, W. (2001), "Recommendation on Adjustment of Census Counts", Cover letter to ESCAP I Report.

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

Coale, A. (1961), "The Design of an Experimental Procedure for Obtaining Accurate Vital Statistics", International Population Conference, Vol 2., London: International Union for the Scientific Study of Population.

Davis, P. (2001), "Accuracy and Coverage Evaluation: Dual System Estimation Results", DSSD Census 2000 Procedures and Operations Memorandum Series B-9*.

Feldpausch (2001) "Census Person Duplication and Corresponding A.C.E. Enumeration Status," ESCAP II Report 6.

Fay (2002), "Evidence of Additional Erroneous Enumerations from the Person Duplication Study", ESCAP II Report 9.

Jabine, T. and Bershad, M., (1968) "Some Comments on the Chandrasekaran-Deming Technique for the Measurement of Population Change", Paper presented at the CENTO Symposium on Demographic Statistics, Karachi, November 1968.

El-Khorazaty, M., Imrey, P., Koch, G. and Wells H. (1977), "Estimating the Total Number of Events with Data from Multiple-Record Systems: A review of Methodological Strategies", International Statistical Review, Vol 45 pp 129-157.

Griffin, R. (2005), "Net Error Estimation for the 2010 Census", DSSD 2010 Census Coverage Measurement Series, # 2010-E-1.

Hogan, H. (1992), "The Post-Enumeration Survey: An Overview", The American Statistician, 46, 261-269.

Hogan, H. (1993), "The Post-Enumeration Survey: Operations and Results", Journal of the American Statistical Association, 88, 423.

Ikeda, M and Porter, E. (2007), “Initial Results from a Nationwide BigMatch Matching of 2000 Census Data”, SRD Research Report Series RRS2007/22

Ikeda, M and Porter, E. (2008), “Additional Results from a Nationwide Matching of 2000 Census Data”, SRD Research Report Series RRS2008/02

Mule (2001), “Person Duplication in Census 2000", ESCAP II Report 20.

Mule (2008) “2010 Census Coverage Measurement Estimation Methodology”, DSSD 2010 Census Coverage Measurement Memorandum Series, #2010-E-18

Seltzer, W. and Adlakha, A. (1974), “On the Effect of Errors in the Application of the Chandrasekar-Deming Technique”. Chapel Hill, N.C., University of North Carolina Press, April 1974, (Laboratories for Population Statistics Reprint Series, No. 14)

Moldoff, M. (2008), “The Design of the Coverage Measurement Program for the 2010 Census”, DSSD 2010 Census Coverage Measurement Memorandum Series, 2010-B-07

Thompson, J., Waite, P.J., and Fay R., “Basis of “Revised Early Approximation” of Undercount Released October 17, 2001", ESCAP II Report 9a, U.S. Census Bureau (2004) “Accuracy and Coverage Evaluation of Census 2000: Design and Methodology”.

Wolter, K. (1983), “Coverage Error Models for Census and Survey Data”, SRD Report Series RR-83-07.

Wolter, K. (1986), “Some Coverage Error Models for Census Data”, Journal of the American Statistical Association, June 1986.

APPENDIX

A.C.E. Revision II Adjustment for Duplicate Links

The A.C.E. Revision II implemented the adjustment based on classifying sample cases with these links based on different "linked situations." The z_t factor in A.C.E. Revision II was the probability that an E-sample case was a correct enumeration given that it was a duplicate to another census enumeration outside the search area.

First, linked situations were identified where one component of the link was thought to be correct and the other incorrect. If a person in a housing unit linked to a person in a group quarter, such as a college dormitory, the person in the housing unit was taken to be incorrect and assigned a z_t of zero. This was "Linked Situation" 1.

Second, the duplicate link identifies that a person has been enumerated in two different households. This person is 18 years or age or older and is listed as a child of the reference person in only *one* of the households. In this situation, this person is assumed to be incorrectly included with their parents and should have been correctly enumerated at their other household. This determination applied unless A.C.E. had already determined the sample enumeration to be an erroneous inclusion. An example of this situation might be a college student that was listed with their parents and also listed as living in an apartment off campus. This was "Linked Situation" 2.

For other "Linked Situations" the choice of which person is correct is not clear. Consider links between whole households where all household members were duplicated. ("Linked Situation" 3). This includes families that might have moved some time around Census Day and were inadvertently included at both places or this might involve households with multiple residences with a helpful, but perhaps, uninformed proxy respondent. Another situation, "Linked Situation" 4., involves children ages 0 to 17, perhaps of divorced parents, that are linked between two different households. The remaining linked cases were grouped together as "Linked Situation 5". For these and all other situations, it is assumed that only half of these census enumerations with duplicate links are correct. To estimate the conditional probability, z_t, that the E-sample person is the correct enumeration, controls cells are defined for "Linked Situations" 3., 4., and 5 by 6 cells based on:

- 3 Race/Hispanic Origin Domains and
- Tenure

Within each control cell the z_t's were determined such that duplicate E-sample cases originally coded correct or unresolved will weight up to one half the number of census duplicates identified including the erroneous enumerations. This was calculated as:

$$\hat{z}_t = \frac{0.5\sum_t W_t\, p_t}{\sum_t W_t\, p_t \Pr(CE)_t}$$

where for person t, W_t is the A.C.E. sampling weight, p_t is the probability that the person has a duplicate link outside the search area, and $Pr(CE)_t$ is the original A.C.E. probability of correct enumeration. The summations are over the links in a control cell.

The residence status of P-sample persons was adjusted for spurious events as well by utilizing both the recoding and the matching results. The residence status of P-sample movers was only adjusted by using the recoding. The computer matching results were not used since outmovers in the P sample were collected by a proxy interview, which made it difficult to obtain date of birth and age information. Since date of birth and age were important characteristics used in the computer matching, the movers were only adjusted based on the recoding. The P-sample nonmovers were adjusted by both the recoding and the matching results. The P-sample nonmovers that linked to census enumerations outside the search area were adjusted by the following.

For the P-sample cases with links outside the search area, the results of the Further Study of Person Duplication did not provide an estimate of the number of correct Census Day residents in the P sample. In order to estimate the probability that a P-sample case was a resident on Census Day given that it links to a census enumeration outside the search area (h_t), it was determined necessary to borrow the resulting census z_t factor from the E-sample links.

First, the P-sample links to census enumerations outside the search area are identified for situations where it can be determined which component of the link is the correct residence. The "Linked Situations" and rules for assigning h_t's were the same as used for the "Linked Situations" of E-sample links stated earlier. For example in "Linked Situation" 2, a P-sample person 18 years of age or older who was listed as a child of the reference person and is a duplicate to the other housing unit (where he is not listed as a child) is assigned an h_t of zero. It is assumed that this person should not have been included as a Census Day resident of the P-sample housing unit.

For the other "Linked Situations" 3., 4., and 5., there once again is no information to determine whether the P sample had the person at the correct location or whether the census had them at the correct location. Additionally, there is no reasonable assumption about how many of these linked P-sample persons were correctly determined at the sample location. To overcome this obstacle, it is assumed that the error in identifying correct residence of the P-sample nonmovers is similar to the error in identifying correct enumeration of E-sample cases for similar situations. Therefore, the h_t for P-sample persons is set equal to the z_t determined for the E sample for comparable linked situations as identified by the same control cells used for the E-sample adjustment.