Census Coverage Measurement Design Alternatives[1]

Michael Beaghen
U.S. Census Bureau
4700 Silver Hill Rd., Washington, DC 20233-9001
(michael.a.beaghen@census.gov)

Key words: *duplication, omission, gross error, any address match, extended search area*

1. INTRODUCTION

This paper introduces design alternatives for the 2010 Census Coverage Measurement (CCM) program. These alternatives address flaws of past census coverage measurement, mainly difficulties in determining Census Day residence, and new goals of future coverage measurement such as the estimation of gross census coverage errors. The paper starts with the Background section, giving a brief overview of past census coverage measurement surveys. More background is provided in the section on the Definitions of Census Coverage Error. The section on Future Census Coverage Measurement Challenges outlines the difficulties faced by past coverage measurement surveys and new goals. The section on Design Alternatives describes alternatives under consideration for meeting the 2010 CCM challenges. The last section discusses some potential risks inherent in adopting these new methods.

2. BACKGROUND

The Census Bureau has conducted coverage measurement programs since the 1940 decennial census. The Census 2000 coverage measurement program was the Accuracy and Coverage Evaluation or A.C.E. (U.S. Census Bureau, 2001). The A.C.E. was flawed because of the ineffectiveness of its interviews in accurately determining people's Census Day residence. As a consequence, the A.C.E. failed to detect a large proportion of the census person duplication and overestimated the number of census omissions (Kostanich 2002, Chapter 1). The A.C.E. Revision II, conducted in 2002, addressed the flaws of the A.C.E. The 1990 census coverage measurement program was the Post-Enumeration Survey (PES), which implemented a design similar to that of the A.C.E.

The A.C.E. comprised two samples, a population (P) sample to measure census omissions and an enumeration (E) sample to measure census erroneous enumerations. The P sample was obtained by independently listing housing units in a sample of block clusters. Later, person interviewing was conducted in those housing units. The E sample consisted of the census enumerations in those sample blocks. For details on the A.C.E. sample see Olson (2003, Chapter 3).

Central to the A.C.E. was a matching operation that compared the P-sample and E-sample records to census records in the block cluster, and a field follow-up to resolve differences (Childers, 2001). People found in both the P sample and the census were called a match. People found only in the P sample and confirmed to be Census Day residents were called non-matches and represent census omissions. E-sample people who match to other census records were duplicated. E-sample people who matched to P-sample people who were residents were correct enumerations. E-sample people not matched to a person in the P-sample were followed-up to determine whether they were correct enumerations or erroneous enumerations. If the non-matched person was found to have existed in the A.C.E. sample cluster on Census Day they were a correct enumeration. If the person was found not to have existed in the A.C.E. sample cluster on Census Day they were an erroneous enumeration. The rate of omissions and the correct enumeration rate, taken with the census count of enumerations eligible to be selected in the E-sample, produced a dual system estimate of the population.

3. DEFINITIONS OF CENSUS COVERAGE ERROR

Census coverage error refers to whether the census correctly captured or enumerated a person. Some basic types of census coverage error are erroneous enumerations, people enumerated in the wrong place, and omissions. These and other census coverage errors are discussed subsequently, along with the distinction between estimating gross census coverage error versus net census coverage error.

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

3.1 Erroneous Enumerations

Enumerations that should not have been recorded are erroneous enumerations. The most common type of erroneous enumeration is duplication. Duplicates are two or more enumerations that refer to the same person who should have been recorded in the census. If both duplicates are enumerated in the correct place we have what is referred to as local duplication. Among local duplication, we distinguish between whole household duplication due to housing unit duplication, whole household duplication due to other causes such as to mail misdelivery, and within household duplication. If one member of a duplicate pair is enumerated in the correct place and another in the wrong place we have long distance duplication.

Other types of erroneous enumerations include those that do not correspond to a person. This could be a pet or a facetious name such as Superman. Further types of erroneous enumerations include people who died before Census Day, people who were born after Census Day, and people who should not be counted as a resident of the United States, such as foreign visitors.

3.2 Counted in the Wrong Place

Someone who should be recorded in the census but is enumerated at an address other than their Census Day residence is counted in the wrong place. In the A.C.E. and PES people enumerated outside of the block cluster and surrounding blocks where their Census Day residence belonged were considered omissions in the block they should have been enumerated in and erroneous enumerations in the block they were enumerated in. People counted in the wrong place can be classified in the following ways:

- in the wrong housing unit, though in the correct block
- in the wrong residence outside the block cluster and surrounding blocks, but in the correct county
- in the wrong county, though in the correct state
- in the wrong state

The first type of counted in the wrong place may be of little practical interest and is listed for completeness.

3.3 Omissions

A person who should have been recorded in the United States but was not is an omission. Note that in the calculation of net census coverage error, depending on the definition of a correct enumeration, a subset of people counted in the wrong place would be considered omissions. Also, in the A.C.E. and PES people enumerated in the census with insufficient information for matching and follow-up or imputed would be counted as omissions when that person was in the P sample.

3.4 Insufficient Information for Matching and Follow-Up

In the A.C.E. and the PES census records of poor data quality, typically with missing or deficient names, were excluded from processing to avoid biases in the dual system estimates. These records might have possessed enough information to match to, but if they did not match they could not be followed-up to determine their enumeration status. Presumably many census records with insufficient information for matching and follow-up corresponded to real people who lived in the address on Census Day, and thus are properly viewed as correct enumerations. However, they were counted as erroneous enumerations for the purposes of the estimation. The dual system estimates of population are otherwise unbiased so long as the census population to which the P-sample people can match is the population of census people meeting the A.C.E. definition of a correct enumeration.

3.5 Non-Data-Defined Enumerations and Imputations

Also to be considered are census enumerations that are not data-defined and census person records that are imputed. In the Census 2000 a non-data-defined enumeration was one which had only a name or no name and only one demographic characteristic. For example, it could be that only the name is captured on the roster of a mail return form, or that someone is listed only as a “boarder.” Demographic characteristics were imputed for non-data-defined enumerations. In contrast to non-data-defined enumerations, an imputed census person is a census person record the census created without reference to a specific person, though a proxy respondent might have given a count of people residing at the housing unit. All demographic characteristics are imputed for these person records. See Chen & Kilmer (2002) for information on census imputation.

Even when such census records correspond to a person who should have been recorded and was not otherwise enumerated, they might arguably be considered census coverage error because the demographic information is typically not correct. In the A.C.E. such records were not counted as erroneous enumerations. However, when a non-data-defined or imputed person corresponded to a person who should have been enumerated but otherwise was not, then the A.C.E. registered an omission.

3.6 Gross Error and Net Error

Net error is the difference between the correct count and the census count for a category, such as a geographical area like the nation or a state or a demographic group such as a race group. For the estimation of net error, the coverage measurement estimates are considered the correct count; thus the net error is estimated as the coverage measurement estimate less the census count. The relationship between net error and gross error for a category is that the net error should equal the omissions less the erroneous enumerations. How we estimate erroneous enumerations and omissions depends on the definition of a correct enumeration we use for people counted in the wrong place. We don't know yet what definition of correct enumeration we will use for the estimation of net error in the 2010 CCM. It is likely that different definitions would lead to different estimates of net error.

4. FUTURE CENSUS COVERAGE MEASUREMENT CHALLENGES

The future census coverage survey faces two major challenges. The first is to confront the flaw of the A.C.E., its failure to correctly determine Census Day residence. In particular, the A.C.E. person and follow-up interviews were ineffective in ascertaining when the interviewed or followed-up person had a Census Day residence other than the one they were interviewed at. In situations where a person's true Census Day residence was an address other than the A.C.E. sample address, the A.C.E. interviews often incorrectly indicated that the A.C.E. sample address was the Census Day residence. As a consequence, the A.C.E. failed to detect a large proportion of the census person duplication (Census Bureau, 2003, Kostanich 2002). For the same reason there were people counted in the wrong place who went undetected. This problem also engendered false P-sample non-matches and inflated the A.C.E. measure of census omissions.

For perspective it should be mentioned that the Census 2000 made many similar such mistakes determining Census Day residence. There were general difficulties with this determination. In part, the problems stem from the difficulty of applying complex residence rules requiring detailed knowledge that respondents, especially proxy respondents, may lack. The first major challenge is to improve the coverage measurement interviews and to design a survey robust to error in the determination of Census Day residence.

The second major challenge is to design a coverage survey that measures gross census coverage errors as well as net census coverage error (Singh 2003). The A.C.E. and the 1990 PES estimated the net census error, but did not directly measure the number of erroneous enumerations or the number of census omissions. Instead, they relied on a balancing of errors to estimate a correct net error while inflating the estimates of omissions and erroneous enumerations. Central to measuring census error will be the definition of what constitutes a census error, as is pointed out in Singh (2003): "what constitutes an error in the census (?) If an enumeration is counted but in the wrong geographic area; i.e., state, should this be considered erroneous and should it be considered an omission in the geographic area that the enumeration should have been counted? We will need to examine different definitions of error to determine which would be appropriate and which can be accurately measured for both gross and net error."

There were two important situations where the A.C.E. inflated the estimates of omissions and erroneous enumerations. First were people enumerated once in the wrong place and that place was outside the block cluster where their Census Day residence belonged. The A.C.E. rules counted such people as omissions in the geography where they should have been enumerated and as erroneous enumerations in the geography where they were enumerated. Such accounting might be plausible if the people were enumerated in wrong state. But most of this type of error was in nearer geography (Mule, 2002), where it is counterintuitive to count it as both an omission and an erroneous enumeration.

The second important situation were the estimated 4.8 million weighted data-defined census records that the A.C.E. removed from its match processing because they did not meet the A.C.E. criterion of sufficient information for matching and follow-up (Feldpausch, 2001). Usually these records had a missing or deficient name. The A.C.E. counted these as erroneous enumerations for the purpose of net error estimation. Presumably, many of them corresponded to people who lived at the address on Census Day and were not otherwise counted; i.e., they could be considered correctly enumerated. Logically, many also corresponded to P-sample non-matches.

5. DESIGN ALTERNATIVES

There are several design alternatives being researched to meet the challenges described in the previous section. It is helpful to start with a summary of them.

- Extend the search area for matches and duplicates to the nation.
- Conduct more probing interviews, including obtaining alternative addresses where a person could have been enumerated.
- Search for matches and duplicates at alternative addresses (any address match).

- Conduct an automated, nationwide search for matches and duplicates.
- Match to census enumerations considered insufficient information for matching and follow-up.
- Implement a method of collecting information on movers that decreases reliance on proxy interviews.

5.1 Approaches to Problems Determining Census Day Residence

There are two general approaches to dealing with the difficulty in determining the true Census Day residence encountered by the A.C.E. The first involves improving the information we collect in the interviews. We anticipate more probing interviews that query respondents specifically about common scenarios where the coverage measurement and census have had difficulty obtaining the information needed to accurately determine Census Day residence. We are encouraged in this attempt by the Measurement Error Reinterview's Evaluation Follow-Up interview, which had more probing questions and detected more erroneous enumerations than did the A.C.E. (Kresja & Raglin, 2001). At the same time research is underway on improved interview questions about residence, and about recall and other reporting error. In particular, there will be cognitive research on the interview addressing questions such as, how are respondents understanding questions? How do respondents think about multiple residences? What is the respondents' understanding of our questions and of the terms we use? Also important, the census is attempting to simplify and improve its residency rules, which would reduce both census error and the potential for the coverage measurement to fail to detect census error.

Another way to improve the information on Census Day residence is to change the way we collect information for movers. Movers are people who moved into the sample address after Census Day but before the coverage measurement interview. The A.C.E. implemented what is known as the PES-C method of collecting mover information (Childers, 2001). In PES-C the information used for matching movers is collected at their Census Day address, that is, at the housing unit that they moved out of. This means that the person interview typically relies on proxy interviews to provide the information. The 1990 PES, on the other hand, implemented PES-B, where the information used for matching movers is collected at their current, interview day address. Thus the person interview can directly interview a member of the mover household, and doesn't need to turn to proxy respondents. Since proxy information proved less reliable (Liu et al, 2001), we propose a return to the PES-B methodology.

The second general approach to dealing with the difficulty of determining Census Day residence is to design the coverage measurement survey such that it either does not rely so much on the determination of Census Day residence, or that it is more robust to errors in that determination. Several suggestions on how this could be achieved have been made. They involve a change in the definition of a census error from 'Not counted in the right place', Definition II (Cowan 1985), which is what census coverage measurement has employed in its 1980, 1990 and 2000 programs, to 'Not counted anywhere', Definition I. To illustrate the advantage of Definition I, consider a child in joint custody whose father lives in Prince George's County Maryland, and whose mother lives in neighboring Montgomery County Maryland. If we define a correct enumeration as being counted within the right state, then we no longer need to make the difficult and error prone determination of the true Census Day residence to establish a correct enumeration.

One method aimed at reducing reliance on the determination of Census Day residence is collecting alternative addresses and conducting 'any address matching' (Hogan, 2001). Collecting alternative addresses means that the person interview obtains all addresses where the P-sample people could be enumerated by the census. For example, a college student interviewed at their apartment in their college town may indicate that they could also have been enumerated at their parents' address. Then, in the any address matching, in addition to searching within the block cluster, a search for matches is conducted at those alternative addresses. Analogously, this design identifies census duplicates by obtaining in the person follow-up interview for people not matched in the E sample all places where they could be enumerated and then searching those addresses for duplicates.

A second method aimed at reducing reliance on the determination of Census Day residence is to incorporate automated, nationwide searches for matches and duplicates into the coverage measurement (Childers and Petroni, 2003). The data capture of names in the Census 2000 presented the opportunity to conduct such searches. The A.C.E. Revision II demonstrated that they can be successfully applied to enhance coverage measurement, though its application was post hoc and left questions unanswered (Kostanich, 2002). Integrating the automated, nationwide searches with the matching and follow-up presents possibilities that include a clerical review of links and a field follow-up of links. A clerical review of duplicate links was successfully conducted as an evaluation of the A.C.E. Revision II links (Byrne et al, 2003).

5.2 Approaches to Better Measuring Gross Census Error

Estimating gross error will require that new information be collected and new processing be conducted that was not done in previous coverage measurement programs. One step necessary to measuring gross error is to establish where a person should have been enumerated when they were enumerated in the wrong place. The A.C.E. and PES interviews collected and processed only the information that a respondent had another residence; they generally did not collect what or where that other address was. A second needed step is to expand to the national level the search for duplicates and matches outside the block cluster. As discussed earlier, such a national search can be conducted based on alternative addresses collected in interviews and based on an automated, nationwide search.

Another step to a better estimate of gross error would be a different treatment of the census enumerations traditionally defined by the A.C.E. and PES as insufficient information for matching and follow-up. An alternative to their exclusion from processing would be to attempt to match them to P-sample people with liberal matching rules that exploit the address matching (Hogan, 2001). The matched census enumerations would be counted as correct enumerations, and the matched P-sample would no longer increase the estimate of omissions. However, the treatment of the non-matched cases presents us with a missing data problem. One option would be to assign the non-matches a probability of correct enumeration such that we obtain the dual system estimate we would have obtained had we not matched to the insufficient information enumerations in the first place. This would be an example of using net error to calculate the gross error. A test match underway on A.C.E. data will give indications as to the feasibility of matching with liberal rules and would facilitate the development of such rules.

5.3 Non-Data-Defined Enumerations and Census Imputations

Another question for the CCM is what to do with non-data-defined enumerations and census imputations. Perhaps comparisons could be made on an aggregate level of how close the imputed counts come to the counts of P-sample people in those same housing units. The larger difficulty with these cases, however, may be conceptual; it is not clear under what conditions we should call them correct enumerations, or if we can even call them enumerations. Can we claim that these reduce the number of omissions when we can't relate them to particular people otherwise omitted?

6. PROSPECTS FOR THE 2010 CENSUS COVERAGE MEASUREMENT

While we believe the design alternatives described in this paper are promising, they are unproven and some discussion of their risks is in order. We will focus on three issues, starting with the question of can we accurately determine Census Day residence? Despite improved interviewing, there may still be problems with determining which address of a pair of duplicates or matches is the correct Census Day residence. Consequently we may find it necessary to adopt a broader definition of a correct enumeration, such as a person is correctly enumerated if counted in the correct state. In the Census 2000 most duplicates were found to be within state (Mule, 2002). If there continue to be more serious problems with determining Census Day residence, then expanding the definition of a correct enumeration to enumerated once anywhere in the nation may be a required step to provide robustness.

Next we consider the collection and processing of alternative addresses and the automated, nationwide searches. Both are approaches to finding matches and duplicates throughout the nation. Evaluations have determined that nearly all matches and duplicates that existed within the sample block cluster were detected by the A.C.E. (Bean, 2002). Ideally, the alternative addresses would detect nearly all matches and duplicates outside the block cluster, as would the automated, nationwide search for matches and duplicates. However, in the search extended to the nation, it is unlikely that either method will come close to that level of success. Future research should determine if a respondent, especially a nonhousehold member, can reliably provide alternative addresses. Research done in the A.C.E. Revision II indicates that the automated, nationwide search for duplicates fell well short of 100 percent detection (Mule, 2002). Perhaps the two approaches will complement each other, and between the two of them the CCM will come acceptably close enough to finding all the duplicates and matches outside the block cluster. Then the CCM may have taken a large problem fatal to coverage measurement and reduced it to a smaller, hopefully tolerable problem. Further, if the alternative addresses and automated, nationwide searches detect distinct matches and duplicates, they can be used to evaluate each other. As a matter of fact, they arguably represent two systems of capture, and a dual system type estimate might be employed to gain insight into how many duplicates and matches both approaches failed to detect.

Lastly we discuss the matching of movers. PES-B, since it is a method that was tried before, may be the least risky of the alternative methods under consideration. However, it requires geocoding the Census Day addresses of people who moved. This difficult process generated a high

unresolved rate in the 1990 PES (Lui et al, 2001), though in A.C.E. movers had an even higher overall unresolved rate because of the necessity of interviewing proxy respondents. It is hoped that with new matching and geographic technologies that the difficulties in 2010 will be less serious than in 1990. Also relevant will be the numbers of movers. Movers constituted about five percent of the P sample in the A.C.E. 2000, which started person interviewing early, in May 2000. If the CCM begins person interviewing later, say in September 2010, which may be necessary to avoid contamination with late census activities, then the proportion of movers is likely to be ten percent or more, making mover matching a greater risk.

7. REFERENCES

Bean, S. 2001: *'ESCAP II: Accuracy and Coverage Evaluation Matching Error.'* Executive Steering Committee For A.C.E. Policy II, Report No. 7., October 12, 2001. Census Bureau, Washington, DC.

Byrne, R., Beaghen, M., and Mulry, M. 2002: *'Clerical Review of Census Duplicates'* DSSD A.C.E. Revision II Memorandum Series #PP-43. Census Bureau, Washington, DC.

Chen, I., Kilmer, A. 2002: *'Census 2000: Overview of Count Imputation – Reissue of Q-2,'* DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES Q-78.

Childers, D. 2001: *'Accuracy and Coverage Evaluation: The Design Document,'* DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1, Revised. Census Bureau, Washington, DC.

Childers, D., and Petroni, R. 2003: *'Topic Report Series, No 4*: *Coverage Measurement from the Perspective of March 2001 A.C.E.'*

Cowan, C. 1985: *'Evaluating Censuses of Population and Housing, Chapter 2,'* Special Training Document ISP-R-5.

Feldpausch, R. 2001: *'E-Sample Erroneous Enumeration Analysis,'* ESCAP II Report #5.

Haines, D. 2002: *'A.C.E. Revision II – Computer Specifications for Defining Full P- and E-Sample Post-Strata,'* DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-26.

Hogan, H. 2001: *'2010 Census Coverage Evaluation Study: Goals, Questions and Principles,'* Memorandum for Jay Waite.

Kostanich, D. 2002: *'A.C.E. Revision II: Design and Methodology,'* DSSD A.C.E. Revision II Memorandum Series #PP-30.

Krejsa, E. and Raglin, D. 2001: *'ESCAP II: Evaluation Results for Changes in A.C.E. Enumeration Status A.C.E.,* Executive Steering Committee For A.C.E. Policy II, Report No. 3., October 15, 2001, Census Bureau, Washington, D.C.

Liu, X., Byrne, R., and Imel, L. 2001: *'Executive Steering Committee on Accuracy and Coverage Evaluation Policy II: Analysis of Movers,'* DSSD Census 2000 Procedures and Operations Memorandum Series #T-10.

Mule, T. 2002: *'Further Study of Person Duplicates.'* DSSD Revised A.C.E. Estimates Memorandum Series #PP-51. Census Bureau, Washington, DC.

Olson, D. 2003: *'Accuracy and Coverage Evaluation: Technical Documentation for March 2001 Estimates,'* DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES # Q-85.

Singh, R. 2003: *'2010 Census Coverage Measurement - Goals and Objectives (Executive Brief),'* DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #A-1.

U.S. Census Bureau 2001: *'Census 2000: A.C.E. Methodology, Volume 1.'*
www.census.gov/dmd/www/pdf/Vol-1.pdf

U.S. Census Bureau 2003: *'Technical Assessment of A.C.E. Revision II'*
www.census.gov/dmd/www/pdf/ACETechAssess.pdf