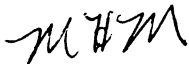




December 18, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP- 8

MEMORANDUM FOR Donna Kostanich
Chair, A.C.E. Revision II Planning Group

From: Mary H. Mulry 
Chair, A.C.E. Revision II Assessment Subgroup

Prepared by: Mary H. Mulry
Principal Researcher
Statistical Research Division

Subject: A.C.E. Revision II – Study Plan for Clerical Review of Census
Duplicates (CRCD)

The primary goal of the Clerical Review of Census Duplicates (CRCD) is for analysts at the National Processing Center (NPC) to examine the quality of the estimates of duplicate enumerations in the census. The Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates will use estimates of duplicate census enumerations produced by the Further Study of Person Duplication. The analysts will review housing units with two or more duplicate links identified by the Further Study of Person Duplication (FSPD) and duplicates identified by another evaluation of FSPD, the Census and Administrative Records Duplication Study (CARDS).

1. BACKGROUND

1.1 A.C.E. Revision II Estimates

The Executive Steering Committee for A.C.E. Policy II (ESCAP II) found that undetected duplication in the census was a major source of error in the A.C.E. estimates. The A.C.E. Revision II estimates will attempt to address this error as well as the measurement error that was detected by the Measurement Error Reinterview (MER). ESCAP II Report 9 Revised (Fay 2002) attempts to combine both sources of additional erroneous enumerations, duplicates and measurement error, to examine the impact on the Dual System Estimates (DSEs). The A.C.E. Revision II operation will extend this work to produce revised estimates that incorporate the effect of erroneous enumerations missed in the original A.C.E. estimates.

1.2 Duplication in the Census

1.2.1. Census 2000 Evaluation O.16

Census 2000 Evaluation O.16: Person Duplication in the Search Area Measured by the Accuracy and Coverage Evaluation (Jones and Feldpausch 2001) found that the estimate of duplicate census enumerations measured by A.C.E. was less than the estimate from the 1990 Post Enumeration Survey (PES).

1.2.2 ESCAP II Report 20

ESCAP II Report 20: Person Duplication in Census 2000 addressed this concern using the results of a computer matching operation to determine the extent of census duplication (Mule 2001). This operation extended the search to include units which were out-of-scope for the A.C.E. but would have been in-scope for the PES. It found an additional 1.2 million duplicate census enumerations in units that were out-of-scope for the A.C.E. but would have been in-scope for the PES. These units are mainly the group quarters residences that PES included, but A.C.E. did not.

The person duplication report also found some patterns of census duplications by race/ethnicity and age/sex groups that parallel previous observations of other types of coverage error. There were higher percentages of duplicate enumerations for the Non-Hispanic Black and the Hispanic domains. These were concentrated outside the one ring of surrounding blocks of a cluster but still within the same county. Duplication for persons 50 years of age or older was seen more in a different state. The 18-29 year-old categories had higher percentages of duplicate enumerations between housing units and group quarters than the other age/sex categories. The female duplication for this age group was predominantly in college dorms while the males were duplicated in college dorms, correctional facilities, and military group quarters.

1.2.3. Further Study of Person Duplication

A similar methodology will be used in the Further Study of Person Duplication (FSPD) to estimate and identify duplication in order to make adjustments for the A.C.E. Revision II estimates. Using a computer matching algorithm, the study will perform a national match of E-sample and P-sample records to census enumerations on the Hundred Percent Census Unedited File (HCUF). The algorithm uses a statistical matching methodology that assigns a probability of linked records being a match. Links with probabilities above specified thresholds are considered duplicates. The thresholds vary by the geographical distance between the pair and are set for four groups: 1) links between enumerations in the same block cluster, 2) links outside the cluster but within the surrounding blocks, 3) links in the same county, 4) links in the same state but different counties, and 5) links in different states. The statistical matching differs from the matching for duplication discussed in ESCAP II Report 20, which was exact matching on name and birth date. The statistical matching will be augmented by exact matching for the A.C.E. Revision II estimation, but those links will not be reviewed in this study unless they also were

discovered by CARDS. (Note: links between the P-sample and the HCUF are referred to as duplicates in this study even though they are really matches between the two different enumeration processes.)

1.3 Census and Administrative Records Duplication Study (CARDS)

CARDS (Bean 2002) is examining the effectiveness of the FSPD methodology with administrative records through the Census Numident File and the Statistical Administrative Records System 2000 (StARS 2000). CARDS is attempting to confirm or deny duplicate links identified by the FSPD. In addition, CARDS can identify duplicates missed by the computer study.

CARDS is the first study in a series of planned research using data from the Administrative Records Duplicate Link Research project. The goals of future research using this data are to analyze the nature of the duplication to reduce census duplication in 2010 and to provide data to StARS 2000 to aid in evaluation of decisions made during the construction of the system.

1.4 Clerical Review of Census Duplicates (CRCD)

The Clerical Review of Census Duplicates (CRCD) examines the effectiveness of the FSPD methodology through a clerical review of the enumerations that the computer designates as duplicates. CRCD also evaluates the accuracy of duplicates identified in CARDS, but missed by FSPD. In addition, CRCD may identify duplicates missed by both FSPD and CARDS. CRCD focuses on housing units where FSPD finds that more than one person is duplicated outside the A.C.E. search area. The requirement for more than one link reduces the workload to a number of households that can be completed within the time frame and concentrates on cases where we believe the analysts have the greatest chance of identifying additional duplicates. In addition, CRCD considers housing units where FSPD found no duplicates, but CARDS found one or more duplicates.

CRCD reviews cases from the block clusters in Evaluation Sample, a subsample of the A. C.E. sample. The results of CRCD will be used to design a more thorough review of additional cases as part of the preparations for the 2010 Census.

2. QUESTIONS TO BE ANSWERED

This study answers the following questions for households where more than one member has a link:

- How much duplication was there in Census 2000?
 - How many duplicates were there overall?

- What was the extent of duplication in all areas outside of the surrounding blocks? outside of the surrounding blocks but within the same county? in a different county but within the same state? beyond the same state?
- If time permits, what were the patterns of duplication for the Race/Ethnicity domains? for the Age/Sex categories?
- Overall, how effective was the methodology used in the FSPD and CARDS?
 - How many FSPD duplicate links can NPC analysts confirm? How many have probabilities of duplication above the threshold for designating a duplicate? How many are below?
 - How many FSPD duplicate links do the NPC analysts determine to be incorrect? How many have FSPD probabilities of duplication above the threshold for designating a duplicate? How many are below?
 - How many FSPD duplicate link do the NPC analysts not have enough information to classify as confirmed or incorrect and thus are undetermined?
 - How many CARDS duplicates missed by the FSPD can NPC analysts confirm?
 - How many CARDS duplicates missed by FSPD do the NPC analysts determine to be incorrect?
 - How many additional duplicates do the NPC analysts find that the FSPD and CARDS missed?

3. METHODOLOGY

3.1 Design matching operation

In addition to reviewing the duplicates, the operation needs to assign a “why” code that indicates the reason for declaring the pair a duplicate or denying the duplication. There are seven categories for the “why” codes: Insufficient Information, Characteristics, Household Composition, Other Residence, Nickname, Duplicate Housing Unit, Other Reason.

The analysts will review the whole households of duplicates. The analysts will have the information for all household members, not just those designated as duplicates. For census enumerations, the information is found on the HCUF. For P-sample cases, the information is collected in the A.C.E. CAPI interview. The images of the census questionnaires for enumerations outside the A.C.E. sample blocks are too difficult to access in the time frame.

The analysts will enter a code indicating whether a pair is a duplicate along with “why” codes and notes, if applicable. For the household members that were not designated as having a duplicate by FSPD or CARDS, the analysts will enter a code indicating whether a duplicate was found. If there is a ‘better’ duplicate in the census household other than the one designated by FSPD or CARDS, the analysts will record a code showing the duplicate was rearranged.

3.2 *Select sample of duplicates*

CRCD will review households with duplicates in the Evaluation Sample clusters. The review will include duplicates from both the E-sample and P-sample. Also included will be pairs that FSPD links but does not declare to be duplicates because the probability of being a duplicate is too low. For the E-sample, the review will be restricted to duplicates between enumerations in the E-sample and census enumerations outside the search area. For the P-sample, the review will be restricted to households with duplicates between nonmovers and census enumerations outside the search area and does not include links to deleted census enumerations.

The review is restricted to households where FSPD finds more than one member is duplicated although households with only CARDS duplicates may have only one. Additional cases from CARDS will not include links to group quarters. The reason for restricting the additional cases to links between housing units is that we believe that few additional duplicates would be found between a household and a group quarters residence. As a result, we will have only an estimate of additional duplicates between housing units with the type of duplication included in the study and other housing units. We will not have no estimate of additional duplicates between housing units and group quarters.

The clerical workload includes a total of 18,713 links in 11,935 housing units (work units). From the E-sample there are 10,248 links in 6,412 housing units while 8,465 links in 5,523 housing units are from the P-sample. We will plan for the contingency that not all of this workload might be completed by selecting a sample on a whole-cluster basis to hold back. This sample contains 3,004 links in 2,031 housing units

3.3 *Review duplicates*

The NPC analysts will determine whether the sets of two enumerations refer to the same people. The analysts will assign a “why” code that indicates the reason for declaring the pair a duplicate, denying the duplication, or not being able to decide. The analysts also will review household members not linked by FSPD to determine if they also have duplicates.

3.4 *Analyze results*

The clerical review results will be classified in the following table:

Clerical Review	FSPD = yes		FSPD = no		Total
	CARDS		CARDS		
	yes	no	yes	no	
Yes					
No					
Undetermined					
Total					

The elements of this table answer the following questions:

- the number and percentage of E-sample enumerations and P-sample nonmovers with duplicate links identified by the FSPD that are correct displayed by whether their probability of duplication is above or below the threshold for designating a duplicate
- the number and percentage of E-sample enumerations and P-sample nonmovers with duplicate links identified by the FSPD that are false displayed by the FSPD that are correct by whether their probability of duplication is above or below the threshold for designating a duplicate
- the number and percentage of E-sample enumerations and P-sample nonmovers with duplicates identified by CARDS that are correct
- the number and percentage of E-sample enumerations and P-sample nonmovers with duplicates identified by CARDS that are false
- the number and percentage of E-sample enumerations and P-sample nonmovers with duplicates missed by FSPD and identified by CARDS that are correct
- the number and percentage of E-sample enumerations and P-sample nonmovers with duplicates missed by FSPD and identified by CARDS that are false
- the number and percentage of P-sample persons and E-sample enumerations that have duplicates but neither FSPD nor CARDS identified them.

We refer to the proportions of correctly identified duplicates produced in the tables above as accuracy rates. We refer to the proportion of valid duplicates that were detected as the efficiency. Using tables of the form described above, we will compute the accuracy rate and efficiency for the FSPD identification of duplicates in the census and between the P-sample nonmatches and the census. We also will examine whether the accuracy rate is different for E-sample correct and erroneous enumerations and for P-sample matches, nonmatches and removed people. Note that we will distinguish whether the cases identified as duplicates by FSPD were found with statistical matching or exact matching, and the tables will be both unweighted and weighted. Unweighted counts give insights into the efficacy of the duplicate identification process while weighted counts give insights into the effect on the A.C.E. Revision II estimates

Furthermore, we will examine the accuracy and efficiency of the FSPD by the geographic distance between the links and the combination of the household size and number of duplicates within the household. The categories for the geographic distance will be: (1) within the county but outside the search area, (2) within the state but outside the county, and (3) outside the state. The categories for the households will be (1) single-person household with a link, (2) multi-person household where all members link, (3) multi-person household where two or more members link, but not all, and (4) multi-person household where only one person links.

4. DATA REQUIREMENTS

- We require the following files from FSPD and CARDS:
 - FSPD Analysis File
 - CARDS Analysis File

- We require the following additional files:
 - File containing information needed for the clerical review for all people in households which contain duplicates to be reviewed
 - CRCD Analysis File

5. DIVISION RESPONSIBILITIES

- The Decennial Statistical Studies Division (DSSD) will be responsible for managing the study, creating training materials, conducting the training, and creating the analysis files.
- The Planning, Research, and Evaluation Division (PRED) and DSSD will be responsible for designing and selecting the sample, creating output files from the administrative records matching process, and providing control counts.
- National Processing Center(NPC) will be responsible for performing the clerical review.
- PRED, DSSD, and the Statistical Research Division (SRD) will collaborate to develop analysis plans, conduct the data analysis, and prepare the report.
- See the milestone schedule for more specific information regarding responsibilities.

6. MILESTONE SCHEDULE

Activity	Person(s) Responsible	Planned Start	Planned Finish
Create match codes & why-codes	All	10/18/02	10/22/02
Identify input data for clerical review	All	10/18/02	10/22/02
Draft document on matching rules	Diane, Damon	10/22/02	10/25/02
Draft study plan	Mary	10/22/02	10/25/02
Decide & document sample design	Susanne, Tom Rita, Mary	10/18/02	10/28/02
Select sample, create input dataset & verify	Susanne, Don, Eli	10/22/02	11/1/02
Develop, test, & verify (prod & QA) Generic Matcher System w/ production input data Deploy s/w & data to NPC.	Rose, Damon, Diane, Norm	10/22/02	11/19/02

Create training materials	Alicia Shermaine	10/28/02	11/19/02
Train analysts	Alicia, Shermaine, Susanne		11/25/02
Conduct review of duplicates w/GMS	Analysts	11/26/02	12/10/02
Transmit & receive final results data set	Rose		12/12/02
Analysis	Michael		
Draft report	Rose	12/12/02	12/23/02
Final report		12/20/02	12/31/02

7. LIMITATIONS

- The study is restricted to households with two or more duplicates in another housing unit. The study does not evaluate duplicates identified in households with only one duplicate in general, only for household where CARDS identified a single duplicate in another housing unit and FSPD found none.
- The study can only find missed duplicates within households where duplicate links were identified by FSPD and/or CARDS.

8. RELATED STUDIES

- CRCD will use a clerical review to examine the effectiveness of methodology used in the Further Study of Person Duplication (FSPD).
- CRCD will use a clerical review to examine the effectiveness of methodology used in Census Administrative Records Duplication Study (CARDS) in identifying census duplicates missed by FSPD.
- CRCD will examine one component error, the estimation of duplicate enumerations, in the A.C.E. Revision II estimates. Additional A.C.E. Revision II evaluation studies will assess other component and relative errors. See Chapter 7 of “A.C.E. Revision II: Design and Methodology” (Kostanich 2003).

9. REFERENCES

Bean, Susanne (2002). Census and Administrative Records Duplication Study Plan. DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-15, U.S. Bureau of the Census, Washington, DC.

Fay, Robert E. (2002). "Evidence of Additional Erroneous Enumerations from the Person Duplication Study," Executive Steering Committee on A.C.E. Policy II Report 9 (Revised). U.S. Census Bureau, Washington, D.C.

Jones, John and Roxanne Feldpausch (2001). "Person Duplication in the Search Area Measured by the Accuracy and Coverage Evaluation," Census 2000 Evaluation O.16. Initial Draft dated July 23, 2001.

Kostanich, D. (2003), "A.C.E. Revision II: Design and Methodology," DSSD A.C.E. REVISION II MEMORANDUM SERIES #PP-30, U.S. Bureau of the Census, Washington, DC.

Mule, Thomas (2001). "Person Duplication in Census 2000," Executive Steering Committee on A.C.E. Policy II Report 20. U.S. Census Bureau, Washington, D.C.