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

MASTER FILE

December 31, 2002

DSSD A.C.E. REVISION II MEMORANDUM SERIES PP#44

PRED CENSUS AND SURVEY MEASUREMENT STAFF MEMORANDUM SERIES:
CSM-A.C.E. Revision II-R6R

MEMORANDUM FOR: Donna Kostanich
Chair, A.C.E. Revision II Planning and Management Group
Decennial Statistical Studies Division

From: Mary H. Mulry ***signed 12/31/02***
Chair, A.C.E. Revision II Quality Indicators Group
Statistical Research Division

Through: David Hubble ***signed 12/31/02***
Assistant Division Chief, Evaluations
Planning, Research, and Evaluation Division

Prepared By: Susanne L. Bean and D. Mark Bauder
Mathematical Statisticians
Planning, Research, and Evaluation Division

Subject: A.C.E. Revision II Report: Census and Administrative Records Duplication Study

Attached is the A.C.E. Revision II Report: Census and Administrative Records Duplication Study. Please direct any comments or questions to Susanne Bean 301-457-3457 or Mark Bauder 301-457-4229.

cc: DSSD A.C.E. REVISION II MEMORANDUM SERIES Distribution List
R. Killion
D. Hubble

A.C.E. Revision II Report
#PP-44
December 31, 2002

Census and Administrative Records Duplication Study

Susanne L. Bean and
D. Mark Bauder

Planning, Research, and
Evaluation Division

USCENSUSBUREAU

Helping You Make Informed Decisions

CONTENTS

LIST OF TABLES

EXECUTIVE SUMMARY

The primary goal of the Census and Administrative Records Duplication Study (CARDS) is to use administrative records to examine the quality of the estimates of duplicate enumerations that were used in the Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates.

The Further Study of Person Duplication in Census 2000 (FSPD) attempted to estimate and identify duplication in order to make adjustments to the A.C.E. Revision II estimates. Using a computer matching algorithm involving a statistical and an exact matching component, the study performed a national match of E-sample and P-sample records to census enumerations on the Hundred Percent Census Unedited File (HCUF). CARDS uses the Statistical Administrative Records System 2000 (StARS 2000) to examine the effectiveness of the FSPD methodology.

Using administrative records, CARDS performed a computer match to attempt to assign each census record (including the E-sample) and P-sample record a Protected Identification Key (PIK). PIKs are used instead of Social Security Numbers (SSNs) for confidentiality. Then CARDS linked E- and P-sample records to census records with the same PIK. These links are then used to create estimates of duplication. In addition, CARDS attempted to confirm or deny links found by FSPD.

In this study, we

- compared estimates of duplication between CARDS and FSPD by geography and by type of census record,
- tested a procedure that combines FSPD and CARDS results to produce estimates of duplication, and
- examined links found in CARDS but not FSPD.

Our key findings and recommendations are as follows:

- **The FSPD process was more effective at finding duplicates that are geographically close**. FSPD found more duplicates within the A.C.E. cluster as well as within the surrounding blocks for all categories of census record except group quarters. This held true both for E-Sample duplicates, and nonmatched (in A.C.E.) P-Sample links.

- **CARDS identified more duplicates that are geographically distant.** As the links got farther apart, CARDS identified relatively more duplicates than FSPD. In different states, CARDS had about twice as many links as FSPD, both for E-sample duplication and for nonmatch P-sample links.

- **CARDS identified more group quarters duplicates.** For group quarters, the FSPD process was limited to its exact matching stage. Thus, we expected that the CARDS process might find more duplicates to group quarters, and the results confirmed this.

- **CARDS links that were geographically more distant were more questionable.** Although CARDS linked many more people in different states, we saw reason to question some of these links. CARDS seems to have identified a large number of duplicates to different states where only one person was linked in a multi-person household. A high percentage of these links were in CARDS but not FSPD (were "CARDS only" links). This raises suspicion about the quality of those links. More research would be needed to determine whether CARDS is finding true duplicates in these cases. In addition, we expect that CARDS duplicate links to be less certain where one or both of the related Numident to HCUF links were done in a matching phase that did not use address data. Among CARDS links between different states where the CARDS matching had not used address data for both links, a high percentage were CARDS only.

- **A combined FSPD/CARDS procedure improved estimates**. We tried a conservative way to incorporate CARDS results into the FSPD process. We used CARDS confirmation and denial of FSPD links to change FSPD duplicate probabilities to one or zero. This process increased the estimates of duplication from FSPD alone, which we believe to be an improvement.

- **We suggest further study of the FSPD and CARDS links.** This is just the beginning of the research that can be done to explore the nature of the duplication found by FSPD and CARDS. Further research is suggested in the CARDS study plan and by the differences found in this report. In addition, the Clerical Review of Census Duplicates (CRCD) can provide further information.

- **We recommend that CARDS style research continue with improved administrative records procedures for detecting duplicates.** Administrative records can be valuable aids for detecting duplicates. CARDS style processes have the potential to identify duplicates that other methods have difficulty detecting – for example, people enumerated with different names, and people whose enumerations have reporting errors. We have seen in this report that CARDS data has been useful in confirmation and denial of FSPD links, and has the potential for finding additional duplicates. But we also have some reason to question some of the CARDS links that were not also found by FSPD. The CARDS process used the results of a match that associated census records with PIKs, which was not initially done for the purpose of detecting duplicates. We believe that a CARDS style process that is developed from the beginning to detect duplicates, and that uses lessons learned from this study, CRCD, and future research, can produce more complete and accurate results.

1. BACKGROUND

The primary goal of the Census and Administrative Records Duplication Study (CARDS) is to use administrative records to examine the quality of the estimates of duplicate enumerations that were used in the Accuracy and Coverage Evaluation (A.C.E.) Revision II estimates.

1.1 A.C.E. Revision II Estimates

Based on findings from the Executive Steering Committee for A.C.E. Policy (ESCAP) reports, duplicates are one of the major sources of error from the A.C.E. which the A.C.E. Revision II estimates will attempt to address. Another source of error identified in the ESCAP reports is measurement error as detected by the Measurement Error Reinterview (MER). ESCAP Report 9 (Revised): Evidence of Additional Erroneous Enumerations from the Person Duplication Study attempted 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 extended 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

Census 2000 Evaluation O.16: Person Duplication in the Search Area Measured by the Accuracy and Coverage Evaluation 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) (Jones, 2003). 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. 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. They 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.

The ESCAP II report also found some intuitive patterns of census duplications by race/ethnicity and age/sex groups. 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.

A similar methodology as used for the ESCAP II report was used in the Further Study of Person Duplication in Census 2000 (FSPD) to estimate and identify duplication in order to make adjustments to the A.C.E. Revision II estimates. Using a computer matching algorithm, the study performed a national match of E-sample and P-sample records to census enumerations on the

Hundred Percent Census Unedited File (HCUF). (Note: In this study we refer to links between the P-sample and the HCUF are referred to as duplicates, although though they are really matches between the two different enumeration processes. When a P-sample person and an HCUF person are linked, it does not mean that the person was in the HCUF twice.)

1.3 Census and Administrative Records Duplication Study (CARDS)

CARDS used the Statistical Administrative Records System 2000 (StARS 2000) to examine the effectiveness of the FSPD methodology. CARDS attempted to confirm or deny duplicate links identified by the FSPD. In addition, CARDS attempted to identify duplicates missed by FSPD.

CARDS is the first study in a series of proposed 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.

2. METHODOLOGY

FSPD performed a computer match to link E- and P-sample records to HCUF records. CARDS used the results of a previous match done by the Administrative Records Research Staff (ARRS) between the HCUF and an administrative records file. A similar match was done for the P-sample for the CARDS project. CARDS then used these results to identify links between sample records and the HCUF.

2.1 Linking Processes

Below are brief descriptions of the FSPD and CARDS linking processes.

2.1.1 FSPD Linking

FSPD used two types of matching to create links and assign probabilities to those links. These types of matching are referred to as statistical matching and exact matching.

The statistical matching had two stages. The first stage was a statistical matching of source (either E- or P-sample) to target (census) records based on name (first name, last name, and middle initial) and age/date of birth (computed age, month of birth, and day of birth). After the first stage identified a person link between two housing units (HU), the second stage performed a statistical match of people in those two HUs. The second stage matching was also based on name and age/date of birth, but used different parameters than those used in the first stage. For links in HUs with 2 or more links (2+ HUs), the statistical matching process assigned a Probability of No Trial Having Observed Outcome called MPROBDUP. MPROBDUP was examined to determine if the link was considered a duplicate. If the link had a MPROBDUP

value over a preset cutoff for the appropriate sample and geography, then it was considered a statistical duplicate and was assigned a final duplicate probability of 1.

The exact matching assigned final duplicate probabilities (between 0 and 1) to links whose MPROBDUP did not meet the statistical matching cutoff, links to group quarters, and links where only one person was linked between the HUs. This matching looked for agreement on first name, last name, month of birth, and day of birth.

For information regarding the FSPD linking process, please refer to Chapter 5 of the A.C.E. Revision II: Design and Methodology. (Fenstermaker, 2003)

2.1.2 CARDS Linking

There are two basic steps in the process which produced CARDS links. First, Protected Identification Keys (PIKs) are assigned to HCUF records and P-Sample records by matching census and A.C.E. files to administrative records in the StARS 2000 database. Then, links are created between records which were assigned the same PIK.

The StARS 2000 database, created by the Administrative Records Research Staff (ARRS) incorporates data from seven administrative record files:

- Internal Revenue Service Individual Master File (1040),
- IRS Information Returns File (W-2 / 1099),
- Department of Housing and Urban Development Tenant Rental Assistance Certification System File,
- Department of Housing and Urban Development's Multifamily Tenant Characteristics System File
- Center for Medicare and Medicaid Services Medicare Enrollment Database File,
- Indian Health Services Patient Registration System File,
- Selective Service System Registration File.

In addition, ARRS maintains a lookup file, called the "Census Numident." This file was created from the Social Security Administration's Numerical Identification File (Numident). The Numident was edited, and for confidentiality reasons a Protected Identification Key (PIK) was created for each Social Security Number (SSN). An additional file was created which also contains all addresses from the IRS 1040 and 1099 files from StARS 2000 for each person. This file is called the Geokey Numident. The geokey is a variable which incorporates address information from the IRS returns file.

In previous work, ARRS had performed a two phase computer match to link Geokey Numident records with HCUF records in order to assign PIKs. In the Geokey Search phase, matching between the files was done based on name, date of birth, and geokey. Additional links were created in the Name Search phase where matching was based on name and date of birth only.

Via this match, PIKs were found for HCUF people and added to HCUF person records. We call the resulting file the HCUF Research File. For the CARDS project, P-Sample people were also linked with Census Numident records using a similar methodology. This associated PIKs with P-Sample records. Note that some person records on the HCUF and the P-sample file had no PIK assigned. This could happen in two ways. If the HCUF record was not linked with any PIK, none could be assigned. In addition, when one HCUF record was linked with more than one PIK, no PIK was assigned to the HCUF record.

Links were created between source (E- or P-sample) and target (census) records with the same PIK. The CARDS process did not assign probabilities, thus each link is considered a duplicate.

2.2 Classifying FSPD Links

We attempted to confirm or deny links of E-Sample and P-Sample records found by FSPD, regardless of the duplicate probabilities assigned in FSPD. Where FSPD created a link of between an E-Sample or P-Sample record and census record (the "FSPD linked person"), we determined whether CARDS had the same PIK for the sample person and the FSPD linked person.

- If the E- or P- sample person and the FSPD linked person had the same PIK (and thus were identified as a CARDS link), we considered the FSPD link to be confirmed.

- If we had a PIK for one, but not for both, of the FSPD linked people, we attempted to confirm the link by performing an address match using all the addresses in the StARS 2000 database. The StARS 2000 addresses for the person with a PIK were matched with the sample or census address for the person without a PIK. If an address match was found, we considered the FSPD link to be confirmed, because we then had evidence from StARS that one person lived at both addresses.

- If the FSPD linked person had a different PIK from the E- or P- sample person, we judged the FSPD link to be denied.

- If we did not have a PIK for either of the records in the linked pair, or we had a PIK for just one and the link was not confirmed in the address match, we called the link undetermined.

2.3 Classifying CARDS Links

We compared links identified by CARDS to those identified by FSPD to determine which links were found by both studies and which were only found by CARDS. (Note: The only links that are considered CARDS links are those where the source and target records were assigned the

same PIK. Thus, FSPD links which were confirmed by the additional StARS 2000 address matching in the second bullet above are not considered CARDS links in this report.)

- If the source and target person had the same PIK and FSPD also identified the link, we classified the CARDS link as found by both CARDS and FSPD.

- If the source and target person had the same PIK but FSPD did not find the link, we called it a CARDS only link.

3. LIMITATIONS

There are several ways in which the process outlined above may fail to confirm or deny FSPD links, or may link false duplicates.

- The match between the Geokey Numident and the HCUF was originally done in order to associate Census race information with Numident records. It was not initially done for the purpose of detecting duplicates. Therefore, decisions about match strategy, and how conservative or liberal to be in accepting links, may not have been optimal for the purposes of identifying duplicates

- In the ARRS HCUF to Numident match, not all HCUF records could be associated with PIKs. Thus, the CARDS process is likely to miss some duplicates, and it left some FSPD links with undetermined status. We found that about 28% of FSPD E-Sample duplicate links, and about 21% of FSPD P-Sample links, could not be confirmed or denied by CARDS. The Clerical Review of Census Duplicates (CRCD) study can provide further information about these undetermined links. (Beaghen and Byrne, 2002)

- Because StARS 2000 is created from administrative records, a person can be duplicated at different addresses, yet StARS 2000 failed to have records from both addresses. In that case, the duplicate is less likely to be detected by CARDS. The duplicate could only be found in the Name Search phase of matching, which requires better person data and a more exact match.

- Some people have two SSNs, and more than one person can have the same SSN. If one sample person has two SSNs, then CARDS may fail to find that person's duplicate. If more than one person has the SSN of a sample person, then CARDS may falsely call them duplicates.

- We were not able to fully investigate the links found in the CARDS process but not in the FSPD process. Without further research, we cannot estimate how many of these are truly duplicates missed by FSPD, and how many are false duplicates. The CRCD study can provide further information about these CARDS only links. (Beaghen and Byrne, 2002)

4. RESULTS

To examine the quality of the estimates of duplicate enumerations that were used in the A.C.E. Revision II estimates, we have computed estimates of duplication based on CARDS to compare to FSPD estimates. These estimates are for the E-sample and for the P-sample nonmover residents. Standard errors were calculated using a simple jacknife method. We also looked at some characteristics of the CARDS links in an attempt to explain some differences between the estimates.

4.1 Comparison of FSPD and CARDS Estimates of E-Sample Duplicates

For comparison with FSPD results, we calculated weighted frequencies of CARDS E-Sample duplicate links. We broke out these frequencies by geographical categories and type of census record.

The geographical categories are:

- within cluster;
- outside of cluster, within surrounding blocks;
- outside of surrounding blocks, within same county;
- outside of surrounding blocks and county, within same state; and
- outside of surrounding blocks, in a different state.

The types of census record are:

- E-Sample eligible;
- Group Quarters;
- Census Reinstate; and
- Census Delete.

Table 1. CARDS Weighted[1] Estimate of E-sample Duplicates by Geography and Census Record Type

Geography	Census Record Type				Total
	E-Sample Eligible	GQ	Reinstate	Delete	
Within Cluster	998,239 (35,162)[2]	107,305 (21,452)	920,405 (42,888)	1,681,962 (82,499)	3,707,911 (113,548)
Surrounding Block	202,741 (15,516)	31,355 (11,686)	22,870 (5,926)	588,300 (48,878)	845,266 (55,656)
Same County	1,145,036 (24,177)	334,983 (47,946)	420,917 (24,624)	187,804 (18,520)	2,088,740 (64,559)
Diff. County, Same State	693,540 (20,531)	307,014 (13,610)	79,986 (10,708)	35,618 (6,734)	1,116,159 (29,646)
Different State	1,183,055 (30,328)	183,917 (10,500)	21,808 (3,276)	32,472 (4,350)	1,421,251 (34,133)
Total	4,222,611 (68,660)	964,574 (57,701)	1,465,986 (52,042)	2,526,156 (102,200)	9,179,326 (169,735)

CARDS identified approximately 9.2 million E-sample duplicates, of which about 4.2 million were to E-sample eligible census records. Within the cluster, CARDS found fewer than one million E-sample duplicates to E-sample eligible records. Therefore, CARDS was not as efficient as the A.C.E. person matching clerical matchers who found about 1.9 million duplicates for this group.

Table 2 presents the E-sample FSPD results. (Note: This table presents the same results as Table 2 in the A.C.E. Revision II report "A.C.E. Revision II Results: Further Study of Person Duplication", but does not have the "Total Records in Census" column.)

[1]This table is weighted by the product of the A.C.E. sampling weight and the multiplicity factor. For more information regarding the multiplicity factor, please see Appendix D of the A.C.E. Revision II Results: Further Study of Person Duplication. (Mule, 2002)

[2]In all tables, standard errors are in parentheses.

Table 2. FSPD Weighted[3] Estimate of E-sample Duplicates by Geography and Census Record Type

Geography	Census Record Type				Total
	E-Sample Eligible	GQ	Reinstate	Delete	
Within Cluster	1,173,344 (47,342)	76,381 (15,753)	1,058,548 (49,236)	1,967,199 (96,051)	4,275,472 (133,999)
Surrounding Block	259,805 (21,849)	25,373 (9,704)	24,751 (6,975)	678,355 (57,807)	988,284 (66,496)
Same County	1,011,920 (25,678)	231,774 (39,853)	482,015 (28,149)	208,246 (20,879)	1,933,956 (61,531)
Diff. County, Same State	563,270 (19,483)	190,417 (9,648)	88,331 (12,594)	35,111 (7,270)	877,129 (27,612)
Different State	527,796 (24,146)	91,793 (7,144)	20,959 (17,317)	16,184 (4,905)	656,732 (34,359)
Total	3,536,136 (71,975)	615,738 (46,326)	1,674,604 (62,097)	2,905,096 (119,206)	8,731,572 (184,528)

FSPD identified approximately 8.7 million E-sample duplicates, which is approximately 0.4 million fewer than CARDS found overall. FSPD also found fewer duplicates to E-sample eligible census records than CARDS (3.5 million versus 4.2 million). However within the cluster, FSPD was more efficient than CARDS in comparison to the A.C.E. clerical person matching to E-sample eligible records.

Two other differences stood out between the CARDS and FSPD E-sample results. CARDS identified more duplicates to group quarters and to census records in different states.

A reason that CARDS could have identified more duplicates to group quarters is that, in FSPD, links to group quarters were assigned final duplicate probabilities using the exact matching process. Because the FSPD exact matching process did not use information from other links within the household, the criteria to link records together were more strict. A more exact on person data was required. CARDS criteria may have been less strict.

Many of the FSPD links to different states were single links – cases where only one person in the HU was linked. Therefore, many of these links were assigned final duplication probabilities in

[3]This table is weighted by the product of the A.C.E. sampling weight, the multiplicity factor, and the final probability of duplication. For more information regarding the multiplicity factor, please see Appendix D of the A.C.E. Revision II Results: Further Study of Person Duplication. (Mule, 2002)

FSPD by the exact matching process. Due to the large geographic distance, many of these links may have been assigned lower probabilities, which would lower the weighted estimates of duplication. However, in CARDS all links were treated as duplicates. (They were treated as if they all have a final duplicate probability of one). So even if there were a lot of overlap between FSPD and CARDS links to different states, the FSPD estimates could be substantially lower.

Therefore, as a second comparison, we recalculated the estimate of FSPD links using duplicate probabilities adjusted based on CARDS results. When CARDS confirmed a duplicate link, the probability was adjusted to 1. When CARDS denied a link, the probability was adjusted to 0. Otherwise, we used the original FSPD duplicate probability.

Table 3. FSPD with CARDS Adjusted Weighted[4] Estimate of E-sample Duplicates by Geography and Census Record Type

Geography	Census Record Type				Total
	E-Sample Eligible	GQ	Reinstate	Delete	
Within Cluster	1,163,024 (47,078)	76,371 (15,754)	1,055,789 (49,113)	1,961,154 (95,809)	4,256,338 (133,593)
Surrounding Block	258,576 (21,473)	25,401 (9,709)	25,009 (6,984)	672,077 (57,504)	981,062 (66,073)
Same County	1,015,854 (25,667)	232,027 (39,918)	483,092 (28,066)	209,560 (20,806)	1,940,533 (61,541)
Diff. County, Same State	552,801 (19,040)	220,536 (10,762)	90,361 (12,784)	37,384 (7,442)	901,082 (27,999)
Different State	602,616 (25,796)	112,363 (8,158)	22,139 (17,337)	20,155 (4,922)	757,273 (35,942)
Total	3,592,871 (72,800)	666,697 (46,811)	1,676,390 (61,993)	2,900,330 (118,935)	8,836,289 (185,516)

This CARDS adjusted weighting increased the FSPD estimate of E-sample duplicates to group quarters in different counties and different states. In total, this increased the count of duplicates to group quarters by 51,000 and to different states by approximately 100,500. However, these are still lower than the CARDS estimates for these groups.

[4]This table is weighted by the product of the A.C.E. sampling weight, the multiplicity factor, and the adjusted final probability of duplication based on the CARDS results. For more information regarding the multiplicity factor, please see Appendix D of the A.C.E. Revision II Results: Further Study of Person Duplication. (Mule, 2002)

4.2 Comparison of FSPD and CARDS Estimates of P-Sample Links

For the P-Sample, we calculated frequencies similar to those for the E-Sample. Our analysis of P-Sample links is restricted to nonmover residents only. (We focused on these records since they were used for the duplicate adjustments to the A.C.E. Revision II estimates.) In addition to the geographical and census type categories used above, we broke out the frequencies by A.C.E. match status, where any record with a match probability greater than 0 is considered a match. Recall that we are using the term "duplicate" to refer to a link between the P-sample and census, even though these are really matches between the two enumeration processes.

The estimates of "duplicates" of nonmatches demonstrate how many more records could have been considered matches if a national search area were used. Excluding the "within cluster" and "surrounding block" rows, the estimates of "duplicates" of matches show the number of records that were matched both within and outside of the A.C.E. search area. These P-sample matches imply duplication within the census. Thus, most of the P-sample analysis will concentrate on the P-sample nonmatches since duplication within the census was discussed in the previous section.

Table 4. CARDS Weighted[5] Estimate of P-sample Nonmover Resident Duplicates by Geography and Census Record Type

Geography	Census Record Type								Total	
	E-Sample Eligible		GQ		Reinstate		Delete			
	Non-match	*Match*	*Non-match*	*Match*	*Non-match*	*Match*	*Non-match*	*Match*	*Non-match*	*Match*
Within Cluster	220,709 (12,449)	187,378,857 (2,110,691)	753 (534)	122,527 (28,940)	402,958 (47,968)	802,936 (40,921)	222,340 (28,771)	1,791,995 (108,474)	846,760 (57,997)	190,096,315 (2,135,807)
Surrounding Block	427,255 (36,377)	8,309,554 (500,758)	7,815 (4,790)	11,348 (5,031)	43,207 (11,347)	59,913 (14,386)	23,553 (6,713)	298,129 (27,899)	501,829 (41,189)	8,678,944 (505,439)
Same County	1,924,193 (106,007)	1,346,069 (35,620)	52,030 (9,194)	182,155 (32,048)	12,246 (3,799)	178,939 (15,558)	40,483 (13,741)	92,615 (13,405)	2,028,953 (113,238)	1,799,778 (54,476)
Diff. County, Same State	452,687 (28,069)	772,124 (26,729)	43,584 (4,926)	111,273 (7,739)	3,144 (1,461)	31,362 (5,939)	8,322 (3,747)	14,219 (3,225)	507,736 (29,394)	928,978 (29,470)
Different State	518,886 (24,609)	1,844,299 (42,488)	23,085 (3,886)	83,502 (6,438)	5,984 (2,212)	15,944 (2,785)	6,283 (1,793)	26,141 (4,296)	554,239 (25,664)	1,969,886 (44,638)
Total	3,543,729 (124,873)	199,650,902 (2,222,443)	127,267 (12,450)	510,805 (45,198)	467,538 (49,652)	1,089,093 (47,133)	300,982 (33,848)	2,223,099 (116,225)	4,439,517 (156,948)	203,473,900 (2,258,863)

CARDS identified approximately 4.4 million nonmatched P-sample duplicates and 203.5 million matched P-sample duplicates. Further, CARDS identified about 3.5 million nonmatched P-sample duplicates to E-sample eligible census records.

Table 5 presents the P-sample FSPD results. (Note: This table, without the "Total" column, corresponds to Table 5 in the A.C.E. Revision II report "A.C.E. Revision II Results: Further Study of Person Duplication".)

[5]This table is weighted by the product of the A.C.E. sampling weight and the A.C.E. residence probability.

Table 5. FSPD Weighted[6] Estimate of P-sample Nonmover Resident Duplicates by Geography and Census Record Type

Geography	Census Record Type								Total	
	E-Sample Eligible		GQ		Reinstate		Delete			
	Non-match	*Match*	*Non-match*	*Match*	*Non-match*	*Match*	*Non-match*	*Match*	*Non-match*	*Match*
Within Cluster	416,280 (18,018)	199,026,173 (2,219,841)	0 (0)	92,379 (22,926)	473,167 (57,809)	912,493 (46,165)	242,867 (33,492)	2,050,732 (119,196)	1,132,314 (70,386)	202,081,778 (2,249,322)
Surrounding Block	512,407 (40,638)	8,886,048 (554,638)	5,158 (2,874)	4,118 (1,669)	50,725 (13,984)	61,334 (14,614)	26,104 (7,482)	323,939 (30,243)	594,394 (45,983)	9,275,439 (558,924)
Same County	2,059,658 (118,086)	1,194,385 (36,534)	39,927 (8,731)	127,393 (25,170)	12,843 (3,966)	195,517 (17,580)	56,759 (24,408)	96,294 (13,677)	2,169,187 (130,288)	1,613,589 (51,900)
Diff. County, Same State	403,823 (28,374)	651,502 (24,389)	29,868 (4,168)	86,527 (6,531)	3,791 (1,732)	39,092 (7,320)	7,676 (3,456)	10,575 (2,931)	445,159 (29,549)	787,696 (27,177)
Different State	268,031 (20,114)	843,350 (25,839)	15,480 (2,318)	102,439 (6,389)	3,851 (2,349)	3,272 (840)	2,871 (1,017)	10,071 (2,577)	290,233 (20,882)	959,132 (27,485)
Total	3,660,200 (136,136)	210,601,459 (2,329,199)	90,433 (10,578)	412,855 (35,784)	544,376 (59,978)	1,211,708 (52,678)	336,277 (43,229)	2,491,612 (127,194)	4,631,286 (178,613)	214,717,634 (2,369,420)

FSPD identified approximately 4.6 million nonmatched P-sample duplicates, which is approximately 0.2 million more than CARDS found. Further, FSPD identified about 0.1 million more nonmatched P-sample duplicates to E-sample eligible census records than CARDS.

Some differences among the nonmatches which stood out between the CARDS and FSPD P-sample results were that CARDS found more duplicates to group quarters and to census records in different states. This is similar to the findings for the E-sample.

As a second comparison, we again used FSPD links, but with probabilities adjusted based on CARDS results as described above.

[6]This table is weighted by the product of the A.C.E. sampling weight, the A.C.E. residence probability, and the final probability of duplication.

Table 6. FSPD with CARDS Adjusted Weighted[7] Estimate of P-sample Nonmover Resident Duplicates by Geography and Census Record Type

Geography	Census Record Type								Total	
	E-Sample Eligible		GQ		Reinstate		Delete			
	Non-match	*Match*	*Non-match*	*Match*	*Non-match*	*Match*	*Non-match*	*Match*	*Non-match*	*Match*
Within Cluster	410,839 (17,938)	198,491,706 (2,215,573)	0 (0)	92,481 (22,963)	469,950 (57,562)	913,369 (46,178)	243,397 (33,490)	2,050,705 (119,182)	1,124,186 (70,056)	201,548,261 (2,245,086)
Surrounding Block	510,610 (40,531)	8,853,657 (550,738)	5,164 (2,876)	3,501 (1,419)	51,464 (14,004)	61,418 (14,664)	26,141 (7,484)	323,486 (29,996)	593,380 (45,822)	9,242,062 (555,203)
Same County	2,048,694 (116,300)	1,169,978 (36,224)	39,647 (8,726)	127,748 (25,205)	12,880 (3,978)	201,481 (17,596)	57,010 (24,414)	100,214 (14,358)	2,158,231 (128,026)	1,599,420 (52,059)
Diff. County, Same State	415,131 (28,772)	597,472 (24,422)	30,492 (4,224)	79,271 (6,325)	3,707 (1,757)	38,230 (7,381)	9,659 (4,039)	12,121 (3,275)	458,989 (30,063)	727,094 (27,218)
Different State	333,422 (21,910)	864,283 (28,819)	15,484 (2,448)	61,373 (5,240)	4,213 (2,400)	4,455 (1,304)	3,811 (1,306)	9,588 (2,989)	356,930 (22,765)	939,699 (30,216)
Total	3,718,696 (135,348)	209,977,095 (2,323,368)	90,787 (10,591)	364,374 (35,582)	542,215 (59,741)	1,218,953 (52,726)	340,018 (43,261)	2,496,114 (127,265)	4,691,717 (177,069)	214,056,537 (2,362,928)

This CARDS adjusted weighting increased the estimate of nonmatched P-sample duplicates to different states, while only slightly increasing the estimate of nonmatched P-sample duplicates to group quarters. The adjustment increased the FSPD estimate of nonmatched P-sample duplicates to different states by approximately 66,700. However, this is still lower than the CARDS estimates for this group.

4.3 Additional Analyses of CARDS E-sample Links

In an attempt to explain some of the differences between the FSPD and CARDS estimates discussed in Section 4.1, we also looked at some characteristics of the CARDS links.

[7]This table is weighted by the product of the A.C.E. sampling weight, the A.C.E. residence probability, and the adjusted final probability of duplication based on the CARDS results.

4.3.1 Household Composition

We examined the CARDS links by household (HH) composition, which looks at size of the sample HH size and HH duplication status (the number of links between the HHs relative to the size of the source HH). We broke this out by whether the link was to the same state (the first four categories of geography in the above tables) or to a different state, since the latter category is where CARDS tended to find more duplication.

The categories of source HH size are:

- 1 person
- 2 or more people

The categories of HH duplication status are:

- *All* – if the link is in a one- person HH or if the number of links equals the HH size
- *Partial with 2 or more links* – if two or more people within the HH linked, but the number of links was less than the HH size
- *Partial with only 1 link* – if one person in the HH was linked, and the HH had two or more people.

Table 7. CARDS Weighted Estimate of CARDS Only E-sample Links by Household Composition and Geography

Household Composition		**Geography**			
HH Size	HH Duplication Status	Same State		Different State	
		% CARDS Only	*Total*	*% CARDS Only*	*Total*
1	*All*	36.0% (1.1)	727,889 (23,908)	54.6% (2.3)	132,379 (7,296)
2+	*All*	2.8% (0.3)	3,052,411 (100,883)	8.7% (1.1)	232,581 (18,014)
	Partial - 2+ links	10.3% (0.5)	2,139,959 (64,818)	39.2% (2.3)	202,463 (11,155)
	Partial - Only 1 link	34.1% (0.7)	1,837,816 (35,799)	64.7% (1.0)	853,828 (19,821)
	Total	13.3% (0.3)	7,030,186 (151,162)	50.6% (1.0)	1,288,872 (31,980)
	Total	15.4% (0.3)	7,758,075 (159,182)	51.0% (1.0)	1,421,251 (34,133)

Approximately 51 percent of the E-sample CARDS links to different states were CARDS only, compared with 15.4 percent of CARDS links to the other geographical distances. This made us question what was unique about these CARDS links to different states.

As expected, when more than one person linked between the HUs more of the CARDS links were also in FSPD (in other words, there were fewer CARDS only links). This general trend held both for links to different states and for links within a state. However, more CARDS links to different states were CARDS only.

There were a greater proportion of HHs with more than two people but only one link (which we call single links) for the different state geographic level (853,828/1,421,251 ≈ 60.1 percent vs 1,837,816/7,758,075 ≈23.7 percent for the other levels). Furthermore, a greater proportion of the single links to different states was found in CARDS only (64.7 percent versus 34.1 percent for the other geographic distances). So like FSPD, CARDS found more single links to different states. However, there was much less overlap between FSPD and CARDS for the single links to different states.

In the FSPD process, the single links to different states tended to receive lower probabilities. However, CARDS treated all links equally. Yet we are less confident that single links to different states are truly duplicates than links in HHs where we were able to find links for all people in a multiple person HH.

4.3.2 Phase of CARDS Matching Process

Recall that the matching process used to assign PIKs to HCUF records had two phases: a Geokey Search phase (using address and person information) and a Name Search phase (using person information only). We are more confident of links created in the Geokey Search phase, because they require similar address data as well as person data. Thus, we broke out the CARDS links by whether the PIKs were assigned to the source and/or target record in the Geokey Search phase or not.

The PIKS assigned in Geokey Search phase categories are:

- Both the source and target PIK assigned in Geokey Search phase
- Only source or target PIK assigned in Geokey Search phase
- Neither source or target PIK assigned in Geokey Search phase

Table 8. CARDS Weighted Estimate of CARDS E-sample Links by Match Phase. Links within a State Only.

PIKs Assigned in Geokey Search Phase	Type of Cards Link					
	Cards Links	Cards Only Links	% of Cards Links That Are Cards Only	% of Total Cards Only Links	Cards Links Also in FSPD	% of Total Cards Links Also in FSPD
Both Source & Target	5,815,854 (134,973)	805,416 (23,770)	13.8% (0.4)	67.4% (0.9)	5,010,438 (125,535)	76.4% (0.6)
Only Source or Target	1,369,758 (35,636)	318,126 (12,317)	23.2% (0.8)	26.6% (0.9)	1,051,632 (32,338)	16.0% (0.5)
Neither Source nor Target	572,463 (25,383)	72,240 (5,502)	12.6% (0.9)	6.0% (0.4)	500,224 (23,521)	7.6% (0.3)
Total	7,758,075 (159,182)	1,195,782 (29,173)	15.4% (0.4)	100.0%	6,562,294 (146,443)	100.0%

Table 9. CARDS Weighted Estimate of CARDS E-sample Links by Match Phase. Links Between States Only.

PIKs Assigned in Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	199,937 (10,740)	58,436 (4,499)	29.2% (1.9)	8.1% (0.6)	141,502 (9,055)	20.3% (1.0)
Only Source or Target	1,092,517 (27,185)	586,635 (15,415)	53.7% (1.1)	81.0% (0.8)	505,882 (19,935)	72.6% (1.1)
Neither Source nor Target	128,797 (6,693)	79,290 (4,948)	61.6% (2.5)	11.0% (0.6)	49,506 (4,275)	7.1% (0.6)
Total	1,421,251 (34,133)	724,362 (17,831)	51.0% (1.0)	100.0%	696,889 (25,540)	100.0%

Table 10. CARDS Weighted Estimate of CARDS E-sample Links by Match Phase. Links within a State and Between States.

PIKs Assigned in Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS Only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	6,015,791 (135,882)	863,852 (24,503)	14.4% (0.4)	45.0% (0.8)	5,151,939 (126,133)	71.0% (0.6)
Only Source or Target	2,462,275 (47,800)	904,761 (20,803)	36.7% (0.7)	47.1% (0.8)	1,557,514 (39,111)	21.5% (0.5)
Neither Source nor Target	701,260 (26,910)	151,530 (7,565)	21.6% (1.0)	7.9% (0.4)	549,730 (24,084)	7.6% (0.3)
Total	9,179,326 (169,734)	1,920,143 (37,380)	20.9% (0.4)	100.0%	7,259,183 (152,077)	100.0%

We see that for links within a state, the phase in which the CARDS linking was done had some relation to the percentage of CARDS links that are also in FSPD. For about 69 percent of CARDS only links, both source and target had been linked to the Numident using Geokey, while the percent was about 78 for the CARDS links also in FSPD. When both HCUF-Numident links were found in the Geokey phase, about 14 percent of the CARDS links are CARDS only links, compared to about 20 percent for the other CARDS links.

However, for links to different states, there is a much more striking relation. About 29 percent of CARDS links where both records were matched in the Geokey phase, were CARDS only. The percentages were higher for CARDS links where either one or zero of the records were matched in the Geokey phase: about 54 percent of those where only the source or target was matched in the Geokey phase were CARDS only cases, and about 62 percent of cases where neither source nor target were matched in the Geokey phase were CARDS only. Because we have some reason to be less confident in the HCUF-Numident links that did not use the Geokey, we believe that more research would be needed to assess these CARDS only links.

4.4 Additional Analyses of CARDS P-sample Links

Similar to Section 4.3, we looked at some characteristics of the nonmatched P-sample nonmover resident links in an attempt to explain some of the differences between the FSPD and CARDS estimates discussed in Section 4.2.

4.4.1 Household Composition

As for the E-sample, we examine the P-sample CARDS links by household (HH) composition and whether the link was to the same or different state.

Table 11. CARDS Weighted Estimate of CARDS Only Nonmatched P-sample Nonmover Resident Links by Household Composition and Geography

Household Composition		Geography			
HH Size	HH Duplication Status	Same State		Different State	
		% CARDS Only	*Total*	*% CARDS Only*	*Total*
1	*All*	33.4% (2.5)	460,110 (32,111)	46.5% (4.7)	67,013 (6,503)
2+	*All*	7.1 (0.5)	2,236,161 (116,268)	32.5 (3.1)	196,319 (15,467)
	Partial - 2+ links	11.8 (1.0)	854,363 (51,012)	47.1 (4.3)	126,205 (11,898)
	Partial - Only 1 link	33.8 (2.1)	334,644 (16,645)	61.8 (3.2)	164,702 (12,041)
	Total	10.9 (0.5)	3,425,168 (140,277)	46.2 (2.3)	487,226 (24,159)
	Total	13.6% (0.6)	3,885,278 (151,036)	46.2% (2.1)	554,239 (25,664)

Approximately 46.2 percent of the nonmatched (in A.C.E.) P-sample nonmover resident CARDS links to different states were CARDS only, versus 13.6 percent for links to the other geographical distances. This suggests that there are differences between links to different states and the other links.

As expected, when there were more people linked between the HUs more of the CARDS links that were also found by FSPD (in other words, fewer CARDS only links). This general trend held for both links to different states and with states. However, there was less overlap between FSPD and CARDS links to different states.

There was a greater proportion of HHs with more than two people but only one link (which I call single links) for the different state geographic level (164,702/554,239 ≈ 29.7 percent vs 334,644/3,885,278 ≈8.6 percent for the other levels). Furthermore, a greater proportion of the single links to different states were found in CARDS only (61.8 percent versus 33.8 percent for the other geographic distances). So like FSPD, CARDS found more single links to different states. However, there was much less overlap between FSPD and CARDS for the single links to different states.

These findings are similar to those found for the E-sample in Section 4.3.1.

4.4.2 Phase of CARDS Matching Process

We broke out the P-sample CARDS links by whether the PIKs were assigned to the source and/or target record in the Geokey Search phase or not and by whether the link was to the same or different state. The results are in Tables 12-14.

Table 12. CARDS Weighted Estimate of CARDS Nonmatched P-sample Nonmover Resident Links by Match Phase. Links within a State Only.

PIKs Assigned In Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS Only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	2,990,769 (130,603)	368,363 (17,559)	12.3% (0.6)	70.0% (1.6)	2,622,405 (124,410)	78.1% (1.1)
Only Source or Target	613,486 (38,289)	126,111 (9,744)	20.6% (1.6)	23.9% (1.5)	487,375 (35,792)	14.5% (1.0)
Neither Source nor Target	281,023 (20,034)	32,091 (4,250)	11.4% (1.5)	6.1% (0.8)	248,932 (19,171)	7.4% (0.5)
Total	3,885,278 (151,036)	526,565 (21,830)	13.6% (0.6)	100.0%	3,358,713 (143,489)	100.0%

Table 13. CARDS Weighted Estimate of CARDS Nonmatched P-sample Nonmover Resident Links by Match Phase. Links Between States Only.

PIKs Assigned In Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS Only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	77,586 (7,477)	24,834 (4,319)	32.0% (4.6)	9.7% (1.6)	52,753 (6,026)	17.7% (1.9)
Only Source or Target	420,408 (22,116)	202,805 (12,976)	48.2% (2.5)	79.1% (2.0)	217,603 (17,063)	73.0% (2.2)
Neither Source nor Target	56,245 (5,569)	28,626 (3,700)	50.9% (4.9)	11.2% (1.4)	27,618 (4,085)	9.3% (1.3)
Total	554,239 (25,664)	256,265 (14,585)	46.2% (2.1)	100.0%	297,973 (19,621)	100.0%

Table 14. CARDS Weighted Estimate of CARDS Nonmatched P-sample Nonmover Resident Links by Match Phase. Links within a State and Links between States.

PIKs Assigned In Geokey Search Phase	Type of CARDS Link					
	CARDS Links	CARDS Only Links	% of CARDS Links That Are CARDS Only	% of Total CARDS Only Links	CARDS Links Also in FSPD	% of Total CARDS Links Also in FSPD
Both Source & Target	3,068,355 (131,090)	393,197 (18,113)	12.8% (0.6)	50.2% (1.5)	2,675,158 (124,628)	73.2% (1.1)
Only Source or Target	1,033,894 (45,621)	328,916 (16,534)	31.8% (1.5)	42.0% (1.5)	704,978 (40,272)	19.3% (1.0)
Neither Source nor Target	337,268 (21,448)	60,717 (5,786)	18.0% (1.6)	7.8% (0.7)	276,551 (19,817)	7.6% (0.5)
Total	4,439,517 (156,948)	782,831 (27,074)	17.6% (0.6)	100.0%	3,656,686 (146,475)	100.0%

We see the same pattern as we saw with the E-sample. For links between different states, the CARDS only cases are disproportionately cases where either source or target was matched in the Name Search phase, not the Geokey phase, of matching. Because we have some reason to be less confident in the HCUF-Numident links that did not use the Geokey, we believe that more research would be needed to assess these CARDS only links.

4.5 CARDS Estimates of FSPD Efficiencies

We used CARDS to provide estimates of efficiency of the FSPD duplicate detection for areas within and outside the block cluster. Table 15 shows the efficiency estimates for the E-sample for different household composition when the duplicate pair has both members in the same state and when they are in different states. We made two estimates of efficiency from CARDS. One is based on the assumption that all the additional duplicates from CARDS are accurate. The other assumes that only 50 percent of the additional duplicates are accurate.

- For single duplicate links within the multi-person households, the efficiency of 20.7% based on the A.C.E. within cluster results was definitely too low.

- For the other household compositions for duplicates in different states, the CARDS efficiency estimates tend to be lower than the estimates from the A.C.E. within cluster clerical search when all the CARDS-only duplicates are assumed to be accurate. However, if only 50% of the CARDS-only duplicates in different states are assumed

accurate, then the efficiency rates are higher for the single links in single-person households and for whole households links in multi-person households.

Table 15. Efficiency Estimates for FSPD for E-sample Based on A.C.E. within Cluster and CARDS by Geography by Accuracy of Additional Duplicates found by CARDS.

Household Composition		Geography				
HH Size	HH Duplication Status	A.C.E.	Same State		Different State	
		Within cluster	*All CARDS Only*	*50% CARDS Only*	*All CARDS Only*	*50% CARDS Only*
1	*All*	45.8%	64.0%	78.1%	45.4%	62.4%
2+	*All (whole HHs)*	93.9%	97.2%	98.6%	91.3%	95.5%
	Partial - 2+ links	98.5%	90.7%	94.6%	60.8%	75.6%
	Partial - Only 1 link	20.7%	65.9%	79.5%	35.3%	54.6%
	Total	86.9%	84.6%	91.7%	49.0%	65.8%

Note: Efficiency estimate based on A.C.E. Within Cluster found in A.C.E. Revision II Report #PP-51 (Mule, 2002). Efficiency estimates for CARDS only are based on Table 7 above.

5. CONCLUSIONS

In this study, we compared the results of two methods for identifying duplicates. The FSPD performed statistical and exact computer matching procedures to link E- and P-sample records to census records. Using administrative records, CARDS performed a computer match to attempt to assign each census record (including the E-sample) and P-sample record a PIK (used in place of a confidential SSN). Then CARDS linked E- and P-sample records to census records with the same PIK.

We examined E-sample duplicate links (cases where E-sample records could be linked with other census records), and links between P-sample nonmover residents and the census. Here are our main conclusions:

- **The FSPD process was more effective at finding duplicates that are geographically close**. Clerical matchers found about 1.9 million E-Sample duplicates to E-Sample eligible records. While both FSPD and CARDS identified significantly fewer such duplicates, FSPD found more (about 1.2 million compared to about 1 million for CARDS). FSPD found more duplicates within the A.C.E. cluster as well as within the surrounding blocks for all categories of census record except group quarters. This held true both for E-Sample duplicates, and nonmatched (in A.C.E.) P-Sample links.

- **CARDS identified more duplicates that are geographically distant.** As the links got farther apart, CARDS identified relatively more duplicates than FSPD. In different states,

CARDS had about twice as many links as FSPD, both for E-sample duplication and for nonmatched P-sample links.

- **CARDS identified more group quarters duplicates.** For group quarters, the FSPD process was limited to its exact matching stage. Thus, we expected that the CARDS process might find more duplicates to group quarters, and the results confirmed this.

- **CARDS links that were geographically more distant were more questionable.** Although CARDS linked many more people to different states, we saw reason to question some of these links. CARDS seems to have identified a large number of duplicates to different states where only one person was linked in a multi-person household. A high percentage (more than 60 percent) of these were CARDS only links. In the FSPD process, the single links to different states tended to receive lower probabilities. However, CARDS treats all links equally. Yet we are less confident that single links to different states are truly duplicates than links in HHs where we were able to find links for all people in a multiple person HH.

  In addition, we are less confident in CARDS links when the source record and/or target record was not matched in the Geokey Search phase of matching in which address data were used. When matching using person characteristics only, finding records with similar names and ages may be coincidental. Among links to different states where one or both PIKs were assigned based on person characteristics (name and date of birth) only, a high percentage were CARDS only links. We believe that in many cases, the CARDS process will have avoided linking different people whose person characteristics were similar. When those characteristics were fairly common, the CARDS matching process is likely to have linked more than one Numident record to one HCUF record. In those cases, CARDS would not have assigned any PIK to the HCUF record. However, we do not know how many false links remained. More research is needed to design methods to adequately address cases in which different people coincidentally have similar person characteristics.

- **A combined FSPD/CARDS procedure improved estimates**. We tried a conservative way to incorporate CARDS results into the FSPD process. We did not use any CARDS only links, but used CARDS confirmation and denial of FSPD links to change FSPD probabilities to one or zero. This process increased the estimates of duplication from FSPD alone, which we believe to be an improvement.

- **We suggest further study of the FSPD and CARDS links.** This is just the beginning of the research that can be done to explore the nature of the duplication found by FSPD and CARDS. There are additional questions in the CARDS study plan that time did not permit analysis of. (Bean and Bauder, 2002) Further, this analysis only begins the exploration of differences between the FSPD and CARDS estimates. In addition, the CRCD (Beaghen and Byrne, 2002) may be able to provide further information about the status of FSPD undetermined duplicates and CARDS only duplicates.

- **We recommend that CARDS style research continue with improved administrative records procedures for detecting duplicates.** Administrative records can be valuable aids for detecting duplicates. CARDS style processes have the potential to identify duplicates that other methods have difficulty detecting – for example, people enumerated with different names, and people whose enumerations have reporting errors. We have seen in this report that CARDS data has been useful in confirmation and denial of FSPD links, and has the potential for finding additional duplicates. But we also have seen reasons here and in CRCD (Beaghen and Byrne, 2002), to question some of the CARDS links that were not also found by FSPD. This limited our ability to draw significant conclusions about duplicates missed by FSPD. However, we believe that administrative records have greater potential to be of value for unduplication research. The CARDS process used the results of an HCUF-Numident match done previously by the Administrative Records Research Staff. That goal of that match was to associate Census race data with Numident records. The match strategy and thresholds were developed with the goal of matching the HCUF as completely as possible, while maintaining a reasonably low false match rate over the whole of the HCUF. However, potential Census duplicates are a small and special subset of the HCUF. In this study, and CRCD, we may be seeing that within this small subset, the matching strategy had a higher false match rate than would be desirable. More research into the CARDS only cases would help determine the quality of the CARDS only links, and may suggest improvements in matching strategies for CARDS style processes. We believe that a CARDS style process that is developed from the beginning to detect duplicates, and that uses lessons learned from this study, CRCD, and further research, can produce more complete and accurate results.

6. REFERENCES

Bean, Susanne L. and D. Mark Bauder (2002). "Census and Administrative Records Duplication Study Plan," DSSD A.C.E. Revision II Estimates Memorandum Series, #PP-15 DRAFT dated September 24, 2002.

Beaghen, Michael and Rose Byrne (2002). "Clerical Review of Census Duplicates," A.C.E. Revision II Report #PP-43. U.S. Census Bureau, Washington, D.C.

Davis, M.C. and P. Biemer (1991). Measurement of the Census Erroneous Enumerations - Clerical Error Made in the Assignment of Enumeration Status. 1990 Coverage Studies and Evaluation Memorandum Series, #L-2 dated July 11, 1991.

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.

Fenstermaker, Debbie (2003). "Chapter 5: Further Study of Person Duplication in Census 2000," A.C.E. Revision II: Design and Methodology. DSSD A.C.E. Revision II Estimates Memorandum Series, #PP-30, DRAFT dated January, 2003.

Jones, John (2003). "Person Duplication in the Search Area Measured by the 2000 Accuracy and Coverage Evaluation," Census 2000 Evaluation O.16. U.S. Census Bureau, Washington D.C.

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.

Mule, Thomas (2002). "A.C.E. Revision II Results: Further Study of Person Duplication," A.C.E. Revision II Report #PP-51. U. S. Census Bureau, Washington, D.C.

Mulry, Mary (2003). "Chapter 7: Assessing the Estimates," A.C.E. Revision II: Design and Methodology. DSSD A.C.E. Revision II Estimates Memorandum Series, #PP-30, DRAFT dated January, 2003.